SDG 4 QUALITY EDUCATION

# INTRODUCTION TO NUMERICAL METHODS

Dr. Gabriel Obed Fosu

Department of Mathematics
Kwame Nkrumah University of Science and Technology
ResearchGate ID: https://www.researchgate.net/profile/Gabriel_Fosu2
Google Scholar: https://scholar.google.com/citations?user=ZJfCMyQAAAAJ&hl=en&oi=ao

# Lecture Outline

# Section 1

# Introduction

# Definition

## Definition

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

1. Such problems originate generally from real-world applications of algebra, geometry, and calculus, and they involve variables which vary continuously.
2. The overall goal of the field is the design and analysis of techniques to give approximate but not exact solutions to problems that do may/may not have analytical solutions.

# Definition

1. When an analytical solution does not exist, numerical techniques are employed to solving hard problems.
2. There are times that an analytical solution to a problem may exist but might be cumbersome and time wasting to find such solution. In such instances, it is better to resort to numerical solutions.

**Most numerical solutions are iterative.**

This implies that a sequence of approximate solution is obtained by repeating a given procedure. These solutions are implemented using computer programs.

### Example

Analytically we can find the roots of

$$x^2 - 1 = 0$$

as

$$x = \pm 1$$

Since the highest degree of $x$ is two, then we have two roots.

### Example

Find all the five roots of the problem

$$x^5 + x - 1 = 0$$

It will be very difficult to find analytical (exact) solution to this problem, hence it would require that we find the approximate solution to the polynomial function using numerical methods.

# Method vs. Algorithm vs. Implementation [1]

## Method

A method in numerical analysis refers to a general mathematical strategy or theoretical framework used to approach and solve a class of numerical problems. Examples:

- The Bisection Method is a root-finding method that relies on the Intermediate Value Theorem.
- The Gauss-Seidel Method is an iterative method used to solve systems of linear equations
- The Simpson's Rule is a method for numerical integration that approximates the integral of a function using parabolic arcs, based on function values at equally spaced intervals.
- The Euler Method is a method for solving ordinary differential equations by approximating the derivative using finite differences.

# Method vs. Algorithm vs. Implementation [2]

## Algorithm

An algorithm is a step-by-step, detailed procedure for executing a numerical method. It defines the logical flow of operations needed to achieve the desired result.

1. We often use a pseudocode to describe algorithms. This provides a way of outlining the sequence of operations, including the form of the input data and the structure of the desired output.
2. Not all numerical methods yield satisfactory output for arbitrarily chosen inputs. Therefore, a stopping technique, typically based on error tolerance or a maximum number of iterations, is incorporated to prevent infinite loops and ensure termination.

# Method vs. Algorithm vs. Implementation [3]

## Implementation

An implementation is the actual programming of an algorithm in a specific computer language such as Python, MATLAB, C++, or Fortran. It transforms the abstract algorithm into executable code that can solve real-world numerical problems on a computer.

An implementation involves:

- Translating the logical steps of the algorithm into syntactically correct program instructions.
- Handling data input/output, memory allocation, and error management.
- Ensuring computational efficiency, numerical stability, and readability of code.

# Numerical focus [1]

People who employ numerical methods for solving problems have the following concerns:

## 1. Accuracy

Accuracy refers to how close the computed solution is to the true solution of the problem. It is a measure of the error in the solution. The higher the accuracy, the smaller the error. Very few of our computations will yield the exact answer to a problem, so we will have to understand how much error is made, and how to control (or even diminish) that error.

## 2. Efficiency

How fast and cheap (memory) can we compute a solution?
Does the algorithm take an excessive amount of computer time?
Efficiency refers to the amount of resources (time, memory, etc.) required to obtain a solution. It is a measure of the computational cost. The higher the efficiency, the less resources are required to obtain a solution.

# Numerical focus [2]

## 3. Completeness

A complete method is one that can find all possible solutions to a problem, without missing any.
This is an important consideration in numerical methods because some methods may miss certain solutions, leading to inaccurate results or incomplete solutions.

## 4. Convergence

Convergence refers to the property that the algorithm produces a sequence of approximations that become increasingly close to the true solution as the number of iterations increases.
Convergence is a critical property of numerical methods, as it ensures that the algorithm will produce an accurate solution if sufficient iterations are performed.
The numerical analyst prefers algorithms that has faster rate of convergence.

# Numerical focus [3]

## 5. Stability

1. Does the method produce similar results for similar data?
2. One criterion we will impose on an algorithm whenever possible is that small changes in the initial data produce correspondingly small changes in the final results. An algorithm that satisfies this property is called stable;
3. otherwise the method is unstable, and unstable methods tend to produce unreliable results.
4. Some algorithms are stable only for certain choices of initial data. These are called conditionally stable.

