

Introduction to Scientific Computing

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Introduction



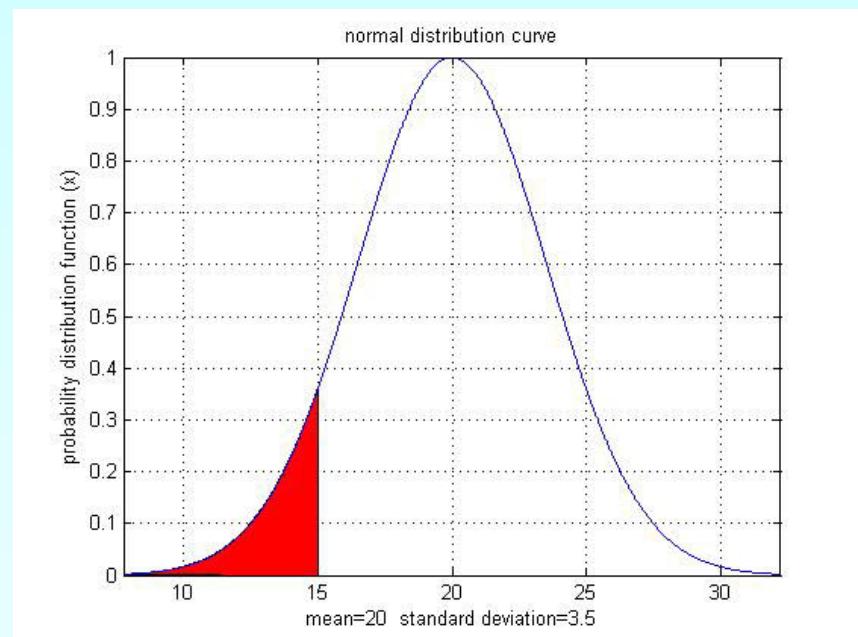
My advice

- *If you don't let a teacher know at what level you are by asking a question, or revealing your ignorance you will not learn or grow.*
- *You can't pretend for long, for you will eventually be found out. Admission of ignorance is often the first step in our education.*
 - Steven Covey—*Seven Habits of Highly Effective People*

Why use Numerical Methods?

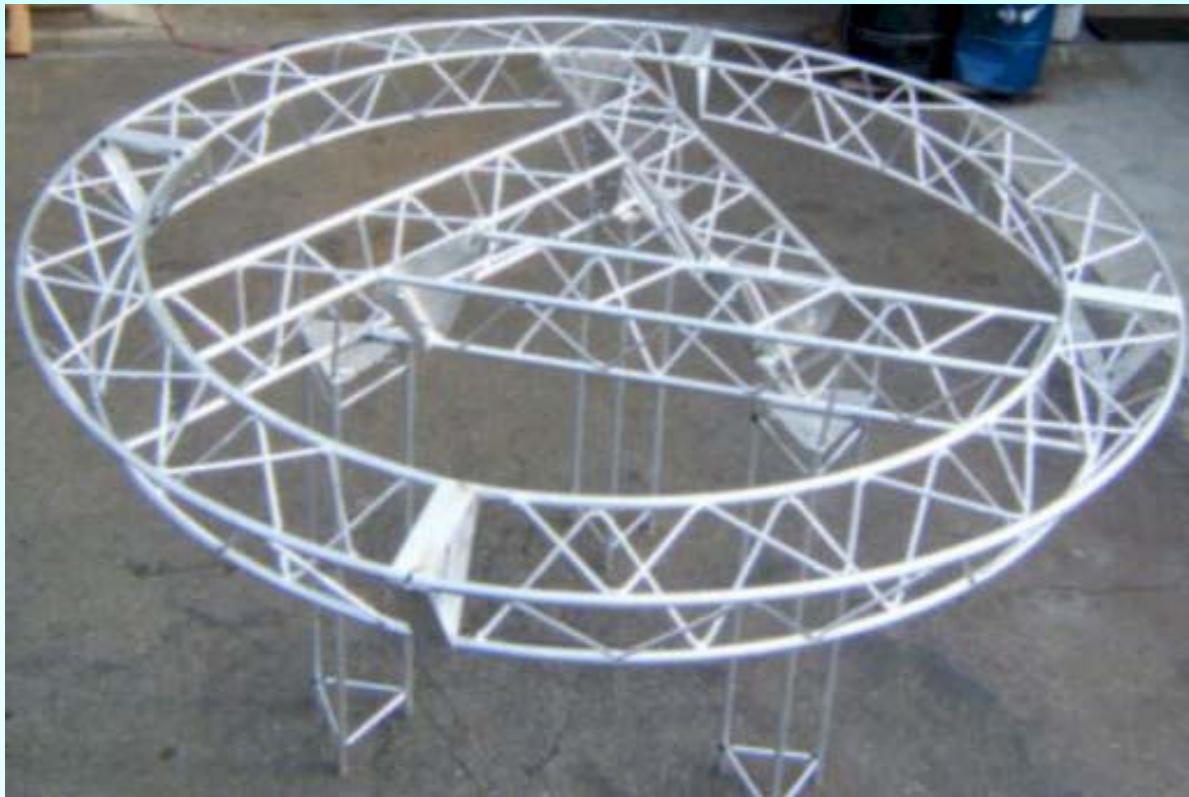
- To solve problems that cannot be solved exactly

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$



Why use Numerical Methods?

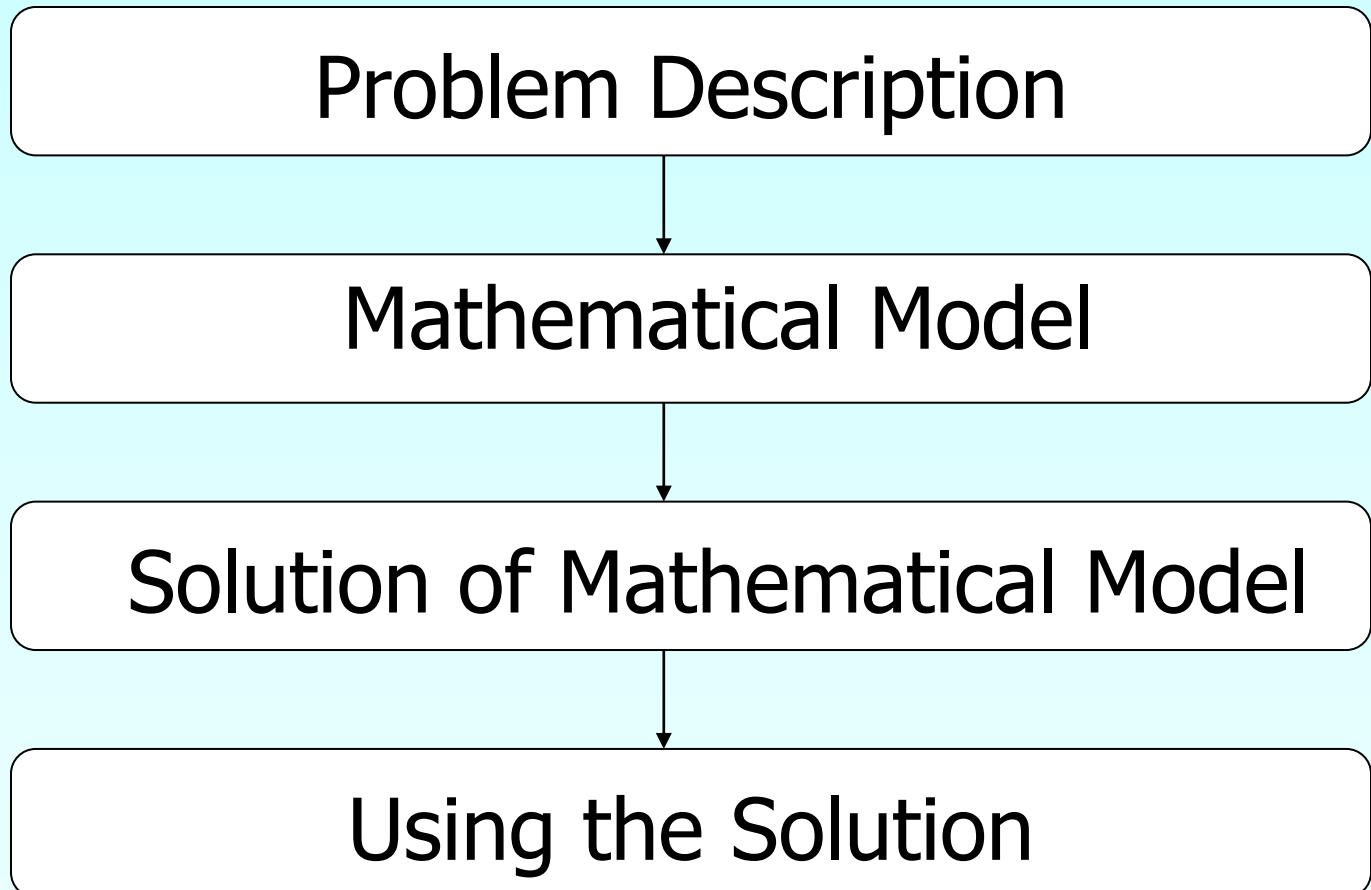
- To solve problems that are intractable!



Steps in Solving an Engineering Problem

<http://numericalmethods.eng.usf.edu>

How do we solve an engineering problem?



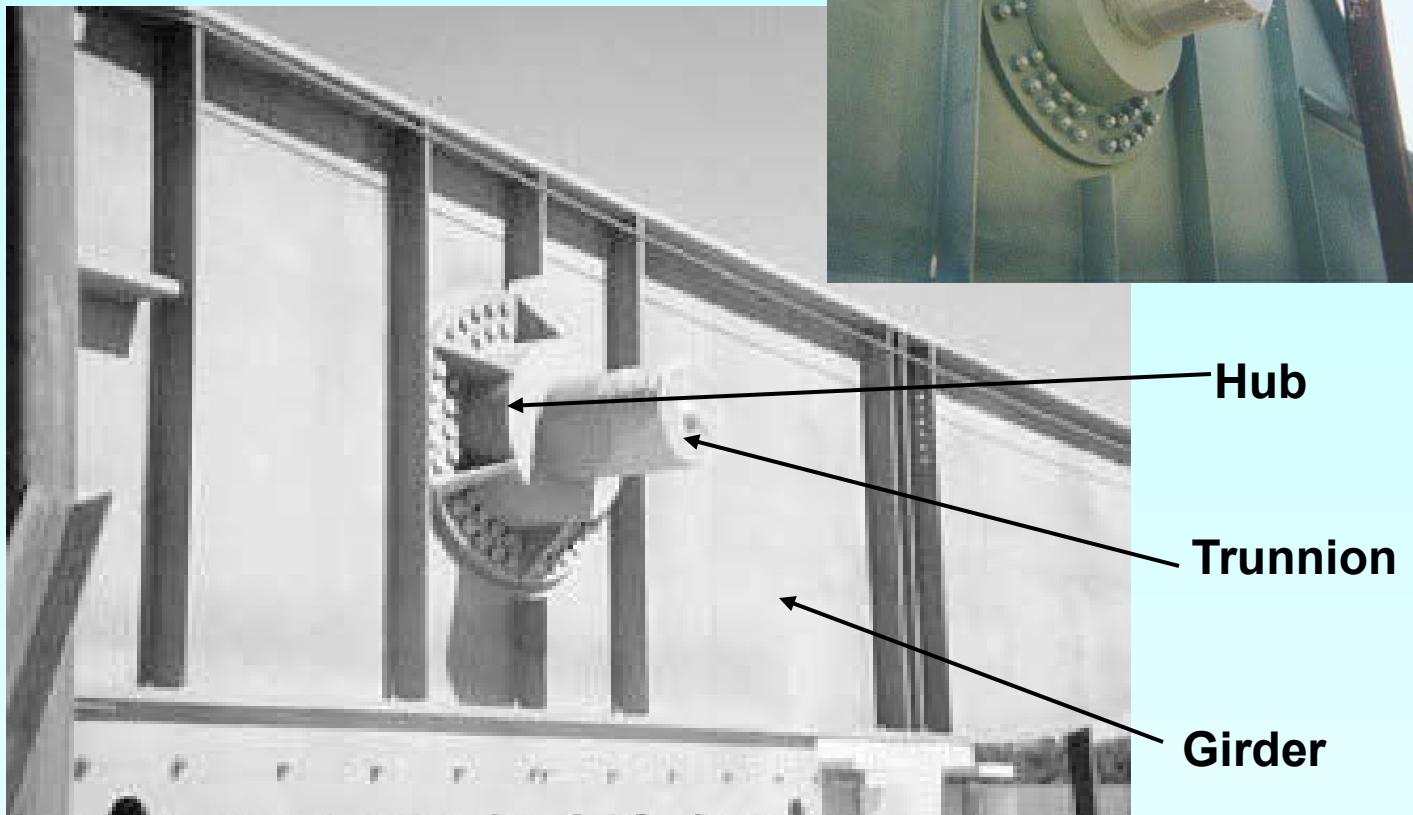
Example of Solving an Engineering Problem



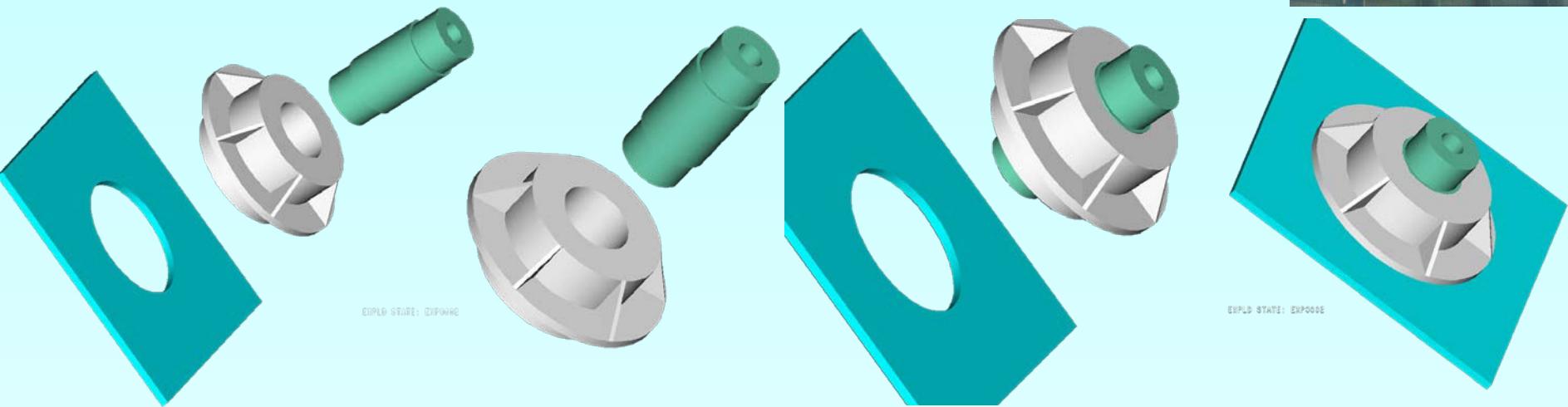
Bascule Bridge THG



Bascule Bridge THG

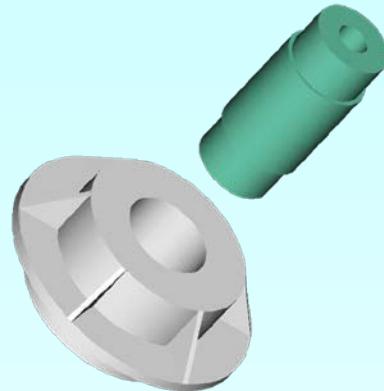


Trunnion-Hub-Girder Assembly Procedure



- Step1.** Trunnion immersed in dry-ice/alcohol
- Step2.** Trunnion warm-up in hub
- Step3.** Trunnion-Hub immersed in
dry-ice/alcohol
- Step4.** Trunnion-Hub warm-up into girder

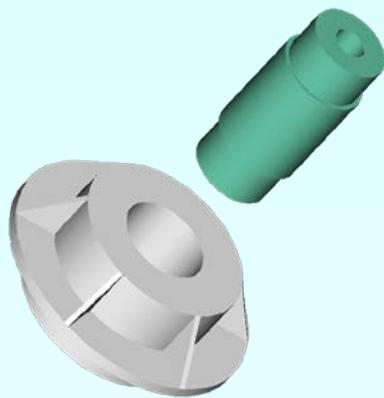
Problem



After Cooling, the Trunnion Got Stuck
in Hub

Why did it get stuck?

Magnitude of contraction needed in the trunnion was 0.015" or more. Did it contract enough?



Video of Assembly Process

Trunnion-Hub-Girder
Assembly of Bascule Bridges

University of South Florida
Tampa

Glen Besterfield (PI)
Autar Kaw (Co-PI)
Roger Crane (Co-PI)
Michael Denninger (Grad Student)
Badri Ratnam (Grad Student)
Sanjeev Nichani (Grad Student)

Trunnion-Hub-Girder
Assembly of Bascule Bridges

University of South Florida
Tampa

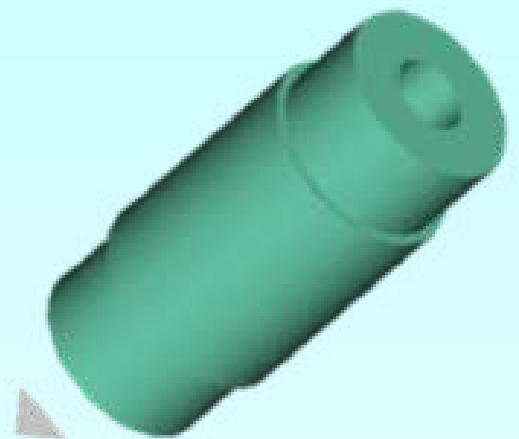
Glen Besterfield (PI)
Autar Kaw (Co-PI)
Roger Crane (Co-PI)
Michael Denninger (Grad Student)
Badri Ratnam (Grad Student)
Sanjeev Nichani (Grad Student)

Unplugged Version

VH1 Version

Consultant calculations

$$\Delta D = D \times \alpha \times \Delta T$$



$$D = 12.363"$$

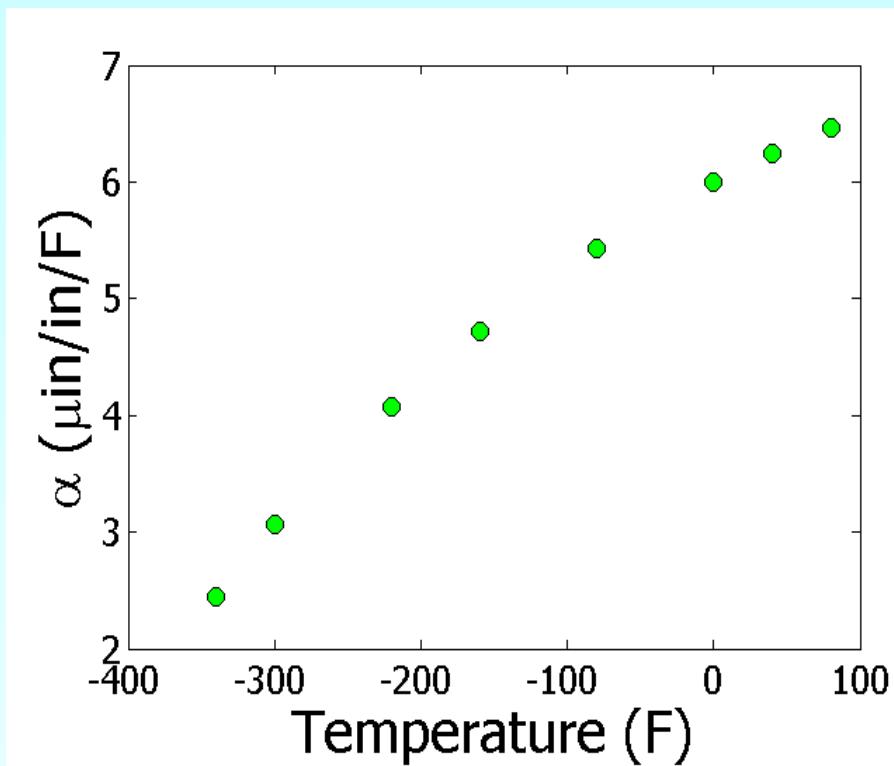
$$\alpha = 6.47 \times 10^{-6} \text{ in/in } /{}^\circ F$$

$$\Delta T = -108 - 80 = -188 {}^\circ F$$

$$\begin{aligned}\Delta D &= (12.363)(6.47 \times 10^{-6})(-188) \\ &= -0.01504"\end{aligned}$$

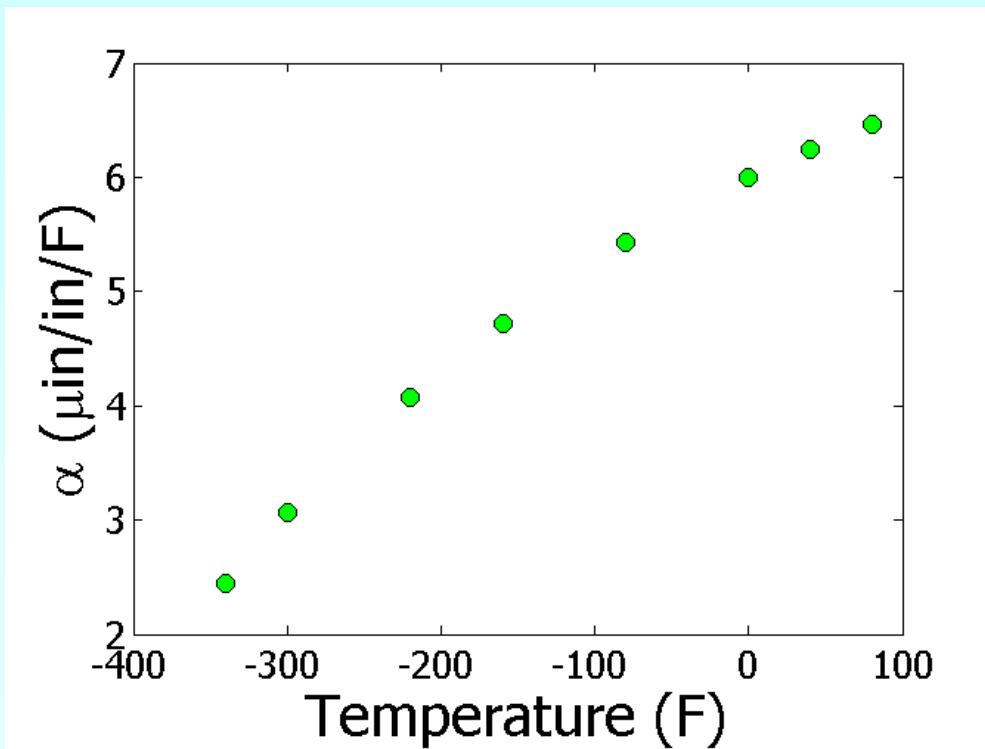
Is the formula used correct?

$$\Delta D = D \times \alpha \times \Delta T$$



$T(^{\circ}\text{F})$	α ($\mu\text{in/in}/^{\circ}\text{F}$)
-340	2.45
-300	3.07
-220	4.08
-160	4.72
-80	5.43
0	6.00
40	6.24
80	6.47

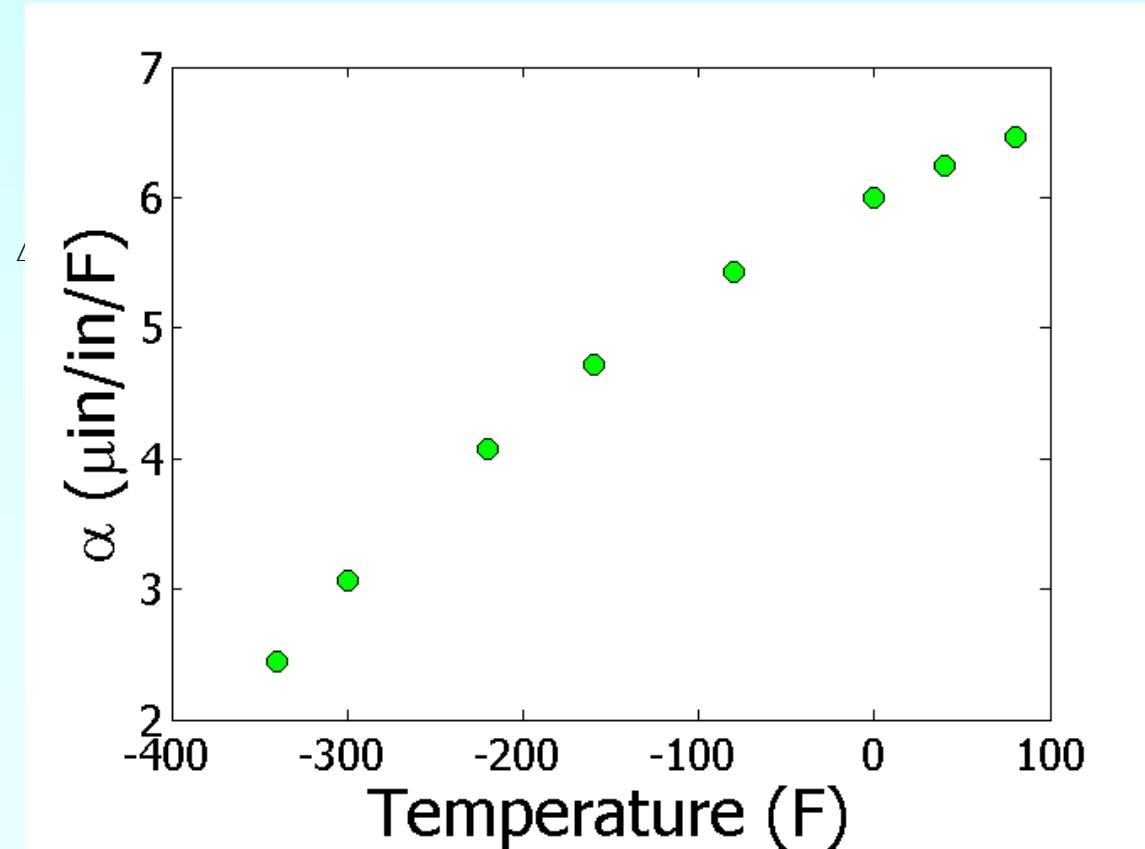
The Correct Model Would Account for Varying Thermal Expansion Coefficient



$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT$$

Can You Roughly Estimate the Contraction?

$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT \quad T_a = 80^\circ\text{F}; T_c = -108^\circ\text{F}; D = 12.363''$$



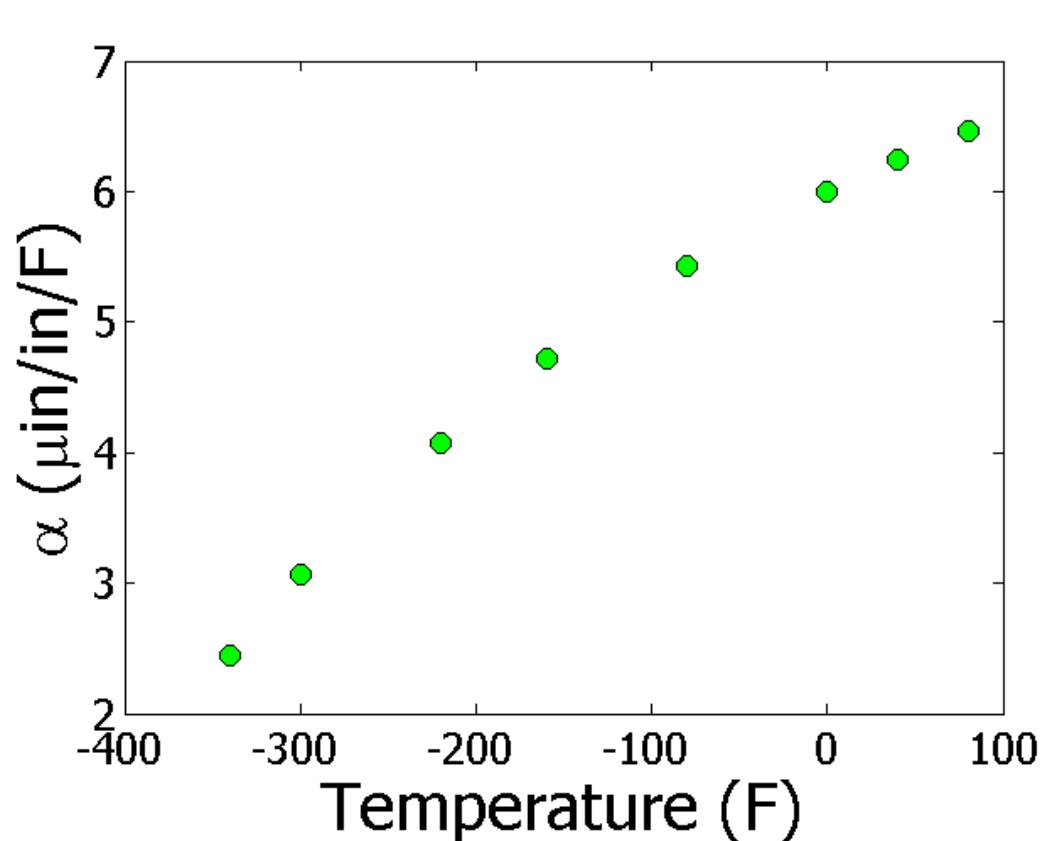
Can You Find a Better Estimate for the Contraction?

$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT$$

$$T_a = 80^{\circ}\text{F}$$

$$T_c = -108^{\circ}\text{F}$$

$$D = 12.363''$$



Estimating Contraction Accurately

Change in diameter (ΔD) by cooling it in dry ice/alcohol is given by

$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT$$

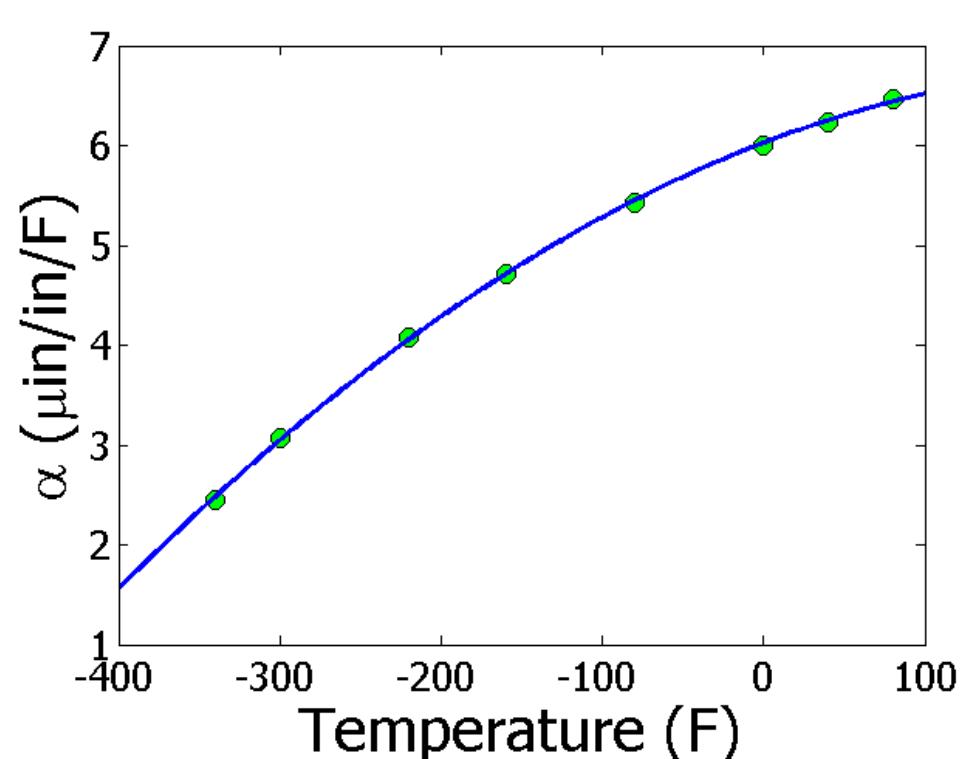
$$T_a = 80^{\circ}\text{F}$$

$$T_c = -108^{\circ}\text{F}$$

$$D = 12.363"$$

$$\alpha = -1.2278 \times 10^{-5} T^2 + 6.1946 \times 10^{-3} T + 6.0150$$

$$\Delta D = -0.0137"$$



So what is the solution to the problem?

One solution is to immerse the trunnion in liquid nitrogen which has a boiling point of -321°F as opposed to the dry-ice/alcohol temperature of -108°F.

$$\Delta D = -0.0244"$$

Revisiting steps to solve a problem

- 1) Problem Statement: Trunnion got stuck in the hub.
- 2) Modeling: Developed a new model

$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT$$

- 3) Solution: 1) Used trapezoidal rule OR b)
Used regression and integration.
- 4) Implementation: Cool the trunnion in liquid nitrogen.

THE END

<http://numericalmethods.eng.usf.edu>

<http://numericalmethods.eng.usf.edu>

23

Introduction to Numerical Methods

Mathematical Procedures

<http://numericalmethods.eng.usf.edu>

Mathematical Procedures

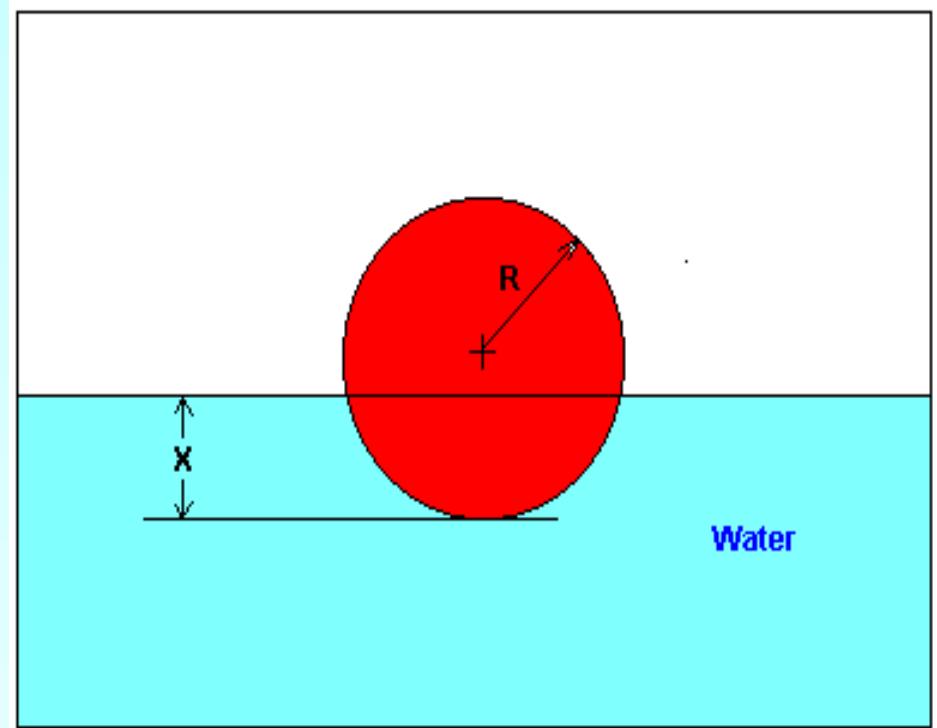
- Nonlinear Equations
- Differentiation
- Simultaneous Linear Equations
- Curve Fitting
 - Interpolation
 - Regression
- Integration
- Ordinary Differential Equations
- Other Advanced Mathematical Procedures:
 - Partial Differential Equations
 - Optimization
 - Fast Fourier Transforms

Nonlinear Equations

How much of the floating ball is under water?

