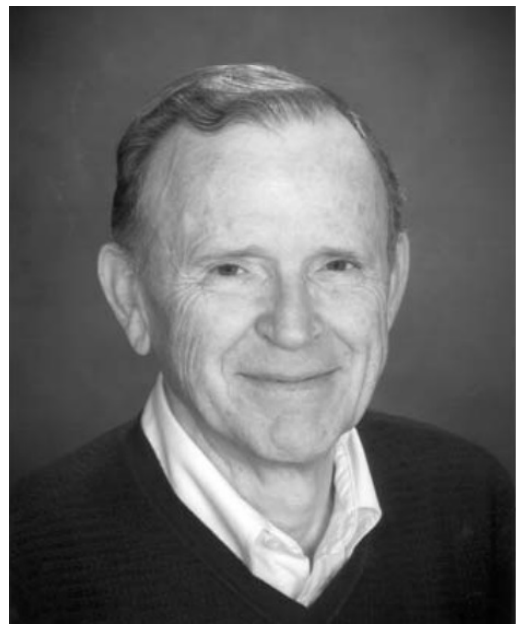
Numerical Analysis

Introduction and Mathematical Preliminaries

Elmer S. Poliquit

Introduction

Numerical analysis is the area of mathematics and computer science that creates, analyzes, and implements algorithms for solving numerically the problems of continuous mathematics.



KENDALL E. ATKINSON

Introduction

- Due to the immense development in the computational technology, numerical approximation has become more popular and a modern tool for scientists and engineers.
- As a result many scientific softwares are developed (for instance, Matlab, Mathematica, Maple, R, etc.) to handle more difficult problems in an efficient and easy way.
- These softwares contain functions that use standard numerical methods, where a user can pass the required parameters and get the results just by a single command without knowing the details of the numerical method.

3 Basic Reasons Why Study Numerical Analysis

1. Learning different numerical methods and their analysis will make a person more familiar with the technique of developing new numerical methods.

2. In many circumstances, one has more methods for a given problem.

3. With a sound background, one can use methods properly and, most importantly, one can understand what is going wrong when results are not as expected.

Obtaining and installing R

You can download and install R from the CRAN (Comprehensive R Archive Network) website at http://cran.r-project.org/. Choose the appropriate link for your operating system (Mac OS X, Windows, or Linux), and follow the (not very complicated) directions. Unless you have some special requirements for customization, you should choose the precompiled binary rather than the source code.

A neater, more streamlined-but perhaps less flexible- integrated development interface can be had by installing the freeware RStudio from https://www.rstudio.com/products/rstudio/download/.

Introduction to R

Conditional execution: if and if else. The *if* statement operates on logical vectors of length 1. A formal construction is

```
x = -3
if(x < -1) {          #if (condition 1) {
y = -1                #result 1
} else if (x < 0) {   #} else if (condition 2) {
y = 0                 #result 2
} else {              #} else {
y = 1                 #result 3
}                     #}
y
```

```
[1] -1
```

Be careful not to insert a line feed between } and else, because R will interpret everything up to the } as a complete command and will return a result prematurely.

Introduction to R

Likewise for more than two choices, with *else if*:

```
x = 0
if (x < 0) 0 else if (x == 0) 0.5 else 1
```

```
[1] 0.5
```

for loop - to repeat an operation through a given range of elements of a vector, use the for construction.

```
x = 0
for (i in 1:10) x = x + i
x
```

```
[1] 55
```

```
sum(1:10)
```

```
[1] 55
```

Introduction to R

Polynomial functions in packages

Use: install.packages{PolynomF}

```
require(PolynomF)
```

```
Loading required package: PolynomF
```

```
x = polynom() # Make x an object of class polynom
p = x^3 - 3*x^2 - 2*x + 7
q = x^2 + 2
p+q;p-q;p*q
```

```
9 - 2*x - 2*x^2 + x^3
```

```
5 - 2*x - 4*x^2 + x^3
```

```
14 - 4*x + x^2 - 3*x^4 + x^5
```

Introduction to R

Polynomial functions in packages

$$p = x^3 - 3x^2 - 2x + 7$$

```
dpdx = deriv(p,"x") # Differentiate p with respect to x
dpdx
```

```
-2 - 6*x + 3*x^2
```

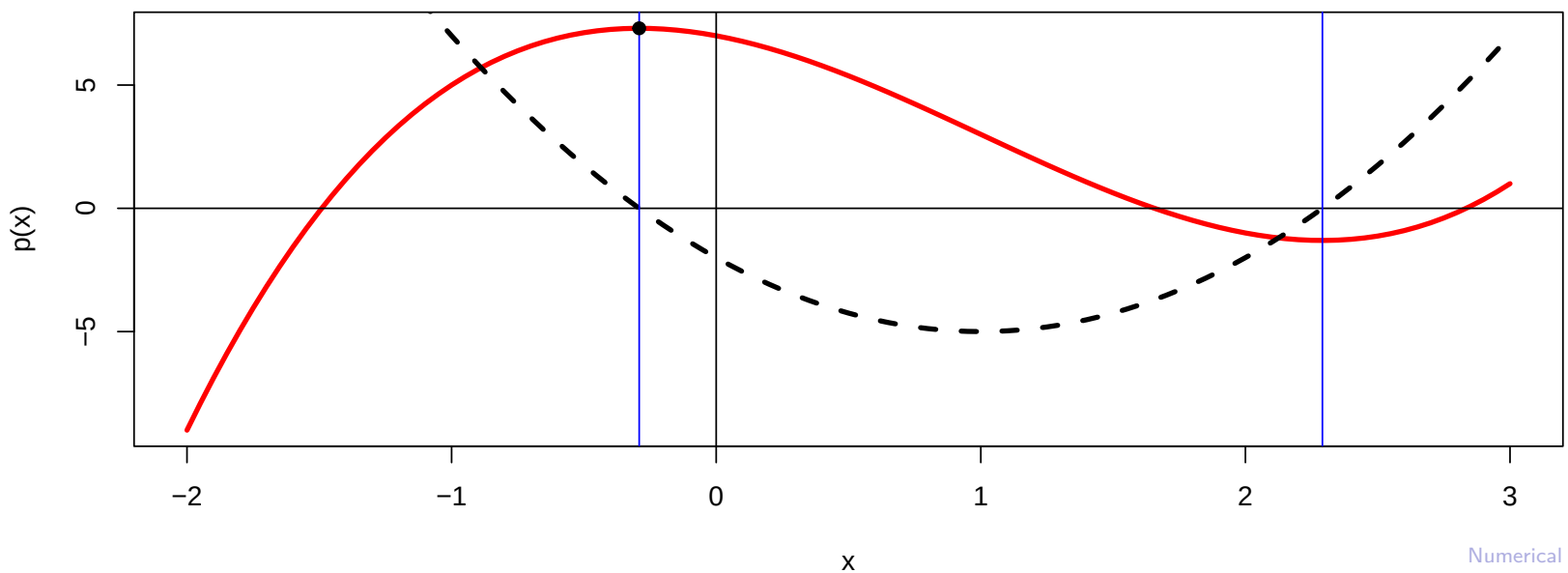
```
dpdx_zeros = solve(dpdx) #solve - finding the roots
dpdx_zeros
```

```
[1] -0.2909944  2.2909944
```

Introduction to R

$$p = x^3 - 3x^2 - 2x + 7$$

```
curve(p,-2,3, col = "red", lwd = 3)
curve(dpdx, -2, 3, lty=2, lwd=3, add = T)
abline(v=0,h=0)
abline(v=c(dpdx_zeros[1],dpdx_zeros[2]), col = "blue")
points(dpdx_zeros[1],p(dpdx_zeros[1]), pch = 19)
```



Polynomial functions in packages

$$\int_{-2}^{3}(dpdx)dx = \int_{-2}^{3}(-2-6x+3x^2)dx$$

```
integral(dpdx)
```

```
-2*x - 3*x^2 + x^3
```

```
integral(dpdx,limits = c(-2,3))
```

```
[1] 10
```

Continuity

A function $f : R \to R$ is said to be *continuous* at a point $x_0 \in R$ if the $lim_{x \to x_0} f(x) = f(x_0)$.

In other words, for any given $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $|x - x_0| < \delta$.

The process here is to work on backwards.

Example 1.1

Prove that $lim_{x \to 1} x^2 = 1$.

It is clear that $f(x) = x^2 \to 1$ when $x \to 1$.
Following the second condition:
This can be considered as your scratch work.

$$|f(x) - f(x_o)| = |x^2 - 1| = |x - 1||x + 1| < \epsilon$$
$$|x - 1| < \frac{\epsilon}{|x + 1|}$$

We are at the phase of saying that $|x - 1| <$ something, where something $= \frac{\epsilon}{|x+1|}$. We want to turn that something into δ.
Since x is approaching 1, we are safe to assume that x is between 0 and 2.
So how to work on this?