On the other hand, robustness refers to the ability of a numerical algorithm to produce accurate and reliable results even in the presence of large perturbations or errors in the input data.

# Well Posed Problem

A well-posed problem is a problem that satisfies three key properties: existence, uniqueness, and stability. These properties were introduced by the mathematician Jacques Hadamard. It provides a framework for assessing the reliability and validity of mathematical models and solutions.

1. Existence: A well-posed problem must have at least one solution. In other words, there should be some solution that satisfies the conditions of the problem.
2. Uniqueness: The solution to the problem should be unique. There should not be multiple solutions that satisfy the given conditions. If there are multiple solutions, the problem is considered ill-posed in terms of uniqueness.
3. Stability (or Continuous Dependence on Data): The solution should depend continuously on the initial conditions or input data.

## Ill-posed problem

A problem that lacks any of these three properties is considered ill-posed. Ill-posed problems can be problematic because they may not have solutions, may have multiple solutions, or may exhibit instability in their solutions.

Section 2

# Accuracy and Precision

# Accuracy and Precision

- In numerical computations, real-world quantities must be represented using a finite number of digits.
- Computers cannot store or manipulate infinitely precise numbers; instead, values are approximated using decimal or floating-point representations.
- The reliability of numerical results depends not only on the correctness of the algorithm but also on how accurately numbers are represented, rounded, and processed.
- This section introduces fundamental concepts such as significant digits, rounding and chopping, and error measurement — all of which form the basis for understanding the accuracy of numerical methods.

# Significant Digits

## Definition

There are digits beginning with the leftmost non-zero digit and ending with the rightmost correct digit, including final zeros that are exact.

1. All non-zero digits are considered significant. For example 91 has two significant figures, likewise 123.45 has five significant figures.
2. Zeros appearing anywhere between two non-zero digits are significant. Example 101.1203 has seven significant figures.
3. Leading zeros are not significant. Example 0.00053 has two significant figures.
4. Trailing zeros in a number containing a decimal point are significant. Example 12.2300 has six significant figures.
5. Again 0.0001200 has four significant figures (the zeros before 1 are not significant).
6. In addition 120.00 has five significant figures since it has three trailing zeroes.

# Accuracy

## Definition

Accurate to $n$ decimal places means that you can trust $n$ digits to the right of the decimal place.
Accurate to $n$ significant digits means that you can trust a total of $n$ digits as being meaningful beginning with the leftmost nonzero digit.

## Example

12.356 has three decimal places but five significant figures.

# Rounding and Chopping

## Definition

We say that a number $x$ is chopped to $n$ digits or figures when all digits that follow the $nth$ digit are discarded and none of the remaining $n$ digits are changed.

Conversely, $x$ is rounded to $n$ digits or figures when $x$ is replaced by an $n$-digit number that approximates $x$ with minimum error.

## Example

The results of rounding some three-decimal numbers to two digits are

$$0.217 \approx 0.22, \qquad 0.365 \approx 0.36, \qquad 0.475 \approx 0.48, \qquad 0.592 \approx 0.59, \tag{1}$$

while chopping them gives

$$0.217 \approx 0.21, \qquad 0.365 \approx 0.36, \qquad 0.475 \approx 0.47, \qquad 0.592 \approx 0.59. \tag{2}$$

# Errors

Suppose that $\alpha$ and $\beta$ are two numbers, of which one is regarded as an approximation to the other.

## Definition

1. The error of $\beta$ as an approximation to $\alpha$ is
$$\text{Error} = \text{exact value - approximate value} = \alpha - \beta \tag{3}$$
2. The absolute error of $\beta$ as an approximation to $\alpha$ is
$$\text{absolute error} = |\text{exact value - approximate value}| = |\alpha - \beta| \tag{4}$$
3. The relative error of $\beta$ as an approximation to $\alpha$ is
$$\text{relative error} = \frac{|\text{exact value - approximate value}|}{|\text{exact value}|} = \frac{|\alpha - \beta|}{|\alpha|} \tag{5}$$

Absolute error and relative error are two important measures of the accuracy of an approximation or numerical solution.

## Example

If $x = 0.00347$ is rounded to $\tilde{x} = 0.0035$, what is its number of significant digits. Again find the absolute error, and relative error. Interpret the results.