Diameter=0.11m

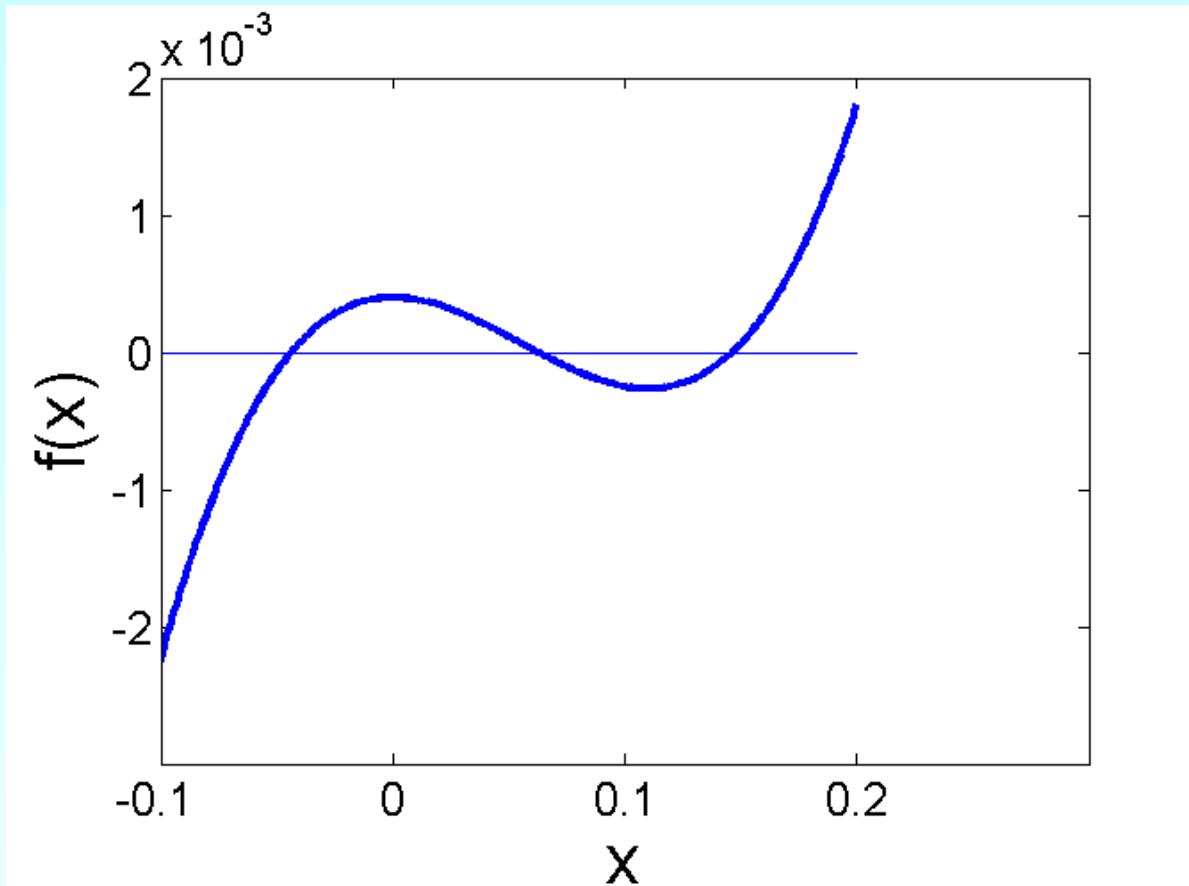
Specific Gravity=0.6



$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Nonlinear Equations

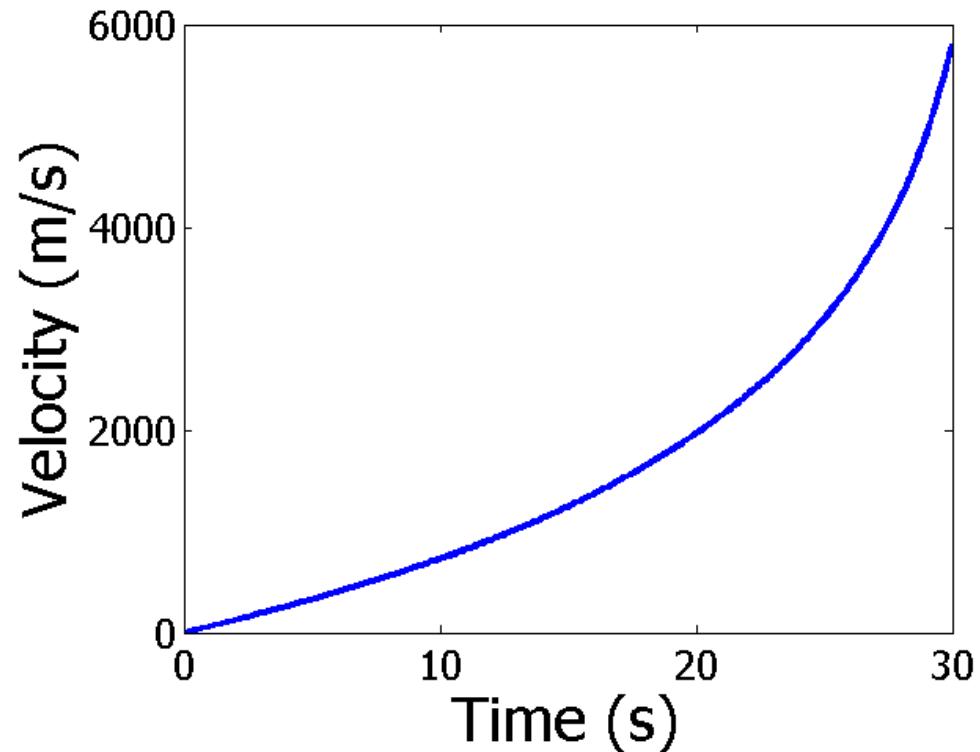
How much of the floating ball is under the water?



$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Differentiation

What is the acceleration at t=7 seconds?



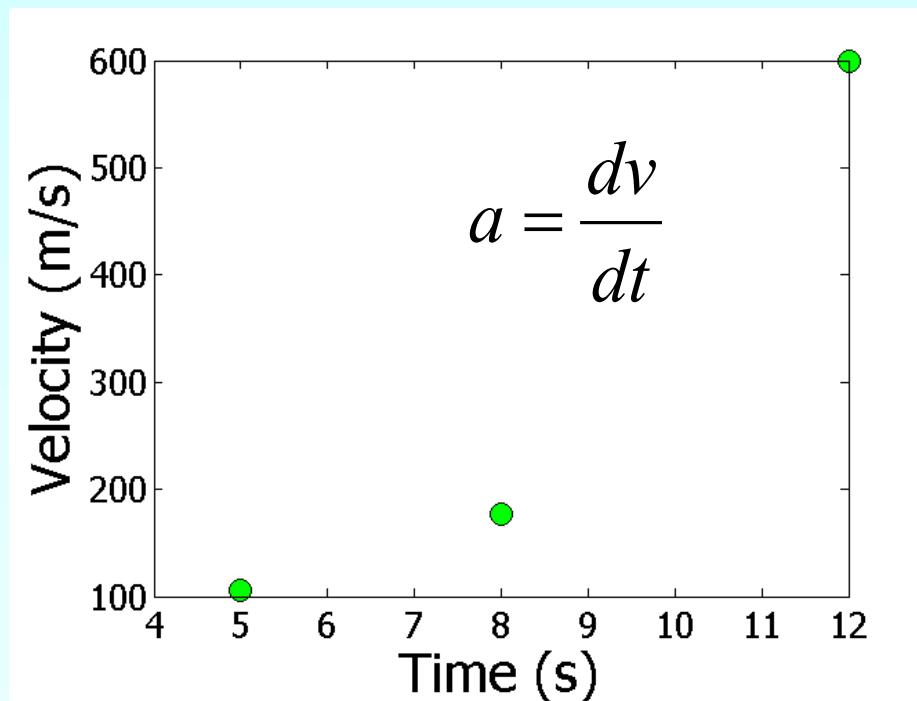
$$v(t) = 2200 \ln\left(\frac{16 \times 10^4}{16 \times 10^4 - 5000t}\right) - 9.8t$$

$$a = \frac{dv}{dt}$$

Differentiation

What is the acceleration at t=7 seconds?

Time (s)	5	8	12
Vel (m/s)	106	177	600



Simultaneous Linear Equations

Find the velocity profile, given

Time (s)	5	8	12
Vel (m/s)	106	177	600

$$v(t) = at^2 + bt + c, \quad 5 \leq t \leq 12$$

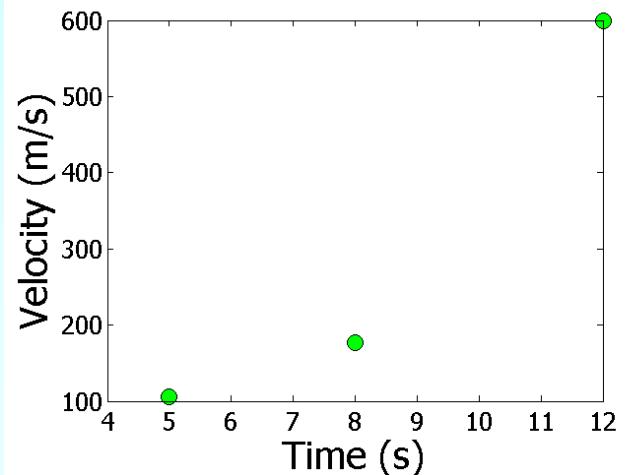


Three simultaneous linear equations

$$25a + 5b + c = 106$$

$$64a + 8b + c = 177$$

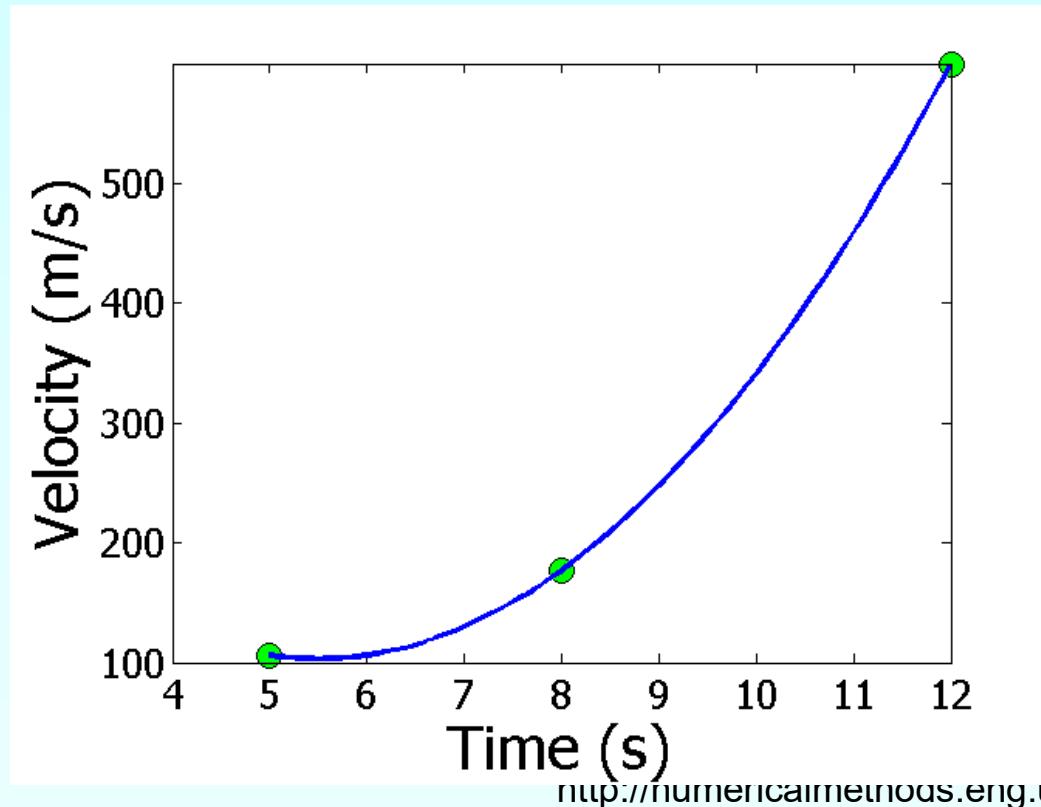
$$144a + 12b + c = 600$$



Interpolation

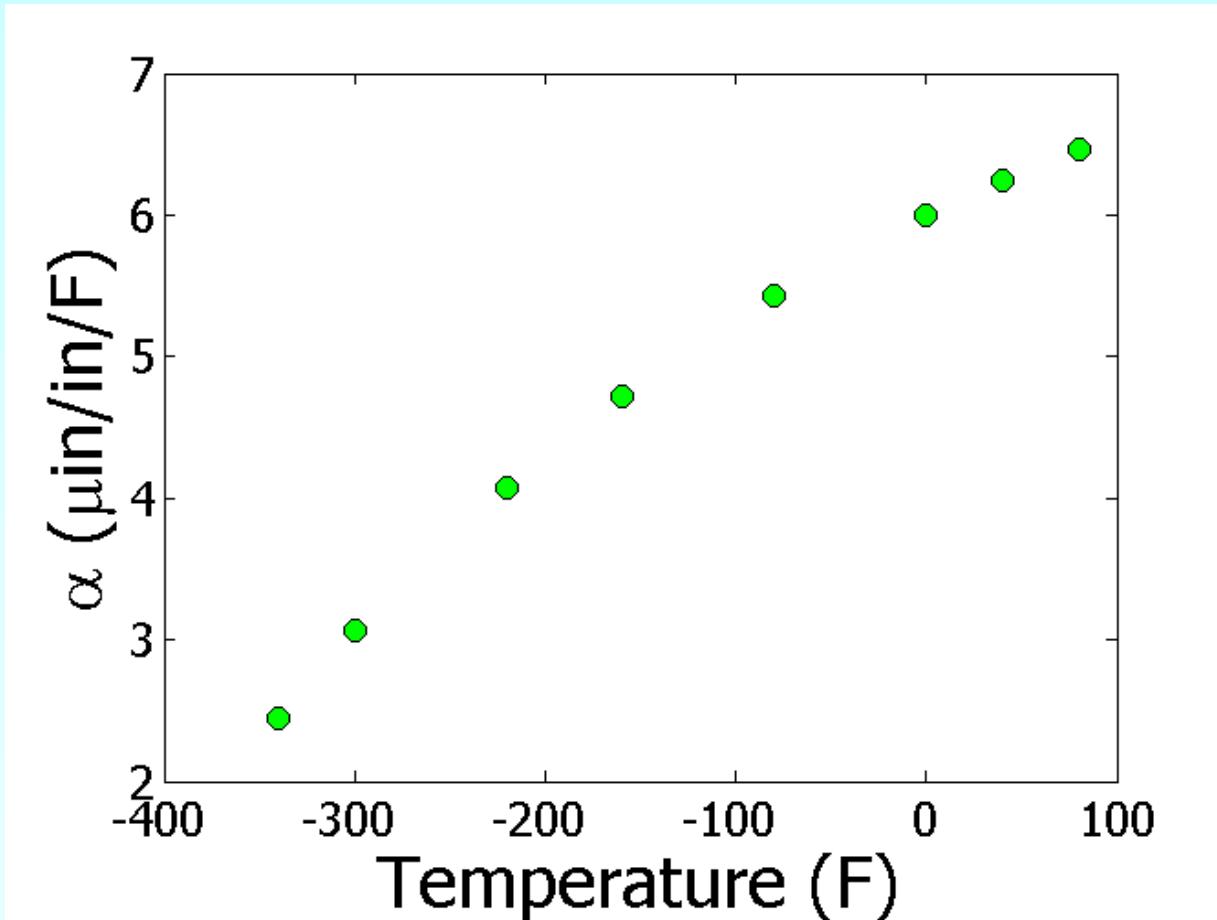
What is the velocity of the rocket at $t=7$ seconds?

Time (s)	5	8	12
Vel (m/s)	106	177	600

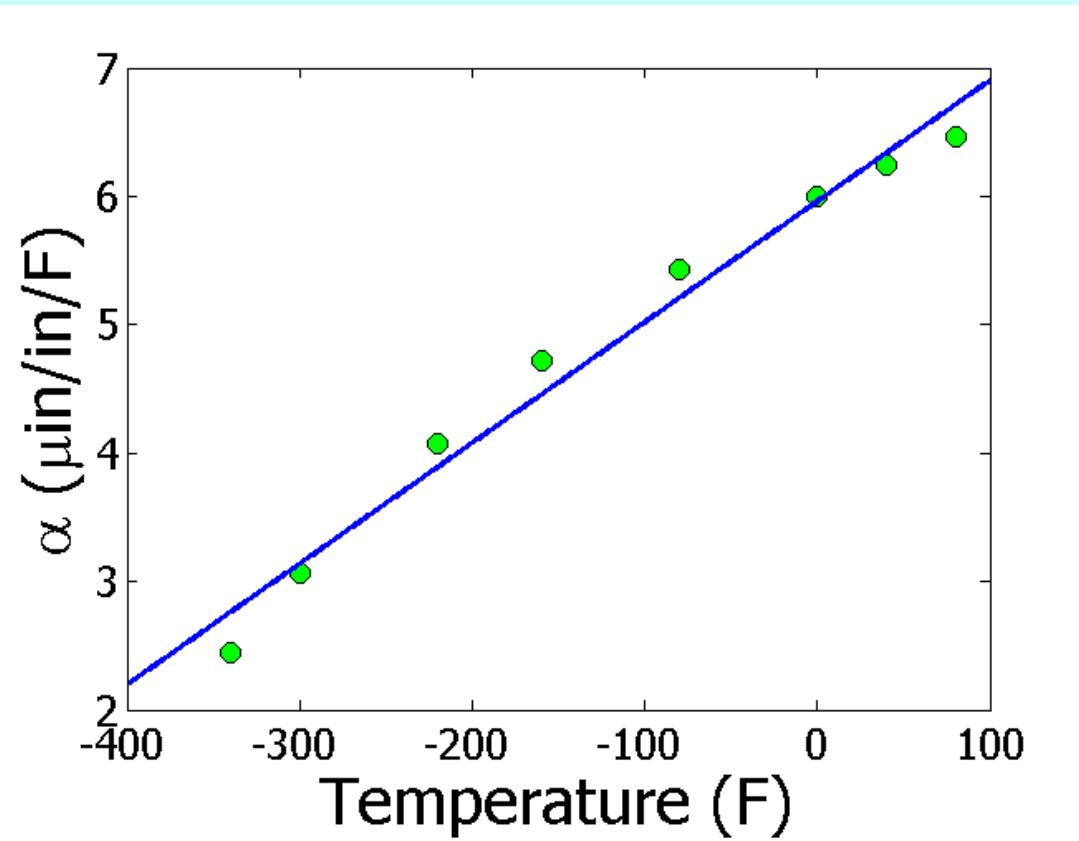


Regression

Thermal expansion coefficient data for cast steel



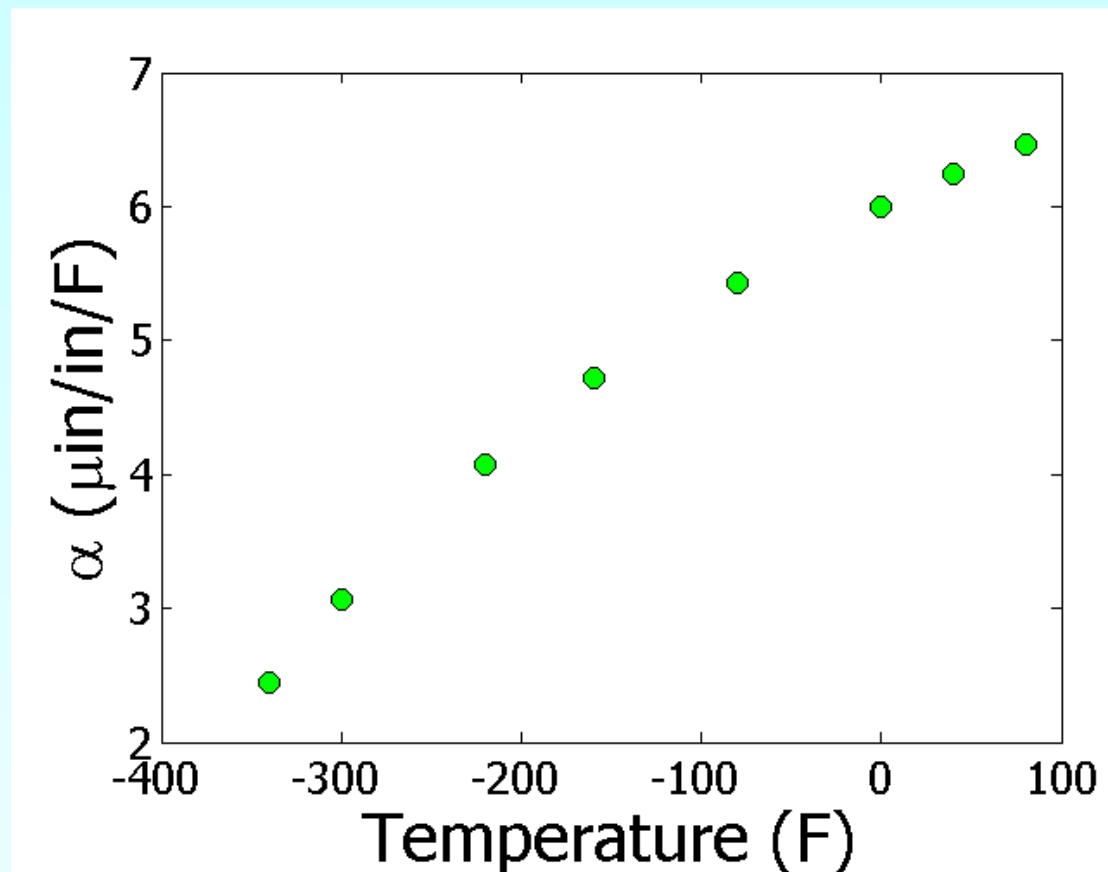
Regression (cont)



Integration

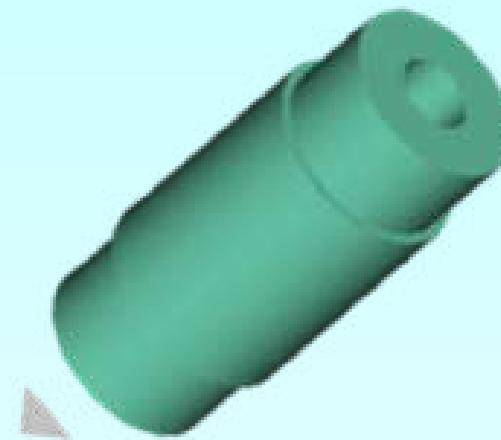
Finding the diametric contraction in a steel shaft when dipped in liquid nitrogen.

$$\Delta D = D \int_{T_{room}}^{T_{fluid}} \alpha \, dT$$



Ordinary Differential Equations

How long does it take a trunnion to cool down?



$$mc \frac{d\theta}{dt} = -hA(\theta - \theta_a), \quad \theta(0) = \theta_{room}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/introduction_numerical.html

THE END

<http://numericalmethods.eng.usf.edu>

Measuring Errors

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Measuring Errors

<http://numericalmethods.eng.usf.edu>

Why measure errors?

- 1) To determine the accuracy of numerical results.
- 2) To develop stopping criteria for iterative algorithms.

True Error

- Defined as the difference between the true value in a calculation and the approximate value found using a numerical method etc.

True Error = True Value – Approximate Value

Example—True Error

The derivative, $f'(x)$ of a function $f(x)$ can be approximated by the equation,

$$f'(x) \approx \frac{f(x + h) - f(x)}{h}$$

If $f(x) = 7e^{0.5x}$ and $h = 0.3$

- a) Find the approximate value of $f'(2)$
- b) True value of $f'(2)$
- c) True error for part (a)

Example (cont.)

Solution:

a) For $x = 2$ and $h = 0.3$

$$\begin{aligned}f'(2) &\approx \frac{f(2 + 0.3) - f(2)}{0.3} \\&= \frac{f(2.3) - f(2)}{0.3} \\&= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3} \\&= \frac{22.107 - 19.028}{0.3} = 10.263\end{aligned}$$

Example (cont.)

Solution:

- b) The exact value of $f'(2)$ can be found by using our knowledge of differential calculus.

$$f(x) = 7e^{0.5x}$$

$$\begin{aligned}f'(x) &= 7 \times 0.5 \times e^{0.5x} \\&= 3.5e^{0.5x}\end{aligned}$$

So the true value of $f'(2)$ is

$$\begin{aligned}f'(2) &= 3.5e^{0.5(2)} \\&= 9.5140\end{aligned}$$

True error is calculated as

$$\begin{aligned}E_t &= \text{True Value} - \text{Approximate Value} \\&= 9.5140 - 10.263 = -0.722\end{aligned}$$

Relative True Error

- Defined as the ratio between the true error, and the true value.

$$\text{Relative True Error} (\epsilon_t) = \frac{\text{True Error}}{\text{True Value}}$$

Example—Relative True Error

Following from the previous example for true error,
find the relative true error for $f(x) = 7e^{0.5x}$ at $f'(2)$
with $h = 0.3$

From the previous example,

$$E_t = -0.722$$

Relative True Error is defined as

$$\begin{aligned}\epsilon_t &= \frac{\text{True Error}}{\text{True Value}} \\ &= \frac{-0.722}{9.5140} = -0.075888\end{aligned}$$

as a percentage,

$$\epsilon_t = -0.075888 \times 100\% = -7.5888\%$$

Approximate Error

- What can be done if true values are not known or are very difficult to obtain?
- Approximate error is defined as the difference between the present approximation and the previous approximation.

Approximate Error (E_a) = Present Approximation – Previous Approximation

Example—Approximate Error

For $f(x) = 7e^{0.5x}$ at $x = 2$ find the following,

- $f'(2)$ using $h = 0.3$
- $f'(2)$ using $h = 0.15$
- approximate error for the value of $f'(2)$ for part b)

Solution:

- For $x = 2$ and $h = 0.3$

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

$$f'(2) \approx \frac{f(2+0.3) - f(2)}{0.3}$$

Example (cont.)

Solution: (cont.)

$$\begin{aligned} &= \frac{f(2.3) - f(2)}{0.3} \\ &= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3} \\ &= \frac{22.107 - 19.028}{0.3} = 10.263 \end{aligned}$$

b) For $x = 2$ and $h = 0.15$

$$\begin{aligned} f'(2) &\approx \frac{f(2 + 0.15) - f(2)}{0.15} \\ &= \frac{f(2.15) - f(2)}{0.15} \end{aligned}$$

Example (cont.)

Solution: (cont.)

$$\begin{aligned} &= \frac{7e^{0.5(2.15)} - 7e^{0.5(2)}}{0.15} \\ &= \frac{20.50 - 19.028}{0.15} = 9.8800 \end{aligned}$$

c) So the approximate error, E_a is

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8800 - 10.263 \\ &= -0.38300 \end{aligned}$$

Relative Approximate Error

- Defined as the ratio between the approximate error and the present approximation.

$$\text{Relative Approximate Error } (\epsilon_a) = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$

Example—Relative Approximate Error

For $f(x) = 7e^{0.5x}$ at $x = 2$, find the relative approximate error using values from $h = 0.3$ and $h = 0.15$

Solution:

From Example 3, the approximate value of $f'(2) = 10.263$ using $h = 0.3$ and $f'(2) = 9.8800$ using $h = 0.15$

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8800 - 10.263 \\ &= -0.38300 \end{aligned}$$

Example (cont.)

Solution: (cont.)

$$\epsilon_a = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$
$$= \frac{-0.38300}{9.8800} = -0.038765$$

as a percentage,

$$\epsilon_a = -0.038765 \times 100\% = -3.8765\%$$

Absolute relative approximate errors may also need to be calculated,

$$|\epsilon_a| = |-0.038765| = 0.038765 \text{ or } 3.8765\%$$

How is Absolute Relative Error used as a stopping criterion?

If $|\epsilon_a| \leq \epsilon_s$ where ϵ_s is a pre-specified tolerance, then no further iterations are necessary and the process is stopped.

If at least m significant digits are required to be correct in the final answer, then

$$|\epsilon_a| \leq 0.5 \times 10^{2-m}\%$$

Table of Values

For $f(x) = 7e^{0.5x}$ at $x = 2$ with varying step size, h

h	$f'(2)$	$ e_a $	m
0.3	10.263	N/A	0
0.15	9.8800	3.877%	1
0.10	9.7558	1.273%	1
0.01	9.5378	2.285%	1
0.001	9.5164	0.2249%	2

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/measuring_errors.html

THE END

<http://numericalmethods.eng.usf.edu>

Sources of Error

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Sources of Error

<http://numericalmethods.eng.usf.edu>

Two sources of numerical error

- 1) Round off error
- 2) Truncation error

Round-off Error

<http://numericalmethods.eng.usf.edu>

Round off Error

- Caused by representing a number approximately

$$\frac{1}{3} \cong 0.333333$$

$$\sqrt{2} \cong 1.4142\dots$$

Problems created by round off error

- 28 Americans were killed on February 25, 1991 by an Iraqi Scud missile in Dhahran, Saudi Arabia.
- The patriot defense system failed to track and intercept the Scud. Why?

Problem with Patriot missile



- Clock cycle of 1/10 seconds was represented in 24-bit fixed point register created an error of 9.5×10^{-8} seconds.
- The battery was on for 100 consecutive hours, thus causing an inaccuracy of

$$\begin{aligned} &= 9.5 \times 10^{-8} \frac{\text{s}}{0.1\text{s}} \times 100\text{hr} \times \frac{3600\text{s}}{1\text{hr}} \\ &= 0.342\text{s} \end{aligned}$$

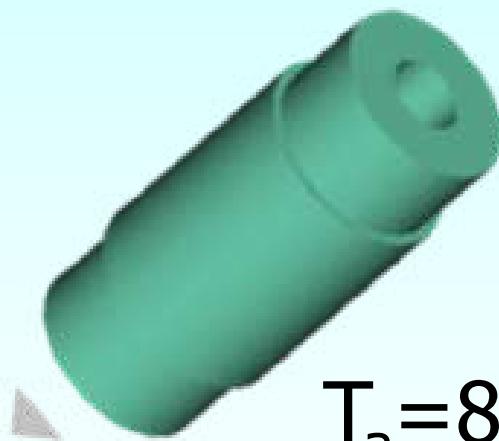
Problem (cont.)

- The shift calculated in the ranging system of the missile was 687 meters.
- The target was considered to be out of range at a distance greater than 137 meters.

Effect of Carrying Significant Digits in Calculations

<http://numericalmethods.eng.usf.edu>

Find the contraction in the diameter

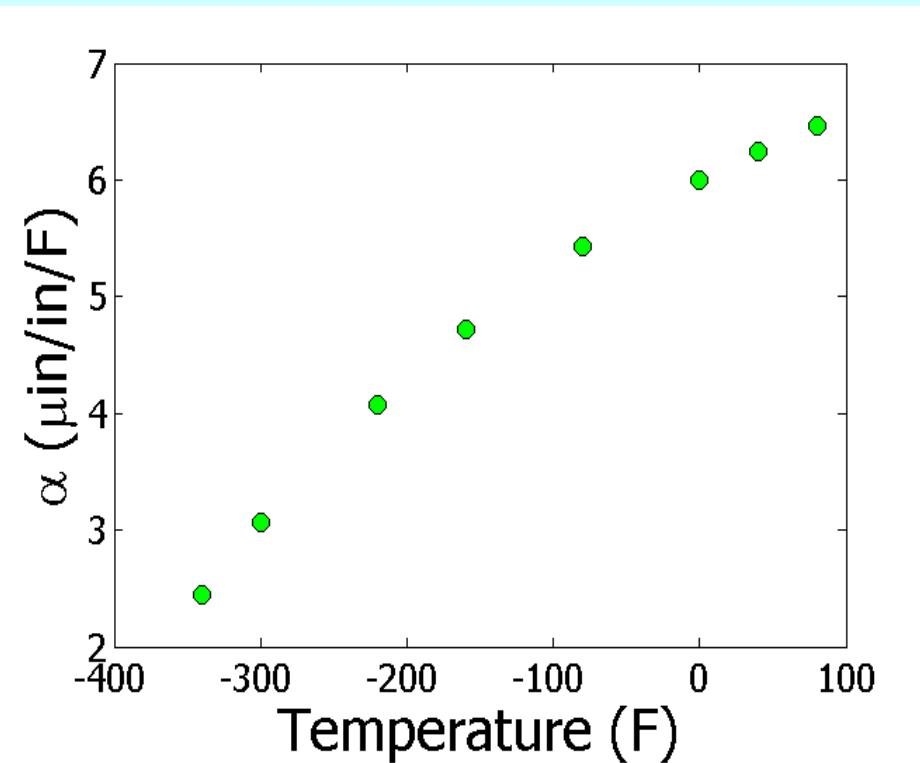


$$\Delta D = D \int_{T_a}^{T_c} \alpha(T) dT$$

$$T_a = 80^\circ\text{F}; T_c = -108^\circ\text{F}; D = 12.363''$$

$$\alpha = a_0 + a_1 T + a_2 T^2$$

Thermal Expansion Coefficient vs Temperature



T($^{\circ}\text{F}$)	α ($\mu\text{in/in}/^{\circ}\text{F}$)
-340	2.45
-300	3.07
-220	4.08
-160	4.72
-80	5.43
0	6.00
40	6.24
80	6.47

Regressing Data in Excel (general format)

$$a = -1E-05T^2 + 0.0062T + 6.0234$$

Observed and Predicted Values

$$\alpha = -1E-05T^2 + 0.0062T + 6.0234$$

T($^{\circ}$ F)	α (μ in/in/ $^{\circ}$ F) Given	α (μ in/in/ $^{\circ}$ F) Predicted
-340	2.45	2.76
-300	3.07	3.26
-220	4.08	4.18
-160	4.72	4.78
-80	5.43	5.46
0	6.00	6.02
40	6.24	6.26
80	6.47	6.46

Regressing Data in Excel (scientific format)

$$a = -1.2360E-05T^2 + 6.2714E-03T + 6.0234$$

Observed and Predicted Values

$$\alpha = -1.2360\text{E-}05T^2 + 6.2714\text{E-}03T + 6.0234$$

T($^{\circ}\text{F}$)	α ($\mu\text{in/in}/{}^{\circ}\text{F}$) Given	α ($\mu\text{in/in}/{}^{\circ}\text{F}$) Predicted
-340	2.45	2.46
-300	3.07	3.03
-220	4.08	4.05
-160	4.72	4.70
-80	5.43	5.44
0	6.00	6.02
40	6.24	6.25
80	6.47	6.45

Observed and Predicted Values

$$\alpha = -1.2360E-05T^2 + 6.2714E-03T + 6.0234$$

$$\alpha = -1E-05T^2 + 0.0062T + 6.0234$$

T(°F)	α ($\mu\text{in/in/}^\circ\text{F}$) Given	α ($\mu\text{in/in/}^\circ\text{F}$) Predicted	α ($\mu\text{in/in/}^\circ\text{F}$) Predicted
-340	2.45	2.46	2.76
-300	3.07	3.03	3.26
-220	4.08	4.05	4.18
-160	4.72	4.70	4.78
-80	5.43	5.44	5.46
0	6.00	6.02	6.02
40	6.24	6.25	6.26
80	6.47	6.45	6.46

THE END

Truncation Error

<http://numericalmethods.eng.usf.edu>

Truncation error

- Error caused by truncating or approximating a mathematical procedure.

Example of Truncation Error

Taking only a few terms of a Maclaurin series to approximate e^x

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

If only 3 terms are used,

$$\text{Truncation Error} = e^x - \left(1 + x + \frac{x^2}{2!} \right)$$

Another Example of Truncation Error

Using a finite Δx to approximate $f'(x)$

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

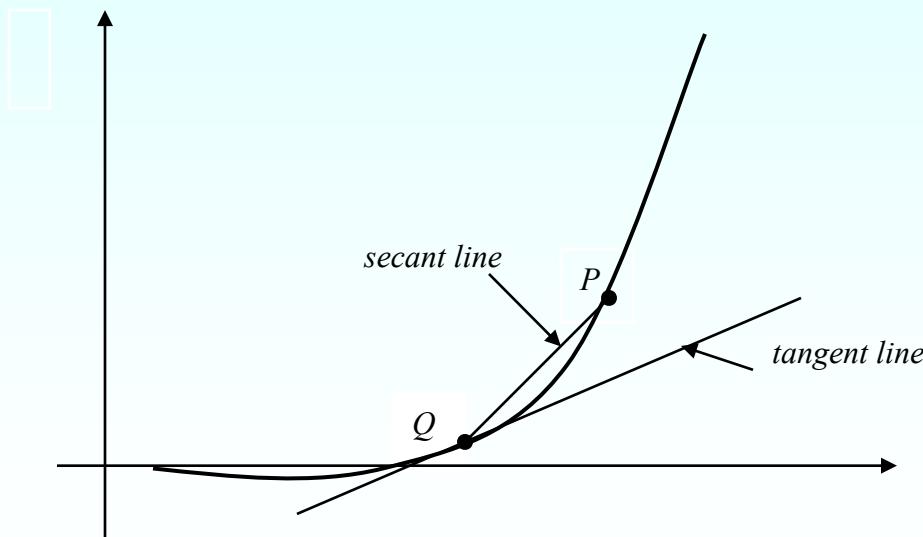
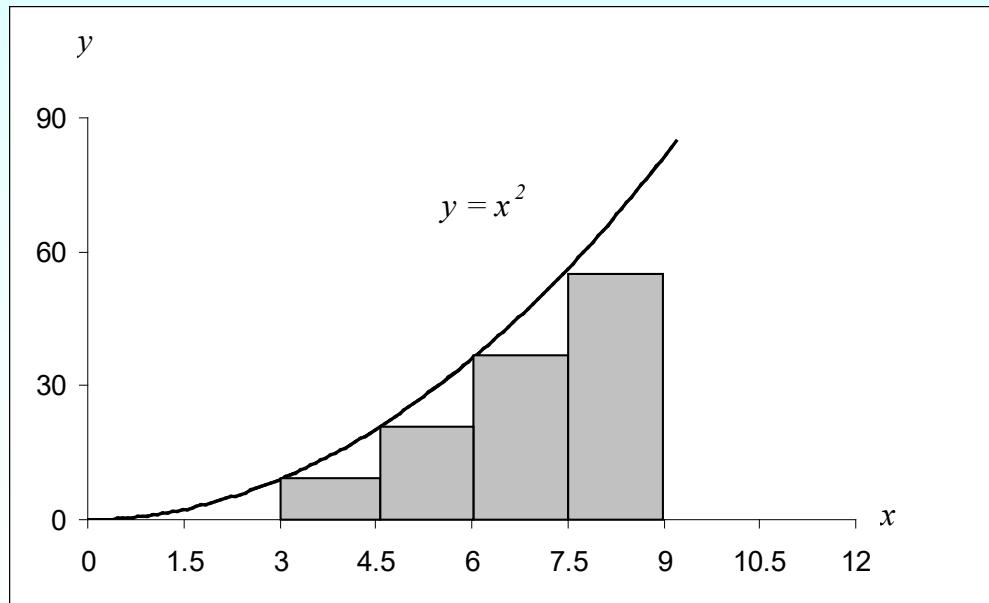


Figure 1. Approximate derivative using finite Δx

Another Example of Truncation Error

Using finite rectangles to approximate an integral.