Example 1.1

We have

$$0 < x < 2$$
$$1 < x + 1 < 3$$
$$|x + 1| < 3$$
$$\frac{1}{3} < \frac{1}{|x + 1|}$$

We wanted

$$|x - 1| < \frac{\epsilon}{|x + 1|}.$$

The above shows that given any x in $[0, 2]$, we know that $|x + 1| < 3 \rightarrow \frac{1}{3} < \frac{1}{|x+1|} \rightarrow \frac{\epsilon}{3} < \frac{\epsilon}{|x+1|}$ (by multiplying ϵ to both sides of the equation).

Example 1.1

So we set $\delta \leq \frac{\epsilon}{3}$. This ends our scratchwork, and we begin the formal proof (which also helps us understand why this was a good choice of δ).

Solution

Given $\epsilon > 0$, let $\delta \leq \frac{\epsilon}{3}$. We want to show that when $|x-1| < \delta$, then $|x^2 - 1| < \epsilon$. We start with $|x-1| < \delta$:

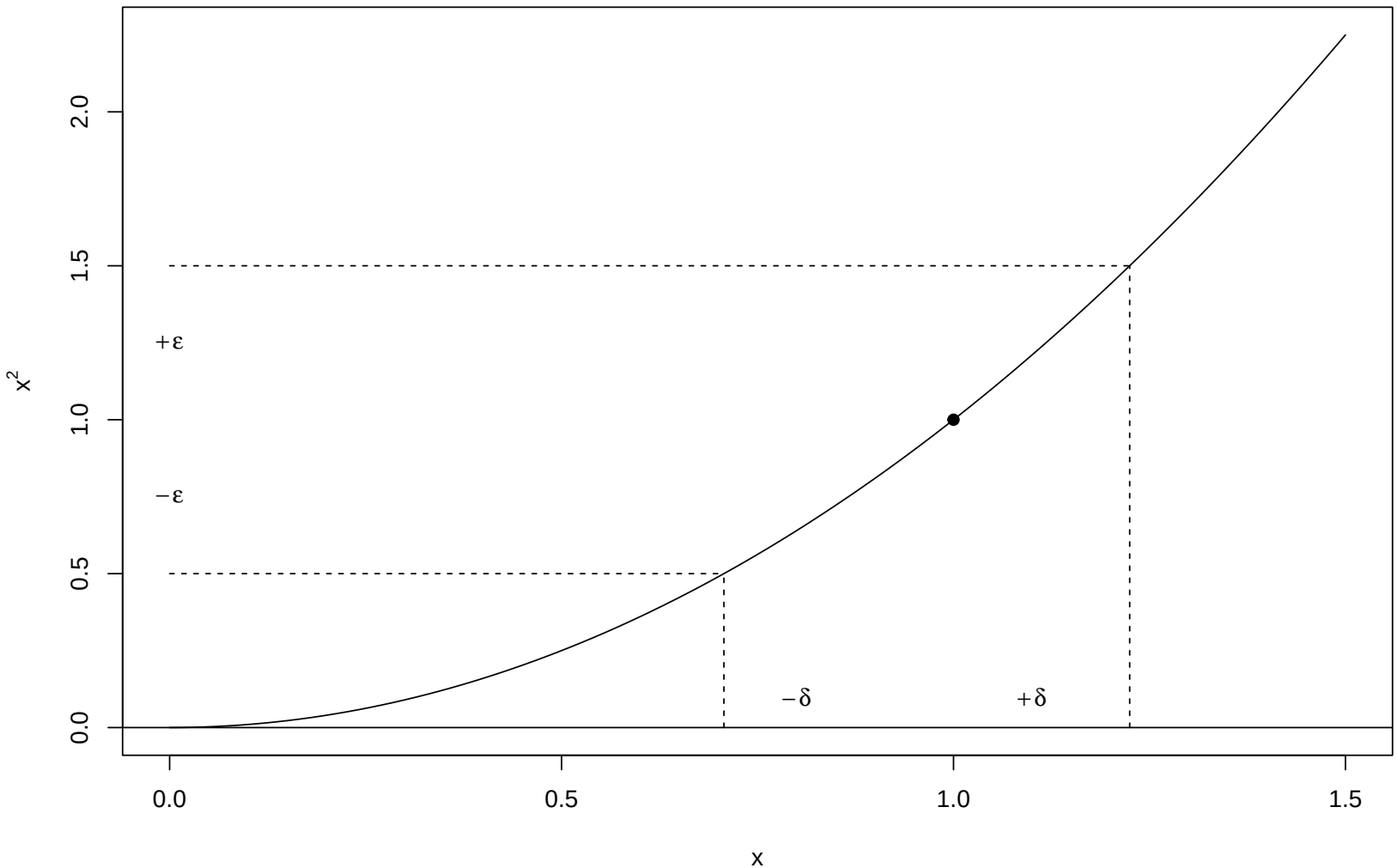
$$|x-1| < \delta$$
$$|x-1| < \frac{\epsilon}{3}$$
$$|x-1| < \frac{\epsilon}{3} < \frac{\epsilon}{|x+1|}$$
$$|x-1||x+1| < \epsilon$$
$$|x^2 - 1| < \epsilon,$$

which is what we wanted to show. Thus, $lim_{x \to 1} x^2 = 1$. $\square$

An illustration of this example is depicted in figure 1.1.

Example 1.1

Figure 1.1



1.1 Mathematical Preliminaries

Intermediate Value Theorem

Let $f(x)$ be continuous on the finite interval $a \leq x \leq b$, and define

$$m = \text{Infimum}_{a \leq x \leq b} f(x), \quad M = \text{Supremum}_{a \leq x \leq b} f(x)$$

Then for any number ζ (zeta) in the interval $[m, M]$, there is at least one point ξ in $[a, b]$ for which $f(\xi) = \zeta$.

In particular, there are points $\underline{x}$ and $\overline{x}$ in $[a, b]$ for which

$$m = f(\underline{x}), \quad M = f(\overline{x}).$$

Example 1.2

Use the *Intermediate Value Theorem* to prove that the equation

$$x^3 - 3x^2 - 2x + 9 = 0$$

is solvable.

Solution
The given equation is $x^3 - 3x^2 - 2x + 9 = 0$. Then we can let the function $f(x) = x^3 + 3x^2 - 2x + 9 = 0$.
Solving using R:

```
f=function(x){x^3-3*x^2-2*x+9}
uniroot(f,c(-2,-1))$root
```

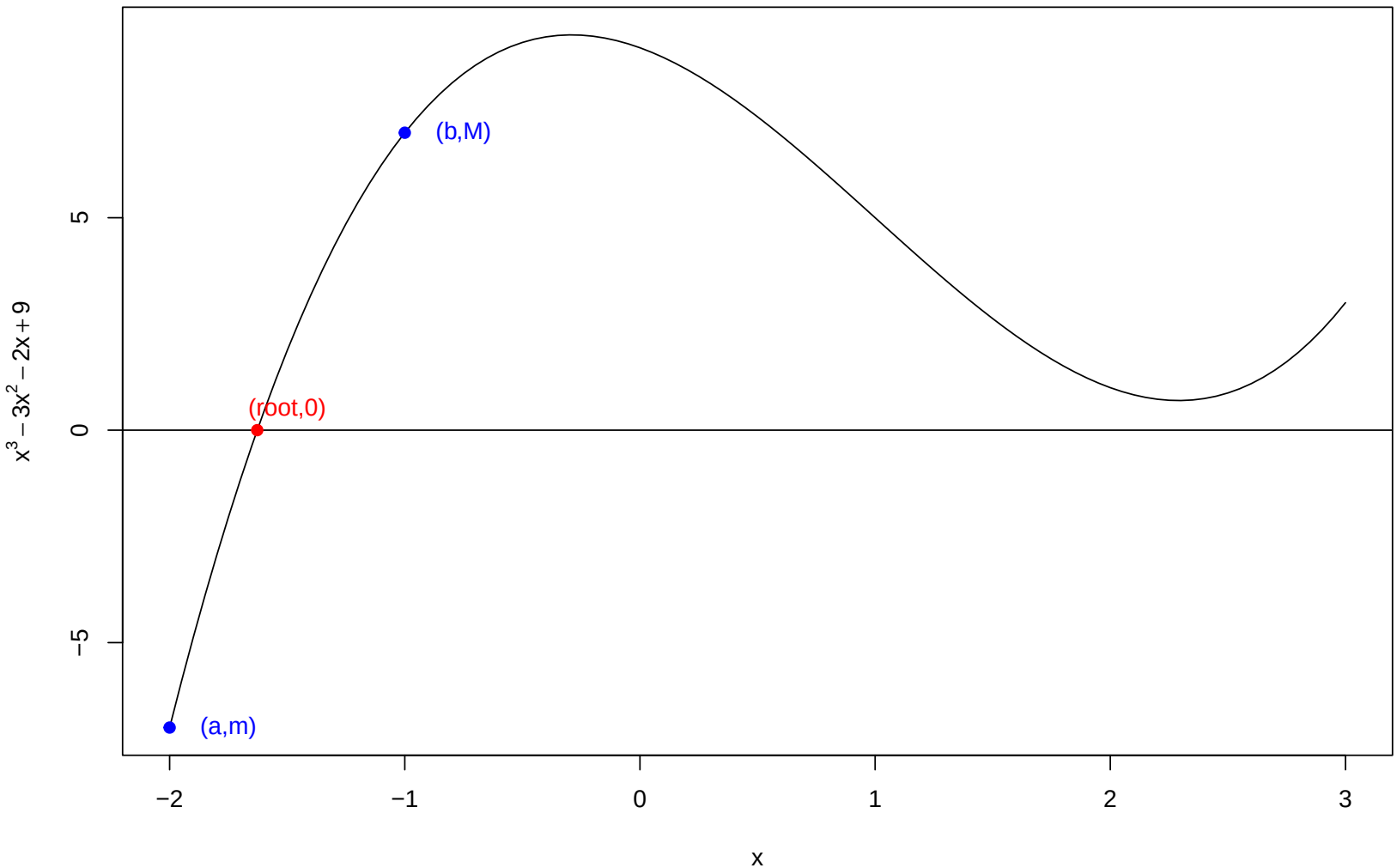
```
[1] -1.627356
```

```
f(uniroot(f,c(-2,-1))$root) #It is not exactly 0.
```

```
[1] 0.0001494108
```