### Solution

$\tilde{x} = 0.0035$ has two significant digits.

$$\begin{aligned} \text{absolute error} &= |\text{exact value - approximate value}| && (6)\\ &= |0.00347 - 0.0035| && (7)\\ &= |-0.00003| && (8)\\ &= 0.00003 && (9)\end{aligned}$$

$$\text{relative error} = \frac{|\text{exact value - approximate value}|}{|\text{exact value}|} = \frac{0.00003}{|0.00347|} = 0.008646 \qquad (10)$$

# Advantages and Disadvantages of Absolute Error

## Advantages

1. Absolute error is a simple and straightforward measure of the magnitude of the error.
2. It is easy to understand and interpret.
3. It is useful when the magnitude of the quantity being measured is important.

## Disadvantages

1. Absolute error does not take into account the scale of the quantity being measured. In other words, a small absolute error for a large quantity may still be significant, while a large absolute error for a small quantity may be negligible.
2. Absolute error is not a normalized measure, so it is not useful for comparing errors across different quantities or situations.

# Advantages and Disadvantages of Relative Error

## Advantages

1. Relative error takes into account the scale of the quantity being measured, making it a more normalized measure.
2. It is useful for comparing errors across different quantities or situations.
3. It can be expressed as a percentage, which makes it easy to understand and compare.

## Disadvantages

1. Relative error can be more difficult to calculate and interpret than absolute error.
2. It may not be useful when the magnitude of the quantity being measured is important, as small relative errors may be insignificant for large quantities.

Section 3

# Review of Calculus

# Review of Calculus

Numerical computation frequently involves approximating functions, solving equations, and analyzing rates of change.

A solid understanding of calculus, especially, Taylor series is essential for:

- Deriving numerical schemes (e.g., finite differences, Newton-Raphson).
- Estimating and controlling error.
- Understanding convergence and accuracy in algorithms.

# Differentiable Functions

## Definition

Assume that $f(x)$ is defined on an open interval containing $x_0$. Then $f$ is said to be differentiable at $x_0$ if

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{11}$$

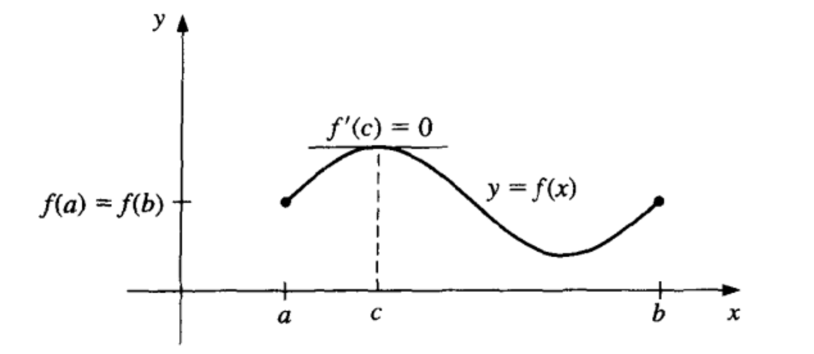
exists. When this limit exists, it is denoted by $f'(x_0)$ called the derivative of $f$ at $x_0$.

An equivalent way to express this limit is to use the $h$-increment notation:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0) \tag{12}$$

# Rolle's Theorem

Suppose $f \in [a,\ b]$ and $f$ is differentiable on $(a, b)$. If $f(a) = f(b)$, then a number $c$ in (a, b) exists with $f'(c) = 0$.



# Taylor Series

## Definition

Taylor series is a representation of a function as an infinite sum of terms that are calculated from the values of the function's derivatives at a specific point.

## Taylor series of a function $f$ at a point $c$

$$f(x) \approx f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 + \cdots \tag{13}$$

$$\approx \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!}(x-c)^n \tag{14}$$

$f', f'', \cdots f^{(n)}$ are the derivatives of the function $f$. Note that the values of these successive derivatives should exist at the point $c$.

That is any smooth function can be locally approximated by a polynomial function.

# Maclaurin Series

In the special case where $c = 0$, then the series is called the Maclaurin series given by:

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots \tag{15}$$

$$\approx \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n \tag{16}$$

## Example

Find the Taylor series expansion of $f(x) = \sin(x)$ about $c = 0$, and use the first three nonzero terms to approximate $\sin(0.5)$.