Example 1 — Maclaurin series

Calculate the value of $e^{1.2}$ with an absolute relative approximate error of less than 1%.

$$e^{1.2} = 1 + 1.2 + \frac{1.2^2}{2!} + \frac{1.2^3}{3!} + \dots$$

n	$e^{1.2}$	E_a	$ E_a \%$
1	1	—	—
2	2.2	1.2	54.545
3	2.92	0.72	24.658
4	3.208	0.288	8.9776
5	3.2944	0.0864	2.6226
6	3.3151	0.020736	0.62550

6 terms are required. How many are required to get at least 1 significant digit correct in your answer?

Example 2 — Differentiation

Find $f'(3)$ for $f(x) = x^2$ using $f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$
and $\Delta x = 0.2$

$$\begin{aligned}f'(3) &= \frac{f(3 + 0.2) - f(3)}{0.2} \\&= \frac{f(3.2) - f(3)}{0.2} = \frac{3.2^2 - 3^2}{0.2} = \frac{10.24 - 9}{0.2} = \frac{1.24}{0.2} = 6.2\end{aligned}$$

The actual value is

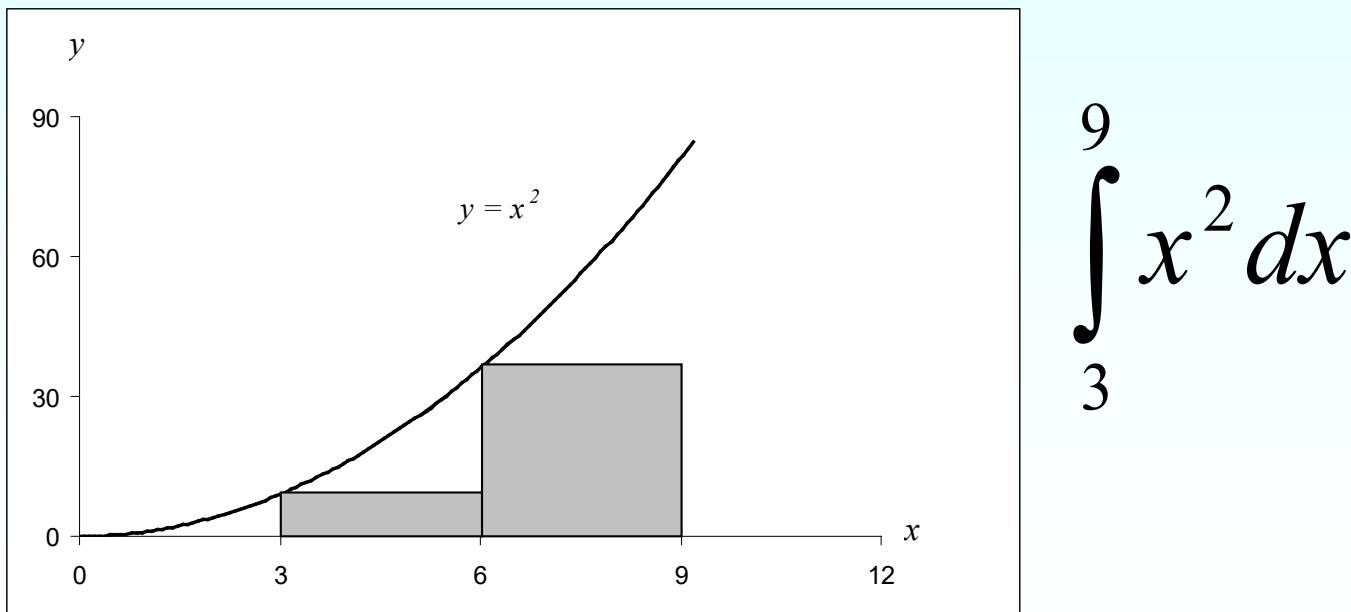
$$f'(x) = 2x, \quad f'(3) = 2 \times 3 = 6$$

Truncation error is then, $6 - 6.2 = -0.2$

Can you find the truncation error with $\Delta x = 0.1$ ²⁴

Example 3 — Integration

Use two rectangles of equal width to approximate the area under the curve for $f(x) = x^2$ over the interval [3,9]



Integration example (cont.)

Choosing a width of 3, we have

$$\begin{aligned}\int_3^9 x^2 dx &= (x^2) \Big|_{x=3} (6-3) + (x^2) \Big|_{x=6} (9-6) \\&= (3^2)3 + (6^2)3 \\&= 27 + 108 = 135\end{aligned}$$

Actual value is given by

$$\int_3^9 x^2 dx = \left[\frac{x^3}{3} \right]_3^9 = \left[\frac{9^3 - 3^3}{3} \right] = 234$$

Truncation error is then

$$234 - 135 = 99$$

Can you find the truncation error with 4 rectangles?

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/sources_of_error.html

THE END

<http://numericalmethods.eng.usf.edu>

Binary Representation

Major: All Engineering Majors

Authors: Autar Kaw, Matthew Emmons

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Binary Representation

<http://numericalmethods.eng.usf.edu>

How a Decimal Number is Represented

$$257.76 = 2 \times 10^2 + 5 \times 10^1 + 7 \times 10^0 + 7 \times 10^{-1} + 6 \times 10^{-2}$$

Base 2

$$(1011.0011)_2 = \left((1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0) + (0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4}) \right)_{10}$$
$$= 11.1875$$

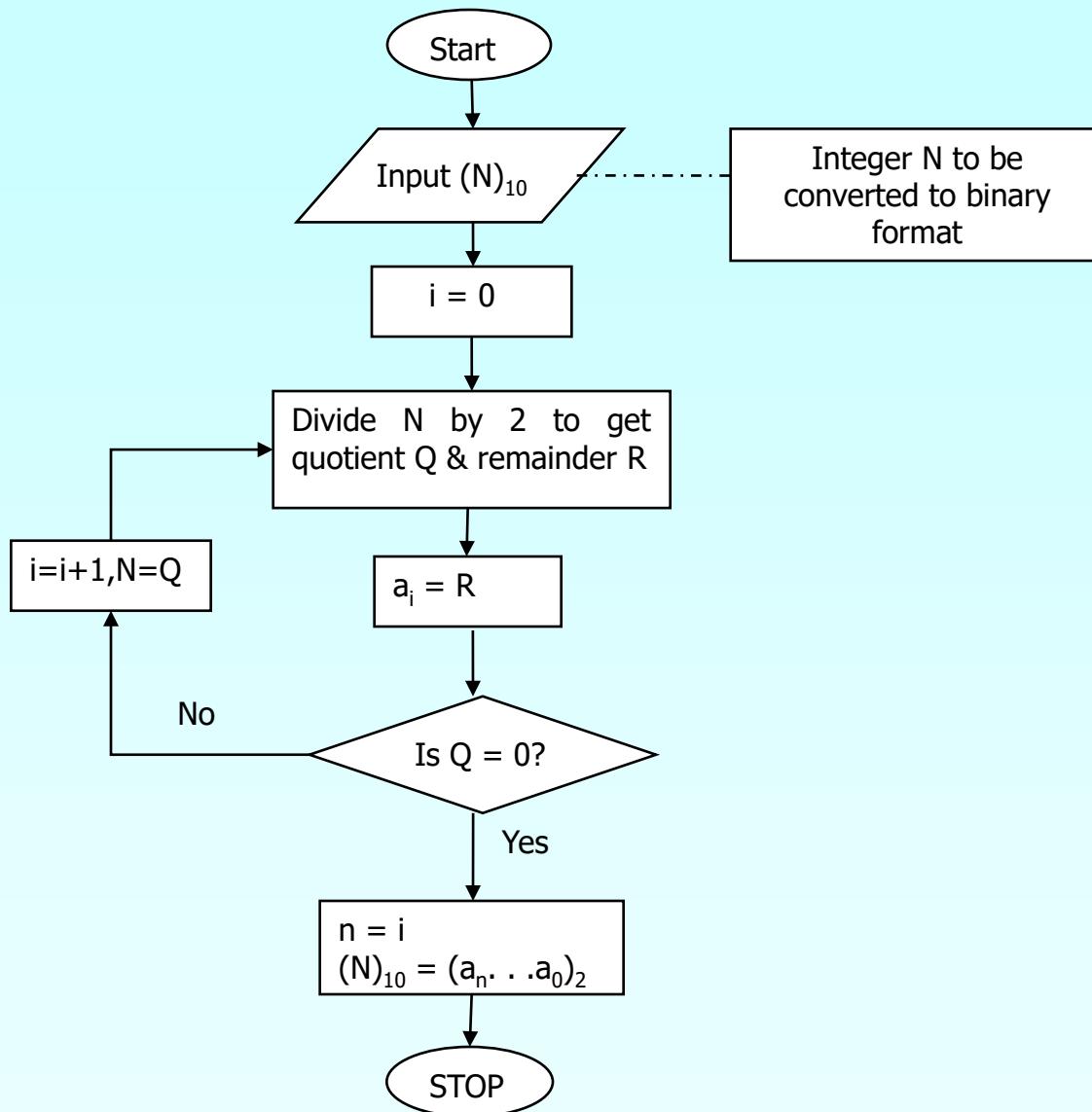
Convert Base 10 Integer to binary representation

Table 1 Converting a base-10 integer to binary representation.

	Quotient	Remainder
11/2	5	$1 = a_0$
5/2	2	$1 = a_1$
2/2	1	$0 = a_2$
1/2	0	$1 = a_3$

Hence

$$\begin{aligned}(11)_{10} &= (a_3 a_2 a_1 a_0)_2 \\ &= (1011)_2\end{aligned}$$



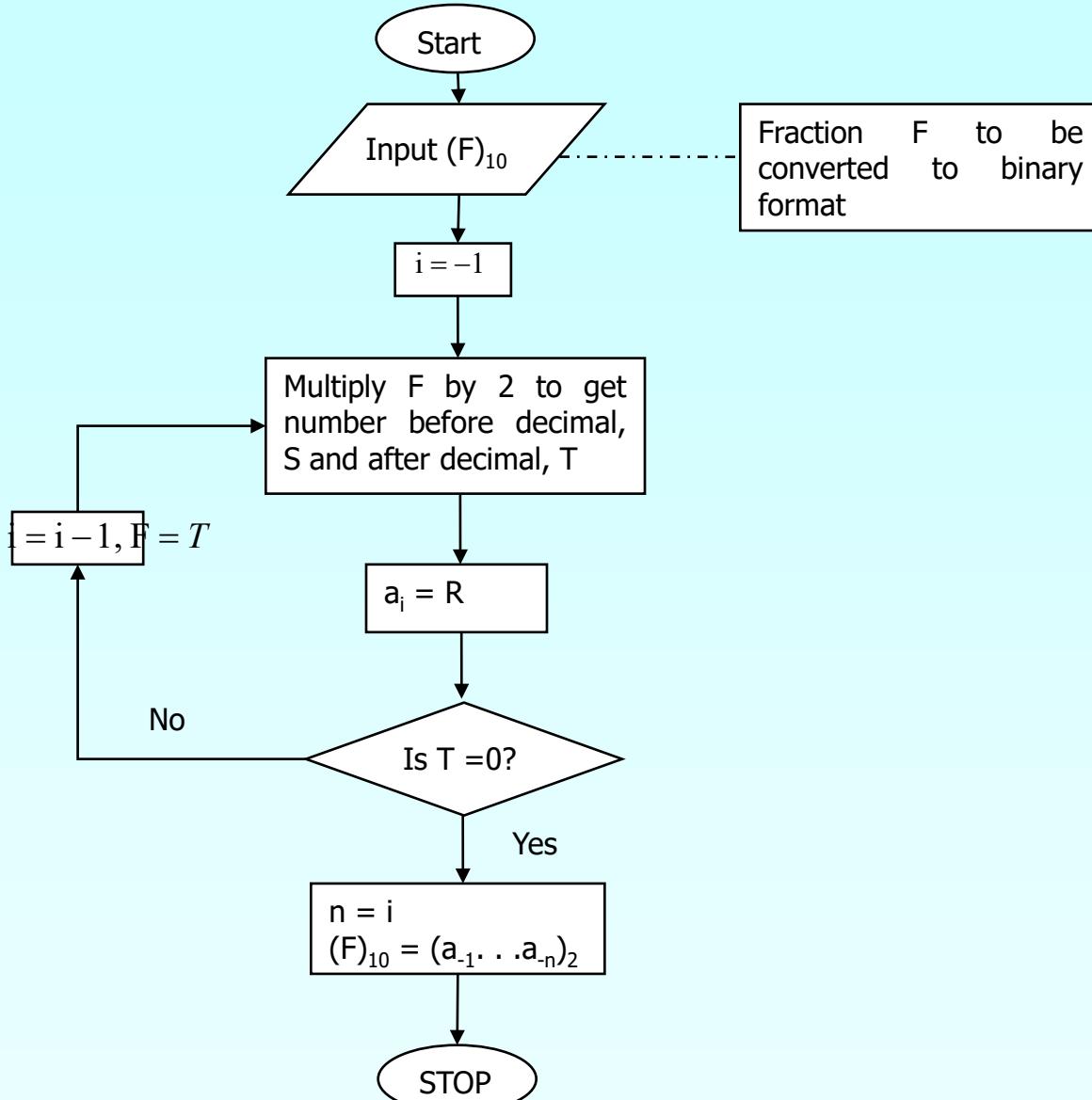
Fractional Decimal Number to Binary

Table 2. Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
0.1875×2	0.375	0.375	$0 = a_{-1}$
0.375×2	0.75	0.75	$0 = a_{-2}$
0.75×2	1.5	0.5	$1 = a_{-3}$
0.5×2	1.0	0.0	$1 = a_{-4}$

Hence

$$\begin{aligned}(0.1875)_{10} &= (a_{-1}a_{-2}a_{-3}a_{-4})_2 \\ &= (0.0011)_2\end{aligned}$$



Decimal Number to Binary

$$(11.1875)_{10} = (\quad ? . ? \quad)_2$$

Since

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2$$

we have

$$(11.1875)_{10} = (1011.0011)_2$$

All Fractional Decimal Numbers Cannot be Represented Exactly

Table 3. Converting a base-10 fraction to approximate binary representation.

	Number	Number after decimal	Number before Decimal
0.3×2	0.6	0.6	$0 = a_{-1}$
0.6×2	1.2	0.2	$1 = a_{-2}$
0.2×2	0.4	0.4	$0 = a_{-3}$
0.4×2	0.8	0.8	$0 = a_{-4}$
0.8×2	1.6	0.6	$1 = a_{-5}$

$$(0.3)_{10} \approx (a_{-1}a_{-2}a_{-3}a_{-4}a_{-5})_2 = (0.01001)_2 = 0.28125$$

Another Way to Look at Conversion

Convert $(11.1875)_{10}$ to base 2

$$(11)_{10} = 2^3 + 3$$

$$= 2^3 + 2^1 + 1$$

$$= 2^3 + 2^1 + 2^0$$

$$= 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$

$$= (1011)_2$$

$$\begin{aligned}(0.1875)_{10} &= 2^{-3} + 0.0625 \\&= 2^{-3} + 2^{-4} \\&= 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} \\&= (.0011)_2\end{aligned}$$

$$(11.1875)_{10} = (1011.0011)_2$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/binary_representation.html

THE END

<http://numericalmethods.eng.usf.edu>

Floating Point Representation

Major: All Engineering Majors

Authors: Autar Kaw, Matthew Emmons

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Floating Point Representation

<http://numericalmethods.eng.usf.edu>

Floating Decimal Point : Scientific Form

256.78 is written as $+2.5678 \times 10^2$

0.003678 is written as $+3.678 \times 10^{-3}$

-256.78 is written as -2.5678×10^2

Example

The form is

$$\text{sign} \times \text{mantissa} \times 10^{\text{exponent}}$$

or

$$\sigma \times m \times 10^e$$

Example: For

$$-2.5678 \times 10^2$$

$$\sigma = -1$$

$$m = 2.5678$$

$$e = 2$$

Floating Point Format for Binary Numbers

$$y = \sigma \times m \times 2^e$$

σ = sign of number (0 for + ve, 1 for - ve)

m = mantissa $[(1)_2 < m < (10)_2]$

1 is not stored as it is always given to be 1.

e = integer exponent

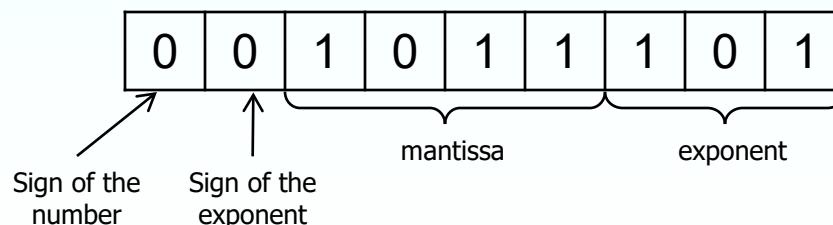
Example

9 bit-hypothetical word

- the first bit is used for the sign of the number,
- the second bit for the sign of the exponent,
- the next four bits for the mantissa, and
- the next three bits for the exponent

$$\begin{aligned}(54.75)_{10} &= (110110.11)_2 = (1.1011011)_2 \times 2^5 \\ &\approx (1.1011)_2 \times (101)_2\end{aligned}$$

We have the representation as



Machine Epsilon

Defined as the measure of accuracy and found by difference between 1 and the next number that can be represented

Example

Ten bit word

- Sign of number
- Sign of exponent
- Next four bits for exponent
- Next four bits for mantissa

$$\boxed{0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0} = (1)_{10}$$

Next number → $\boxed{0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1} = (1.0001)_2 = (1.0625)_{10}$

$$\epsilon_{mach} = 1.0625 - 1 = 2^{-4}$$

Relative Error and Machine Epsilon

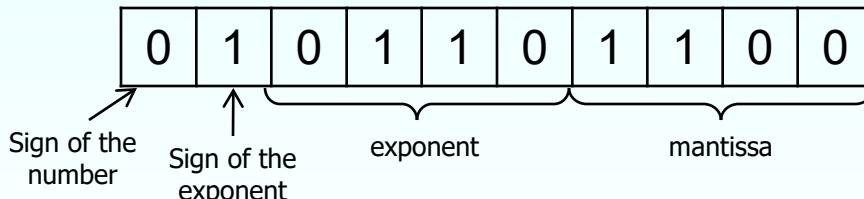
The absolute relative true error in representing a number will be less than the machine epsilon

Example

$$(0.02832)_{10} \cong (1.1100)_2 \times 2^{-5}$$

$$= (1.1100)_2 \times 2^{-(0110)_2}$$

10 bit word (sign, sign of exponent, 4 for exponent, 4 for mantissa)



$$(1.1100)_2 \times 2^{-(0110)_2} = 0.0274375$$

$$\epsilon_a = \left| \frac{0.02832 - 0.0274375}{0.02832} \right|$$

$$= 0.034472 < 2^{-4} = 0.0625$$

IEEE 754 Standards for Single Precision Representation

<http://numericalmethods.eng.usf.edu>

IEEE-754 Floating Point Standard

- Standardizes representation of floating point numbers on different computers in single and double precision.
- Standardizes representation of floating point operations on different computers.

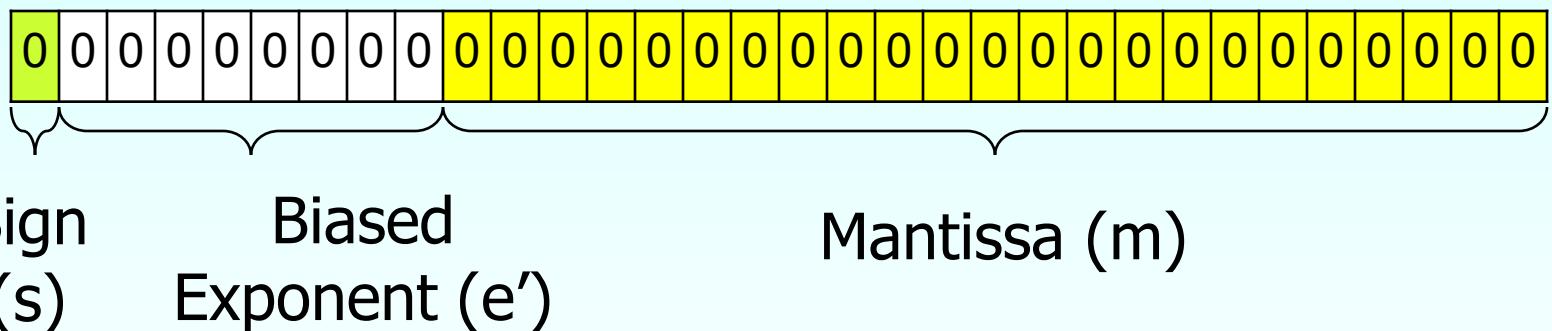
One Great Reference

What every computer scientist (and even if you are not) should know about floating point arithmetic!

<http://www.validlab.com/goldberg/paper.pdf>

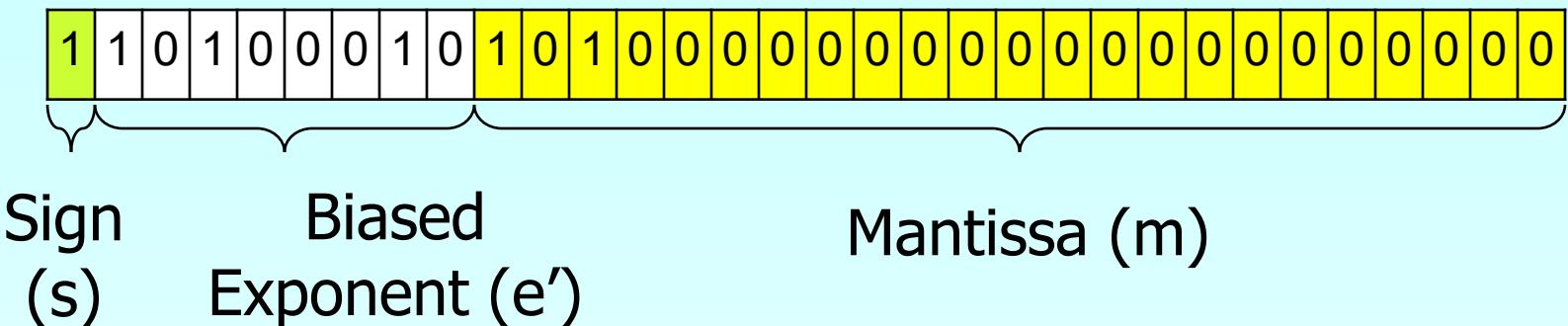
IEEE-754 Format Single Precision

32 bits for single precision



$$\text{Value} = (-1)^s \times (1.m)_2 \times 2^{e'-127}$$

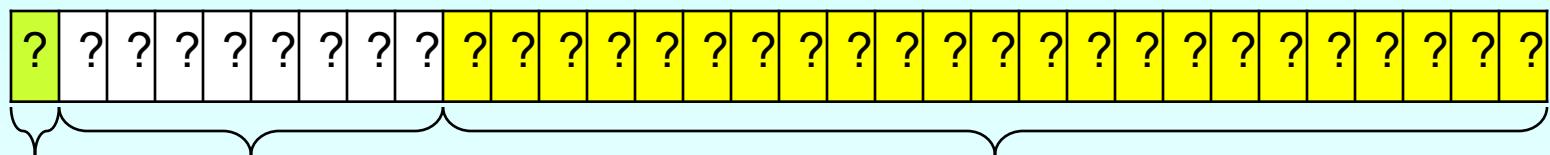
Example#1



$$\begin{aligned}\text{Value} &= (-1)^s \times (1.m)_2 \times 2^{e'-127} \\ &= (-1)^1 \times (1.10100000)_2 \times 2^{(10100010)_2 - 127} \\ &= (-1) \times (1.625) \times 2^{162-127} \\ &= (-1) \times (1.625) \times 2^{35} = -5.5834 \times 10^{10}\end{aligned}$$

Example#2

Represent -5.5834×10^{10} as a single precision floating point number.



Sign Biased
(s) Exponent (e') Mantissa (m)

$$-5.5834 \times 10^{10} = (-1)^1 \times (1.? \times 2^{\pm?})$$

Exponent for 32 Bit IEEE-754

8 bits would represent

$$0 \leq e' \leq 255$$

Bias is 127; so subtract 127 from representation

$$-127 \leq e \leq 128$$

Exponent for Special Cases

Actual range of e'

$$1 \leq e' \leq 254$$

$e' = 0$ and $e' = 255$ are reserved for special numbers

Actual range of e

$$-126 \leq e \leq 127$$

Special Exponents and Numbers

$e' = 0$ — all zeros

$e' = 255$ — all ones

s	e'	m	Represents
0	all zeros	all zeros	0
1	all zeros	all zeros	-0
0	all ones	all zeros	∞
1	all ones	all zeros	$-\infty$
0 or 1	all ones	non-zero	NaN

IEEE-754 Format

The largest number by magnitude

$$(1.1\ldots\ldots 1)_2 \times 2^{127} = 3.40 \times 10^{38}$$

The smallest number by magnitude

$$(1.00\ldots\ldots 0)_2 \times 2^{-126} = 2.18 \times 10^{-38}$$

Machine epsilon

$$\varepsilon_{mach} = 2^{-23} = 1.19 \times 10^{-7}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/floatingpoint_representation.html

THE END

<http://numericalmethods.eng.usf.edu>

Propagation of Errors

Major: All Engineering Majors

Authors: Autar Kaw, Matthew Emmons

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Propagation of Errors

<http://numericalmethods.eng.usf.edu>

Propagation of Errors

In numerical methods, the calculations are not made with exact numbers. How do these inaccuracies propagate through the calculations?

Example 1:

Find the bounds for the propagation in adding two numbers. For example if one is calculating $X + Y$ where

$$X = 1.5 \quad 0.05$$

$$Y = 3.4 \quad 0.04$$

Solution

Maximum possible value of $X = 1.55$ and $Y = 3.44$

Maximum possible value of $X + Y = 1.55 + 3.44 = 4.99$

Minimum possible value of $X = 1.45$ and $Y = 3.36$.

Minimum possible value of $X + Y = 1.45 + 3.36 = 4.81$

Hence

$$4.81 \leq X + Y \leq 4.99.$$

Propagation of Errors In Formulas

If f is a function of several variables $X_1, X_2, X_3, \dots, X_{n-1}, X_n$ then the maximum possible value of the error in f is

$$\Delta f \approx \left| \frac{\partial f}{\partial X_1} \Delta X_1 \right| + \left| \frac{\partial f}{\partial X_2} \Delta X_2 \right| + \dots + \left| \frac{\partial f}{\partial X_{n-1}} \Delta X_{n-1} \right| + \left| \frac{\partial f}{\partial X_n} \Delta X_n \right|$$

Example 2:

The strain in an axial member of a square cross-section is given by

$$\epsilon = \frac{F}{h^2 E}$$

Given

$$F = 72 \pm 0.9 \text{ N}$$

$$h = 4 \pm 0.1 \text{ mm}$$

$$E = 70 \pm 1.5 \text{ GPa}$$

Find the maximum possible error in the measured strain.



Example 2:

Solution

$$\begin{aligned}\epsilon &= \frac{72}{(4 \times 10^{-3})^2 (70 \times 10^9)} \\ &= 64.286 \times 10^{-6} \\ &= 64.286 \mu\end{aligned}$$

$$\Delta \epsilon = \left| \frac{\partial \epsilon}{\partial F} \Delta F \right| + \left| \frac{\partial \epsilon}{\partial h} \Delta h \right| + \left| \frac{\partial \epsilon}{\partial E} \Delta E \right|$$



Example 2:

$$\frac{\partial \epsilon}{\partial F} = \frac{1}{h^2 E} \quad \frac{\partial \epsilon}{\partial h} = -\frac{2F}{h^3 E} \quad \frac{\partial \epsilon}{\partial E} = -\frac{F}{h^2 E^2}$$

Thus

$$\begin{aligned}\Delta E &= \left| \frac{1}{h^2 E} \Delta F \right| + \left| \frac{2F}{h^3 E} \Delta h \right| + \left| \frac{F}{h^2 E^2} \Delta E \right| \\ &= \left| \frac{1}{(4 \times 10^{-3})^2 (70 \times 10^9)} \times 0.9 \right| + \left| \frac{2 \times 72}{(4 \times 10^{-3})^3 (70 \times 10^9)} \times 0.0001 \right| \\ &\quad + \left| \frac{72}{(4 \times 10^{-3})^2 (70 \times 10^9)^2} \times 1.5 \times 10^9 \right| \\ &= 5.3955 \mu\end{aligned}$$

Hence

$$\epsilon = (64.286 \mu \pm 5.3955 \mu)$$

Example 3:

Subtraction of numbers that are nearly equal can create unwanted inaccuracies. Using the formula for error propagation, show that this is true.

Solution

Let

$$z = x - y$$

Then

$$\begin{aligned} |\Delta z| &= \left| \frac{\partial z}{\partial x} \Delta x \right| + \left| \frac{\partial z}{\partial y} \Delta y \right| \\ &= |(1)\Delta x| + |(-1)\Delta y| \\ &= |\Delta x| + |\Delta y| \end{aligned}$$

So the relative change is

$$\left| \frac{\Delta z}{z} \right| = \frac{|\Delta x| + |\Delta y|}{|x - y|}$$

Example 3:

For example if

$$x = 2 \pm 0.001$$

$$y = 2.003 \pm 0.001$$

$$\left| \frac{\Delta z}{z} \right| = \frac{|0.001| + |0.001|}{|2 - 2.003|}$$

$$= 0.6667$$

$$= 66.67\%$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics-propagation_of_errors.html

THE END

<http://numericalmethods.eng.usf.edu>

Taylor Series Revisited

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Taylor Series Revisited

<http://numericalmethods.eng.usf.edu>

What is a Taylor series?

Some examples of Taylor series which you must have seen

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$$

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

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

General Taylor Series

The general form of the Taylor series is given by

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \dots$$

provided that all derivatives of $f(x)$ are continuous and exist in the interval $[x, x+h]$

What does this mean in plain English?

As Archimedes would have said, "*Give me the value of the function at a single point, and the value of all (first, second, and so on) its derivatives at that single point, and I can give you the value of the function at any other point*" (*fine print excluded*)

Example—Taylor Series

Find the value of $f(6)$ given that $f(4)=125$, $f'(4)=74$,
 $f''(4)=30$, $f'''(4)=6$ and all other higher order derivatives
of $f(x)$ at $x=4$ are zero.

Solution:

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + \dots$$

$$x = 4$$

$$h = 6 - 4 = 2$$

Example (cont.)

Solution: (cont.)

Since the higher order derivatives are zero,

$$f(4+2) = f(4) + f'(4)2 + f''(4)\frac{2^2}{2!} + f'''(4)\frac{2^3}{3!}$$

$$\begin{aligned}f(6) &= 125 + 74(2) + 30\left(\frac{2^2}{2!}\right) + 6\left(\frac{2^3}{3!}\right) \\&= 125 + 148 + 60 + 8 \\&= 341\end{aligned}$$

Note that to find $f(6)$ exactly, we only need the value of the function and all its derivatives at some other point, in this case $x = 4$

Derivation for Maclaurin Series for e^x

Derive the Maclaurin series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

The Maclaurin series is simply the Taylor series about the point $x=0$

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + f''''(x)\frac{h^4}{4!} + f'''''(x)\frac{h^5}{5!} + \dots \\ f(0+h) &= f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4!} + f'''''(0)\frac{h^5}{5!} + \dots \end{aligned}$$

Derivation (cont.)

Since $f(x) = e^x$, $f'(x) = e^x$, $f''(x) = e^x$, ..., $f^n(x) = e^x$ and $f^n(0) = e^0 = 1$

the Maclaurin series is then

$$\begin{aligned}f(h) &= (e^0) + (e^0)h + \frac{(e^0)}{2!}h^2 + \frac{(e^0)}{3!}h^3 \dots \\&= 1 + h + \frac{1}{2!}h^2 + \frac{1}{3!}h^3 \dots\end{aligned}$$

So,

$$f(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Error in Taylor Series

The Taylor polynomial of order n of a function $f(x)$ with $(n+1)$ continuous derivatives in the domain $[x, x+h]$ is given by

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + \cdots + f^{(n)}(x)\frac{h^n}{n!} + R_n(x)$$

where the remainder is given by

$$R_n(x) = \frac{(x-h)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

where

$$x < c < x+h$$

that is, c is some point in the domain $[x, x+h]$

Example—error in Taylor series

The Taylor series for e^x at point $x = 0$ is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

It can be seen that as the number of terms used increases, the error bound decreases and hence a better estimate of the function can be found.

How many terms would it require to get an approximation of e^1 within a magnitude of true error of less than 10^{-6} .

Example—(cont.)

Solution:

Using $(n+1)$ terms of Taylor series gives error bound of

$$R_n(x) = \frac{(x-h)^{n+1}}{(n+1)!} f^{(n+1)}(c) \quad x = 0, h = 1, f(x) = e^x$$

$$\begin{aligned} R_n(0) &= \frac{(0-1)^{n+1}}{(n+1)!} f^{(n+1)}(c) \\ &= \frac{(-1)^{n+1}}{(n+1)!} e^c \end{aligned}$$

Since

$$x < c < x + h$$

$$0 < c < 0 + 1$$

$$0 < c < 1$$

$$\frac{1}{(n+1)!} < |R_n(0)| < \frac{e}{(n+1)!}$$

Example—(cont.)

Solution: (cont.)

So if we want to find out how many terms it would require to get an approximation of e^1 within a magnitude of true error of less than 10^{-6} ,

$$\frac{e}{(n+1)!} < 10^{-6}$$

$$(n+1)! > 10^6 e$$

$$(n+1)! > 10^6 \times 3$$

$$n \geq 9$$

So 9 terms or more are needed to get a true error less than 10^{-6}

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/taylor_series.html

THE END

<http://numericalmethods.eng.usf.edu>

Differentiation-Continuous Functions

Major: All Engineering Majors

Authors: Autar Kaw, Sri Harsha Garapati

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Differentiation – Continuous Functions

<http://numericalmethods.eng.usf.edu>

Forward Difference Approximation

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

For a finite ' Δx '

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Graphical Representation Of Forward Difference Approximation

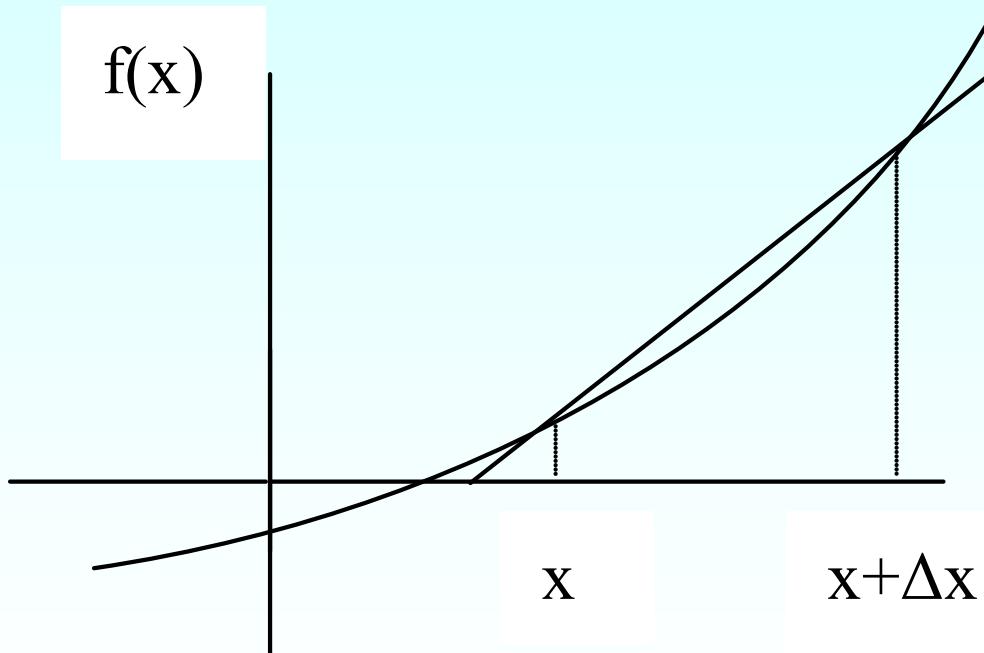


Figure 1 Graphical Representation of forward difference approximation of first derivative.

