Constructor University Bremen Spring Semester 2024 –

Dr. D. Sinden

CTMS-MAT-13: Numerical Methods

Summary from 6 August 2025

[6 August 2025 – Version 0.1.12]

Contents

Recommended Reading

- J. F. Epperson “*An Introduction to Numerical Methods and Analysis*”, Wiley 2^{nd} Edition (2013).
- R. L. Burden and J. D. Faires “*Numerical Analysis*”, Brooks/Cole 9^{th} Edition (2011).

1 Taylor Series

Chapter abstract: This chapter is a foundational concept numerical methods. The Taylor series and the remainder theorem allow estimates of errors, as well as estimates of derivatives which shall be used in the solution of numerical solutions to differential equations.

The Taylor series, or the Taylor expansion of a function, is defined as

Definition 1: Taylor Series
For a function $f : \mathbb{R} \mapsto \mathbb{R}$ which is infinitely differentiable at a point c, the Taylor series of $f(c)$ is given by

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!} (x-c)^k$$

where $f^{(k)} = \dfrac{\mathrm{d}^k f}{\mathrm{d}x^k}$ is the k^{th} derivative.

This is a power series, which is convergent for some radius.

Theorem 1: Taylor's Theorem
For a function $f \in C^{n+1}([a,b])$, i.e. f is $(n+1)$-times continuously differentiable in the interval $[a,b]$, then for some c in the interval, the function can be written as

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}$$

for some value $\xi \in [a,b]$ where

$$\lim_{\xi \to c} \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1} = 0.$$

Example 1.
With $f(x) = \sin(x)$ around $c = 0$. Thus, as $f' = \cos(x)$, it can be shown that

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \dots$$

Note that in this example, only odd powers of x contribute to the expansion.

2 Errors

Chapter abstract: This chapter is a foundational concept numerical methods. The Taylor series and the remainder theorem allow estimates of errors, as well as estimates of derivatives which shall be used in the solution of numerical solutions to differential equations.

2.1 Errors

Definition 2: Absolute and Relative Errors

Let $\tilde{a}$ be an approximation to a, then the **absolute error** is given by

$$|\tilde{a} - a|.$$

If $|a| \neq 0$, the **relative error** may be given by

$$\left|\frac{\tilde{a} - a}{a}\right|.$$

The error bound is the magnitude of the admissible error.

Theorem 2:

For both addition and subtraction the bounds for the *absolute errors* are added.
In division and multiplication the bounds for the *relative errors* are added.

Definition 3: Linear Sensitivity to Uncertainties

If $y(x)$ is a smooth function, i.e. is differentiable, then $|y'|$ can be interpreted as the **linear sensitivity** of $y(x)$ to uncertainties in x.
For functions of several variables, i.e. $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then

$$|\Delta y| \leq \sum_{i=1}^{n} \left|\frac{\partial y}{\partial x_i}\right| |\Delta x_i|$$

where $|\Delta x_i| = |\tilde{x}_i - x_i|$ for an approximation $\tilde{x}_i$.

2.2 Number Representations

Definition 4: Base Representation
Every number $x \in \mathbb{N}_0$ can be written as a unique expansion with respect to base $b \in \mathbb{N} \backslash \{1\}$ as

$$(x)_b = a_0 b^0 + a_1 b^1 + \ldots + a_n b^n = \sum_{i=0}^{n} a_i b^i.$$

A number can be written in a nested form:

$$\begin{aligned} (x)_b &= a_0 b^0 + a_1 b^1 + \ldots a_n b^n \\ &= a_0 + b\left(a_1 + b\left(a_2 + b\left(a_3 + \ldots + b a_n\right) \ldots\right)\right. \end{aligned}$$

with $a_i < \mathbb{N}_0$ and $a_i < b$, i.e. $a_i \in \{0, \ldots, b-1\}$.

For a real number, $x \in \mathbb{R}$, write

$$\begin{aligned} x &= \sum_{i=0}^{n} a_i b^i + \sum_{i=1}^{\infty} \alpha_i b^{-i} \\ &= a_n \ldots a_0 \cdot \alpha_1 \alpha_2 \ldots \end{aligned}$$

There are two issues: finding n maybe difficult and for large values of b^i division maybe computationally costly. Horner's algorithm seeks to overcome these issues.

Definition 5: Normalized Floating Point Representations
Normalized floating point representations with respect to some base b, store a number x as

$$x = 0 \cdot a_1 \ldots a_k \times b^n$$

where the $a_i \in \{0, 1, \ldots b-1\}$ are called the **digits**, k is the **precision** and n is the **exponent**. The set $a_1, \ldots, a_k$ is called the **mantissa**. Impose that $a_1 \neq 0$, it makes the representation unique.

Theorem 3: This
Let x and y be two normalized floating point numbers with $x > y > 0$ and base $b = 2$. If there exists integers p and $q \in \mathbb{N}_0$ such that

$$2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$$

then, at most p and at least q significant bits (i.e. significant figures written in base 2) are lost during subtraction.