### Step 1: Compute derivatives at $c = 0$

$$
\begin{aligned}
f(x) &= \sin(x) && \Rightarrow f(0) = 0 \\
f'(x) &= \cos(x) && \Rightarrow f'(0) = 1 \\
f''(x) &= -\sin(x) && \Rightarrow f''(0) = 0 \\
f^{(3)}(x) &= -\cos(x) && \Rightarrow f^{(3)}(0) = -1 \\
f^{(4)}(x) &= \sin(x) && \Rightarrow f^{(4)}(0) = 0 \\
f^{(5)}(x) &= \cos(x) && \Rightarrow f^{(5)}(0) = 1
\end{aligned}
$$

## Step 2: Substitute into Taylor series formula at $c = 0$

$$
\begin{aligned}
\sin(x) &\approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \frac{f^{(5)}(0)}{5!}x^5 + \cdots \\
&= 0 + x + 0 - \frac{x^3}{6} + 0 + \frac{x^5}{120} + \cdots \\
&= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots
\end{aligned}
$$

## Step 3: Approximate $\sin(0.5)$ using first 3 nonzero terms

$$
\sin(0.5) \approx 0.5 - \frac{(0.5)^3}{6} + \frac{(0.5)^5}{120} = 0.5 - 0.020833 + 0.000260 = \boxed{0.479427}
$$

Actual value: $\sin(0.5) \approx 0.479426 \Rightarrow$ very accurate!

## Example

Find the Taylor series of the following function using the first four terms of the series.

$$f(x) = e^{2x}, \qquad \text{where } c = 1$$

**Solution**

i. For the Taylor series we obtain

$$f(x) = e^{2x} \qquad f(1) = e^2 \tag{17}$$

$$f'(x) = 2e^{2x} \qquad f'(1) = 2e^2 \tag{18}$$

$$f''(x) = 4e^{2x} \qquad f''(1) = 4e^2 \tag{19}$$

$$f'''(x) = 8e^{2x} \qquad f'''(1) = 8e^2 \tag{20}$$

$$
\begin{aligned}
f(x) &\approx f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 && (21)\\
&\approx f(1) + f'(1)(x-1) + \frac{f''(1)}{2!}(x-1)^2 + \frac{f'''(1)}{3!}(x-1)^3 && (22)\\
&= e^2 + 2e^2(x-1) + \frac{4e^2}{2}(x-1)^2 + \frac{8e^2}{6}(x-1)^3 && (23)\\
&= e^2\left[1 + 2x - 2 + 2x^2 - 4x + 2 + \frac{4}{3}(x^3 - 3x^2 + 3x - 1)\right] && (24)\\
&= e^2\left(\frac{7}{3} + 2x - 2x^2 + \frac{4}{3}x^3\right) && (25)
\end{aligned}
$$

Some standard Taylor series at $c = 0$ include:

1. $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$
2. $\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$
3. $\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$
4. $\cosh x = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots$
5. $\sinh x = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots$
6. $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots, \qquad |x| < 1$
7. $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \qquad -1 < x <\leq 1$

# Taylor's Theorem for $f(x)$

## Definition

If the function $f$ possesses continuous derivatives of order $0, 1, 2, \cdots, (n+1)$ in a closed interval $I = [a, b]$, then for any $c$ and $x$ in $I$,

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}(x-c)^k + E_{n+1} \tag{26}$$

where the error term (reminder) $E_{n+1}$ can be given in the form

$$E_{n+1} = \frac{f^{n+1}(\epsilon)}{(n+1)!}(x-c)^{n+1} \tag{27}$$

for some $\epsilon$ between $c$ and $x$ and depends on both.

In practical computations with Taylor series, it is usually necessary to truncate the series because it is not possible to carry out an infinite number of additions, and it is made possible by this theorem.

## Example

Approximate $\ln(1.2)$ using the Taylor polynomial of degree 2 centered at $c = 0$, and estimate the error using Taylor's Theorem.

### Step 1: Compute derivatives of $f(x) = \ln(1+x)$ at $c = 0$

$$
\begin{aligned}
f(x) &= \ln(1+x) & f(0) &= 0 \\
f'(x) &= \frac{1}{1+x} & f'(0) &= 1 \\
f''(x) &= -\frac{1}{(1+x)^2} & f''(0) &= -1 \\
f^{(3)}(x) &= \frac{2}{(1+x)^3} & f^{(3)}(0) &= 2
\end{aligned}
$$

## Step 2: Write the Taylor polynomial of degree 2 at $c = 0$

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 = 0 + x - \frac{x^2}{2}$$

$$\ln(1.2) \approx 1.2 - \frac{(1.2)^2}{2} = 1.2 - 0.72 = 0.48$$

## Step 3: Estimate the error using Taylor's Theorem

$$E_3 = \frac{f^{(3)}(\epsilon)}{3!}(x-0)^3 = \frac{2}{6(1+\epsilon)^3}(1.2)^3 \quad \text{for } \epsilon \in (0, 1.2)$$

$$\text{Max error (where epsilons is smallest)}: |E_3| \leq \frac{2}{6(1)^3}(1.728) = \frac{1.728}{3} = 0.576$$