Example 1

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where ' v ' is given in m/s and ' t ' is given in seconds.

- Use forward difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16s$. Use a step size of $\Delta t = 2s$.
- Find the exact value of the acceleration of the rocket.
- Calculate the absolute relative true error for part (b).

Example 1 Cont.

Solution

$$a(t_i) \approx \frac{v(t_{i+1}) - v(t_i)}{\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned}t_{i+1} &= t_i + \Delta t \\&= 16 + 2 \\&= 18\end{aligned}$$

$$a(16) \approx \frac{v(18) - v(16)}{2}$$

Example 1 Cont.

$$\begin{aligned}v(18) &= 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) \\&= 453.02 \text{m/s}\end{aligned}$$

$$\begin{aligned}v(16) &= 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) \\&= 392.07 \text{m/s}\end{aligned}$$

Hence

$$a(16) \approx \frac{v(18) - v(16)}{2}$$

Example 1 Cont.

$$\approx \frac{453.02 - 392.07}{2}$$

$$\approx 30.474 \text{m/s}^2$$

- b) The exact value of $a(16)$ can be calculated by differentiating

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

as

$$a(t) = \frac{d}{dt}[v(t)]$$

Example 1 Cont.

Knowing that

$$\frac{d}{dt}[\ln(t)] = \frac{1}{t} \quad \text{and} \quad \frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2}$$

$$\begin{aligned} a(t) &= 2000 \left(\frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \frac{d}{dt} \left(\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right) - 9.8 \\ &= 2000 \left(\frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) (-1) \left(\frac{14 \times 10^4}{(14 \times 10^4 - 2100t)^2} \right) (-2100) - 9.8 \\ &= \frac{-4040 - 29.4t}{-200 + 3t} \end{aligned}$$

Example 1 Cont.

$$a(16) = \frac{-4040 - 29.4(16)}{-200 + 3(16)}$$
$$= 29.674 \text{m/s}^2$$

The absolute relative true error is

$$|\epsilon_t| = \left| \frac{\text{True Value} - \text{Approximate Value}}{\text{True Value}} \right| \times 100$$

$$= \left| \frac{29.674 - 30.474}{29.674} \right| \times 100$$

$$= 2.6967\%$$

Backward Difference Approximation of the First Derivative

We know

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

For a finite ' Δx ',

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

If ' Δx ' is chosen as a negative number,

$$f'(x) \approx \frac{f(x - \Delta x) - f(x)}{-\Delta x}$$

$$= \frac{f(x) - f(x - \Delta x)}{\Delta x}$$

Backward Difference Approximation of the First Derivative Cont.

This is a backward difference approximation as you are taking a point backward from x . To find the value of $f'(x)$ at $x = x_i$, we may choose another point ' Δx ' behind as $x = x_{i-1}$. This gives

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{\Delta x}$$

$$= \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

where

$$\Delta x = x_i - x_{i-1}$$

Backward Difference Approximation of the First Derivative Cont.

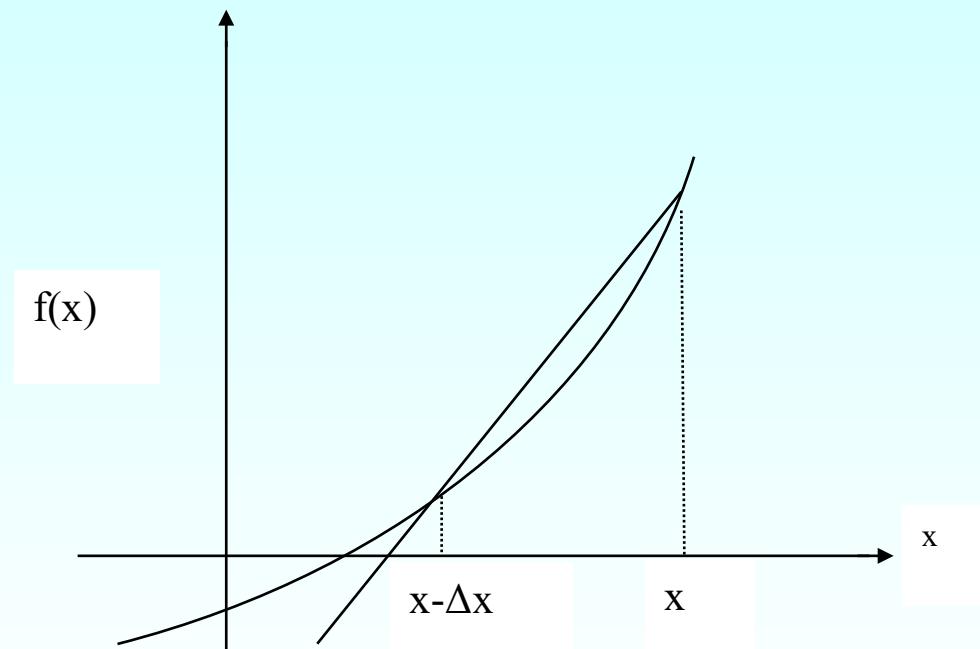


Figure 2 Graphical Representation of backward difference approximation of first derivative

Example 2

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where ' v ' is given in m/s and ' t ' is given in seconds.

- Use backward difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16\text{ s}$. Use a step size of $\Delta t = 2\text{ s}$.
- Find the absolute relative true error for part (a).

Example 2 Cont.

Solution

$$a(t) \approx \frac{v(t_i) - v(t_{i-1})}{\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i-1} &= t_i - \Delta t \\ &= 16 - 2 \\ &= 14 \end{aligned}$$

$$a(16) \approx \frac{v(16) - v(14)}{2}$$

Example 2 Cont.

$$\nu(16) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16)$$
$$= 392.07 \text{ m/s}$$

$$\nu(14) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14)$$
$$= 334.24 \text{ m/s}$$

$$a(16) \approx \frac{\nu(16) - \nu(14)}{2}$$
$$= \frac{392.07 - 334.24}{2}$$
$$\approx 28.915 \text{ m/s}^2$$

Example 2 Cont.

The exact value of the acceleration at $t = 16\text{ s}$ from Example 1 is

$$a(16) = 29.674\text{m/s}^2$$

The absolute relative true error is

$$|e_t| = \left| \frac{29.674 - 28.915}{29.674} \right| \times 100$$

$$= 2.5584\%$$

Derive the forward difference approximation from Taylor series

Taylor's theorem says that if you know the value of a function ' f ' at a point x_i and all its derivatives at that point, provided the derivatives are continuous between x_i and x_{i+1} , then

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \dots$$

Substituting for convenience $\Delta x = x_{i+1} - x_i$

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{\Delta x} - \frac{f''(x_i)}{2!}(\Delta x) + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{\Delta x} + O(\Delta x)$$

Derive the forward difference approximation from Taylor series Cont.

The $(0\Delta x)$ term shows that the error in the approximation is of the order of (Δx) . Can you now derive from Taylor series the formula for backward divided difference approximation of the first derivative?

As shown above, both forward and backward divided difference approximation of the first derivative are accurate on the order of $(0\Delta x)$.

Can we get better approximations? Yes, another method to approximate the first derivative is called the **Central difference approximation of the first derivative**.

Derive the forward difference approximation from Taylor series Cont.

From Taylor series

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 + \dots$$

$$f(x_{i-1}) = f(x_i) - f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 - \frac{f'''(x_i)}{3!}(\Delta x)^3 + \dots$$

Subtracting equation (2) from equation (1)

$$f(x_{i+1}) - f(x_{i-1}) = f'(x_i)(2\Delta x) + \frac{2f'''(x_i)}{3!}(\Delta x)^3 + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2\Delta x} - \frac{f'''(x_i)}{3!}(\Delta x)^2 + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2\Delta x} + O(\Delta x)^2$$

Central Divided Difference

Hence showing that we have obtained a more accurate formula as the error is of the order of $O(\Delta x)^2$.

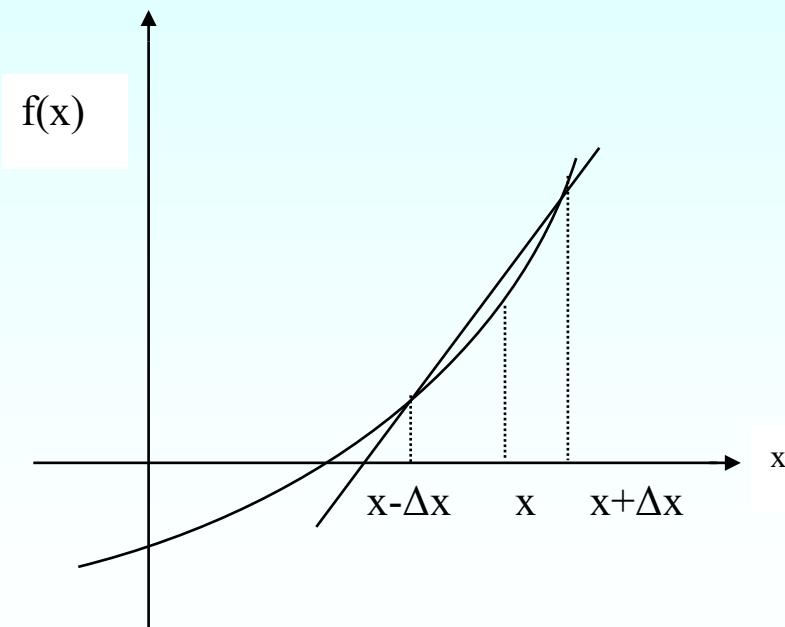


Figure 3 Graphical Representation of central difference approximation of first derivative

Example 3

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where ' v ' is given in m/s and ' t ' is given in seconds.

- Use central divided difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16s$. Use a step size of $\Delta t = 2s$.
- Find the absolute relative true error for part (a).

Example 3 cont.

Solution

$$a(t_i) \approx \frac{v(t_{i+1}) - v(t_{i-1})}{2\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned}t_{i+1} &= t_i + \Delta t \\&= 16 + 2 \\&= 18\end{aligned}$$

$$\begin{aligned}t_{i-1} &= t_i - \Delta t \\&= 16 - 2 \\&= 14\end{aligned}$$

$$\begin{aligned}a(16) &\approx \frac{v(18) - v(14)}{2(2)} \\&\approx \frac{v(18) - v(14)}{4}\end{aligned}$$

Example 3 cont.

$$\nu(18) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18)$$
$$= 453.02 \text{m/s}$$

$$\nu(14) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14)$$
$$= 334.24 \text{m/s}$$

$$a(16) \approx \frac{\nu(18) - \nu(14)}{4}$$
$$\approx \frac{453.02 - 334.24}{4}$$
$$\approx 29.694 \text{m/s}^2$$

Example 3 cont.

The exact value of the acceleration at $t = 16 \text{ s}$ from Example 1 is

$$a(16) = 29.674 \text{ m/s}^2$$

The absolute relative true error is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{29.674 - 29.694}{29.674} \right| \times 100 \\ &= 0.069157\% \end{aligned}$$

Comparision of FDD, BDD, CDD

The results from the three difference approximations are given in Table 1.

Table 1 Summary of a (16) using different divided difference approximations

Type of Difference Approximation	$a(16)$ (m/s^2)	$ \epsilon_t \%$
Forward	30.475	2.6967
Backward	28.915	2.5584
Central	29.695	0.069157

Finding the value of the derivative within a prespecified tolerance

In real life, one would not know the exact value of the derivative – so how would one know how accurately they have found the value of the derivative.

A simple way would be to start with a step size and keep on halving the step size and keep on halving the step size until the absolute relative approximate error is within a pre-specified tolerance.

Take the example of finding $v'(t)$ for

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

at $t = 16$ using the backward divided difference scheme.

Finding the value of the derivative within a prespecified tolerance Cont.

Given in Table 2 are the values obtained using the backward difference approximation method and the corresponding absolute relative approximate errors.

Table 2 First derivative approximations and relative errors for different Δt values of backward difference scheme

Δt	$v'(t)$	$ e_a \%$
2	28.915	
1	29.289	1.2792
0.5	29.480	0.64787
0.25	29.577	0.32604
0.125	29.625	0.16355

Finding the value of the derivative within a prespecified tolerance Cont.

From the above table, one can see that the absolute relative approximate error decreases as the step size is reduced. At $\Delta t = 0.125$ the absolute relative approximate error is 0.16355%, meaning that at least 2 significant digits are correct in the answer.

Finite Difference Approximation of Higher Derivatives

One can use Taylor series to approximate a higher order derivative.

For example, to approximate $f''(x)$, the Taylor series for

$$f(x_{i+2}) = f(x_i) + f'(x_i)(2\Delta x) + \frac{f''(x_i)}{2!}(2\Delta x)^2 + \frac{f'''(x_i)}{3!}(2\Delta x)^3 + \dots$$

where

$$x_{i+2} = x_i + 2\Delta x$$

$$f(x_{i+1}) = f(x_i) + f'(x_i)(\Delta x) + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 \dots$$

where

$$x_{i-1} = x_i - \Delta x$$

Finite Difference Approximation of Higher Derivatives Cont.

Subtracting 2 times equation (4) from equation (3) gives

$$f(x_{i+2}) - 2f(x_{i+1}) + f(x_i) = -f''(x_i)(\Delta x)^2 + f'''(x_i)(\Delta x)^3 \dots$$

$$f''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{(\Delta x)^2} - f'''(x_i)(\Delta x) + \dots$$

$$f''(x_i) \approx \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{(\Delta x)^2} + O(\Delta x) \quad (5)$$

Example 4

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

Use forward difference approximation of the second derivative $v''(t)$ of to calculate the jerk at $t = 16s$. Use a step size of $\Delta t = 2s$.

Example 4 Cont.

Solution

$$j(t_i) \approx \frac{v(t_{i+2}) - 2v(t_{i+1}) + v(t_i)}{(\Delta t)^2}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned}t_{i+1} &= t_i + \Delta t \\&= 16 + 2 \\&= 18\end{aligned}$$

$$\begin{aligned}t_{i+2} &= t_i + 2(\Delta t) \\&= 16 + 2(2) \\&= 20\end{aligned}$$

$$j(16) \approx \frac{v(20) - 2v(18) + v(16)}{(2)^2}$$

Example 4 Cont.

$$v(20) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(20)} \right] - 9.8(20)$$
$$= 517.35 \text{ m/s}$$

$$v(18) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18)$$
$$= 453.02 \text{ m/s}$$

$$v(16) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16)$$
$$= 392.07 \text{ m/s}$$

Example 4 Cont.

$$j(16) \approx \frac{517.35 - 2(453.02) + 392.07}{4}$$
$$\approx 0.84515 \text{m/s}^3$$

The exact value of $j(16)$ can be calculated by differentiating

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

twice as

$$a(t) = \frac{d}{dt}[v(t)] \quad \text{and} \quad j(t) = \frac{d}{dt}[a(t)]$$

Example 4 Cont.

Knowing that

$$\frac{d}{dt}[\ln(t)] = \frac{1}{t} \quad \text{and} \quad \frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2}$$

$$\begin{aligned} a(t) &= 2000 \left(\frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \frac{d}{dt} \left(\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right) - 9.8 \\ &= 2000 \left(\frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) (-1) \left(\frac{14 \times 10^4}{(14 \times 10^4 - 2100t)^2} \right) (-2100) - 9.8 \\ &= \frac{-4040 - 29.4t}{-200 + 3t} \end{aligned}$$

Example 4 Cont.

Similarly it can be shown that

$$\begin{aligned} j(t) &= \frac{d}{dt}[a(t)] \\ &= \frac{18000}{(-200 + 3t)^2} \end{aligned}$$

$$\begin{aligned} j(16) &= \frac{18000}{[-200 + 3(16)]^2} \\ &= 0.77909 \text{m/s}^3 \end{aligned}$$

The absolute relative true error is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{0.77909 - 0.84515}{0.77909} \right| \times 100 \\ &= 8.4797 \% \end{aligned}$$

Higher order accuracy of higher order derivatives

The formula given by equation (5) is a forward difference approximation of the second derivative and has the error of the order of (Δx) . Can we get a formula that has a better accuracy? We can get the central difference approximation of the second derivative.

The Taylor series for

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 + \frac{f''''(x_i)}{4!}(\Delta x)^4 \dots \quad (6)$$

where

$$x_{i+1} = x_i + \Delta x$$

Higher order accuracy of higher order derivatives Cont.

$$f(x_{i-1}) = f(x_i) - f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 - \frac{f'''(x_i)}{3!}(\Delta x)^3 + \frac{f''''(x_i)}{4!}(\Delta x)^4 \dots \quad (7)$$

where

$$x_{i-1} = x_i - \Delta x$$

Adding equations (6) and (7), gives

$$f(x_{i+1}) + f(x_{i-1}) = 2f(x_i) + f''(x_i)(\Delta x)^2 + f'''(x_i) \frac{(\Delta x)^4}{12}$$

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} - \frac{f''''(x_i)(\Delta x)^2}{12}$$

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} + O(\Delta x)^2$$

Example 5

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

Use central difference approximation of second derivative of $v(t)$ to calculate the jerk at $t = 16s$. Use a step size of $\Delta t = 2s$.

Example 5 Cont.

Solution

$$a(t_i) \approx \frac{v(t_{i+1}) - 2v(t_i) + v(t_{i-1})}{(\Delta t)^2}$$
$$t_i = 16$$

$$\Delta t = 2$$

$$t_{i+1} = t_i + \Delta t$$
$$= 16 + 2$$
$$= 18$$

$$t_{i-1} = t_i - \Delta t$$
$$= 16 - 2$$
$$= 14$$

$$j(16) \approx \frac{v(18) - 2v(16) + v(14)}{(2)^2}$$

Example 5 Cont.

$$v(18) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18)$$
$$= 453.02 \text{ m/s}$$

$$v(16) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16)$$
$$= 392.07 \text{ m/s}$$

$$v(14) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14)$$
$$= 334.24 \text{ m/s}$$

Example 5 Cont.

$$\begin{aligned} j(16) &\approx \frac{\nu(18) - 2\nu(16) + \nu(14)}{(2)^2} \\ &\approx \frac{453.02 - 2(392.07) + 334.24}{4} \\ &\approx 0.77969 \text{m/s}^3 \end{aligned}$$

The absolute relative true error is

$$|\epsilon_t| = \left| \frac{0.77908 - 0.78}{0.77908} \right| \times 100$$

$$= 0.077992\%$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/continuous_02_dif.html

THE END

<http://numericalmethods.eng.usf.edu>

Differentiation-Discrete Functions

Major: All Engineering Majors

Authors: Autar Kaw, Sri Harsha Garapati

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Differentiation –Discrete Functions

<http://numericalmethods.eng.usf.edu>

Forward Difference Approximation

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

For a finite ' Δx '

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Graphical Representation Of Forward Difference Approximation

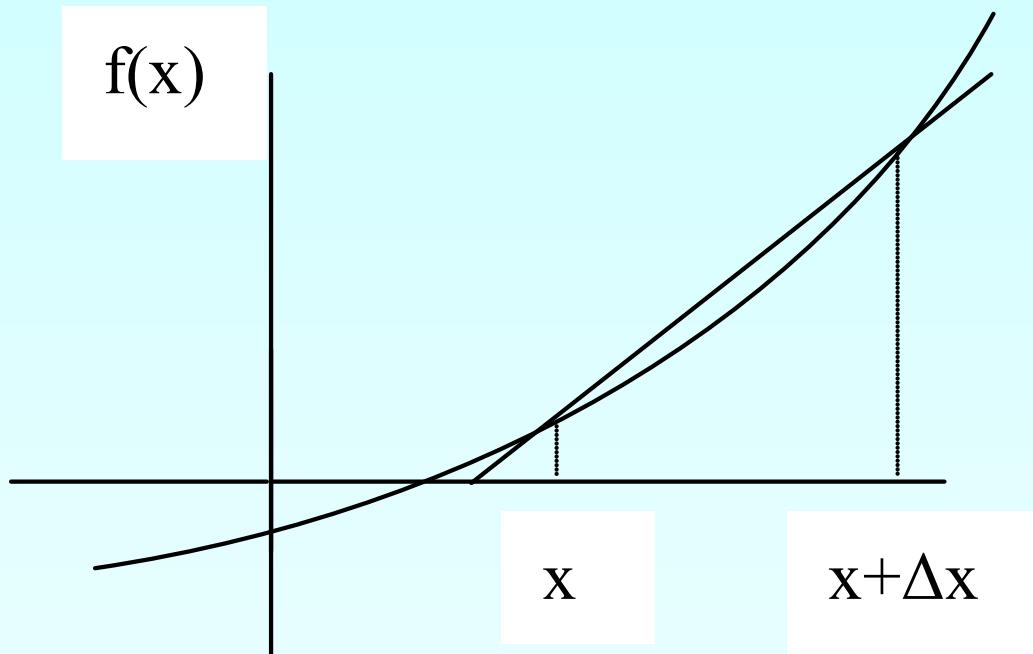


Figure 1 Graphical Representation of forward difference approximation of first derivative.

Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

Table 1 Velocity as a function of time

t	v(t)
s	m/s
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Using forward divided difference, find the acceleration of the rocket at $t = 16$ s .

Example 1 Cont.

Solution

To find the acceleration at $t = 16\text{s}$, we need to choose the two values closest to $t = 16\text{s}$, that also bracket $t = 16\text{s}$ to evaluate it. The two points are $t = 15\text{s}$ and $t = 20\text{s}$.

$$a(t_i) \approx \frac{v(t_{i+1}) - v(t_i)}{\Delta t}$$

$$t_i = 15$$

$$t_{i+1} = 20$$

$$\begin{aligned}\Delta t &= t_{i+1} - t_i \\ &= 20 - 15 \\ &= 5\end{aligned}$$

Example 1 Cont.

$$\begin{aligned}a(16) &\approx \frac{\nu(20) - \nu(15)}{5} \\&\approx \frac{517.35 - 362.78}{5} \\&\approx 30.914 \text{ m/s}^2\end{aligned}$$

Direct Fit Polynomials

In this method, given ' $n + 1$ ' data points $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
one can fit a n^{th} order polynomial given by

$$P_n(x) = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n$$

To find the first derivative,

$$P'_n(x) = \frac{dP_n(x)}{dx} = a_1 + 2a_2x + \dots + (n-1)a_{n-1}x^{n-2} + na_nx^{n-1}$$

Similarly other derivatives can be found.

Example 2-Direct Fit Polynomials

The upward velocity of a rocket is given as a function of time in Table 2.

Table 2 Velocity as a function of time

t	v(t)
s	m/s
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Using the third order polynomial interpolant for velocity, find the acceleration of the rocket at $t = 16$ s .

Example 2-Direct Fit Polynomials cont.

Solution

For the third order polynomial (also called cubic interpolation), we choose the velocity given by

$$v(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

Since we want to find the velocity at $t = 16$ s, and we are using third order polynomial, we need to choose the four points closest to $t = 16$ s and that also bracket $t = 16$ s to evaluate it.

The four points are $t_o = 10$, $t_1 = 15$, $t_2 = 20$, and $t_3 = 22.5$.

$$t_o = 10, \quad v(t_o) = 227.04$$

$$t_1 = 15, \quad v(t_1) = 362.78$$

$$t_2 = 20, \quad v(t_2) = 517.35$$

$$t_3 = 22.5, \quad v(t_3) = 602.97$$

Example 2-Direct Fit Polynomials cont.

such that

$$v(10) = 227.04 = a_0 + a_1(10) + a_2(10)^2 + a_3(10)^3$$

$$v(15) = 362.78 = a_0 + a_1(15) + a_2(15)^2 + a_3(15)^3$$

$$v(20) = 517.35 = a_0 + a_1(20) + a_2(20)^2 + a_3(20)^3$$

$$v(22.5) = 602.97 = a_0 + a_1(22.5) + a_2(22.5)^2 + a_3(22.5)^3$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & 10 & 100 & 1000 \\ 1 & 15 & 225 & 3375 \\ 1 & 20 & 400 & 8000 \\ 1 & 22.5 & 506.25 & 11391 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \\ 602.97 \end{bmatrix}$$

Example 2-Direct Fit Polynomials cont.

Solving the above four equations gives

$$a_0 = -4.3810$$

$$a_1 = 21.289$$

$$a_2 = 0.13065$$

$$a_3 = 0.0054606$$

Hence

$$\begin{aligned}v(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\&= -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, \quad 10 \leq t \leq 22.5\end{aligned}$$

Example 2-Direct Fit Polynomials cont.

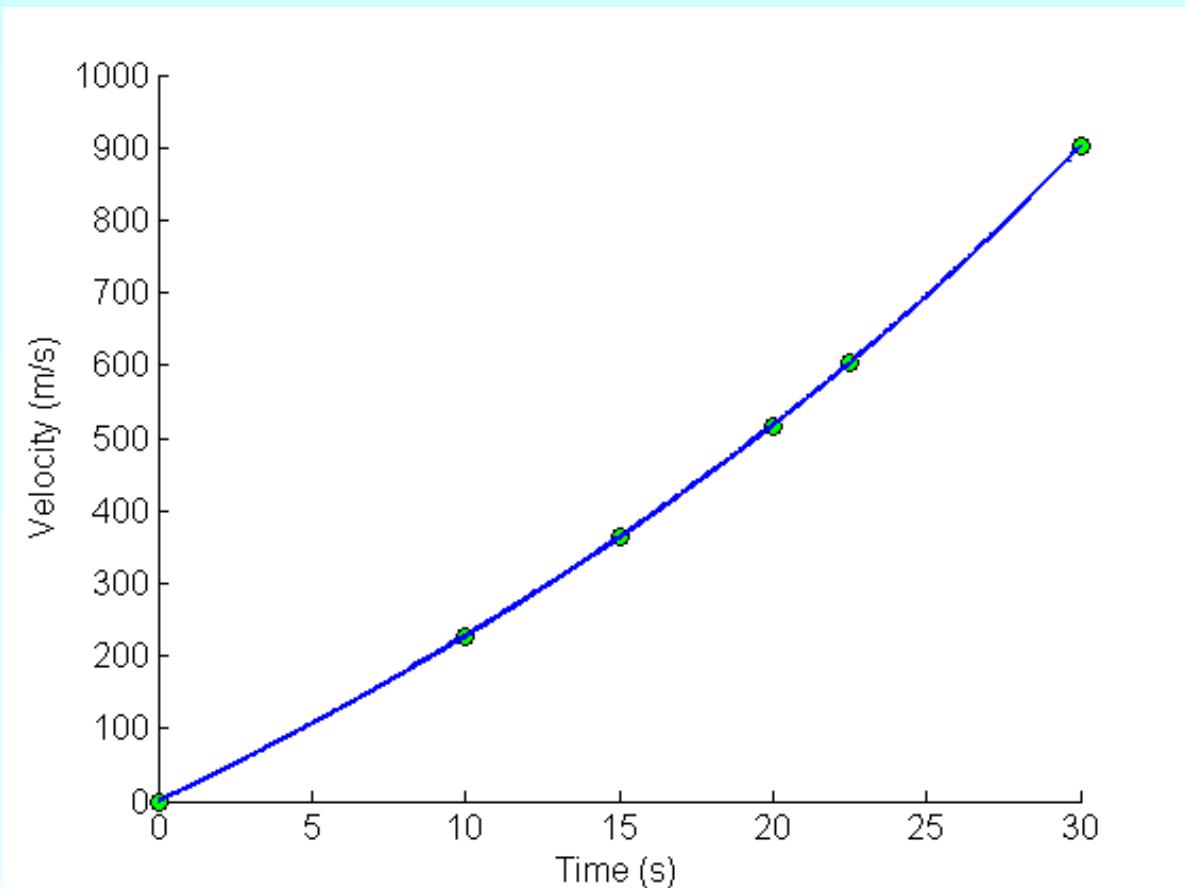


Figure 1 Graph of upward velocity of the rocket vs. time.

Example 2-Direct Fit Polynomials cont.

The acceleration at t=16 is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16}$$

Given that

$$v(t) = -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, 10 \leq t \leq 22.5$$

$$\begin{aligned} a(t) &= \frac{d}{dt} v(t) \\ &= \frac{d}{dt} (-4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3) \\ &= 21.289 + 0.26130t + 0.016382t^2, \quad 10 \leq t \leq 22.5 \end{aligned}$$

$$\begin{aligned} a(16) &= 21.289 + 0.26130(16) + 0.016382(16)^2 \\ &= 29.664 \text{m/s}^2 \end{aligned}$$

Lagrange Polynomial

In this method, given $(x_1, y_1), \dots, (x_n, y_n)$, one can fit a $(n-1)^{th}$ order Lagrangian polynomial given by

$$f_n(x) = \sum_{i=0}^n L_i(x) f(x_i)$$

where 'n' in $f_n(x)$ stands for the n^{th} order polynomial that approximates the function $y = f(x)$ given at $(n+1)$ data points as $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, and

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

$L_i(x)$ a weighting function that includes a product of $(n-1)$ terms with terms of $j = i$ omitted.

Lagrange Polynomial Cont.

Then to find the first derivative, one can differentiate $f_n(x)$ once, and so on for other derivatives.

For example, the second order Lagrange polynomial passing through $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ is

$$f_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f(x_2)$$

Differentiating equation (2) gives

Lagrange Polynomial Cont.

$$f_2'(x) = \frac{2x - (x_1 + x_2)}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \frac{2x - (x_0 + x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + \frac{2x - (x_0 + x_1)}{(x_2 - x_0)(x_2 - x_1)} f(x_2)$$

Differentiating again would give the second derivative as

$$f_2''(x) = \frac{2}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \frac{2}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + \frac{2}{(x_2 - x_0)(x_2 - x_1)} f(x_2)$$

Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

Table 3 Velocity as a function of time

t	v(t)
s	m/s
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the value of the acceleration at $t = 16$ s using the second order Lagrangian polynomial interpolation for velocity.

Example 3 Cont.

Solution

$$v(t) = \left(\frac{t - t_1}{t_0 - t_1} \right) \left(\frac{t - t_2}{t_0 - t_2} \right) v(t_0) + \left(\frac{t - t_0}{t_1 - t_0} \right) \left(\frac{t - t_2}{t_1 - t_2} \right) v(t_1) + \left(\frac{t - t_0}{t_2 - t_0} \right) \left(\frac{t - t_1}{t_2 - t_1} \right) v(t_2)$$

$$a(t) = \frac{2t - (t_1 + t_2)}{(t_0 - t_1)(t_0 - t_2)} v(t_0) + \frac{2t - (t_0 + t_2)}{(t_1 - t_0)(t_1 - t_2)} v(t_1) + \frac{2t - (t_0 + t_1)}{(t_2 - t_0)(t_2 - t_1)} v(t_2)$$

$$\begin{aligned} a(16) &= \frac{2(16) - (15 + 20)}{(10 - 15)(10 - 20)} (227.04) + \frac{2(16) - (10 + 20)}{(15 - 10)(15 - 20)} (362.78) + \frac{2(16) - (10 + 15)}{(20 - 10)(20 - 15)} (517.35) \\ &= -0.06(227.04) - 0.08(362.78) + 0.14(517.35) \\ &= 29.784 \text{m/s}^2 \end{aligned}$$

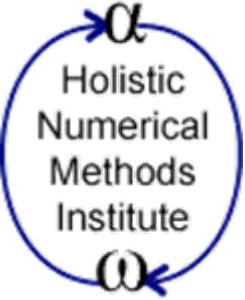
Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

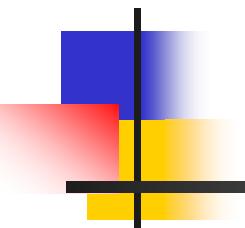
http://numericalmethods.eng.usf.edu/topics/discrete_02_dif.html

THE END

<http://numericalmethods.eng.usf.edu>



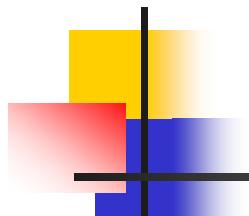
Forward Divided Difference



Topic: Differentiation

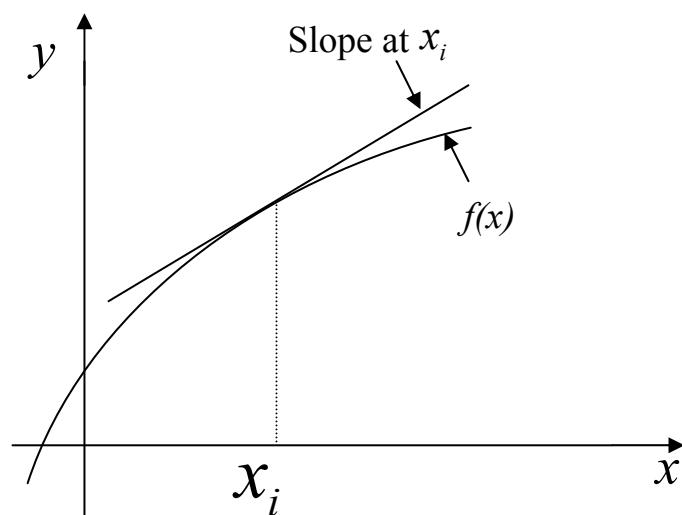
Major: General Engineering

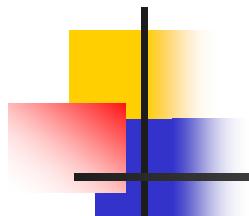
Authors: Autar Kaw, Sri Harsha Garapati



Definition

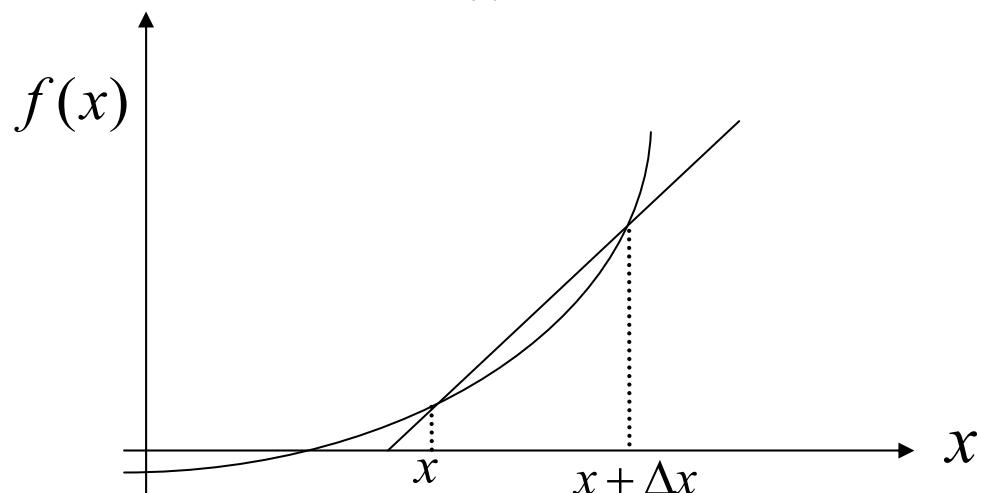
$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



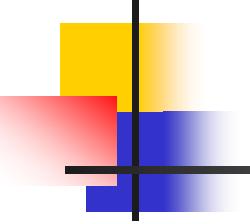


Forward Divided Difference

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



$$f'(x_i) \approx \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x}$$



Example

Example:

The velocity of a rocket is given by

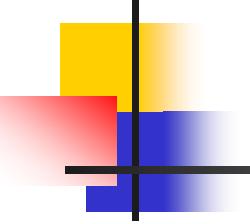
$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, 0 \leq t \leq 30$$

where v given in m/s and t is given in seconds. Use forward difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16s$. Use a step size of $\Delta t = 2s$.

Solution:

$$a(t_i) \cong \frac{v(t_{i+1}) - v(t_i)}{\Delta t}$$

$$t_i = 16$$



Example (contd.)

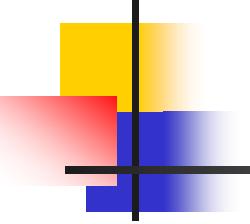
$$\Delta t = 2$$

$$t_{i+1} = t_i + \Delta t = 16 + 2 = 18$$

$$a(16) = \frac{v(18) - v(16)}{2}$$

$$v(18) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) = 453.02 \text{ m/s}$$

$$v(16) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) = 392.07 \text{ m/s}$$



Example (contd.)

Hence

$$a(16) = \frac{v(18) - v(16)}{2} = 453.02 - 392.07 = 30.475 \text{ m/s}^2$$

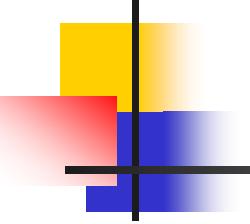
The exact value of $a(16)$ can be calculated by differentiating

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

as

$$a(t) = \frac{d}{dt}[v(t)] = \frac{-4040 - 29.4t}{-200 + 3t}$$

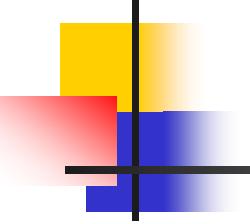
$$a(16) = 29.674 \text{ m/s}^2$$



Example (contd.)