Example 1.2

We can see the function graphically as shown.



Example 1.2

When $x = -2$, then we can have
$f(-2) = (-2)^3 - 3(-2)^2 - 2(-2) + 9 = -7(m) < 0$ and when
$x = -1$, $f(-1) = (-1)^3 - 3(-1)^2 - 2(-1) + 9 = 7(M) > 0$.
We have shown that $f(-2) = -7 < \zeta < f(-1) = 7$. That is, ζ is between $f(-2)$ and $f(-1)$. Then we can choose the interval to be $[-2, -1]$.
The assumptions of the Intermediate Value Theorem is met, so we can conclude that there is some number ξ in the interval $[-2, -1]$ which satisfies $f(\xi) = \zeta$, that is $x^3 - 3x^2 - 2x + 9 = 0$. Thus, the equation is solvable. ■

```
f=function(x){x^3-3*x^2-2*x+9}
f(uniroot(f,c(-2,-1))$root)
```

```
[1] 0.0001494108
```

1.1 Mathematical Preliminaries

Differentiation

A function $f : (a, b) \to R$ is said to be *differentiable* at a point $c \in (a, b)$ if the limit

$$\lim_{h \to 0} \frac{f(c+h)-f(c)}{h} \text{ exists.}$$

In this case, the value of the limit is denoted by $f'(c)$ and is called the *derivative* of f at c. The function f is said to be differentiable in (a, b) if it is differentiable at every point in (a, b).

Write

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

Rolle's Theorem

Let $f(x)$ be continuous on the bounded interval $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then $f'(\xi) = 0$, for some $\xi \in (a, b)$.

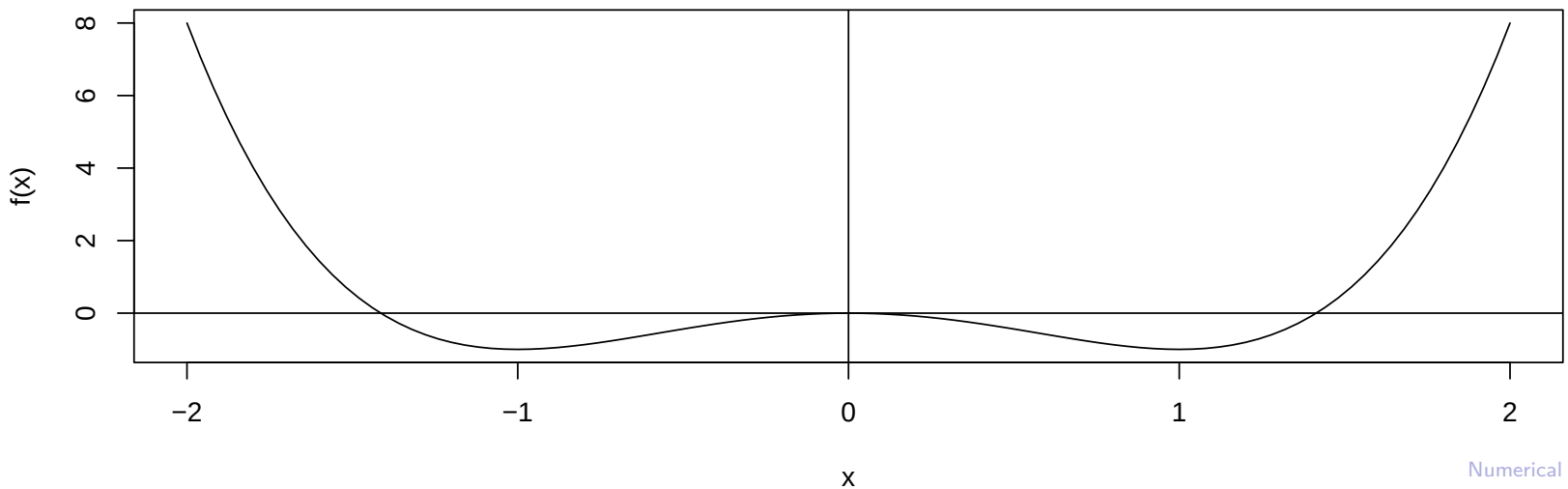
Example 1.3

Suppose we are asked to determine whether Rolle's theorem can be applied to $f(x) = x^4 - 2x^2$ on the closed interval $[-2, 2]$. And if so, find all values of ξ in the interval that satisfy the theorem's conclusion.

Solution

Since $f(x) = x^4 - 2x^2$ is a polynomial function, then $f(x)$ is continuous and differentiable.

Graphically, we can check $f(-2) = f(2)$.



Example 1.3

```
f=function(x) {x^4-2*x^2}
f(c(-2,2))
```

```
[1] 8 8
```

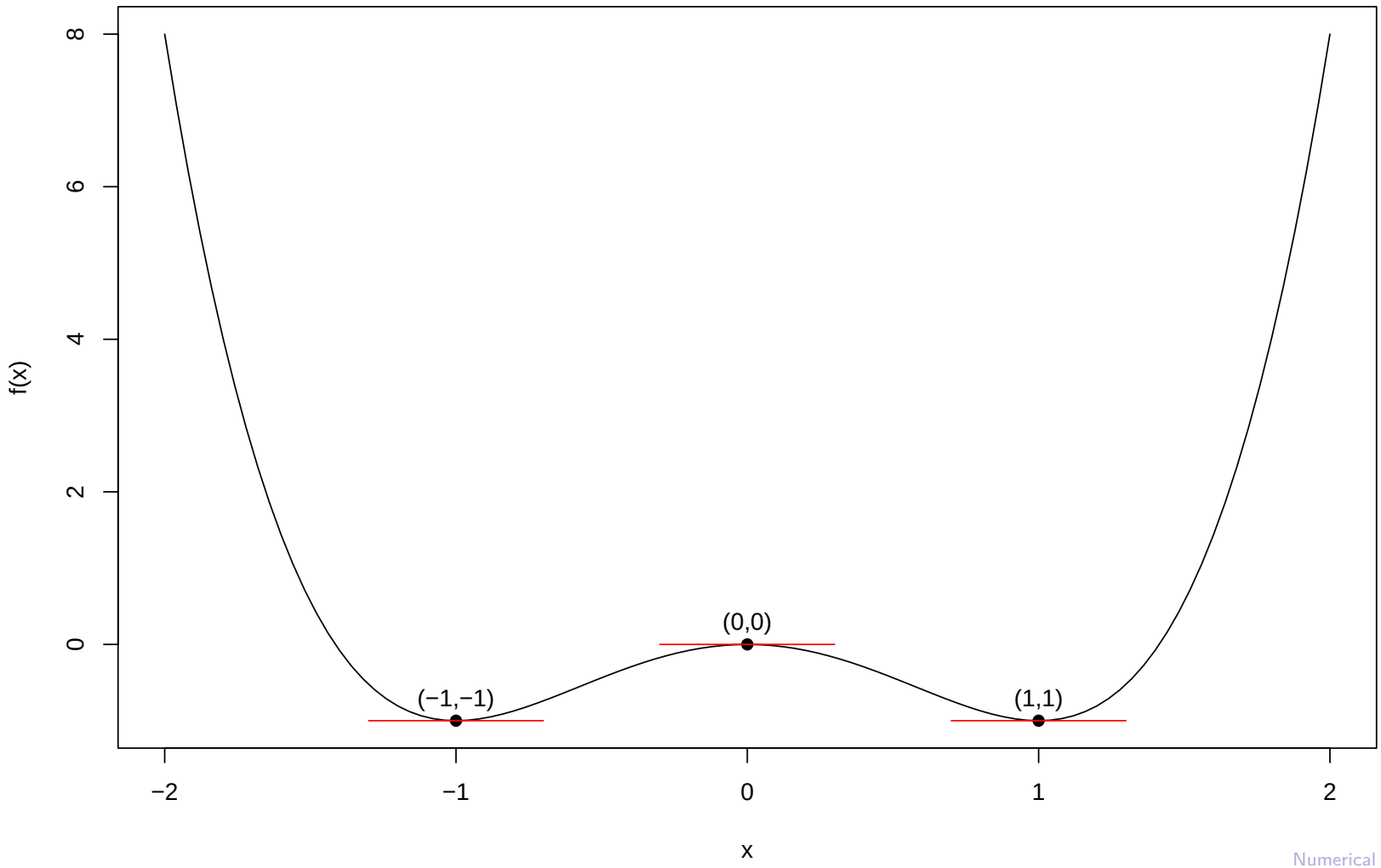
Because they both yield the same y-value of 8, we know that all requirements are satisfied, which means we can now find all values of ξ within the open interval $(-2, 2)$ where $f'(\xi) = 0$.