Actual value: $\ln(1.2) \approx 0.182 \rightarrow$ Our 2-term estimate (0.48) has large error, confirmed by $E_3$

# Taylor's Theorem for $f(x+h)$

Taylor's theorem for $f(x)$ provides an approximation of a function around a specific point $c$, while the formulation for $f(x+h)$ offers an approximation of the same function, but with a shifted argument by a value $h$. This is a reformulation of the standard Taylor expansion and is particularly important in numerical methods.

## Definition

If the function $f$ possesses continuous derivatives of order $0, 1, 2, \cdots, (n+1)$ in a closed interval $I = [a, b]$, then for any $x$ in $I$,

$$f(x+h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + E_{n+1} \tag{28}$$

where $h$ is any value such that $x+h$ is in $I$ and where

$$E_{n+1} = \frac{f^{n+1}(\epsilon)}{(n+1)!} h^{n+1}, \qquad \text{for some } \epsilon \in [x,\ x+h]. \tag{29}$$

# Big O notation

- The error term $E_{n+1}$ in a Taylor series approximation depends on $h$ both through the explicit term $h^{n+1}$, and through the choice of the point $\epsilon$ .
- We use the notation

$$E_{n+1} = \mathcal{O}(h^{n+1}) \tag{30}$$

  to convey a qualitative relationship. It tells us how the error term $E_{n+1}$ behaves as $h$ approaches zero.
- This notation, often called big O notation, is a compact way of describing how one quantity grows or decreases relative to another. Here, $\mathcal{O}(h^{n+1})$ indicates that the error $E_{n+1}$ behaves similarly to $h^{n+1}$ as $h$ becomes very small.
- Equation (30) can also be written as $|E_{n+1}| \le C|h|^{n+1}$. It tells us that the magnitude of the error $|E_{n+1}|$ is always less than or equal to some constant $C$ times $|h|^{n+1}$.
- The constant $C$ represents a limit on how much $E_{n+1}$ can deviate from $h^{n+1}$. It remains fixed as $h$ approaches zero, providing a clear boundary for the error's behavior.

- As the interval size $h$ approaches zero, it means that we are making our approximation finer and finer. In other words, we're zooming in on the point of interest. This is a common practice in calculus and numerical analysis when we want to approximate functions with higher accuracy.
- The error term $E_{n+1}$ represents the difference between the actual value of the function and its Taylor series approximation truncated at $n$ terms. As $h$ gets smaller, $E_{n+1}$ decreases, meaning our approximation becomes more accurate.
- The statement suggests that the rate at which $E_{n+1}$ decreases is similar to the rate at which $h^{n+1}$ decreases. This means that the error diminishes at a similar pace as we refine our approximation by reducing $h$.
- For large values of $n$, this convergence is quite rapid. This indicates that with more terms in the Taylor series (larger $n$), our approximation becomes increasingly accurate as $h$ approaches zero, and the error $E_{n+1}$ diminishes rapidly.
- For example, if you have a numerical method with an error that is proportional to $\mathcal{O}(h^2)$, you might say that the method is $\mathcal{O}(h^2)$. This means that as you decrease the step size $h$ by a factor of $k$, the error decreases by a factor of $k^2$.

# Some Derived Equations from Equation(28)

## $n = 1$

$$f(x+h) \approx f(x) + f'(x)h + \frac{1}{2!}f''(\epsilon_2)h^2 \tag{31}$$

$$= f(x) + f'(x)h + \mathcal{O}(h^2) \tag{32}$$

## $n = 2$

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + \frac{1}{3!}f'''(\epsilon_3)h^3 \tag{33}$$

$$= f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + \mathcal{O}(h^3) \tag{34}$$

| Order | Error when $h = 1$ | Error when $h = 0.5$ |
|---|---|---|
| $\mathcal{O}(h^2)$ | 1 | 0.25 |
| $\mathcal{O}(h^3)$ | 1 | 0.125 |
| $\mathcal{O}(h^4)$ | 1 | 0.0625 |

Table 1: Effect of halving $h$ on different Taylor error orders

## Exercise

*Find the Taylor series and the Maclaurin series of the following function using the first four terms of the series.*

1. $$f(x) = \sin(2x), \qquad \text{where } c = \pi$$

2. $$f(x) = \cosh(3x), \qquad \text{where } c = 2$$

# END OF LECTURE