The absolute relative true error is

$$\begin{aligned} |\varepsilon_t| &= \left| \frac{\text{TrueValue} - \text{ApproximateValue}}{\text{TrueValue}} \right| \times 100 \\ &= \left| \frac{29.674 - 30.475}{29.674} \right| \times 100 \\ &= 2.6993\% \end{aligned}$$



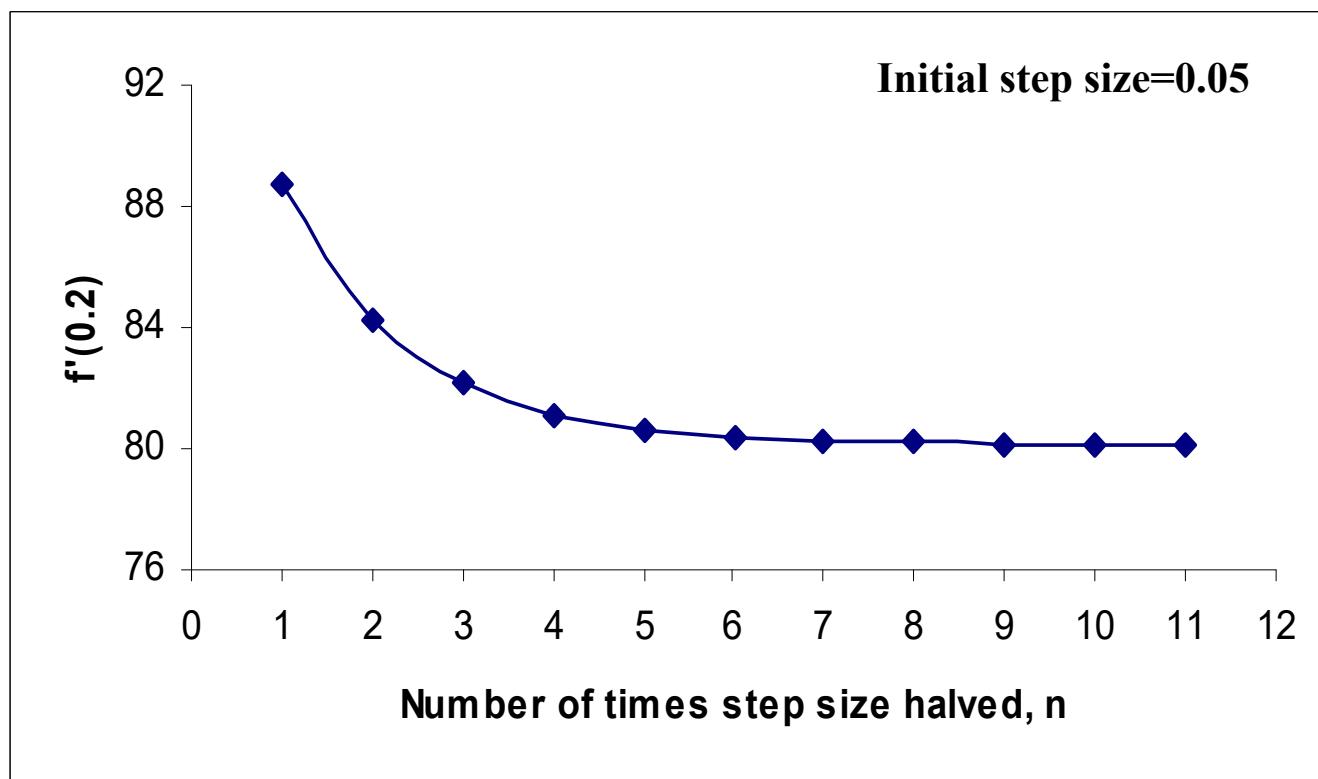
Effect Of Step Size

$$f(x) = 9e^{4x}$$

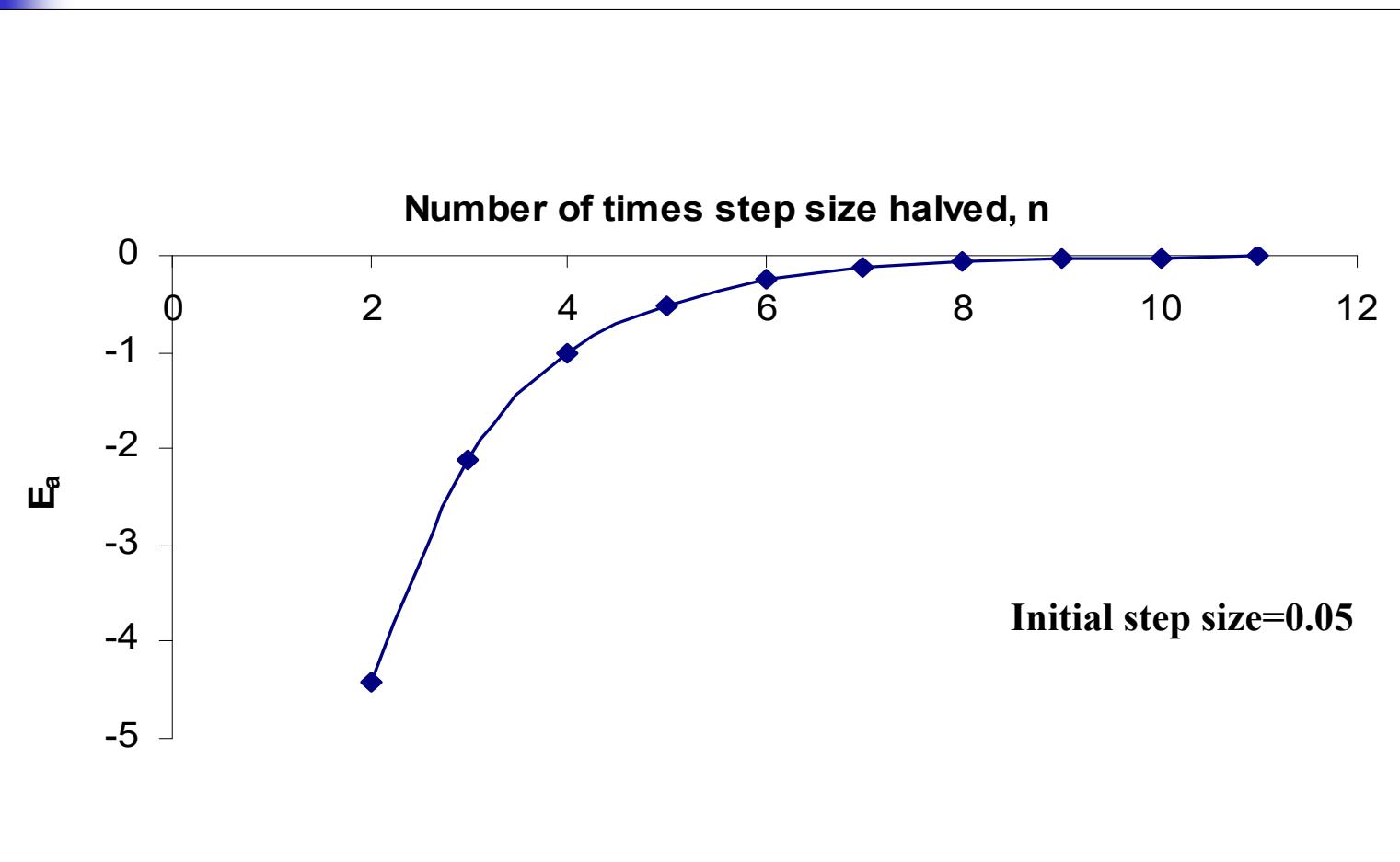
Value of $f'(0.2)$ Using forward difference method.

h	$f'(0.2)$	E_a	$ \varepsilon_a \%$	Significant digits	E_t	$ \varepsilon_t \%$
0.05	88.69336				-8.57389	10.70138
0.025	84.26239	-4.430976	5.258546	0	-4.14291	5.170918
0.0125	82.15626	-2.106121	2.563555	1	-2.03679	2.542193
0.00625	81.12937	-1.0269	1.265756	1	-1.00989	1.260482
0.003125	80.62231	-0.507052	0.628923	1	-0.50284	0.627612
0.001563	80.37037	-0.251944	0.313479	2	-0.25090	0.313152
0.000781	80.24479	-0.125579	0.156494	2	-0.12532	0.156413
0.000391	80.18210	-0.062691	0.078186	2	-0.06263	0.078166
0.000195	80.15078	-0.031321	0.039078	3	-0.03130	0.039073
9.77E-05	80.13512	-0.015654	0.019535	3	-0.01565	0.019534
4.88E-05	80.12730	-0.007826	0.009767	3	-0.00782	0.009766

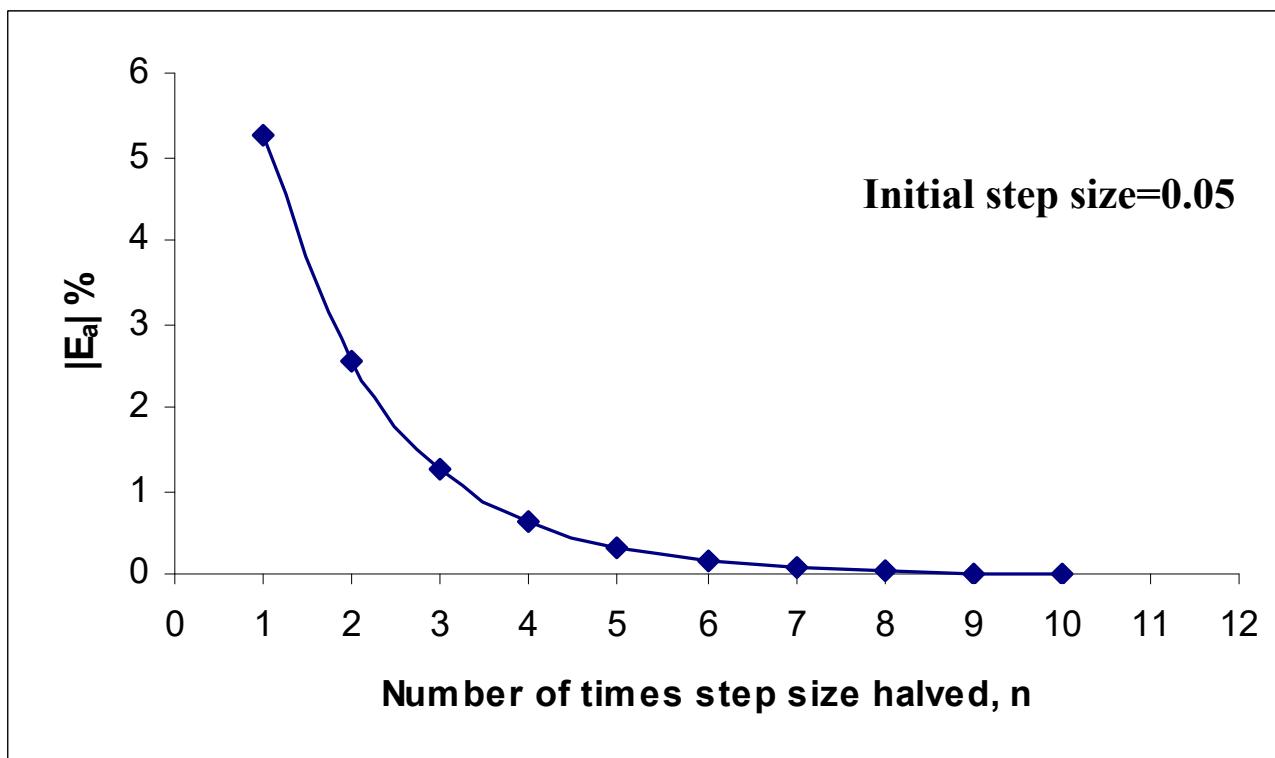
Effect of Step Size in Forward Divided Difference Method



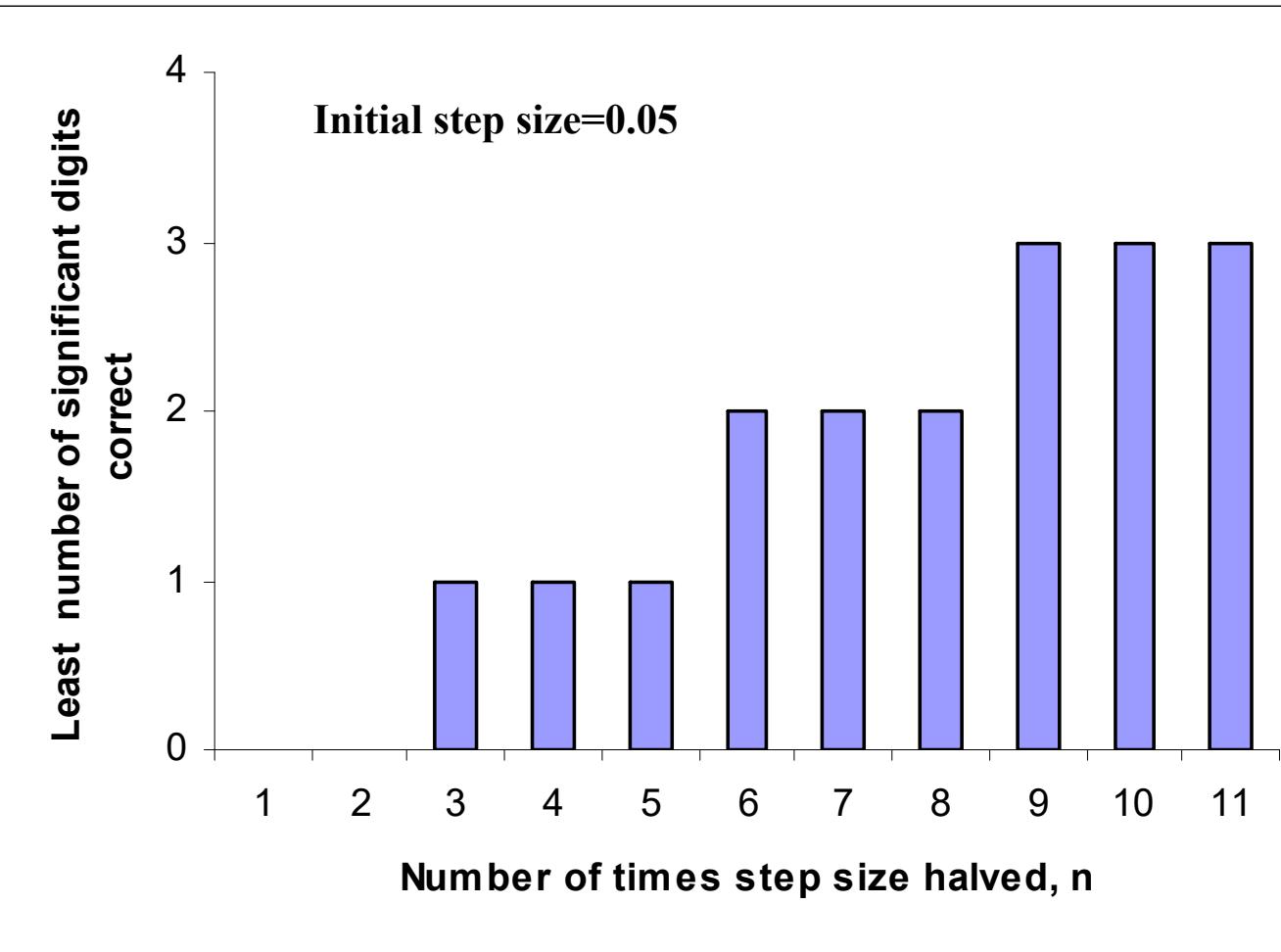
Effect of Step Size on Approximate Error



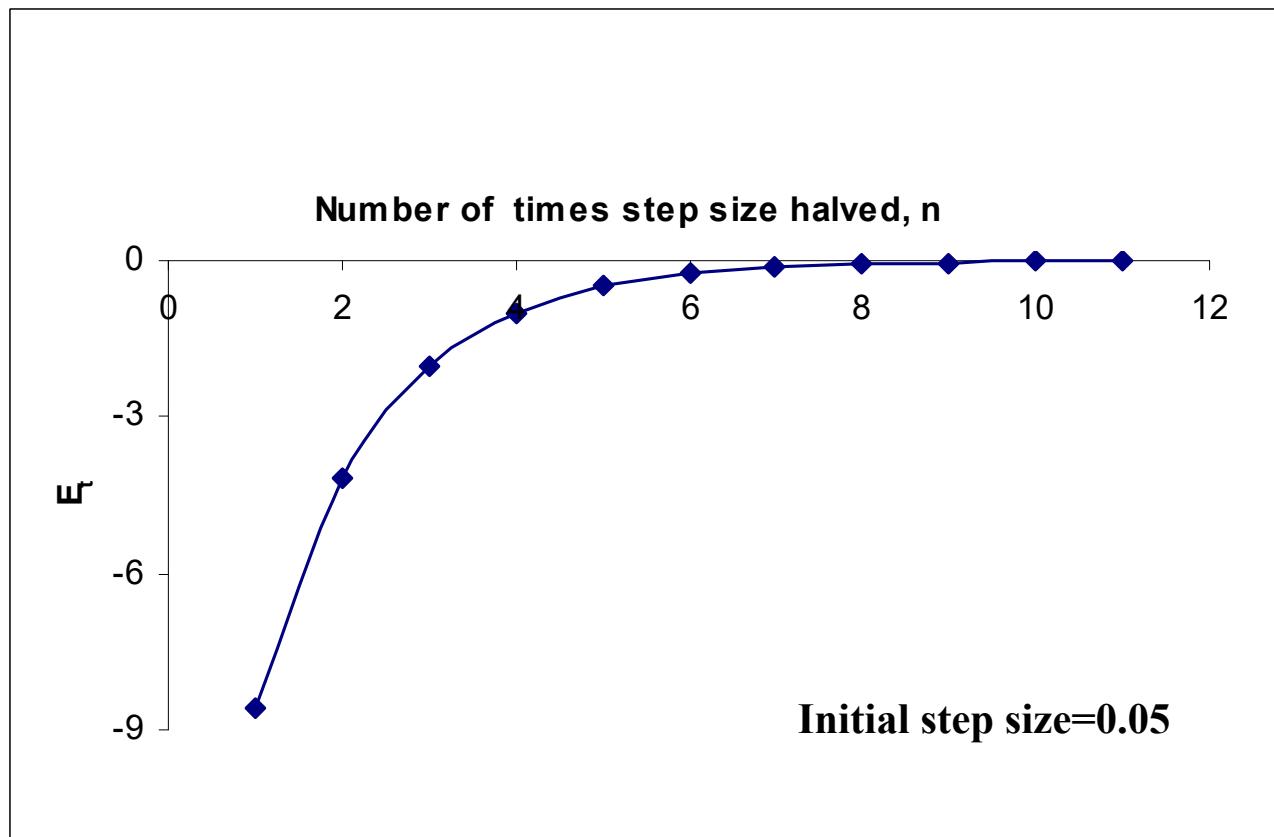
Effect of Step Size on Absolute Relative Approximate Error



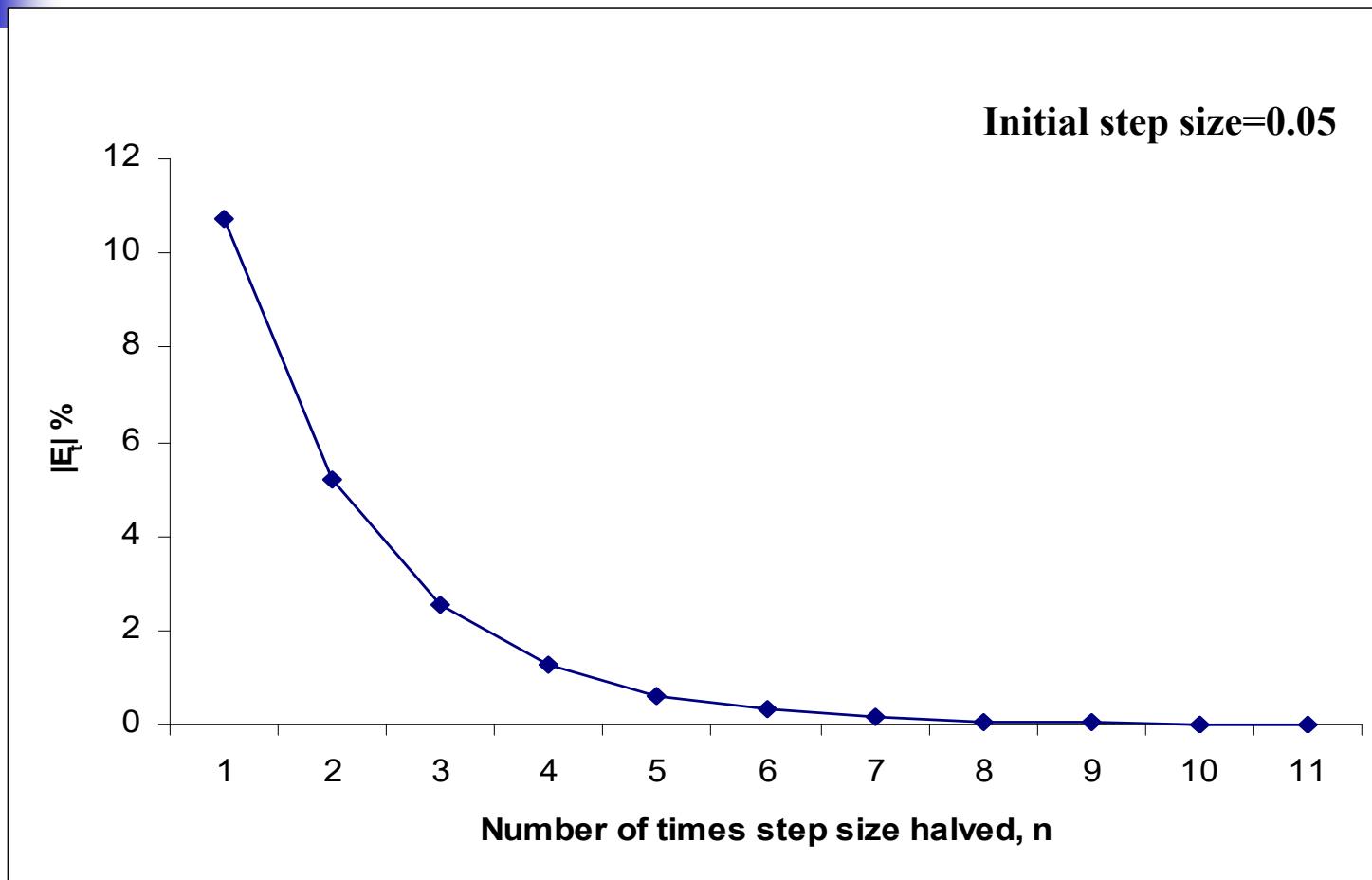
Effect of Step Size on Least Number of Significant Digits Correct

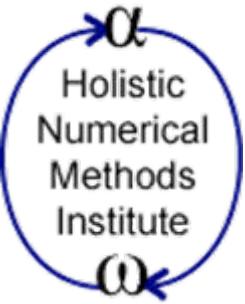


Effect of Step Size on True Error

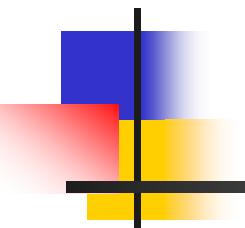


Effect of Step Size on Absolute Relative True Error





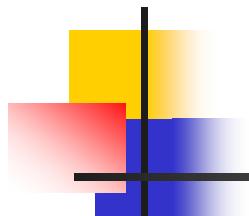
Backward Divided Difference



Topic: Differentiation

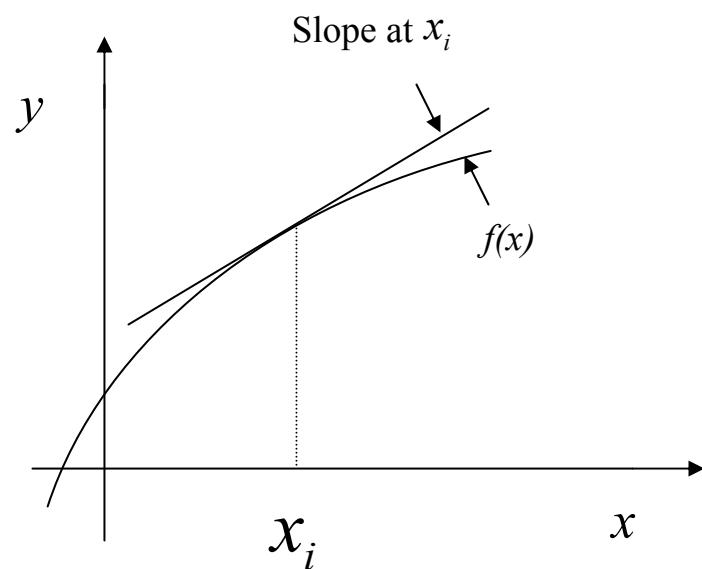
Major: General Engineering

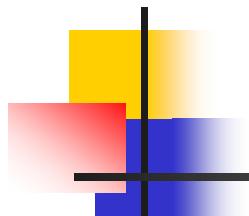
Authors: Autar Kaw, Sri Harsha Garapati



Definition

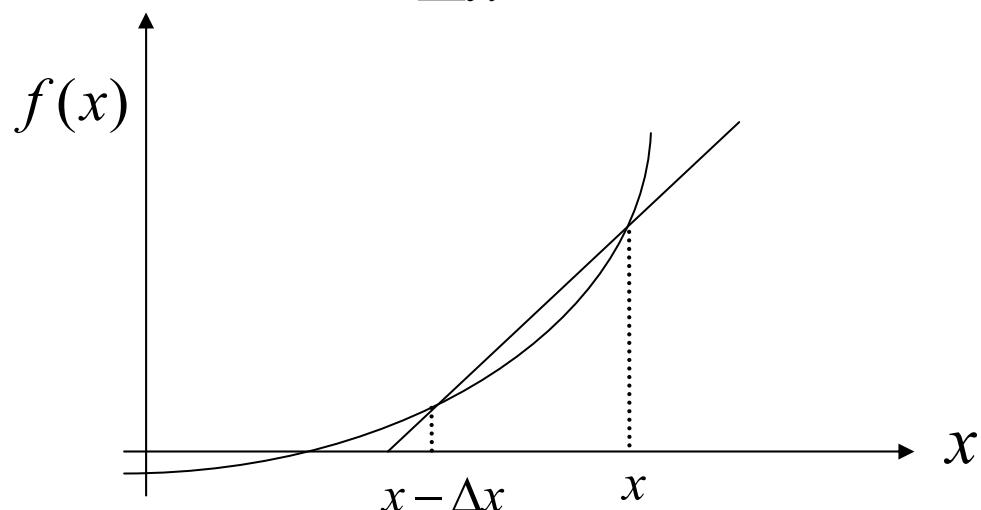
$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x) - f(x - \Delta x)}{\Delta x}$$



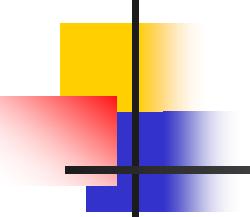


Backward Divided Difference

$$f'(x) \approx \frac{f(x) - f(x - \Delta x)}{\Delta x}$$



$$f'(x_i) \approx \frac{f(x_i) - f(x_i - \Delta x)}{\Delta x}$$



Example

Example:

The velocity of a rocket is given by

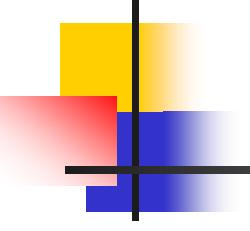
$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where v given in m/s and t is given in seconds. Use backward difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16s$. Use a step size of $\Delta t = 2s$.

Solution:

$$a(t_i) \cong \frac{v(t_i) - v(t_{i-1})}{\Delta t}$$

$$t_i = 16$$



Example (contd.)

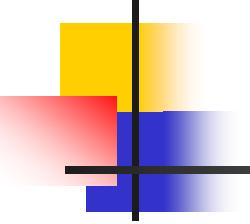
$$\Delta t = 2$$

$$t_{i-1} = t_i - \Delta t = 16 - 2 = 14$$

$$a(16) = \frac{v(16) - v(14)}{2}$$

$$v(16) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) = 392.07 \text{ m/s}$$

$$v(14) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14) = 334.24 \text{ m/s}$$



Example (contd.)

Hence

$$a(16) = \frac{v(16) - v(14)}{2} = 392.07 - 334.24 = 28.915 \text{ m/s}^2$$

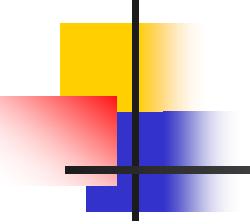
The exact value of $a(16)$ can be calculated by differentiating

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

as

$$a(t) = \frac{d}{dt}[v(t)] = \frac{-4040 - 29.4t}{-200 + 3t}$$

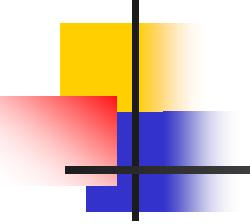
$$a(16) = 29.674 \text{ m/s}^2$$



Example (contd.)

The absolute relative true error is

$$\begin{aligned} |\varepsilon_t| &= \left| \frac{\text{TrueValue} - \text{ApproximateValue}}{\text{TrueValue}} \right| \times 100 \\ &= \left| \frac{29.674 - 28.915}{29.674} \right| \times 100 \\ &= 2.557\% \end{aligned}$$



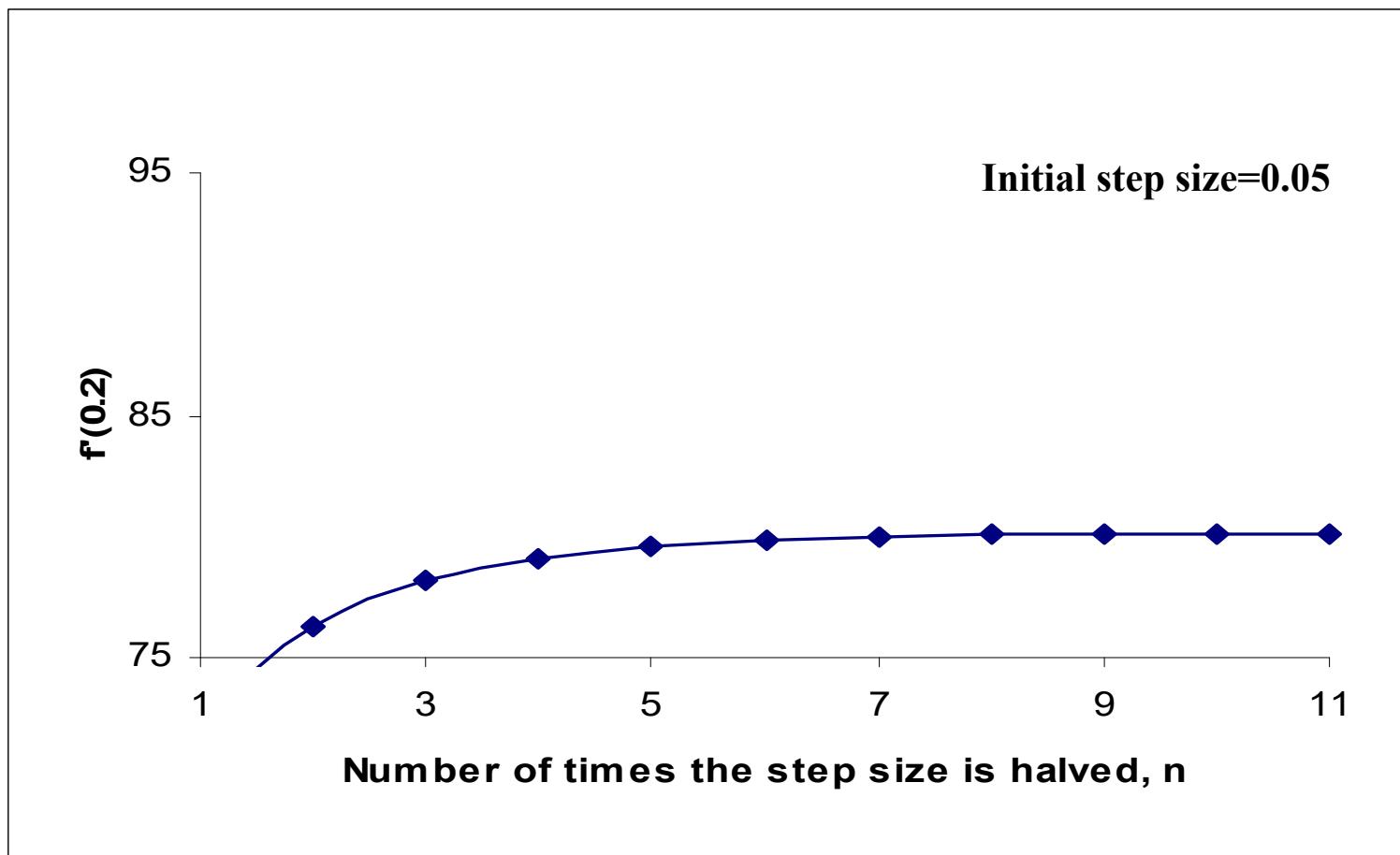
Effect Of Step Size

$$f(x) = 9e^{4x}$$

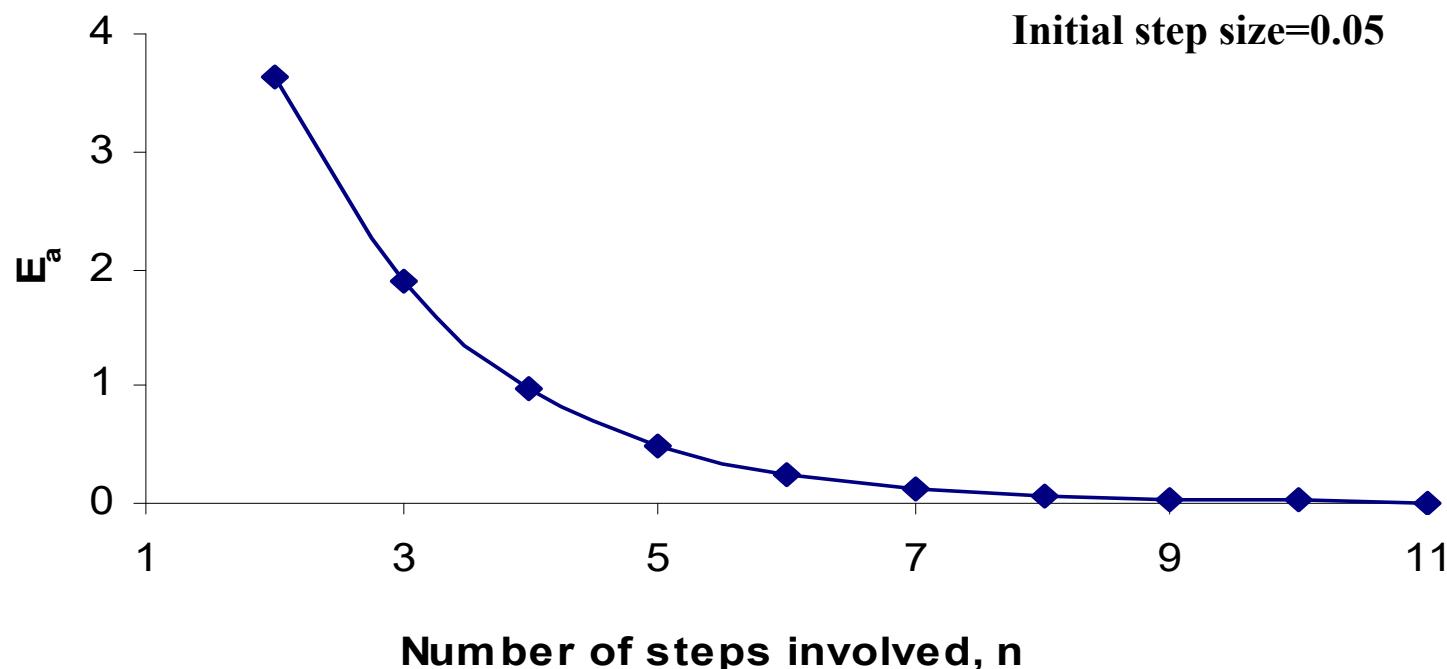
Value of $f'(0.2)$ Using backward Divided difference method.

h	$f'(0.2)$	E_a	$ \varepsilon_a \%$	Significant digits	E_t	$ \varepsilon_t \%$
0.05	72.61598				7.50349	9.365377
0.025	76.24376	3.627777	4.758129	1	3.87571	4.837418
0.0125	78.14946	1.905697	2.438529	1	1.97002	2.458849
0.00625	79.12627	0.976817	1.234504	1	0.99320	1.239648
0.003125	79.62081	0.494533	0.62111	1	0.49867	0.622404
0.001563	79.86962	0.248814	0.311525	2	0.24985	0.31185
0.000781	79.99442	0.124796	0.156006	2	0.12506	0.156087
0.000391	80.05691	0.062496	0.078064	2	0.06256	0.078084
0.000195	80.08818	0.031272	0.039047	3	0.03129	0.039052
9.77E-05	80.10383	0.015642	0.019527	3	0.01565	0.019529
4.88E-05	80.11165	0.007823	0.009765	3	0.00782	0.009765