$$f'(\xi) = 4\xi^3 - 4\xi$$
$$0 = 4\xi^3 - 4\xi$$
$$0 = 4\xi(\xi^2 - 1)$$
$$\xi = -1, 0, 1$$

Therefore, the derivative of the function is 0 when x is either -1, 0, or 1.

Example 1.3

Using this process, we have found three values where the slope of the tangent *line* is zero within the interval. ■



Mean Value Theorem

Let $f(x)$ be continuous for $a \leq x \leq b$, and let it be differentiable for $a < x < b$. Then there is at least one point ξ in (a, b) for which

$$f(b) - f(a) = f'(\xi)(b - a).$$

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Example 1.4

Determine all the numbers ξ which satisfy the conclusions of the *Mean Value Theorem* for the following function.

$$f(x) = x^3 + 2x^2 - x \text{ on } [-1, 2].$$

Solution

Since $f(x)$ is a polynomial, then it is both continuous and differentiable, so $f'(x) = 3x^2 + 4x - 1$.

Now, $f'(\xi) = \frac{f(2)-f(-1)}{2-(-1)}$.

$$3\xi^2 + 4\xi - 1 = \frac{f(2) - f(-1)}{2 - (-1)}$$

$$3\xi^2 + 4\xi - 1 = \frac{14 - 2}{3}$$

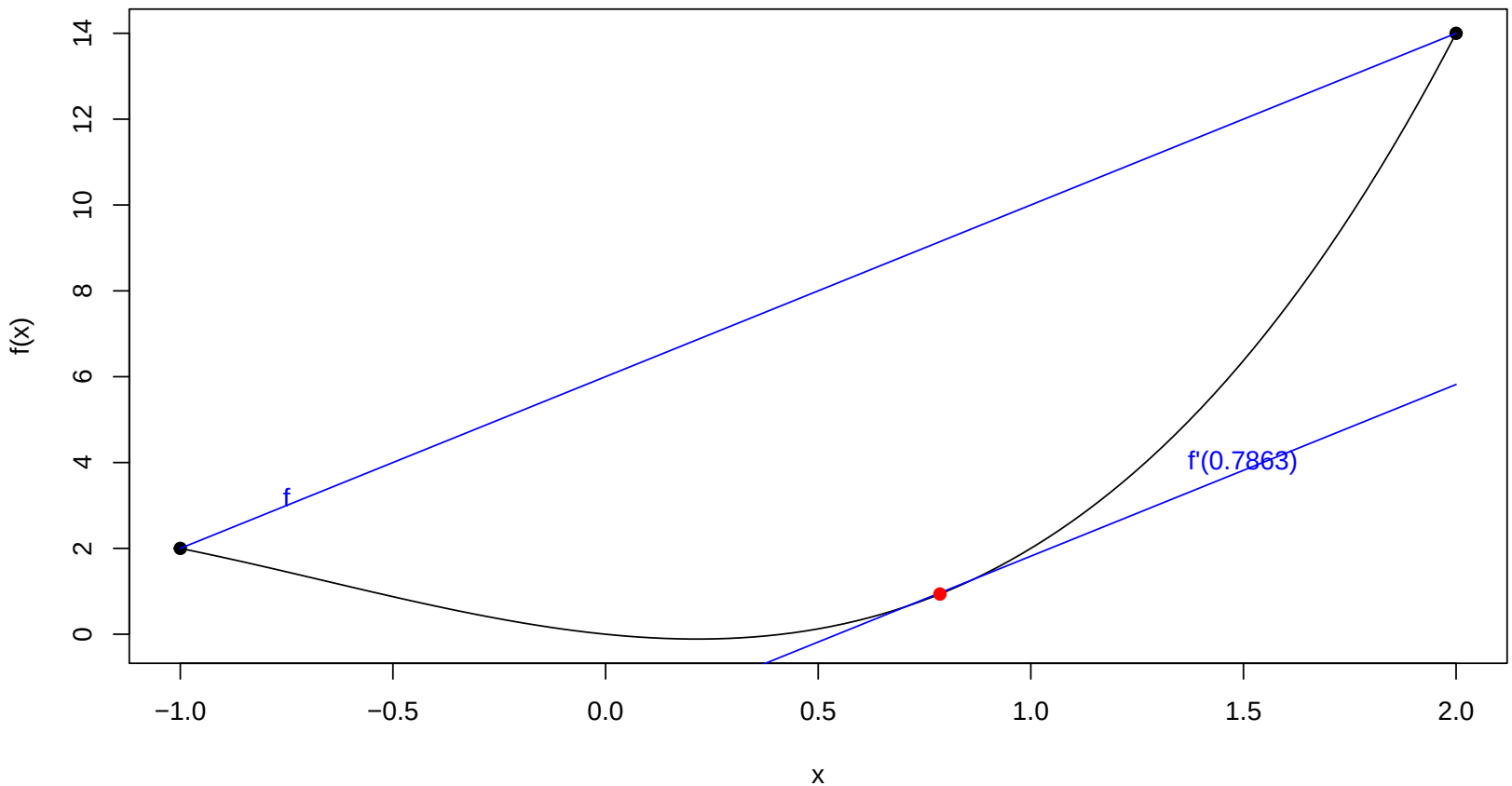
$$\xi = \frac{-4 \pm \sqrt{76}}{6}.$$

Example 1.4

Only one is a solution on $[-1, 2]$ which is

$\xi = \frac{-4+\sqrt{76}}{6} \approx 0.7863 \in [-1, 2]$. ■

Graphically, we can see that there is only one solution.



1.1 Mathematical Preliminaries

Integral Mean Value Theorem

Let $w(x)$ be nonnegative and integrable on $[a, b]$, and let $f(x)$ be continuous on $[a, b]$. Then

$$\int_a^b w(x)f(x)dx = f(\xi)\int_a^b w(x)dx$$

for some $\xi \in [a, b]$.

Note that one way to think of this theorem is the following. First rewrite the result as,

$$\frac{1}{b-a}\int_a^b f(x)dx = f(\xi) \rightarrow \int_a^b f(x)dx = f(\xi)(b-a)$$

and from this we can see that this theorem is telling us that there is a number $a < \xi < b$ such that $f_{average} = f(\xi)$. Or, in other words, if $f(x)$ is a continuous function then somewhere in $[a, b]$ the function will take on its average value.

Example 1.5

Find the point ξ that satisfies the mean value theorem for integrals on the interval $[1, 4]$. The function is $f(x) = 3x^2 - 2x$.