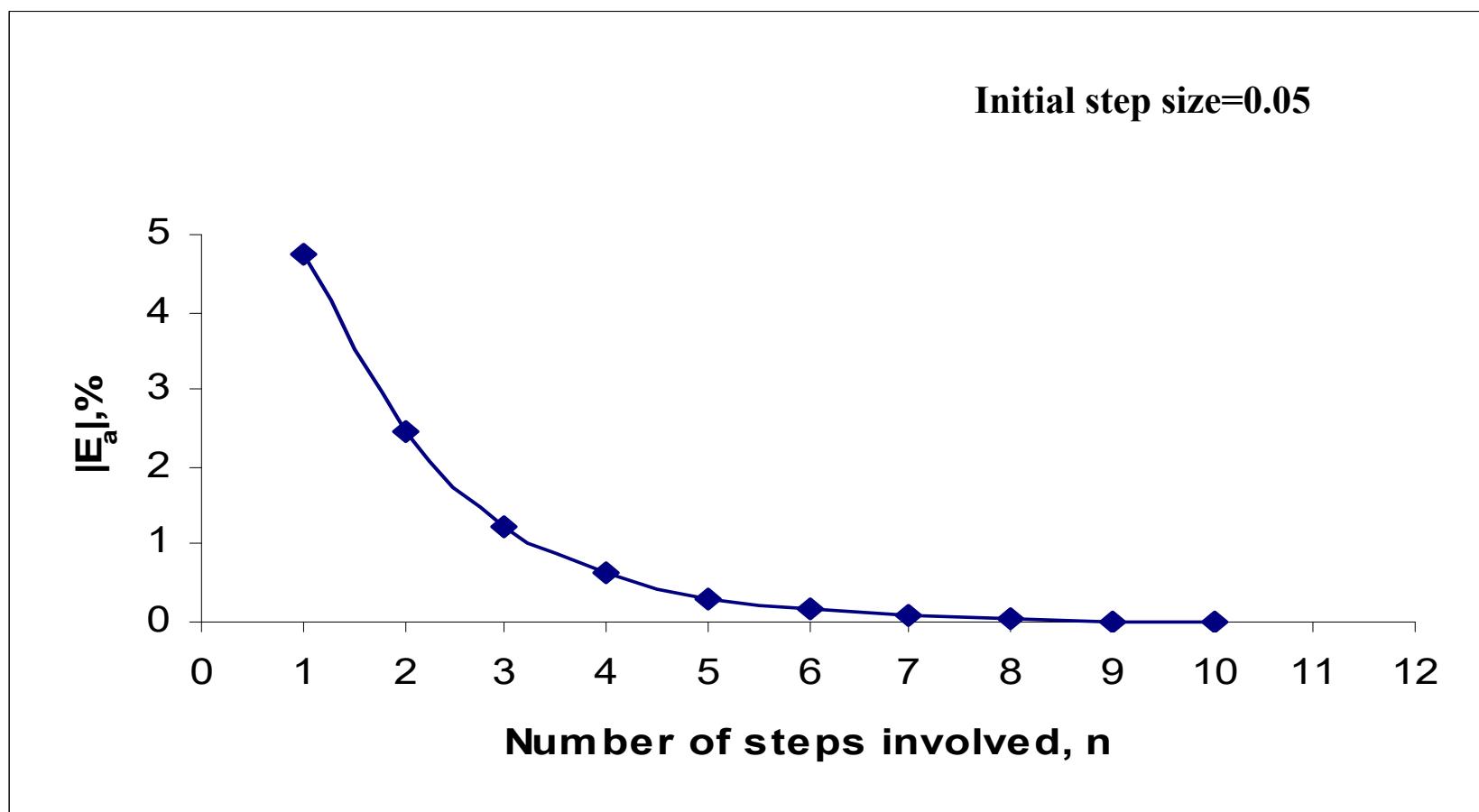
Effect of Step Size in Backward Divided Difference Method



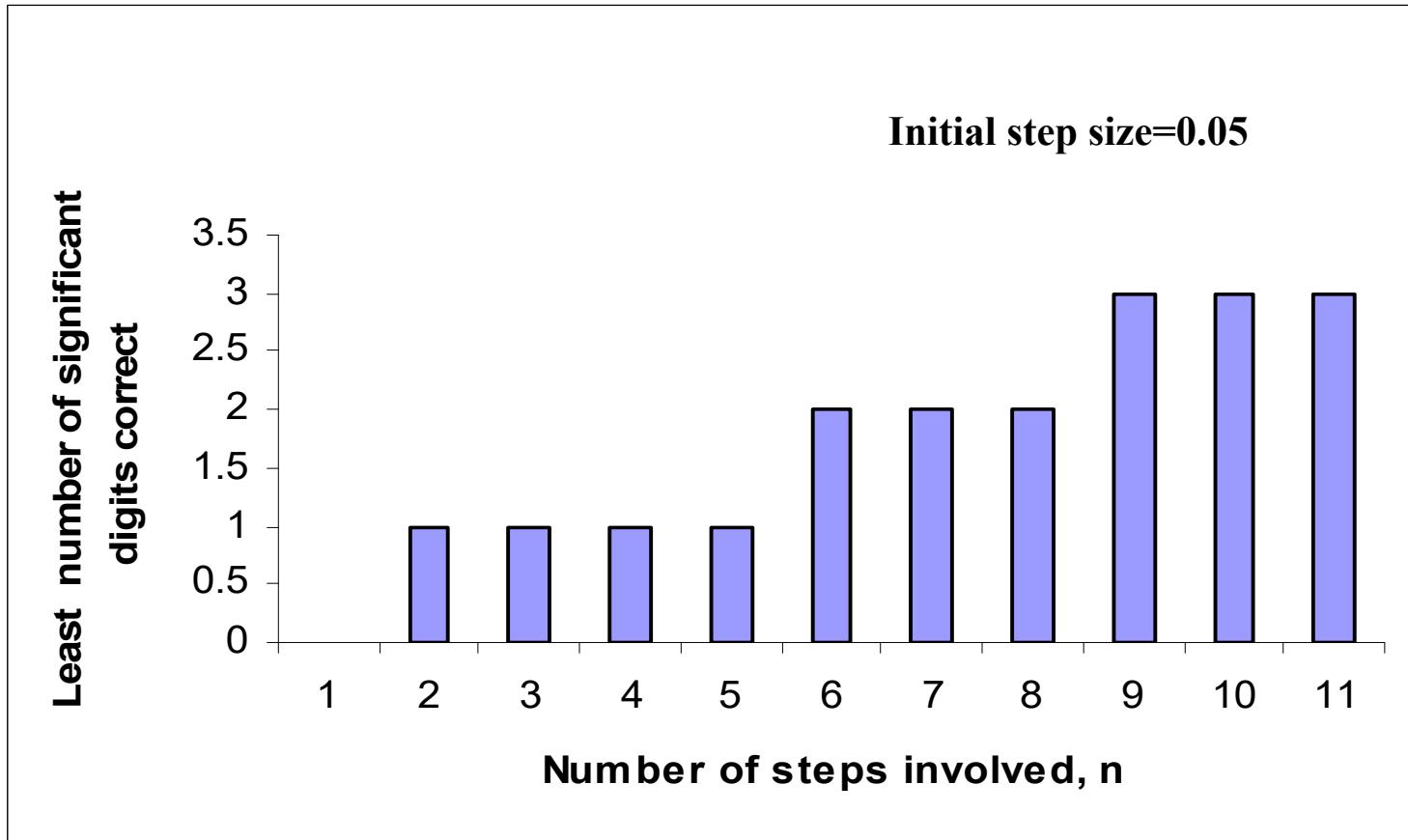
Effect of Step Size on Approximate Error



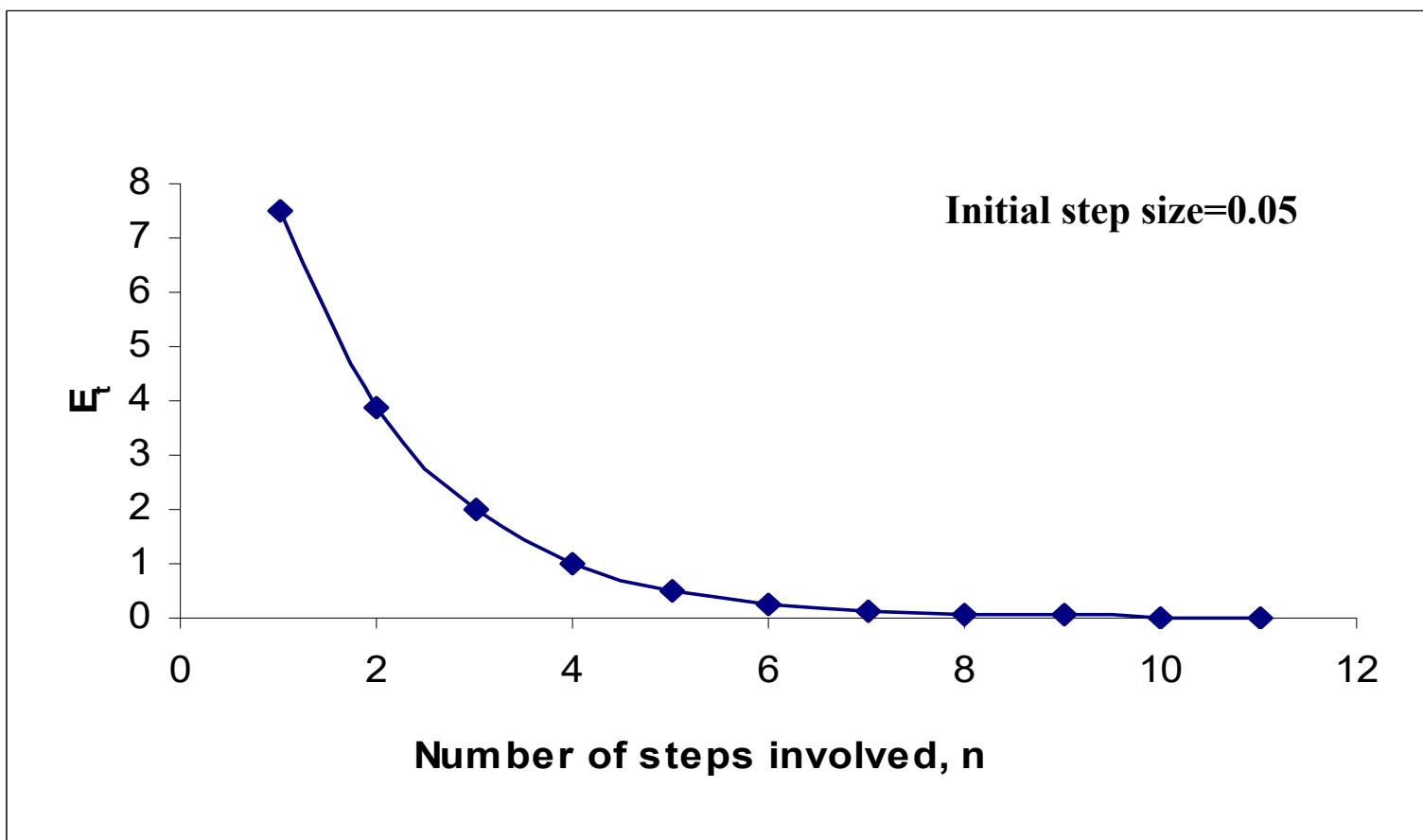
Effect of Step Size on Absolute Relative Approximate Error



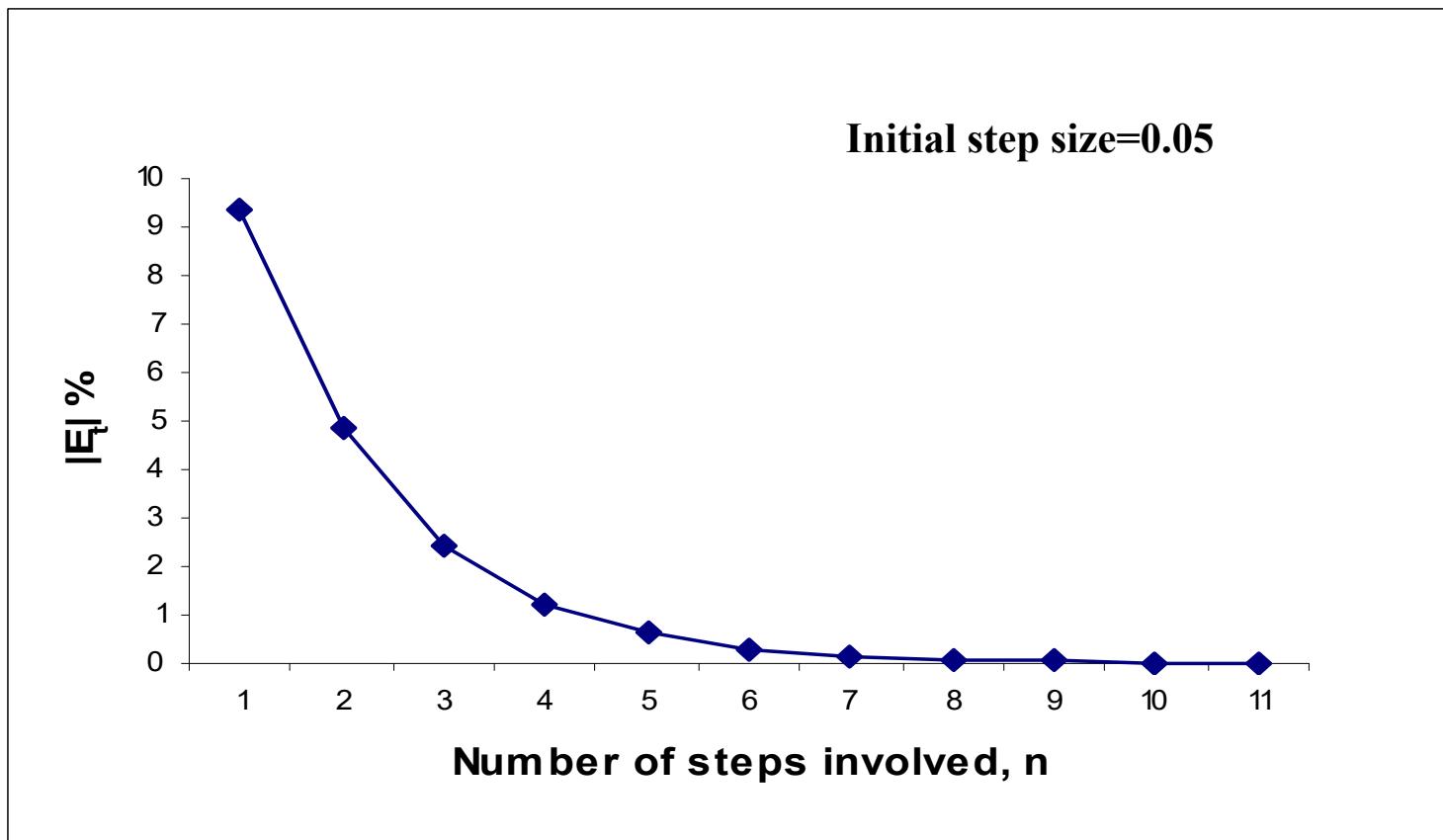
Effect of Step Size on Least Number of Significant Digits Correct

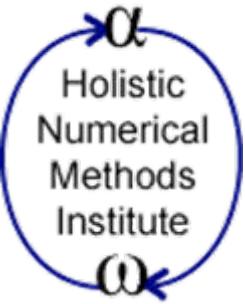


Effect of Step Size on True Error



Effect of Step Size on Absolute Relative True Error





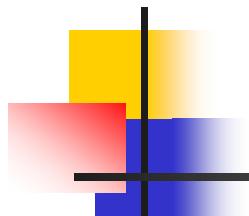
Central Divided Difference



Topic: Differentiation

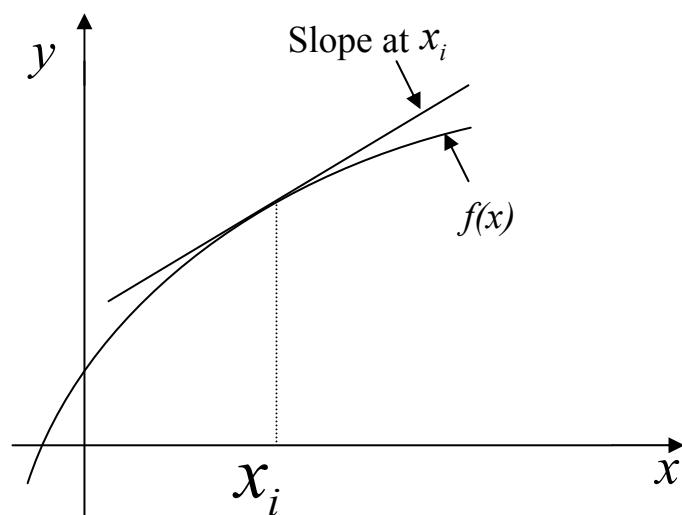
Major: General Engineering

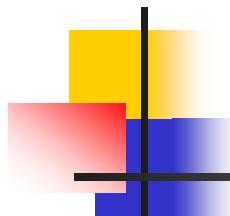
Authors: Autar Kaw, Sri Harsha Garapati



Definition

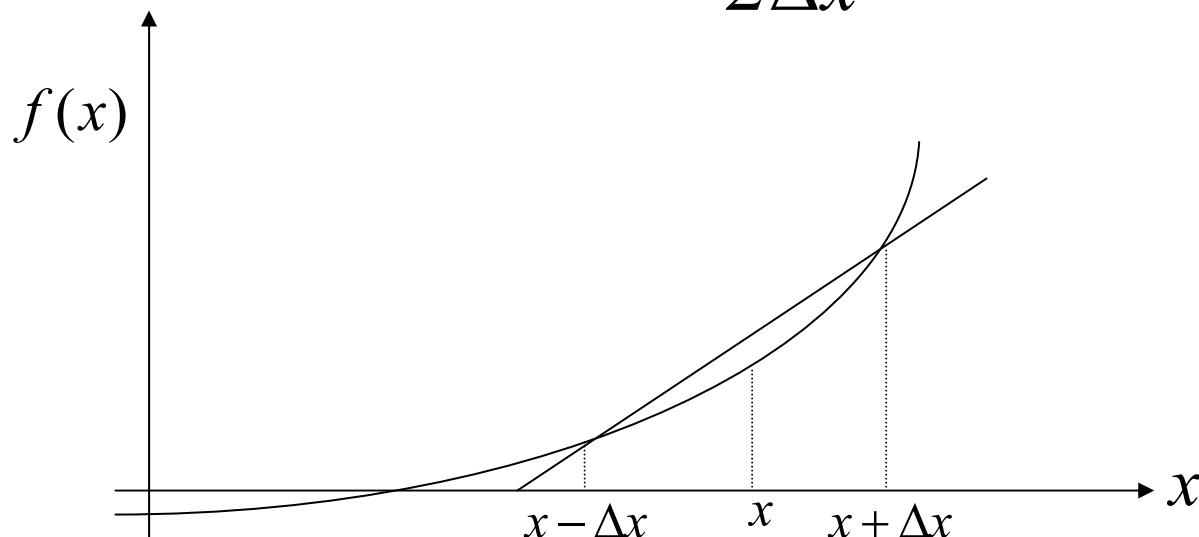
$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x) - f(x - \Delta x)}{\Delta x}$$



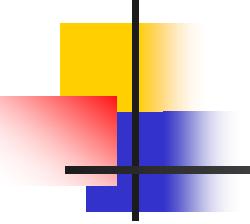


Central Divided Difference

$$f'(x) \cong \frac{f(x + \Delta x) - f(x - \Delta x)}{2\Delta x}$$



$$f'(x_i) \cong \frac{f(x_i + \Delta x) - f(x_i - \Delta x)}{2\Delta x}$$



Example

Example:

The velocity of a rocket is given by

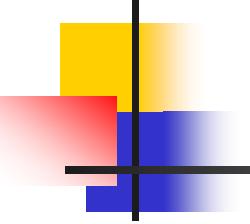
$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where v given in m/s and t is given in seconds. Use central difference approximation of the first derivative of $v(t)$ to calculate the acceleration at $t = 16s$. Use a step size of $\Delta t = 2s$.

Solution:

$$a(t_i) \approx \frac{v(t_{i+1}) - v(t_{i-1})}{2\Delta t}$$

$$t_i = 16$$



Example (contd.)

$$\Delta t = 2$$

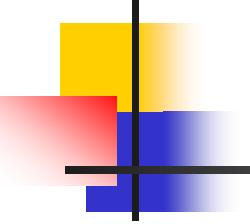
$$t_{i+1} = t_i + \Delta t = 16 + 2 = 18$$

$$t_{i-1} = t_i - \Delta t = 16 - 2 = 14$$

$$a(16) = \frac{\nu(18) - \nu(14)}{2(2)} = \frac{\nu(18) - \nu(14)}{4}$$

$$\nu(18) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) = 453.02 \text{ m/s}$$

$$\nu(14) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14) = 334.24 \text{ m/s}$$



Example (contd.)

Hence

$$a(16) = \frac{v(18) - v(14)}{4} = 453.02 - 334.24 = 29.695 \text{ m/s}^2$$

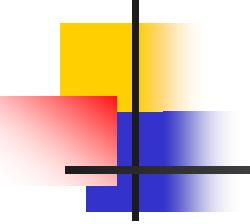
The exact value of $a(16)$ can be calculated by differentiating

$$v(t) = 2000 \ln \left[\frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

as

$$a(t) = \frac{d}{dt}[v(t)] = \frac{-4040 - 29.4t}{-200 + 3t}$$

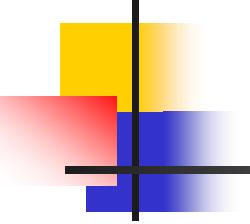
$$a(16) = 29.674 \text{ m/s}^2$$



Example (contd.)

The absolute relative true error is

$$\begin{aligned} |\varepsilon_t| &= \left| \frac{\text{TrueValue} - \text{ApproximateValue}}{\text{TrueValue}} \right| \times 100 \\ &= \left| \frac{29.674 - 29.695}{29.674} \right| \times 100 \\ &= 0.070769 \% \end{aligned}$$



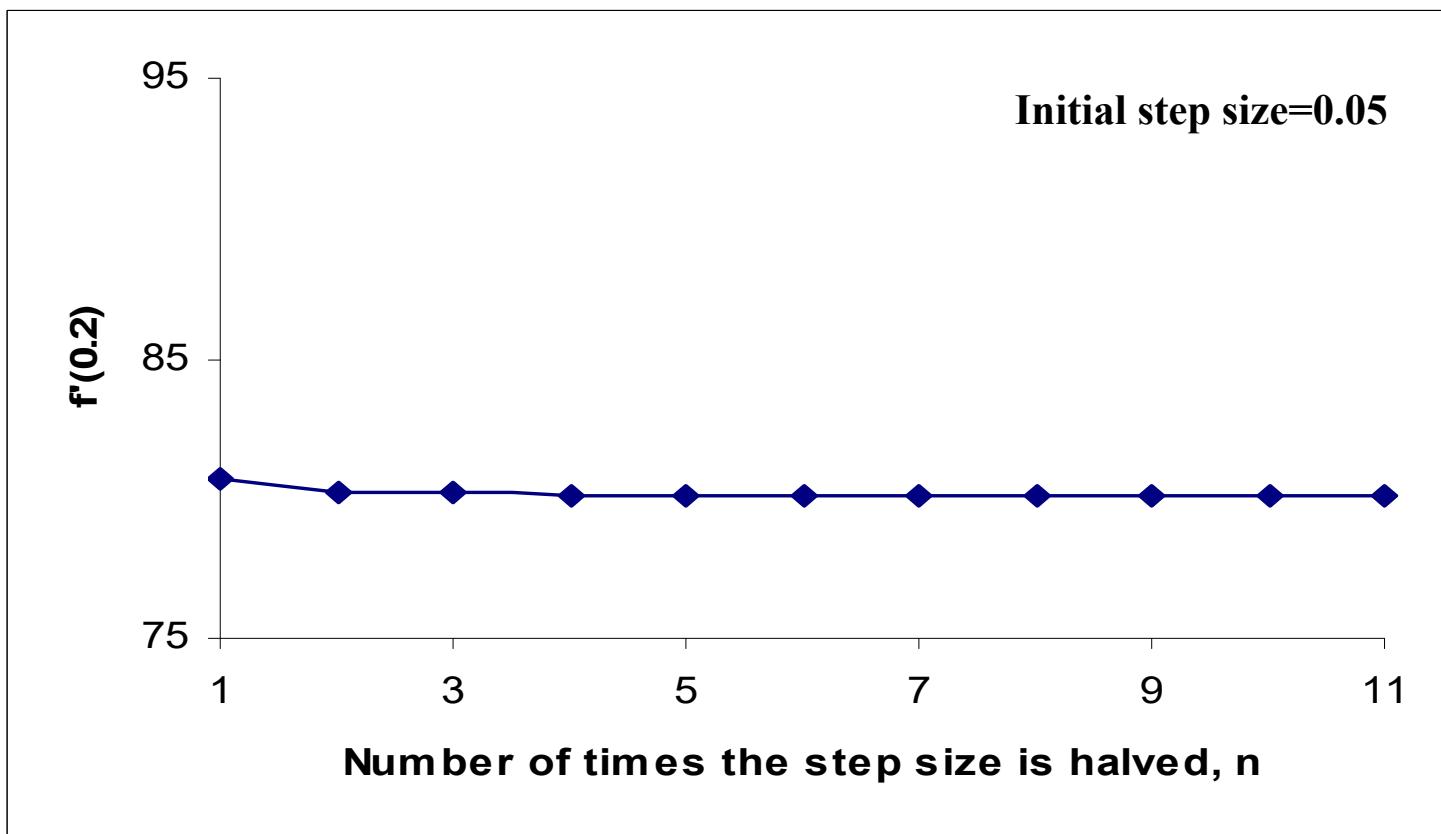
Effect Of Step Size

$$f(x) = 9e^{4x}$$

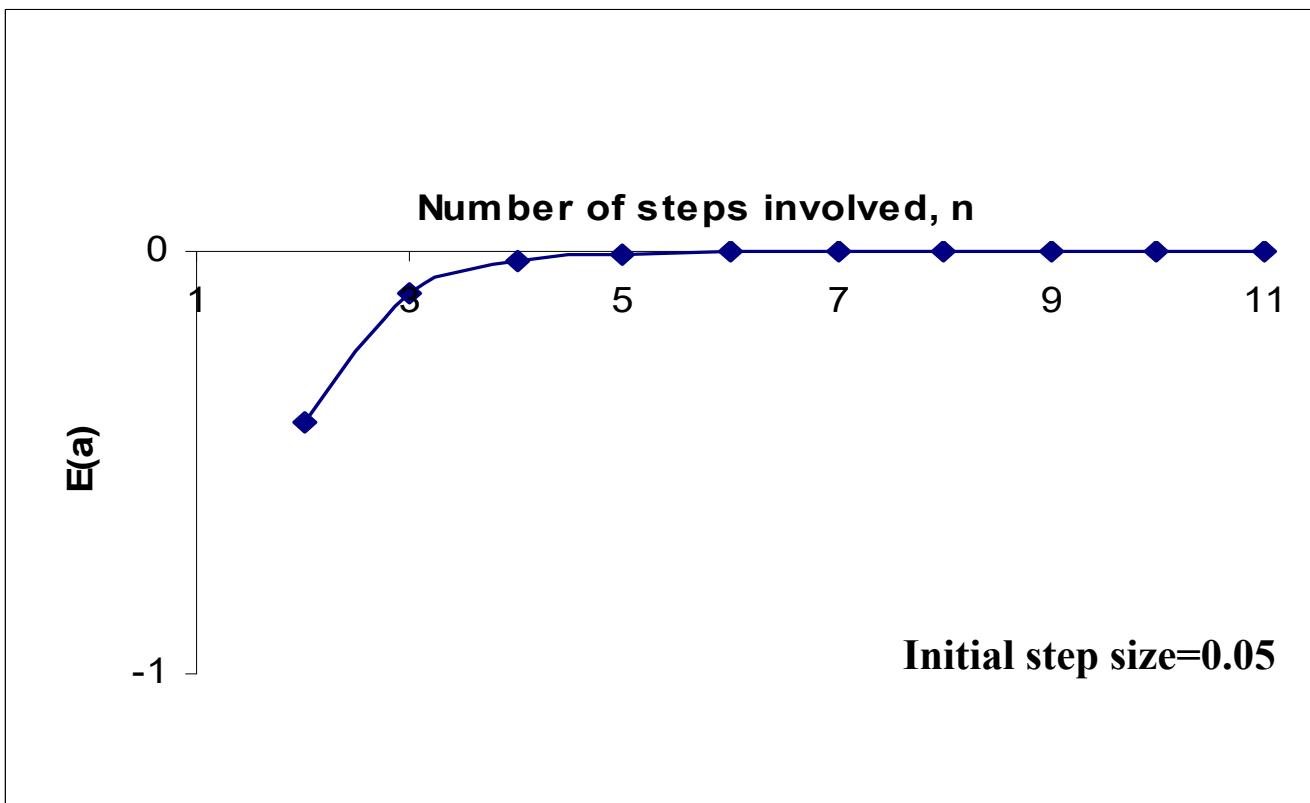
Value of $f'(0.2)$ Using Central Divided Difference difference method.

h	$f'(0.2)$	E_a	$ E_a \%$	Significant digits	E_t	$ E_t \%$
0.05	80.65467				-0.53520	0.668001
0.025	80.25307	-0.4016	0.500417	1	-0.13360	0.16675
0.0125	80.15286	-0.100212	0.125026	2	-0.03339	0.041672
0.00625	80.12782	-0.025041	0.031252	3	-0.00835	0.010417
0.003125	80.12156	-0.00626	0.007813	3	-0.00209	0.002604
0.001563	80.12000	-0.001565	0.001953	4	-0.00052	0.000651
0.000781	80.11960	-0.000391	0.000488	5	-0.00013	0.000163
0.000391	80.11951	-9.78E-05	0.000122	5	-0.00003	4.07E-05
0.000195	80.11948	-2.45E-05	3.05E-05	6	-0.00001	1.02E-05
9.77E-05	80.11948	-6.11E-06	7.63E-06	6	0.00000	2.54E-06
4.88E-05	80.11947	-1.53E-06	1.91E-06	7	0.00000	6.36E-07

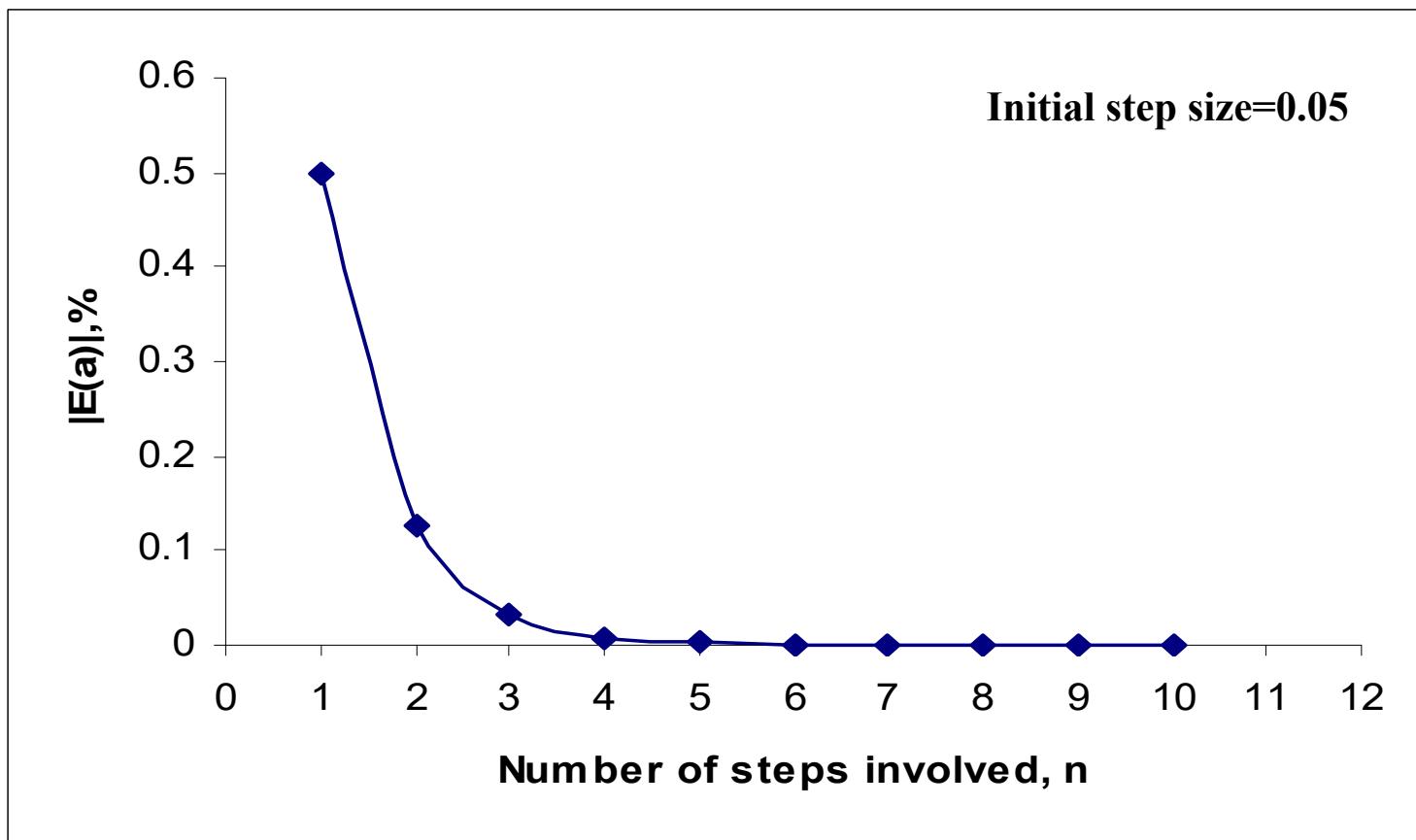
Effect of Step Size in Central Divided Difference Method



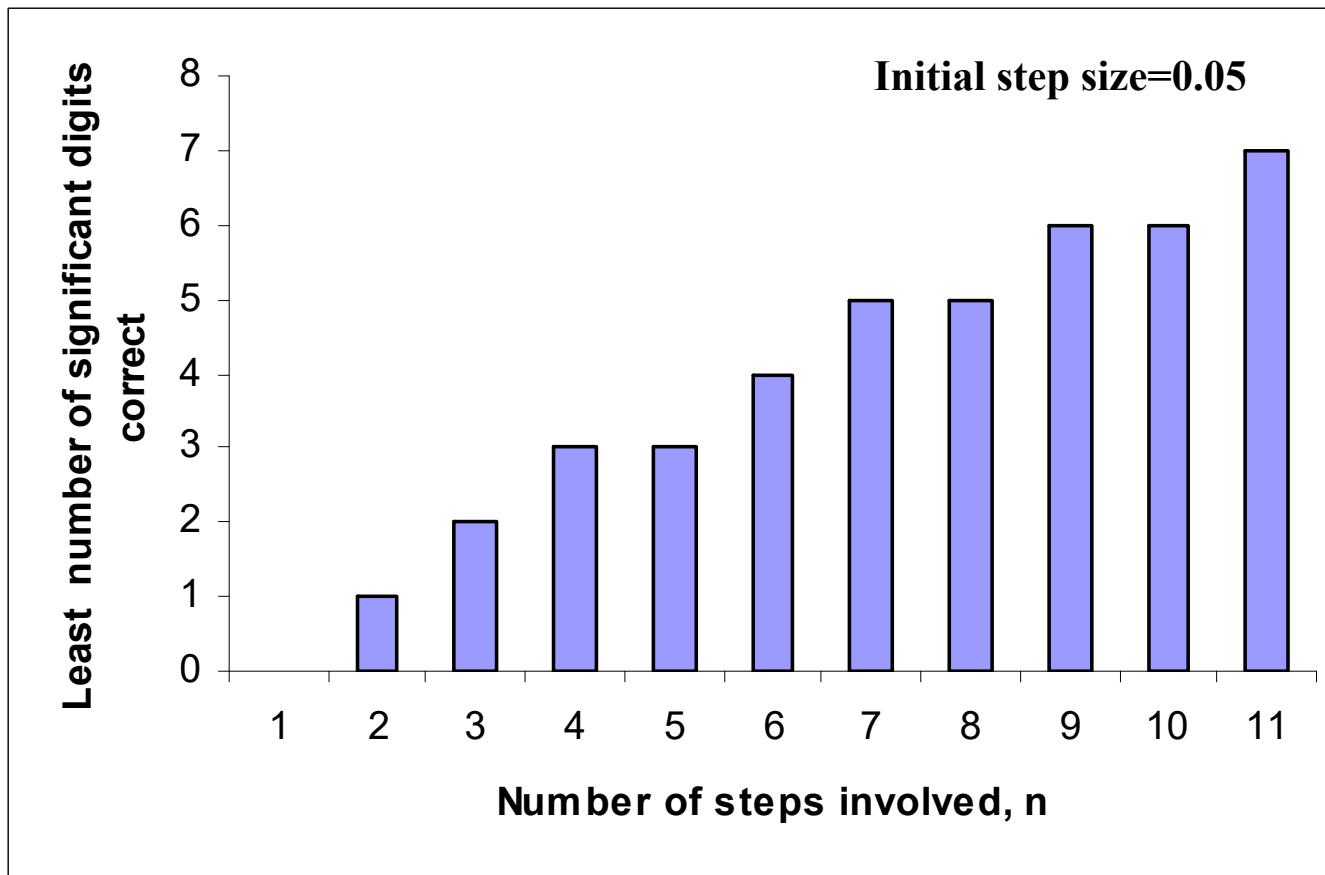
Effect of Step Size on Approximate Error