Solution

The function is a polynomial and is therefore continuous.

$$\int_1^4 (3x^2 - 2x)dx = (3\xi^2 - 2\xi)(4 - 1)$$

$$\int_1^4 3x^2 - 2xdx = 9\xi^2 - 6\xi$$

$$x^3 - x^2\Big]_1^4 = 9\xi^2 - 6\xi$$

$$48 = 9\xi^2 - 6\xi$$

$$\xi = \frac{8}{3}, -2.$$

Example 1.5

Since $\xi = \frac{8}{3} \in [1, 4]$, so $\xi = \frac{8}{3}$ satisfies the mean value theorem for integrals on the interval $[1, 4]$. ■

```
## define the integrated function
f <- function(x) {3*x^2-2*x}
## integrate the function from 1 to 4
integrate(f, lower = 1, upper = 4)
```

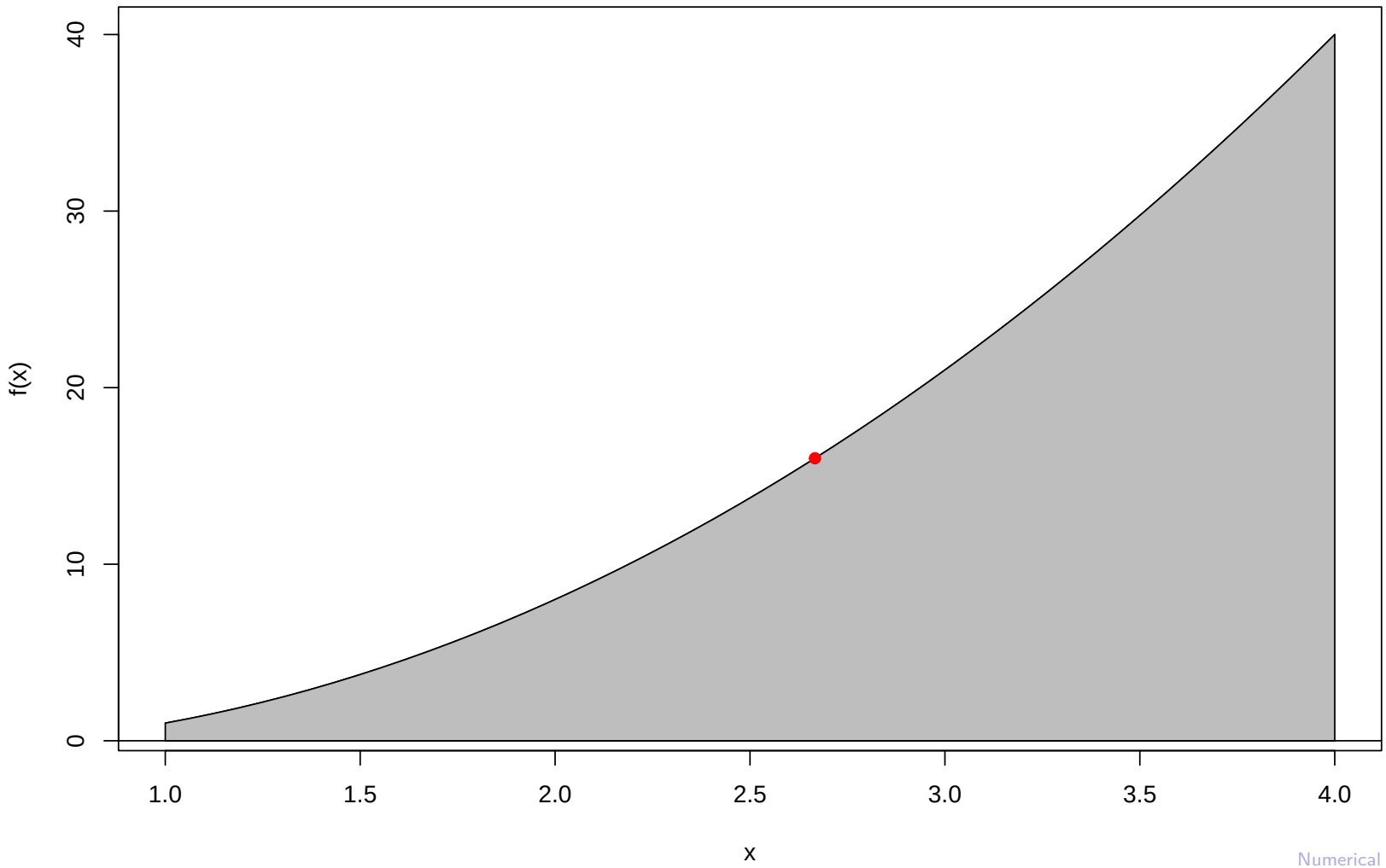
```
48 with absolute error < 5.3e-13
```

```
f(8/3)*(4-1)
```

```
[1] 48
```

Example 1.5

Since $\xi = \frac{8}{3} \in [1, 4]$, so $\xi = \frac{8}{3}$ satisfies the mean value theorem for integrals on the interval $[1, 4]$. ■



1.1 Mathematical Preliminaries

Taylor's Theorem

Let $f(x)$ have $n+1$ continuous derivatives on $[a, b]$ for some $n \geq 0$, and let x, $x_0 \in [a, b]$. Then

$$f(x) = p_n(x) + R_{n+1}(x)$$

where

$$p_n(x) = f(x_0) + \frac{x - x_0}{1!} f'(x_0) + \cdots + \frac{(x - x_0)^n}{n!} f^n(x_0)$$

and

$$R_{n+1}(x) = \frac{1}{n!} \int_{x_0}^{x} (x - t)^n f^{n+1}(t) dt = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{n+1}(\xi)$$

for some ξ between x_0 and x.

Remark

Note that as $x \to x_0$, the remainder $R_{n+1}(x) \to 0$. Thus, the Taylor's formula (without remainder) can be used to get an approximate value of f at any point x in a small neighborhood of x_0, once the values of f and all its n derivatives are known at x_0.

Taylor's theorem can be written as

$$f(x) = \left[\sum_{k=0}^{n} \frac{f^k(x_0)}{k!}(x - x_0)^k\right] + R_{n+1}(x).$$

Example 1.6

Consider the function $f(x) = \sqrt[3]{x}$.

1. Find the first and second Taylor polynomials for f at $x = 8$. Use a graphing utility to compare these polynomials with f near $x = 8$.
2. Use these two polynomials to estimate $\sqrt[3]{11}$.
3. Use Taylor's theorem to bound the error.

Example 1.6

Solution

$$p_n(x) = f(x_0) + \frac{x - x_0}{1!} f'(x_0) + \cdots + \frac{(x - x_0)^n}{n!} f^n(x_0)$$

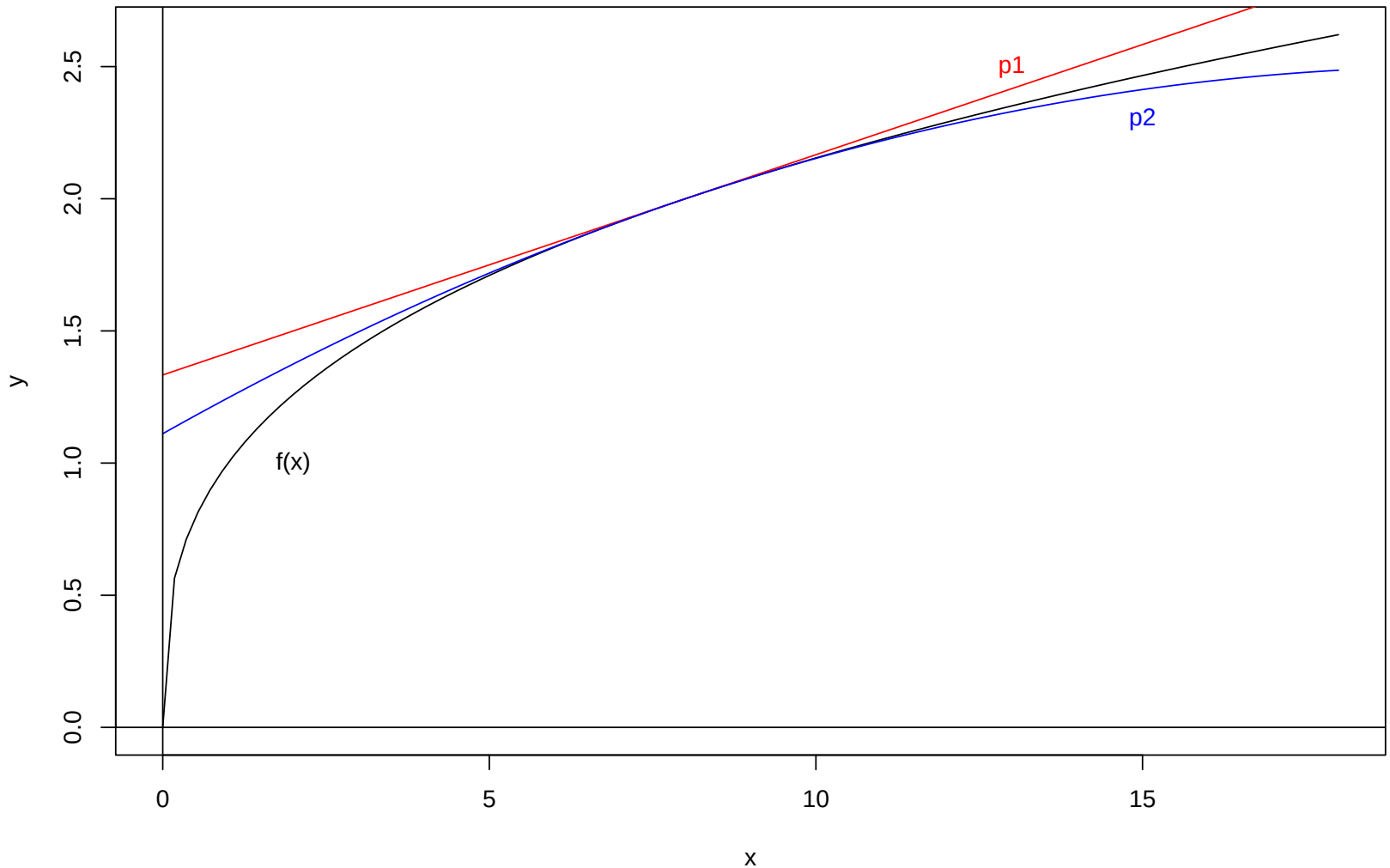
1.
Consider the function $f(x) = \sqrt[3]{x}$. For $f(x) = \sqrt[3]{x}$, the values of the function and its first three derivatives at $x = 8$ are as follows:

- $f(x) = \sqrt[3]{x} \to f(8) = 2$
- $f'(x) = \frac{1}{3x^{2/3}} \to f'(8) = \frac{1}{12}$
- $f''(x) = \frac{-2}{9x^{5/3}} \to f''(8) = -\frac{1}{144}$
- $f'''(x) = \frac{10}{27x^{8/3}} \to f'''(8) = \frac{5}{3456}$

Thus, the first and second Taylor polynomials at $x = 8$ are given by
$p_1 = f(8) + f'(8)(x - 8) = 2 + \frac{1}{12}(x - 8)$
$p_2 = f(8) + f'(8)(x-8) + \frac{f''(8)}{2!}(x-8)^2 = 2 + \frac{1}{12}(x-8) - \frac{1}{288}(x-8)^2$

Example 1.6

The graph of the function with the Taylor polynomials are shown below.



Example 1.6

2. *Solution*. Using the first Taylor polynomial at $x = 8$, we can estimate
$\sqrt[3]{11} \approx p_1(11) = 2 + \frac{1}{12}(11 - 8) = 2.25$.
Using the second Taylor polynomial at $x = 8$, we can estimate
$\sqrt[3]{11} \approx p_2(11) = 2 + \frac{1}{12}(11 - 8) - \frac{1}{288}(11 - 8)^2 = 2.21875$.

```
2+(1/12)*(11-8)
```

```
[1] 2.25
```

```
2+(1/12)*(11-8)-(1/288)*(11-8)^2
```

```
[1] 2.21875
```

```
11^(1/3)
```

```
[1] 2.22398
```

```
abs(11^(1/3)-2+(1/12)*(11-8)-(1/288)*(11-8)^2) #error
```

```
[1] 0.4427301
```

Example 1.6

3. *Solution*

$$R_{n+1}(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{n+1}(\xi)$$

By Taylor's Theorem with Remainder, there exists a ξ in the interval $(8, 11)$ such that the remainder when approximating $\sqrt[3]{11}$ by the *first* Taylor polynomial satisfies

$$R_1(11) = \frac{f''(\xi)}{2!}(11 - 8)^2$$

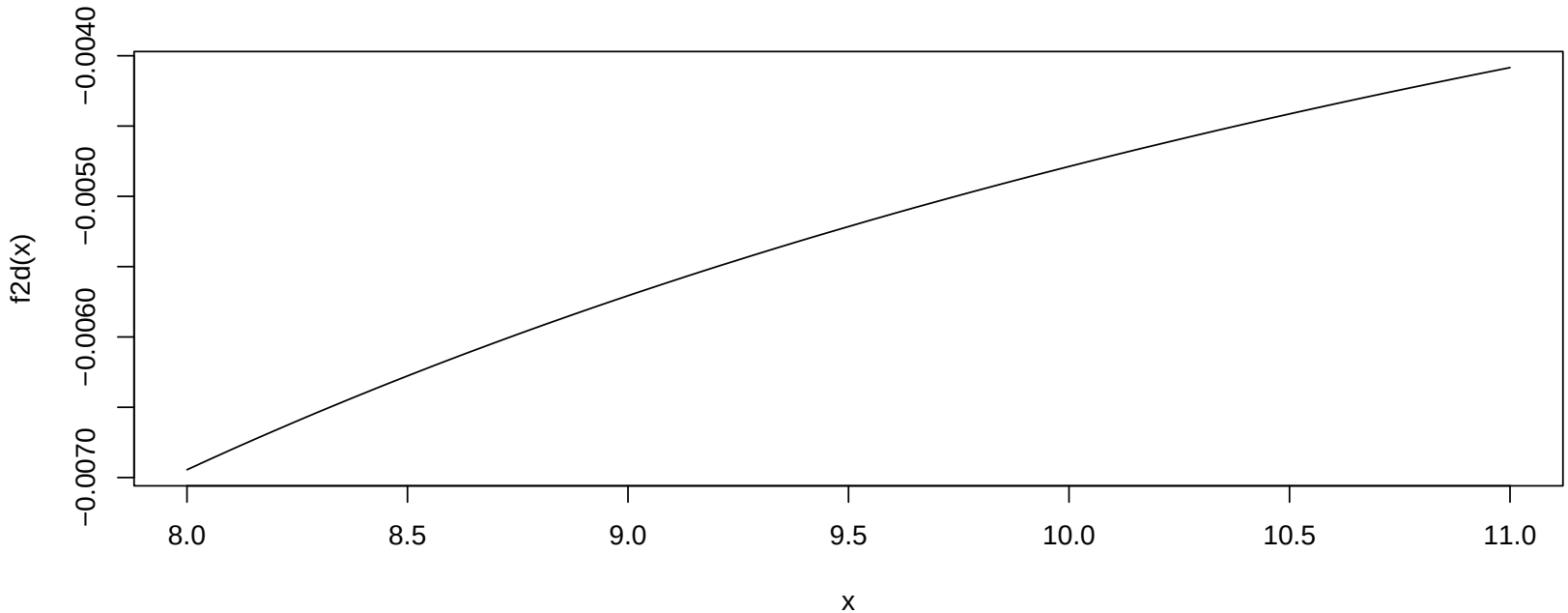
We do not know the exact value of c, so we find an upper bound on $R_1(11)$ by determining the maximum value of f'' on the interval $(8, 11)$. Since $f''(x) = \frac{-2}{9x^{5/3}}$, the largest value for $|f''(x)|$ on that interval occurs at $x = 8$. Using the fact that

```
f2d=function(x){-2/(9*x^(5/3))}
f2d(c(8,11))
```

```
[1] -0.006944444 -0.004084445
```

Example 1.6

```
f2d=function(x){-2/(9*x^(5/3))}
curve(f2d,8,11)
```




```
R1=(-0.004084445)/2*(11-8)^2; R1
```

```
[1] -0.01838
```

Example 1.6

Thus,

$$R_1(11) \leq \frac{-0.004084445}{2!}(11-8)^2 \approx -0.01838.$$

$$\sqrt[3]{11} \approx p_1(11) - R_1(11) = 2 + \frac{1}{12}(11-8) - 0.01838 = 2.23162$$

```
2+(1/12)*(11-8)-0.01838
```

```
[1] 2.23162
```

```
11^(1/3)
```

```
[1] 2.22398
```

Example 1.6

For $R_2(11)$:

$$p_2 = f(8)+f'(8)(x-8)+\frac{f''(8)}{2!}(x-8)^2 = 2+\frac{1}{12}(x-8)-\frac{1}{288}(x-8)^2$$

```
2+(1/12)*(11-8)-(1/288)*(11-8)^2
```

```
[1] 2.21875
```

```
11^(1/3)
```

```
[1] 2.22398
```

```
f3d=function(x){10/(27*x^(8/3))}
f3d(c(8,11))
```

```
[1] 0.0014467593 0.0006188552
```

Example 1.6

For $R_2(11)$:

```
(0.0014467593)/6*(11-8)^3
```

```
[1] 0.006510417
```

```
2+(1/12)*(11-8)-(1/288)*(11-8)^2+0.006510417
```

```
[1] 2.22526
```

```
11^(1/3)
```

```
[1] 2.22398
```

$p_2 = f(8)+f'(8)(x-8)+\frac{f''(8)}{2!}(x-8)^2 = 2+\frac{1}{12}(x-8)-\frac{1}{288}(x-8)^2$
Since $f'''(x) = \frac{10}{27x^{8/3}}$, the maximum value of f''' on the interval $(8, 11)$ is $f'''(8) \approx 0.0014468$. Therefore, we have

$$R_2(11) \leq \frac{0.0014467593}{3!}(11-8)^3 \approx 0.006510417.$$

Taylor's Theorem with Remainder

If there exists a real number M such that $|f^{n+1}(x)| \leq M$ for all $x \in I$ then

$$|R_n(x)| \leq \frac{M}{(n+1)!}|x - x_0|^{n+1}.$$

Error Analysis

A real number x can have infinitely many digits. But a digital calculating device can hold only a finite number of digits and therefore, after a finite number of digits (depending on the capacity of the calculating device), the rest should be discarded in some sense.