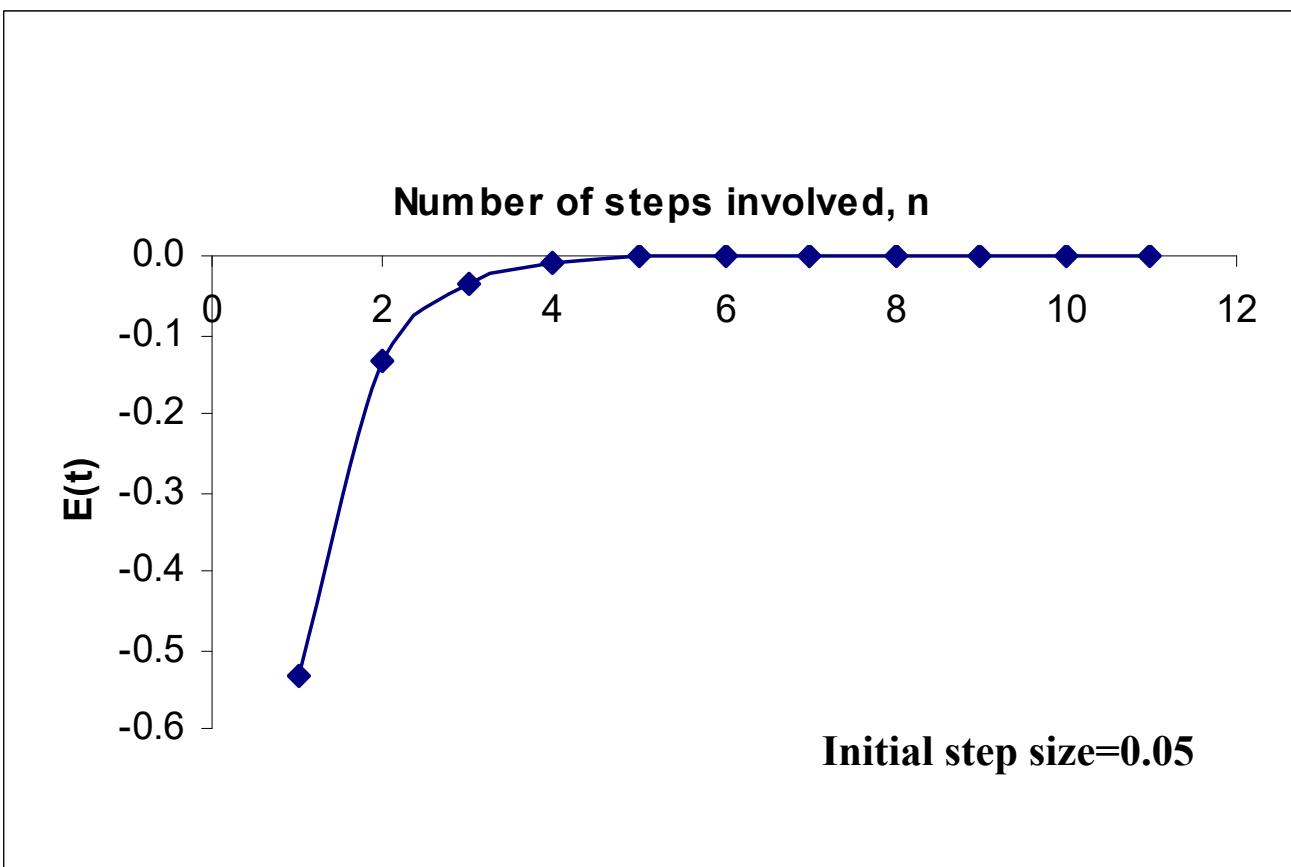
Effect of Step Size on Absolute Relative Approximate Error



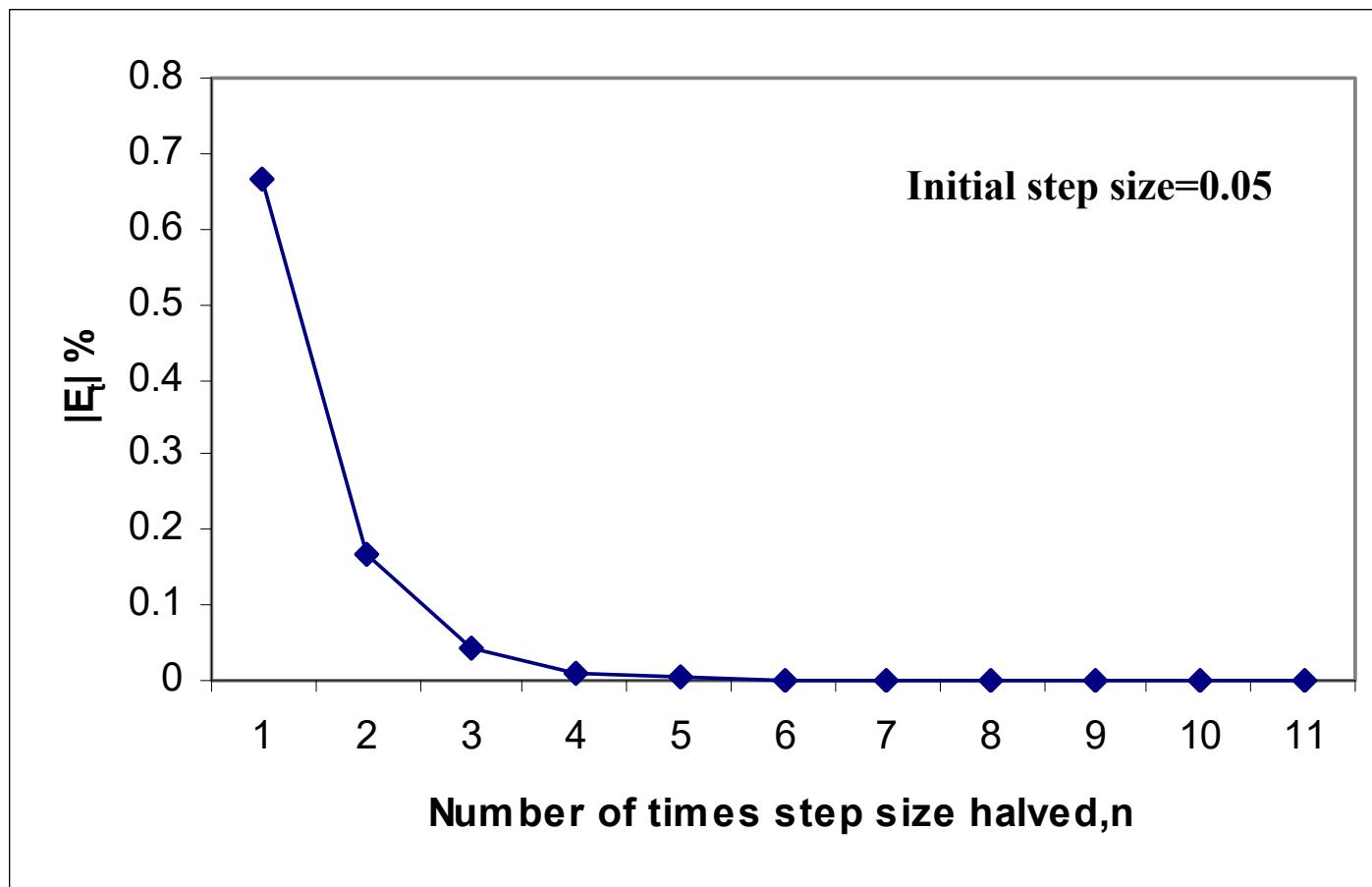
Effect of Step Size on Least Number of Significant Digits Correct



Effect of Step Size on True Error



Effect of Step Size on Absolute Relative True Error



Bisection Method

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Bisection Method

<http://numericalmethods.eng.usf.edu>

Basis of Bisection Method

Theorem An equation $f(x)=0$, where $f(x)$ is a real continuous function, has at least one root between x_l and x_u if $f(x_l) f(x_u) < 0$.

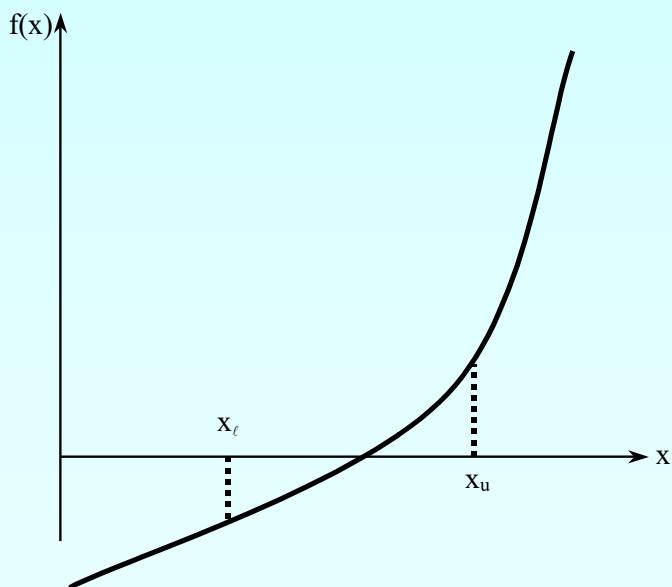


Figure 1 At least one root exists between the two points if the function is real, continuous, and changes sign.

Basis of Bisection Method

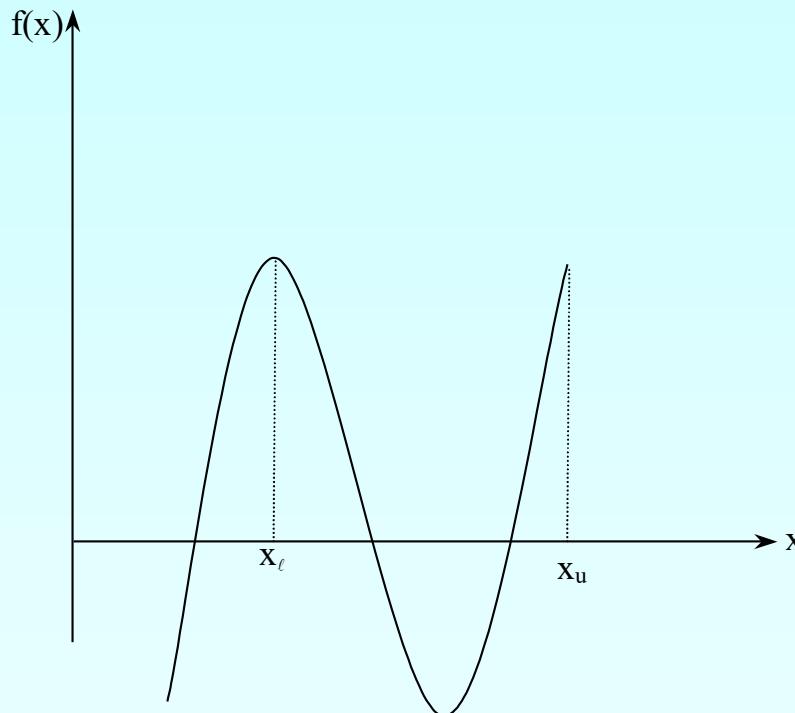


Figure 2 If function $f(x)$ does not change sign between two points, roots of the equation $f(x)=0$ may still exist between the two points.

Basis of Bisection Method

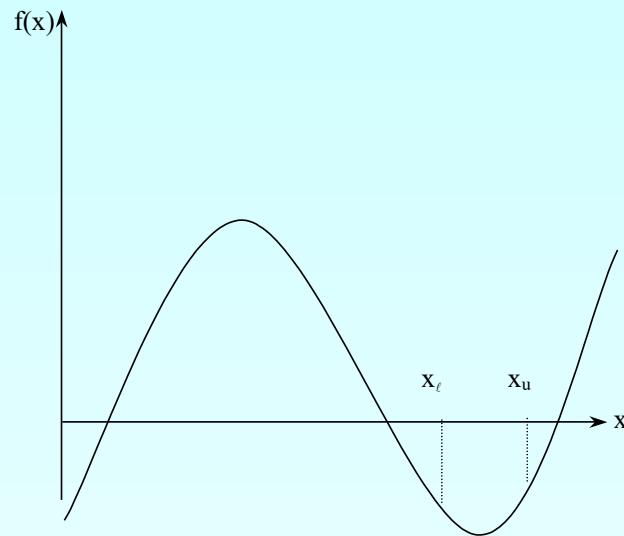
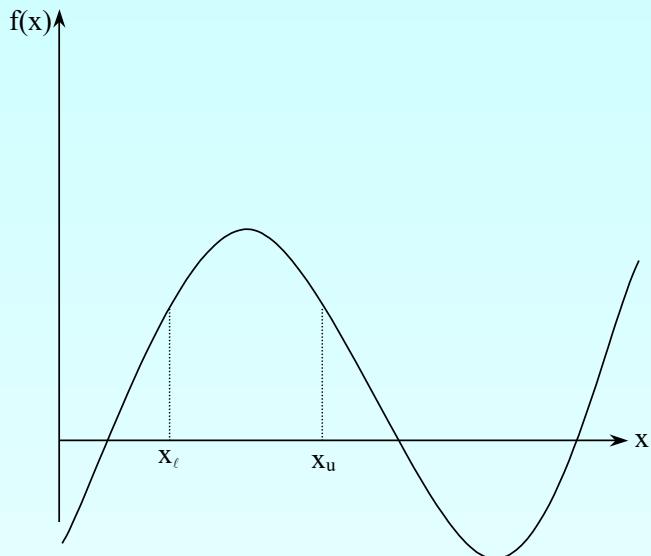


Figure 3 If the function $f(x)$ does not change sign between two points, there may not be any roots for the equation $f(x)=0$ between the two points.

Basis of Bisection Method

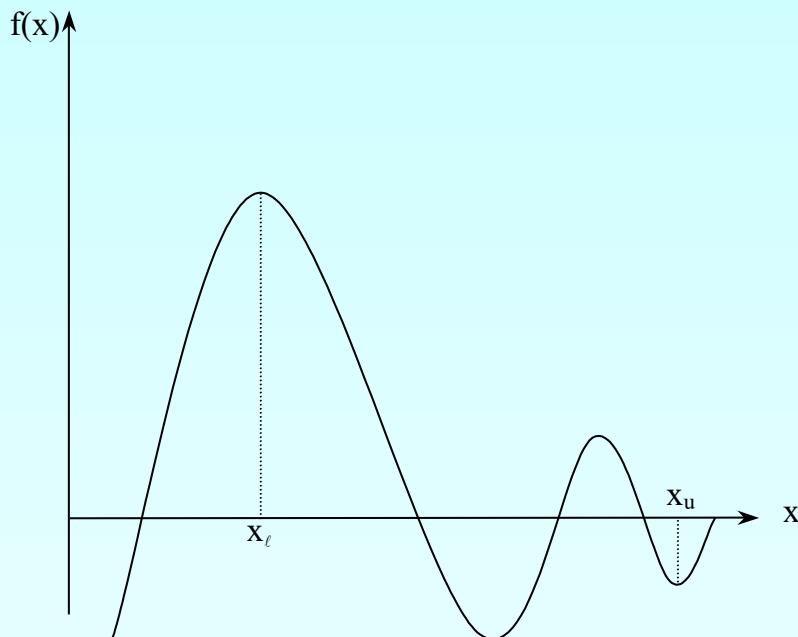


Figure 4 If the function $f(x)$ changes sign between two points, more than one root for the equation $f(x)=0$ may exist between the two points.

Algorithm for Bisection Method

Step 1

Choose x_ℓ and x_u as two guesses for the root such that $f(x_\ell) f(x_u) < 0$, or in other words, $f(x)$ changes sign between x_ℓ and x_u . This was demonstrated in Figure 1.

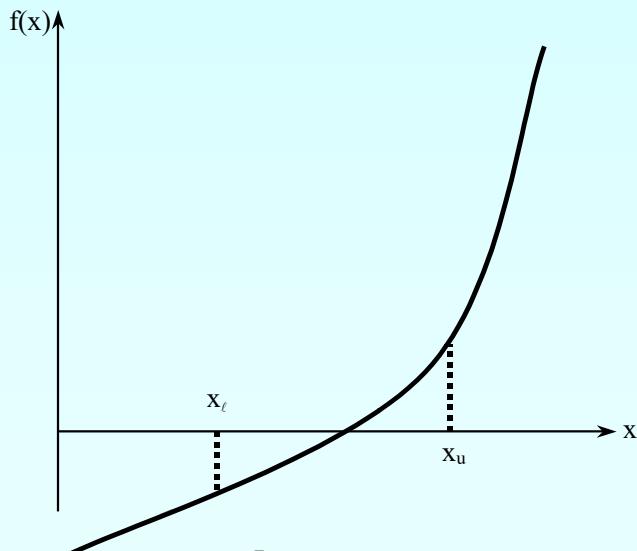


Figure 1

Step 2

Estimate the root, x_m of the equation $f(x) = 0$ as the mid point between x_ℓ and x_u as

$$x_m = \frac{x_\ell + x_u}{2}$$

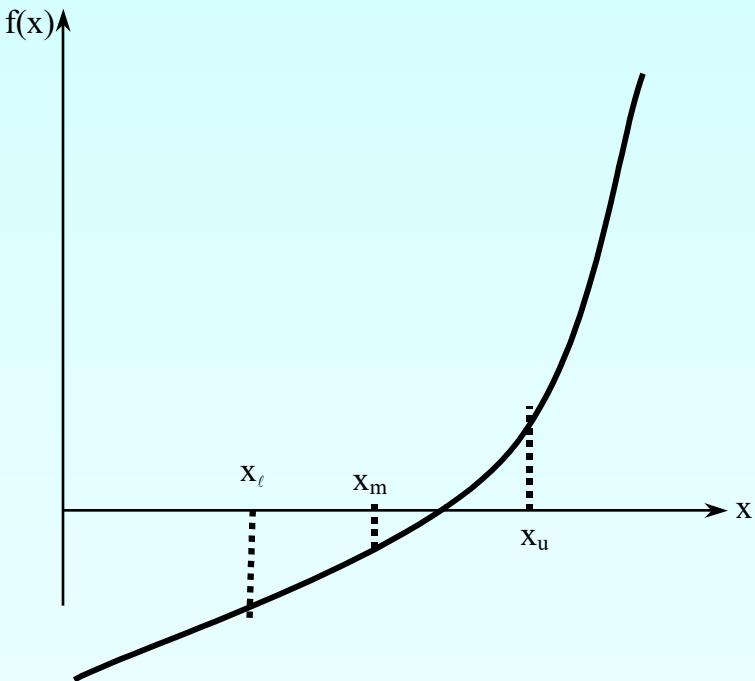


Figure 5 Estimate of x_m

Step 3

Now check the following

- a) If $f(x_l)f(x_m) < 0$, then the root lies between x_l and x_m ; then $x_l = x_l$; $x_u = x_m$.
- b) If $f(x_l)f(x_m) > 0$, then the root lies between x_m and x_u ; then $x_l = x_m$; $x_u = x_u$.
- c) If $f(x_l)f(x_m) = 0$; then the root is x_m . Stop the algorithm if this is true.

Step 4

Find the new estimate of the root

$$x_m = \frac{x_\ell + x_u}{2}$$

Find the absolute relative approximate error

$$|\epsilon_a| = \left| \frac{x_m^{new} - x_m^{old}}{x_m^{new}} \right| \times 100$$

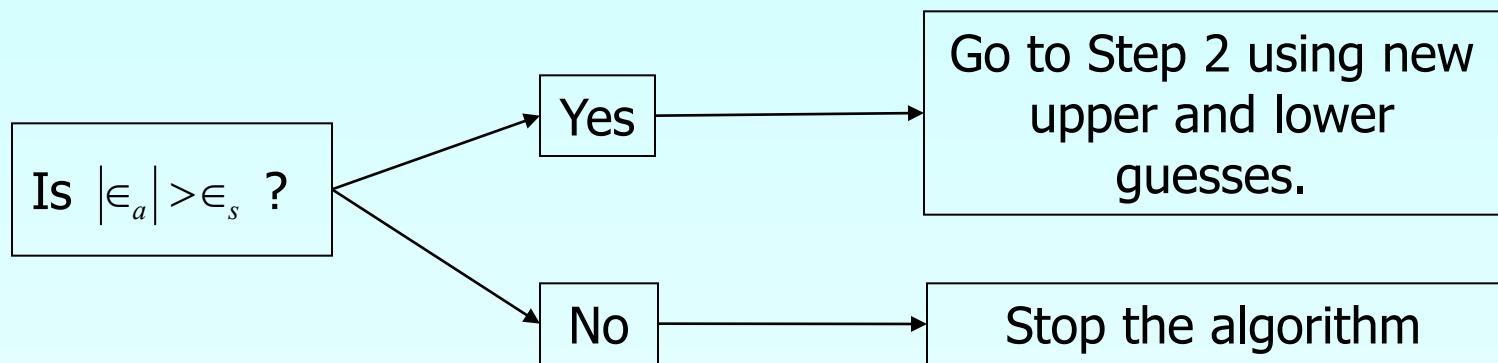
where

x_m^{old} = previous estimate of root

x_m^{new} = current estimate of root

Step 5

Compare the absolute relative approximate error $|\epsilon_a|$ with the pre-specified error tolerance ϵ_s .



Note one should also check whether the number of iterations is more than the maximum number of iterations allowed. If so, one needs to terminate the algorithm and notify the user about it.

Example 1

You are working for 'DOWN THE TOILET COMPANY' that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

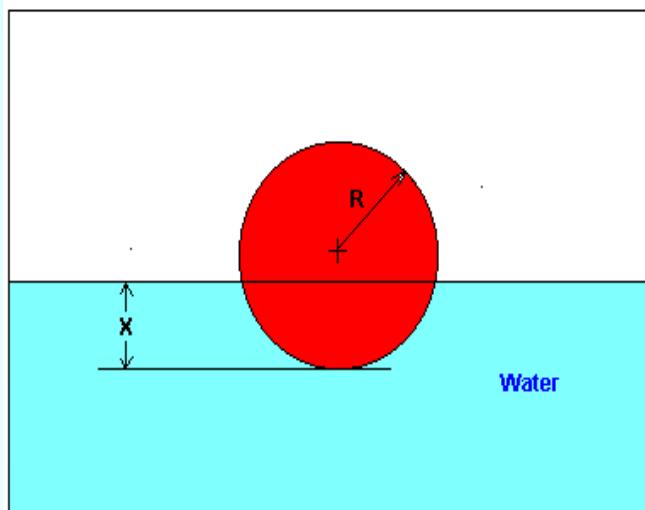


Figure 6 Diagram of the floating ball

Example 1 Cont.

The equation that gives the depth x to which the ball is submerged under water is given by

$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

- Use the bisection method of finding roots of equations to find the depth x to which the ball is submerged under water. Conduct three iterations to estimate the root of the above equation.
- Find the absolute relative approximate error at the end of each iteration, and the number of significant digits at least correct at the end of each iteration.

Example 1 Cont.

From the physics of the problem, the ball would be submerged between $x = 0$ and $x = 2R$,

where R = radius of the ball,

that is

$$0 \leq x \leq 2R$$

$$0 \leq x \leq 2(0.055)$$

$$0 \leq x \leq 0.11$$

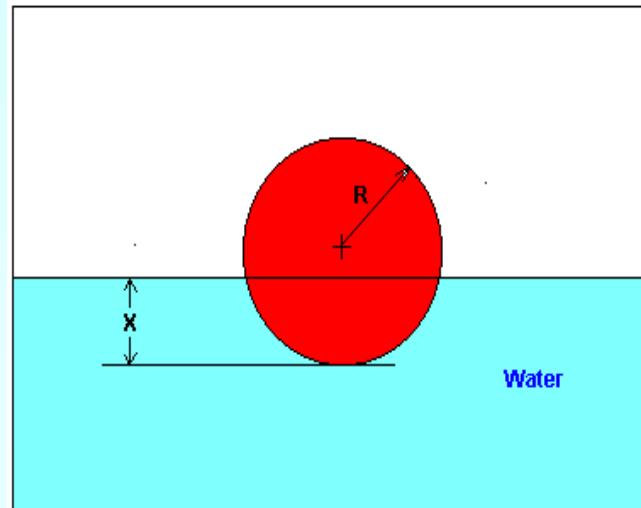


Figure 6 Diagram of the floating ball

Example 1 Cont.

Solution

To aid in the understanding of how this method works to find the root of an equation, the graph of $f(x)$ is shown to the right,

where

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

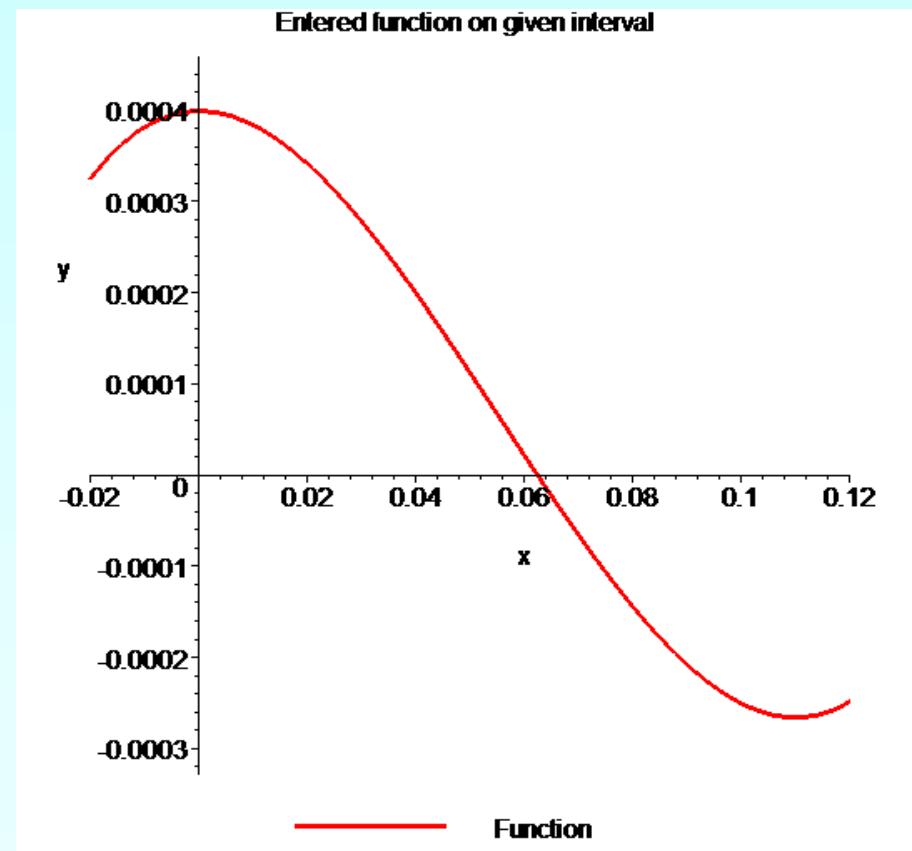


Figure 7 Graph of the function $f(x)$

Example 1 Cont.

Let us assume

$$x_l = 0.00$$

$$x_u = 0.11$$

Check if the function changes sign between x_l and x_u .

$$f(x_l) = f(0) = (0)^3 - 0.165(0)^2 + 3.993 \times 10^{-4} = 3.993 \times 10^{-4}$$

$$f(x_u) = f(0.11) = (0.11)^3 - 0.165(0.11)^2 + 3.993 \times 10^{-4} = -2.662 \times 10^{-4}$$

Hence

$$f(x_l)f(x_u) = f(0)f(0.11) = (3.993 \times 10^{-4})(-2.662 \times 10^{-4}) < 0$$

So there is at least one root between x_l and x_u , that is between 0 and 0.11

Example 1 Cont.

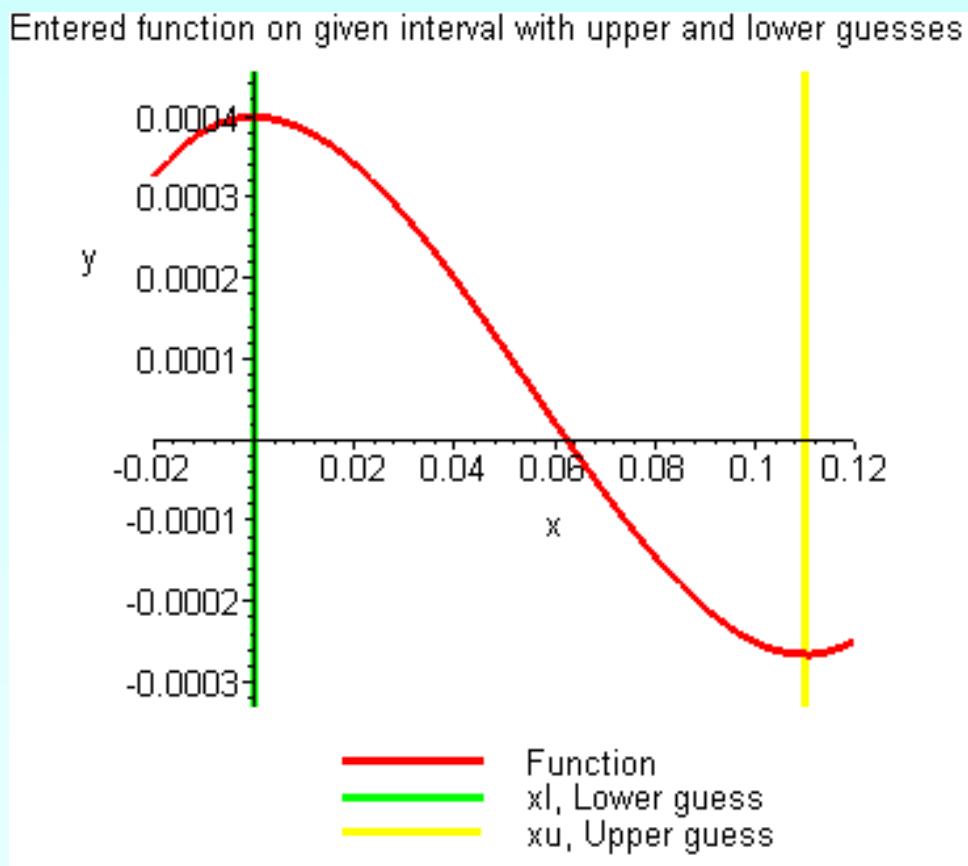


Figure 8 Graph demonstrating sign change between initial limits

Example 1 Cont.

Iteration 1

The estimate of the root is $x_m = \frac{x_l + x_u}{2} = \frac{0 + 0.11}{2} = 0.055$

$$f(x_m) = f(0.055) = (0.055)^3 - 0.165(0.055)^2 + 3.993 \times 10^{-4} = 6.655 \times 10^{-5}$$
$$f(x_l)f(x_m) = f(0)f(0.055) = (3.993 \times 10^{-4})(6.655 \times 10^{-5}) > 0$$

Hence the root is bracketed between x_m and x_u , that is, between 0.055 and 0.11. So, the lower and upper limits of the new bracket are

$$x_l = 0.055, x_u = 0.11$$

At this point, the absolute relative approximate error $|e_a|$ cannot be calculated as we do not have a previous approximation.

Example 1 Cont.

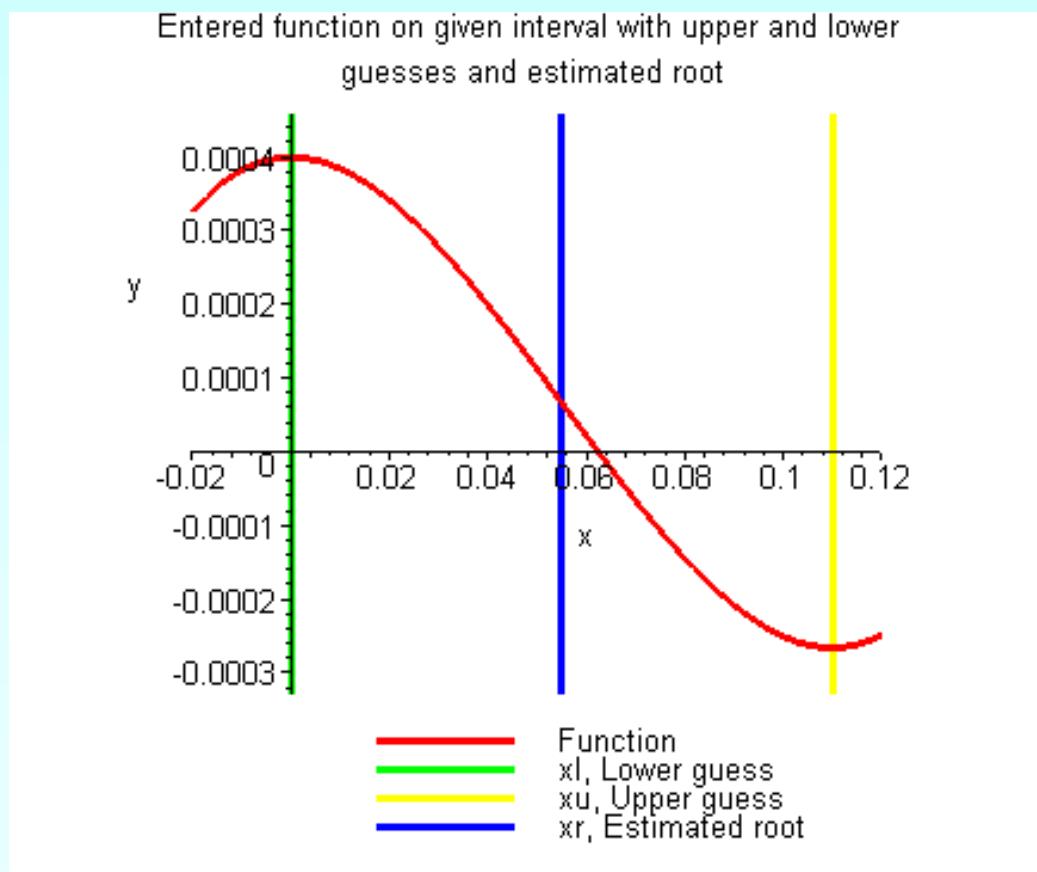


Figure 9 Estimate of the root for Iteration 1

Example 1 Cont.

Iteration 2

The estimate of the root is $x_m = \frac{x_\ell + x_u}{2} = \frac{0.055 + 0.11}{2} = 0.0825$

$$f(x_m) = f(0.0825) = (0.0825)^3 - 0.165(0.0825)^2 + 3.993 \times 10^{-4} = -1.622 \times 10^{-4}$$
$$f(x_l)f(x_m) = f(0.055)f(0.0825) = (-1.622 \times 10^{-4})(6.655 \times 10^{-5}) < 0$$

Hence the root is bracketed between x_ℓ and x_m , that is, between 0.055 and 0.0825. So, the lower and upper limits of the new bracket are

$$x_l = 0.055, x_u = 0.0825$$

Example 1 Cont.