In base 2, the digits are 0 and 1. As an example of the conversion of a base 2 number to decimal, we have

$$(11011.01)_2 = 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2}$$

```
1*2^4 +1*2^3 +0*2^2 +1*2^1 +1*2^0 +0*2^(-1) +1*2^(-2)
```

```
[1] 27.25
```

We study on this chapter how a real number can be represented on a computing device.

Floating-Point Form of Numbers

Floating-Point Form

Let x be a non-zero real number. An n-digit floating-point number in base β has the form

$$fl(x) = (-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e$$

where

$$(.d_1 d_2 \cdots d_n)_\beta = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^2}$$

is a β-fraction called the **mantissa** or **significand**, $s = 1$ or 0 is called the **sign** and e is an integer called the **exponent**. The number β is also called the **radix** and the point preceding d_1 is called the **radix point**.

Floating-Point Form of Numbers

When $\beta = 2$, the floating-point representation is called the binary floating-point representation and when $\beta = 10$, it is called the **decimal** floating-point representation.

Remark

Note that there are only finite number of digits in the floating-point representation, where as a real number can have infinite sequence of digits for instance, $1/3 = 0.33333\cdots$. Therefore, the representation is only an approximation to a real number.

Normalization

Normalization

A floating-point number is said to be normalized if either $d_1 \neq 0$ or $d_1 = d_2 = \cdots = d_n = 0$.
The following are examples of real numbers in the decimal floating point representation.

1. The real number $x = 6.238$ can be represented as $6.238 = (-1)^0 \times 0.6238 \times 10^1$, in which case, we have $s = 0$, $\beta = 10$, $e = 1$, $d_1 = 6$, $d_2 = 2$, $d_3 = 3$ and $d_4 = 8$. Note that this representation is the *normalized floating-point* representation.
2. The real number $x = -0.0014$ can be represented in the decimal float-point representation as $-0.0014 = (-1)^1 \times 0.0014 \times 10^1$, which is not in the normalized form. The normalized representation is $x = (-1)^1 \times 0.14 \times 10^{-2}$.

Overflow and Underflow

The exponent e is limited to a range

$$m < e < M.$$

During the calculation, if some computed number has an exponent $e > M$ then we say, the memory *overflow* or if $e < m$, we say the memory *underflow*.

Remarks

- In the case of overflow, computer will usually produce meaningless results or simply prints the symbol NaN, which means, the quantity obtained due to such a calculation is 'not a number'. The symbol ∞ is also denoted as NaN on some computers. The underflow is less serious because in this case, a computer will simply consider the number as zero.
- The floating-point representation of a number has two restrictions, one is the number of digits n in the mantissa and the second is the range of e. The number n is called the **precision** or **length** of the floating point representation.

In R, we have

```
0/0;10^1000000000000000000000000000000 #overflow
```

```
[1] NaN
```

```
[1] Inf
```

Example 1.7

The IEEE (Institute of Electrical and Electronics Engineers) standard for floating-point arithmetic (IEEE 754) is the most widely-used standard for floating-point computation, and is followed by many hardware (CPU and FPU), including intel processors, and software implementations. Many computer languages allow or require that some or all arithmetic be carried out using IEEE 754 formats and operations. The IEEE 754 floating-point representation for a binary number x is given by

$$fl(x) = (-1)^s \times (1.a_1 a_2 \cdots a_n)_2 \times 2^e,$$

where $a_1, \cdots, a_n$ are either 1 or 0. The IEEE 754 standard always uses binary operations.

Example 1.7

The **IEEE single precision** floating-point format uses 4 bytes (32 bits) to store a number. Out of these 32 bits, 24 are allocated for storing mantissa (one binary digit needs 1 bit storage space), 1 bit for s (sign) and remaining 8 bits for the exponent. The storage scheme is given by

$$|(sign)b_1|(exponent)b_2b_3\cdots b_9|(mantissa)b_{10}b_{11}\cdots b_{32}|$$

Note here that there are only 23 bits used for mantissa. This is because, the digit 1 before the binary point is not stored in the memory and will be inserted at the time of calculation.

Chopping and Rounding a Number

Any real number x can be represented exactly as

$$x = (-1)^s \times (.d_1 d_2 \cdots d_n d_{n+1} \cdots)_\beta \times \beta^e$$

with $d_1 \neq 0$ or $d_2 = d_3 = \cdots = 0$, $s = 0$ or 1, and e satisfies (Def. 1.5), for which the floating-point form (Def. 1.3) is an approximate representation. Let us denote this approximation of x by $fl(x)$. There are two ways to produce $fl(x)$ from x as defined below.

Chopped and Rounded Numbers

The **chopped** machine approximation of x is given by

$$fl(x) = (-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e$$

The **rounded** machine approximation of x is given by

$$fl(x) = \begin{cases} (-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e, & 0 \leq d_{n+1} < \frac{\beta}{2} \\ (-1)^s \times (.d_1 d_2 \cdots d_{n+1})_\beta \times \beta^e, & \frac{\beta}{2} \leq d_{n+1} < \beta \end{cases}$$

Suppose the number is 0.33378.
- Chopping the number to 3 decimal digits is 0.333.
- Rounding the number to 3 decimal digits is 0.334.

Different Type of Errors

The **error** in a computed quantity is defined as

$$\textbf{Error} = \textbf{True Value} - \textbf{Approximate Value}.$$

The **absolute error** is the absolute value of the error defined above. The **relative error** is a measure of the error in relation to the size of the true value as given by

$$\textbf{Relative Error} = \textbf{Error}/\textbf{True Value}.$$

The **percentage error** is defined as 100 times the relative error. The term **truncation error** is used to denote error, which result from approximating a smooth function by truncating its Taylor series representation to a finite number of terms.

Example 1.8

Suppose that the true value $x_T = e = 2.7182818$ and the approximate value $x_A = 2.7142857$. Then the Error(x_A) and the Relative Error (x_A):

```
Ex_A=2.7182818-2.7142857
Relx_A=(2.7182818-2.7142857)/2.7182818
Ex_A;Relx_A
```

```
[1] 0.0039961
```

```
[1] 0.001470083
```

Remark

Let x_A be the approximation of the real number x. Then
$E(x_A) = \text{Error}(x_A) = x - x_A$
$E_a(x_A) = \text{Absolute Error}(x_A) = |E(x_A)|$
$E_r(x_A) = \text{Relative Error}(x_A) = \frac{E(x_A)}{x}$

Example 1.9
If we denote the relative error in $fl(x)$ as $\epsilon > 0$, then we have

$$fl(x) = (1 - \epsilon)x,$$

where x is a real number.

Loss of Significant Digits

Significant Digits

In place of relative error, we often use the concept of significant digits.
If x_A is an approximation to x, then we say that x_A approximates x to r significant β-digits if

$$|x - x_A| \leq \frac{1}{2}\beta^{s-r+1}$$

with s the largest integer such that $\beta^s \leq |x|$.

Example 1.10

a. For $x = 1/3$, the approximate number $x_A = 0.333$ has three significant digits, since $|x - x_A| \approx .00033 < 0.005 = \frac{1}{2} \times 10^{-3}$. But $10^{-1} < 0.333\cdots = x$. Therefore, in this case $s = -1$ and hence $r = 3$.
$|x - x_A| = 0.000333\cdots \leq \frac{1}{2}\beta^{-1-3+1} = \frac{1}{2} \times 10^{-3} = 0.0005$.

b. For $x = 0.02138$, the approximate number $x_A = 0.02144$ has the absolute error $|x - x_A| = 0.00006 < 0.0005 = 0.5 \times 10^{-3}$. But $10^{-2} < 0.02138 = x$. Therefore, in this case $s = -2$ and therefore, the number x_A has only two significant digits, but not three, with respect to x.
$|x - x_A| = 0.00006 \leq \frac{1}{2}\beta^{-2-2+1} = \frac{1}{2} \times 10^{-3} = 0.0005$.

Example 1.11

Let us consider two real numbers

$$x = 7.6545428 = 0.76545428 \times 10^1,$$

$$y = 7.6544201 = 0.76544201 \times 10^1.$$

The numbers

$$x_A = 7.6545421 = 0.76545421 \times 10^1,$$

$$y_A = 7.6544200 = 0.76544200 \times 10^1.$$

are approximation to x and y, correct to six and seven significant digits, respectively. In eight-digit floating-point arithmetic,

$$z_A = x_A - y_A = 0.12210000 \times 10^{-3}$$

is the exact difference between x_A and y_A and

$$z = x - y = 0.12270000 \times 10^{-3}$$

is the exact difference between x and y.

Example 1.12

Therefore,

$$z - z_A = 0.6 \times 10^{-6} < 0.5 \times 10^{-5}$$

and hence z_A has only three significant digits with respect to z as $10^{-3} < z = 0.0001227$.
A simple calculation shows that

$$E_r(z_A) \approx 53736 \times E_r(x_A),$$

and similarly for y. Loss of significant digits is therefore dangerous if we wish to minimize the relative error. The loss of significant digits in the process of calculation is refered to as **Loss of Significance**.

Example 1.13

Consider the function $f(x) = x(\sqrt{x+1} - \sqrt{x})$. Using a 12-digit calculator, $f(100000) \approx 158.11348772$. Using R, we have

```
f=function(x){x*(sqrt(x+1)-sqrt(x))}
f(100000)
```

```
[1] 158.1135
```

```
round(f(100000),3)
```

```
[1] 158.113
```

Propagated Error in Function Evaluation

Once an error is committed, it affects subsequent results as this error propagates through subsequent calculations.

Consider evaluating $f(x)$ at the approximate value x_A rather than at x. Then consider how well does $f(x_A)$ approximate $f(x)$. Using the mean-value theorem, we get

$$f(x) - f(x_A) = f'(\xi)(x - x_A),$$

where ξ is an unknown point between x and x_A. The relative error of $f(x)$ with respect to $f(x_A)$ is given by

$$E_r(f(x)) = \frac{f'(\xi)}{f(x)}(x - x_A) = \frac{f'(\xi)}{f(x)} x E_x(x).$$

Since x_A and x are assumed to be very close to each other and ξ lies between x and x_A, we make the approximation

$$f(x) - f(x_A) \approx f'(x)(x - x_A) \approx f'(x_A)(x - x_A).$$

Condition Number of a Function

Condition number of a function

The condition number of a function f at a point $x = c$ is given by

$$\left| \frac{f'(c)}{f(c)} c \right|.$$

Example 1.14

Consider the function $f(x) = \sqrt{x}$, for all $x \in [0, \infty)$. Then

$$f'(x) = \frac{1}{2\sqrt{x}}, \text{for all } x \in [0, \infty).$$

The condition number of f is

$$\left| \frac{f'(x)}{f(x)} x \right| = \left| \frac{\frac{1}{2\sqrt{x}}}{\sqrt{x}} x \right| = \frac{1}{2} \text{ for all } x \in [0, \infty).$$

From relative error of $f(x)$ with respect to $f(x_A)$ we see that taking square roots is a **well-conditioned** process since it actually reduces the relative error.

Example 1.15

Consider the function

$$f(x) = \frac{10}{1-x^2}, \text{ for all } x \in R.$$

Then $f'(x) = -\frac{20x}{(1-x^2)^2}$, so that

$$\left|\frac{f'(x)}{f(x)}x\right| = \left|\frac{-\frac{20x}{(1-x^2)^2}}{\frac{10}{1-x^2}}x\right| = \left|\frac{2x^2}{(1-x^2)}\right| = \frac{2x^2}{|1-x^2|}$$

and this number can be quite large for $|x|$ near 1. Thus, for x near 1 or -1, this function is **ill-conditioned**, as it magnifies the relative error.

Stability and Instability

Stability and Instability in Evaluating a Function

Suppose there are n steps to evaluate a function $f(x)$. Then the total process of evaluating this function is said to have **instability** if atleast one step is ill-conditioned. If all the steps are well-conditioned, then the process is said to be **stable**.

Example 1.16

Consider the function

$$f(x) = \sqrt{x+1} - \sqrt{x} \text{ for all } x \in [0, \infty)$$

```
f=function(x){sqrt(x+1)-sqrt(x)}
f(12345) #consider as true value
```

```
[1] 0.004500033
```

```
#If we calculate f(12345) in three digit rounding
round(f(12346),3)-round(f(12345),3)
```

```
[1] -0.001
```

```
#Relative error has 22% error
(abs(0.004-0.005)/f(12345))*100
```

```
[1] 22.22206
```

Convergence

Convergent definition in mathematics is a property (displayed by certain innumerable series and functions) of approaching a limit more and more explicitly as an argument (variable) of the function increases or decreases or as the number of terms of the series gets increased.

If the sequence of partial sums is a convergent sequence (i.e. its limit exists and is finite) then the series is also called **convergent** and in this case if $lim_{n\to\infty} s_n = s$ then $\sum_{i=1}^{\infty} a_i = s$.

Likewise, if the sequence of partial sums is a divergent sequence (i.e. its limit doesn't exist or is plus or minus infinity) then the series is also called **divergent**.

Example 1.17

Determine if the following series is convergent or divergent. If it converges determine its value.

$$\sum_{n=1}^{\infty} n$$

To determine if the series is convergent we first need to get our hands on a formula for the general term in the sequence of partial sums. This is a known series and its value can be shown to be,

$$s_n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Taking the limit,

$$lim_{n\to\infty} \frac{n(n+1)}{2} = \infty.$$

Therefore, the sequence of partial sums diverges to ∞ and so the series also diverges.

Example 1.18

Determine if the following series converges or diverges. If it converges determine its sum.

$$\sum_{n=2}^{\infty} \frac{1}{n^2-1}$$

The general formula for the partial sums is,

$$s_n = \sum_{i=2}^{n} \frac{1}{i^2-1} = \frac{3}{4} - \frac{1}{2n} - \frac{1}{2(n+1)}$$

Taking the limit,

$$lim_{n\to\infty} \left(\frac{3}{4} - \frac{1}{2n} - \frac{1}{2(n+1)} \right) = \frac{3}{4}$$

The sequence of partial sums converges and so the series converges also and its value is,

$$\sum_{n=2}^{\infty} \frac{1}{n^2-1} = \frac{3}{4}.$$

```
f=function(x)2*exp(x)
curve(f, -1, 1)
```

Problem Set 1

1. Prove that $lim_{x \to -1} 2x + 1 = -1$.
2. Determine all the numbers c which satisfy the conclusions of the Mean Value Theorem for the following function and graph using R with the point/s identified. $f(x) = x^3 - 4x^2 - 2x - 5$ on $[-10, 10]$.
3. Find the point c that satisfies the mean value theorem for integrals on the interval $[-1, 1]$. The function is $f(x) = 2e^x$.
4. Consider the function $f(x) = \cos(x/2)$.
   a. Find the fourth Taylor polynomial for f at $x = \pi$.
   b. Use the fourth Taylor polynomial to approximate $\cos(\pi/2)$.
   c. Use the fourth Taylor polynomial to bound the error.
5. If $fl(x)$ is the machine approximated number of a real number x and ϵ is the corresponding relative error, then show that $fl(x) = (1 - \epsilon)x$.

Problem Set 1

6. For the following numbers x and their corresponding approximations x_A, find the number of significant digits in x_A with respect to x and find the relative error.
   a. $x = 451.01, x_A = 451.023$
   b. $x = -0.04518, x_A = -0.045113$
   c. $x = 23.4604, x_A = 23.4213$
7. Find the condition number for the following functions
   a. $f(x) = 2x^2$
   b. $f(x) = 2\pi^x$
   c. $f(x) = 2b^x$
8. Determine if the following series converges or diverges. If it converges determine its sum.

$$\sum_{n=1}^{\infty} \frac{1}{2^n}$$

References

- Atkinson, K.E. (1989). An Introduction to Numerical Analysis. John Wiley and Sons, New York.
- Gerald, C.F. and Wheatly, P.O. (2004). Applied Numerical Analysis. Pearson Education, Inc.
- Kreyszig, H. (2011). Advanced Engineering Mathematics. John Wiley & Sons, Inc.
- Sastry, S.S. (2012). Introductory Methods of Numerical Analysis. Rajkamal Electric Press
- Bloomfield, V. A (2014). Using R for Numerical Analysis in Science and Engineering. Taylor & Francis Group, LLC

Thank You!