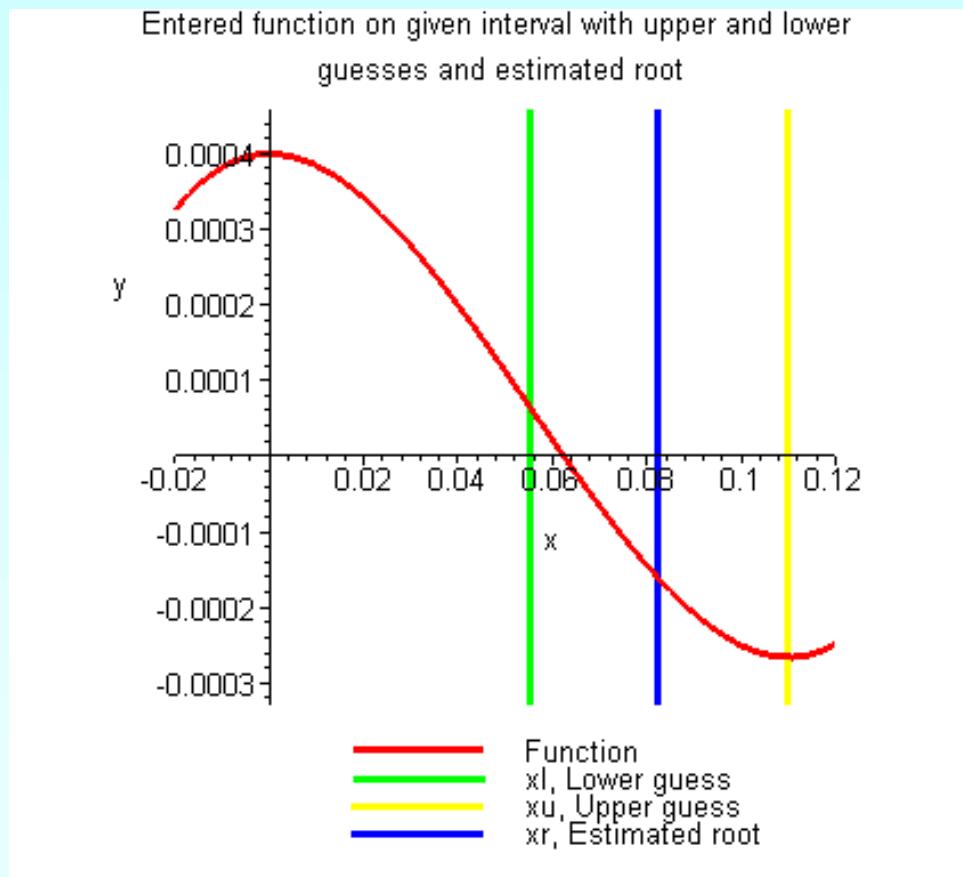


Figure 10 Estimate of the root for Iteration 2

Example 1 Cont.

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 2 is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_m^{new} - x_m^{old}}{x_m^{new}} \right| \times 100 \\ &= \left| \frac{0.0825 - 0.055}{0.0825} \right| \times 100 \\ &= 33.333\% \end{aligned}$$

None of the significant digits are at least correct in the estimate root of $x_m = 0.0825$ because the absolute relative approximate error is greater than 5%.

Example 1 Cont.

Iteration 3

The estimate of the root is $x_m = \frac{x_\ell + x_u}{2} = \frac{0.055 + 0.0825}{2} = 0.06875$

$$f(x_m) = f(0.06875) = (0.06875)^3 - 0.165(0.06875)^2 + 3.993 \times 10^{-4} = -5.563 \times 10^{-5}$$
$$f(x_l)f(x_m) = f(0.055)f(0.06875) = (6.655 \times 10^{-5})(-5.563 \times 10^{-5}) < 0$$

Hence the root is bracketed between x_ℓ and x_m , that is, between 0.055 and 0.06875. So, the lower and upper limits of the new bracket are

$$x_l = 0.055, x_u = 0.06875$$

Example 1 Cont.

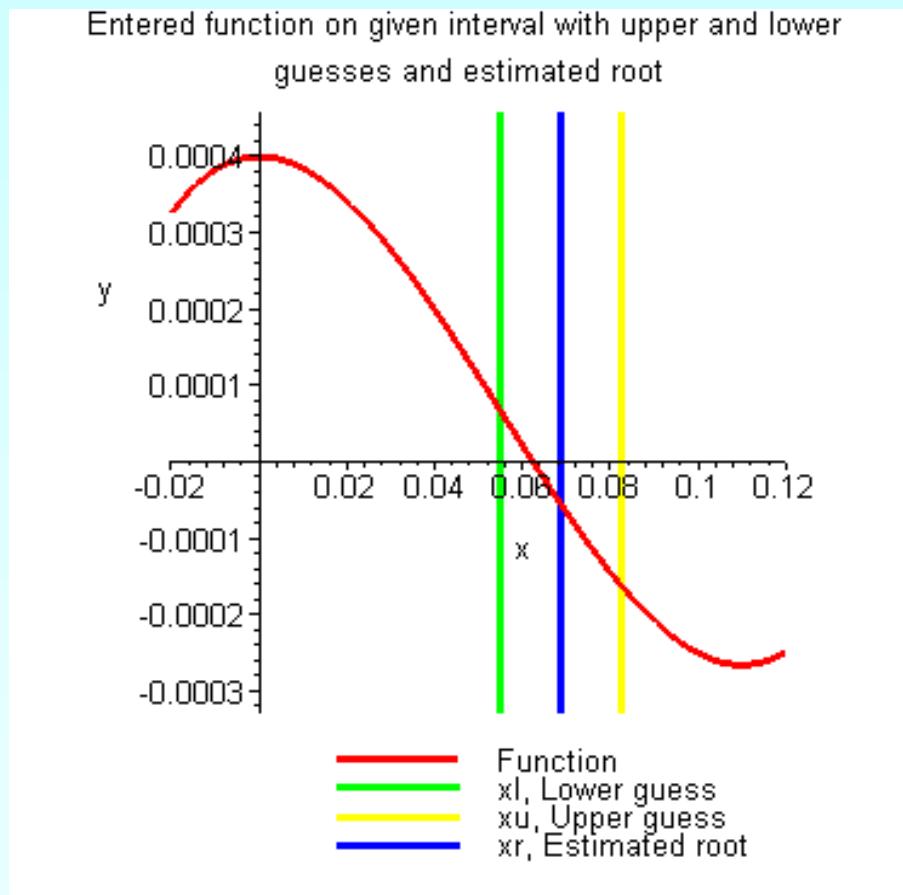


Figure 11 Estimate of the root for Iteration 3

Example 1 Cont.

The absolute relative approximate error $|e_a|$ at the end of Iteration 3 is

$$\begin{aligned}|e_a| &= \left| \frac{x_m^{new} - x_m^{old}}{x_m^{new}} \right| \times 100 \\ &= \left| \frac{0.06875 - 0.0825}{0.06875} \right| \times 100 \\ &= 20\%\end{aligned}$$

Still none of the significant digits are at least correct in the estimated root of the equation as the absolute relative approximate error is greater than 5%.

Seven more iterations were conducted and these iterations are shown in Table 1.

Table 1 Cont.

Table 1 Root of $f(x)=0$ as function of number of iterations for bisection method.

Iteration	x_l	x_u	x_m	$ e_a \%$	$f(x_m)$
1	0.00000	0.11	0.055	-----	6.655×10^{-5}
2	0.055	0.11	0.0825	33.33	-1.622×10^{-4}
3	0.055	0.0825	0.06875	20.00	-5.563×10^{-5}
4	0.055	0.06875	0.06188	11.11	4.484×10^{-6}
5	0.06188	0.06875	0.06531	5.263	-2.593×10^{-5}
6	0.06188	0.06531	0.06359	2.702	-1.0804×10^{-5}
7	0.06188	0.06359	0.06273	1.370	-3.176×10^{-6}
8	0.06188	0.06273	0.0623	0.6897	6.497×10^{-7}
9	0.0623	0.06273	0.06252	0.3436	-1.265×10^{-6}
10	0.0623	0.06252	0.06241	0.1721	-3.0768×10^{-7}

Table 1 Cont.

Hence the number of significant digits at least correct is given by the largest value or m for which

$$|\epsilon_a| \leq 0.5 \times 10^{2-m}$$

$$0.1721 \leq 0.5 \times 10^{2-m}$$

$$0.3442 \leq 10^{2-m}$$

$$\log(0.3442) \leq 2 - m$$

$$m \leq 2 - \log(0.3442) = 2.463$$

So

$$m = 2$$

The number of significant digits at least correct in the estimated root of 0.06241 at the end of the 10th iteration is 2.

Advantages

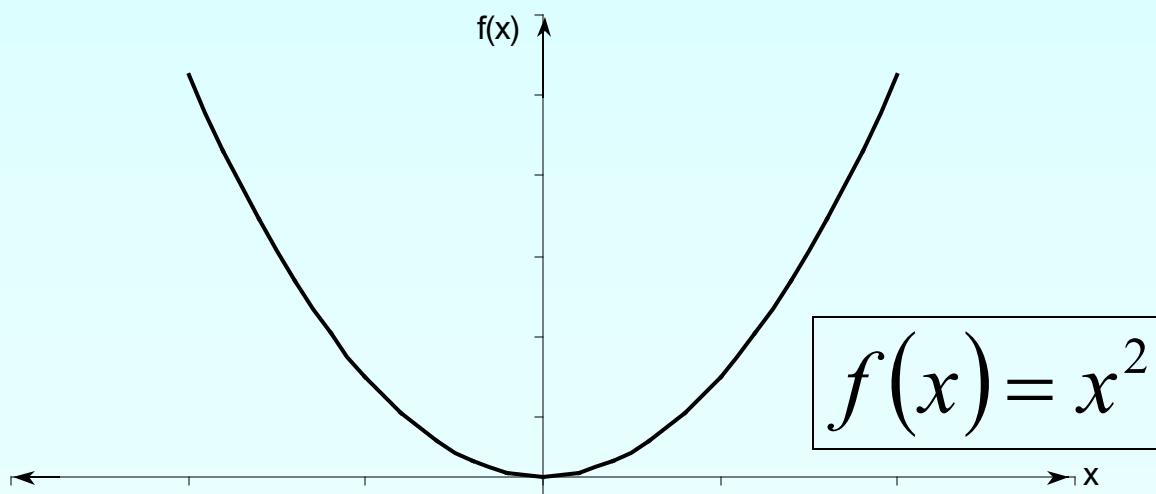
- Always convergent
- The root bracket gets halved with each iteration - guaranteed.

Drawbacks

- Slow convergence
- If one of the initial guesses is close to the root, the convergence is slower

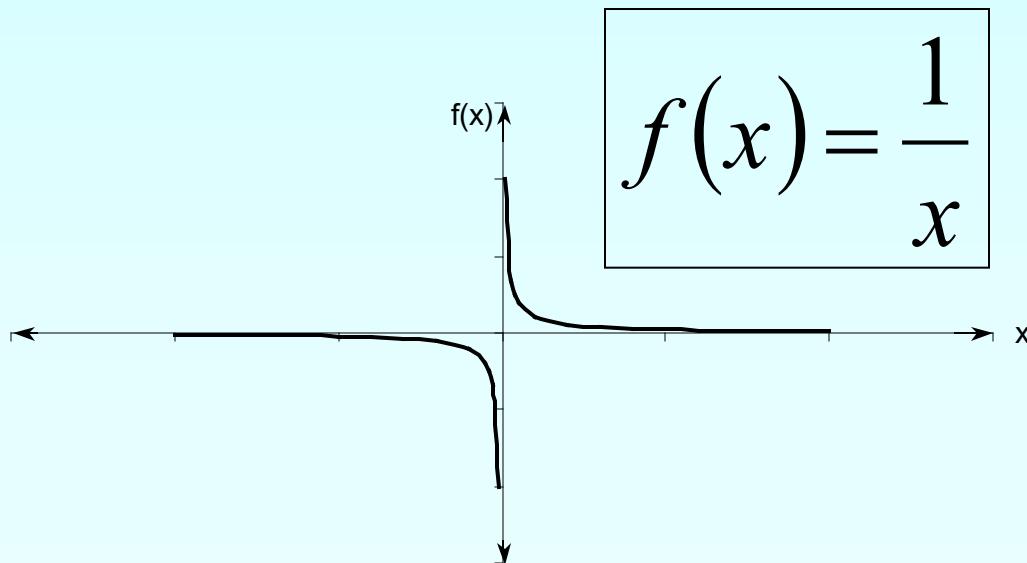
Drawbacks (continued)

- If a function $f(x)$ is such that it just touches the x -axis it will be unable to find the lower and upper guesses.



Drawbacks (continued)

- Function changes sign but root does not exist



Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

[http://numericalmethods.eng.usf.edu/topics/bisection
method.html](http://numericalmethods.eng.usf.edu/topics/bisection_method.html)

THE END

<http://numericalmethods.eng.usf.edu>

Newton-Raphson Method

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Newton-Raphson Method

<http://numericalmethods.eng.usf.edu>

Newton-Raphson Method

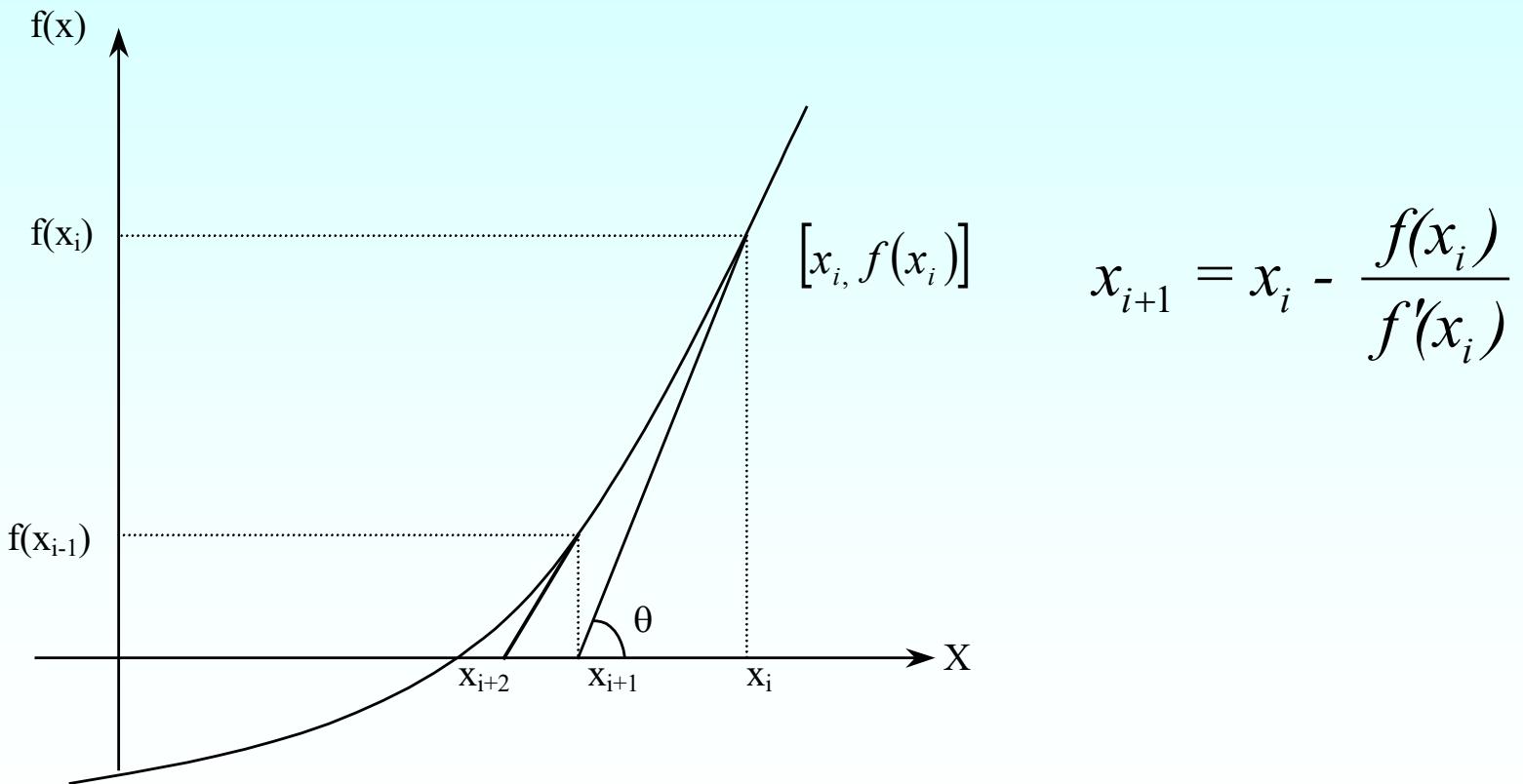
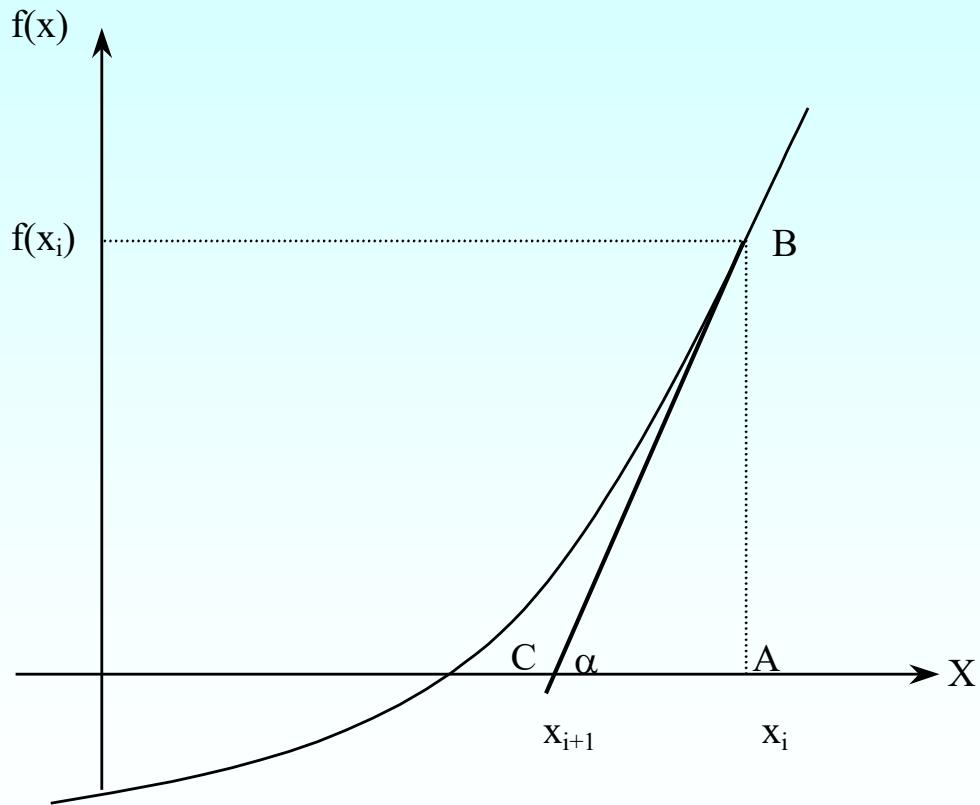


Figure 1 Geometrical illustration of the Newton-Raphson method.

Derivation



$$\tan(\alpha) = \frac{AB}{AC}$$

$$f'(x_i) = \frac{f(x_i)}{x_i - x_{i+1}}$$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Figure 2 Derivation of the Newton-Raphson method.

Algorithm for Newton-Raphson Method

Step 1

Evaluate $f'(x)$ symbolically.

Step 2

Use an initial guess of the root, x_i , to estimate the new value of the root, x_{i+1} , as

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

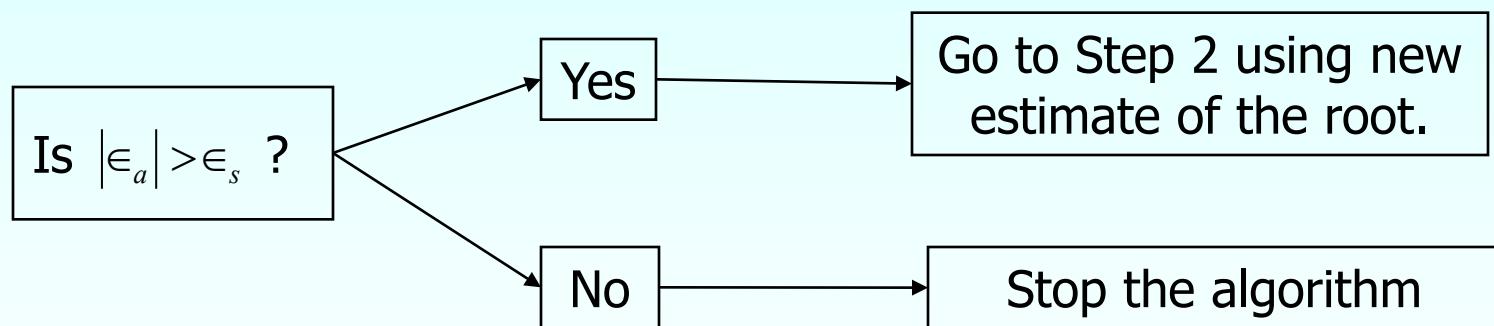
Step 3

Find the absolute relative approximate error $|\epsilon_a|$ as

$$|\epsilon_a| = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| \times 100$$

Step 4

Compare the absolute relative approximate error with the pre-specified relative error tolerance ϵ_s .



Also, check if the number of iterations has exceeded the maximum number of iterations allowed. If so, one needs to terminate the algorithm and notify the user.

Example 1

You are working for 'DOWN THE TOILET COMPANY' that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

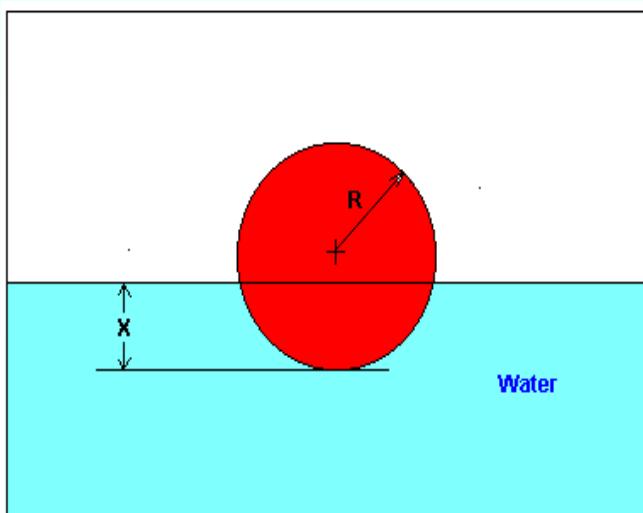


Figure 3 Floating ball problem.

Example 1 Cont.

The equation that gives the depth x in meters to which the ball is submerged under water is given by

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

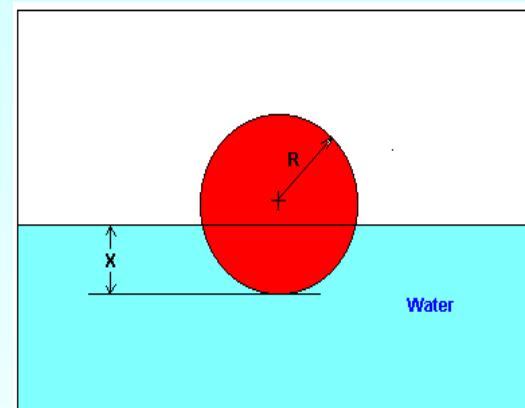


Figure 3 Floating ball problem.

Use the Newton's method of finding roots of equations to find

- the depth 'x' to which the ball is submerged under water. Conduct three iterations to estimate the root of the above equation.
- The absolute relative approximate error at the end of each iteration, and
- The number of significant digits at least correct at the end of each iteration.

Example 1 Cont.

Solution

To aid in the understanding of how this method works to find the root of an equation, the graph of $f(x)$ is shown to the right,

where

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

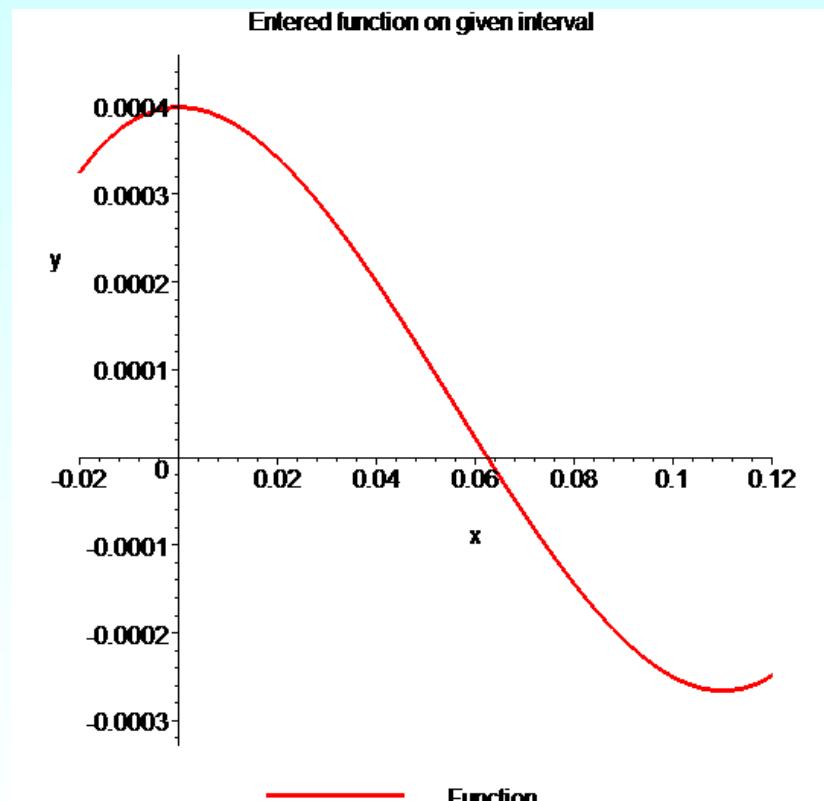


Figure 4 Graph of the function $f(x)$

Example 1 Cont.

Solve for $f'(x)$

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

$$f'(x) = 3x^2 - 0.33x$$

Let us assume the initial guess of the root of $f(x) = 0$ is $x_0 = 0.05\text{m}$. This is a reasonable guess (discuss why $x = 0$ and $x = 0.11\text{m}$ are not good choices) as the extreme values of the depth x would be 0 and the diameter (0.11 m) of the ball.

Example 1 Cont.

Iteration 1

The estimate of the root is

$$\begin{aligned}x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \\&= 0.05 - \frac{(0.05)^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}}{3(0.05)^2 - 0.33(0.05)} \\&= 0.05 - \frac{1.118 \times 10^{-4}}{-9 \times 10^{-3}} \\&= 0.05 - (-0.01242) \\&= 0.06242\end{aligned}$$

Example 1 Cont.

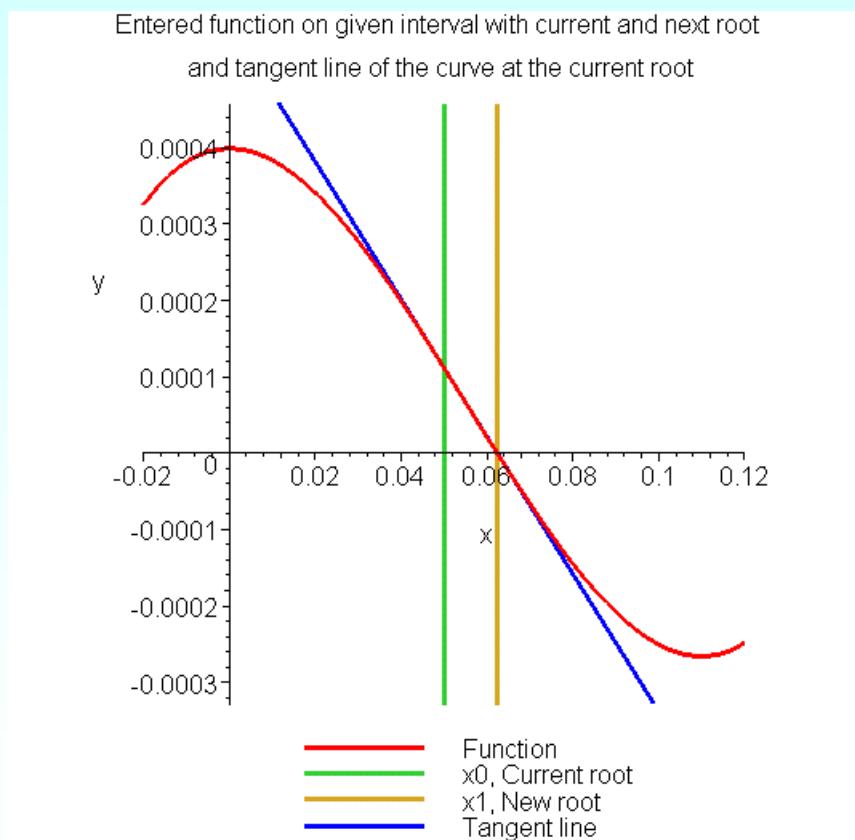


Figure 5 Estimate of the root for the first iteration.

Example 1 Cont.

The absolute relative approximate error $|e_a|$ at the end of Iteration 1 is

$$\begin{aligned}|e_a| &= \left| \frac{x_1 - x_0}{x_1} \right| \times 100 \\&= \left| \frac{0.06242 - 0.05}{0.06242} \right| \times 100 \\&= 19.90\%\end{aligned}$$

The number of significant digits at least correct is 0, as you need an absolute relative approximate error of 5% or less for at least one significant digits to be correct in your result.

Example 1 Cont.

Iteration 2

The estimate of the root is

$$\begin{aligned}x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} \\&= 0.06242 - \frac{(0.06242)^3 - 0.165(0.06242)^2 + 3.993 \times 10^{-4}}{3(0.06242)^2 - 0.33(0.06242)} \\&= 0.06242 - \frac{-3.97781 \times 10^{-7}}{-8.90973 \times 10^{-3}} \\&= 0.06242 - (4.4646 \times 10^{-5}) \\&= 0.06238\end{aligned}$$

Example 1 Cont.

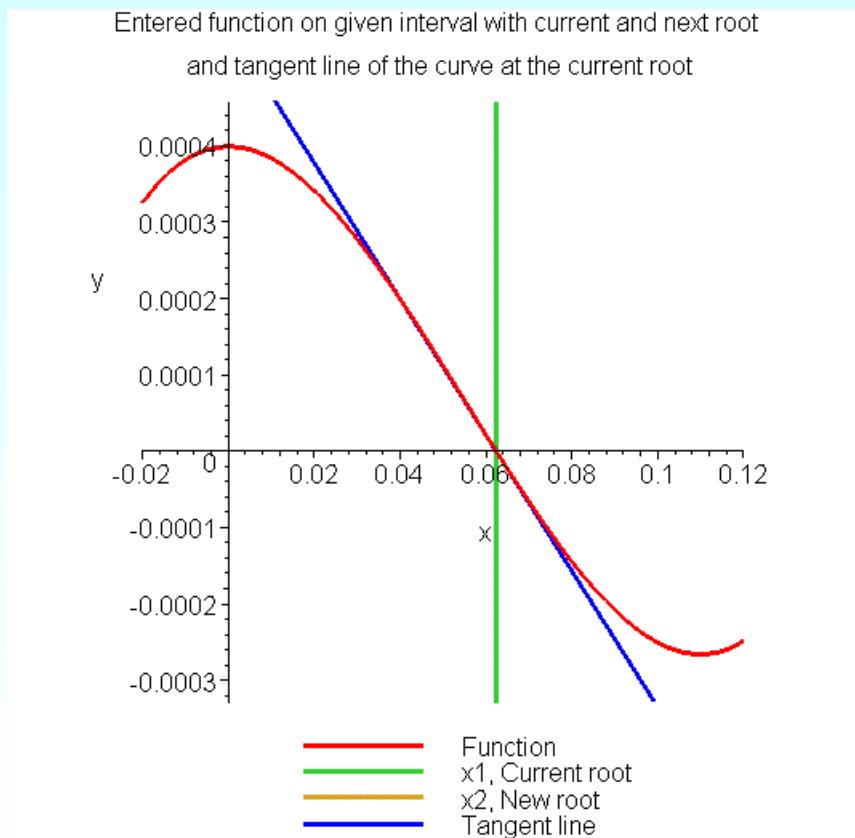


Figure 6 Estimate of the root for the Iteration 2.

Example 1 Cont.

The absolute relative approximate error $|e_a|$ at the end of Iteration 2 is

$$\begin{aligned}|e_a| &= \left| \frac{x_2 - x_1}{x_2} \right| \times 100 \\&= \left| \frac{0.06238 - 0.06242}{0.06238} \right| \times 100 \\&= 0.0716\%\end{aligned}$$

The maximum value of m for which $|e_a| \leq 0.5 \times 10^{2-m}$ is 2.844. Hence, the number of significant digits at least correct in the answer is 2.

Example 1 Cont.

Iteration 3

The estimate of the root is

$$\begin{aligned}x_3 &= x_2 - \frac{f(x_2)}{f'(x_2)} \\&= 0.06238 - \frac{(0.06238)^3 - 0.165(0.06238)^2 + 3.993 \times 10^{-4}}{3(0.06238)^2 - 0.33(0.06238)} \\&= 0.06238 - \frac{4.44 \times 10^{-11}}{-8.91171 \times 10^{-3}} \\&= 0.06238 - (-4.9822 \times 10^{-9}) \\&= 0.06238\end{aligned}$$

Example 1 Cont.

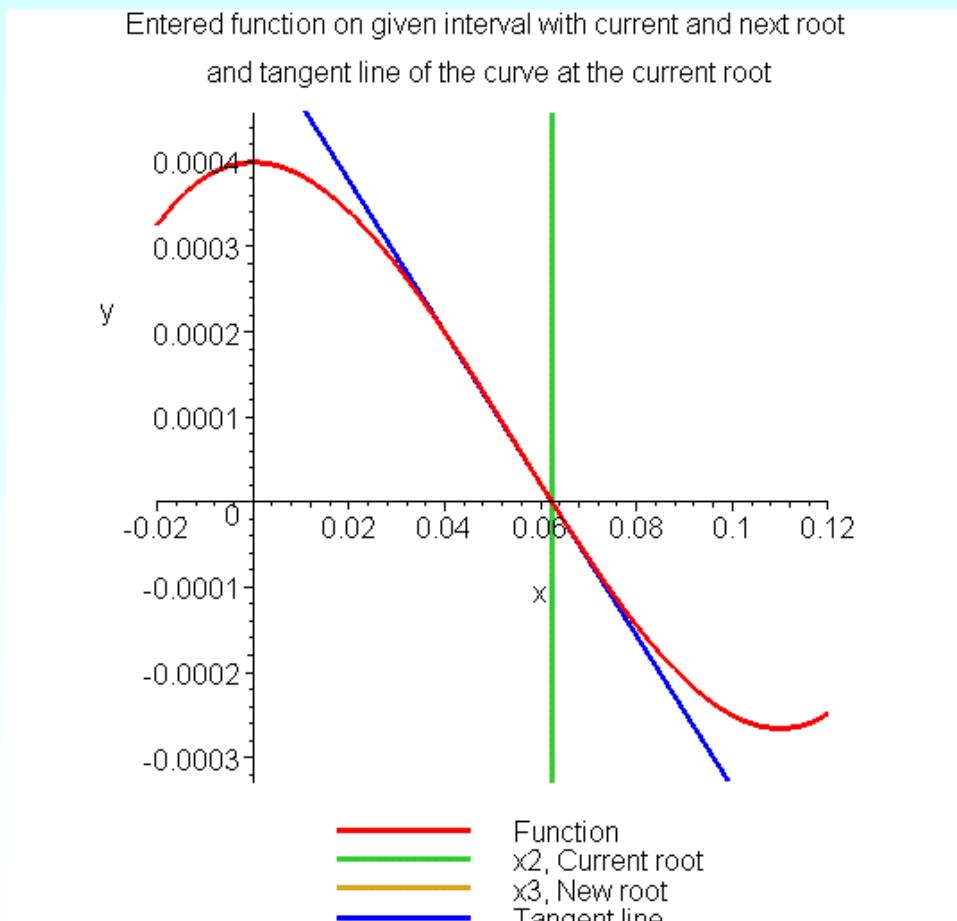


Figure 7 Estimate of the root for the Iteration 3.

Example 1 Cont.

The absolute relative approximate error $|e_a|$ at the end of Iteration 3 is

$$\begin{aligned}|e_a| &= \left| \frac{x_2 - x_1}{x_2} \right| \times 100 \\&= \left| \frac{0.06238 - 0.06238}{0.06238} \right| \times 100 \\&= 0\%\end{aligned}$$

The number of significant digits at least correct is 4, as only 4 significant digits are carried through all the calculations.

Advantages and Drawbacks of Newton Raphson Method

<http://numericalmethods.eng.usf.edu>

Advantages

- Converges fast (quadratic convergence), if it converges.
- Requires only one guess

Drawbacks

1. Divergence at inflection points

Selection of the initial guess or an iteration value of the root that is close to the inflection point of the function $f(x)$ may start diverging away from the root in the Newton-Raphson method.

For example, to find the root of the equation $f(x) = (x - 1)^3 + 0.512 = 0$.

The Newton-Raphson method reduces to $x_{i+1} = x_i - \frac{(x_i^3 - 1)^3 + 0.512}{3(x_i - 1)^2}$.

Table 1 shows the iterated values of the root of the equation.

The root starts to diverge at Iteration 6 because the previous estimate of 0.92589 is close to the inflection point of $x = 1$.

Eventually after 12 more iterations the root converges to the exact value of $x = 0.2$.

Drawbacks – Inflection Points

Table 1 Divergence near inflection point.

Iteration Number	x_i
0	5.0000
1	3.6560
2	2.7465
3	2.1084
4	1.6000
5	0.92589
6	-30.119
7	-19.746
18	0.2000

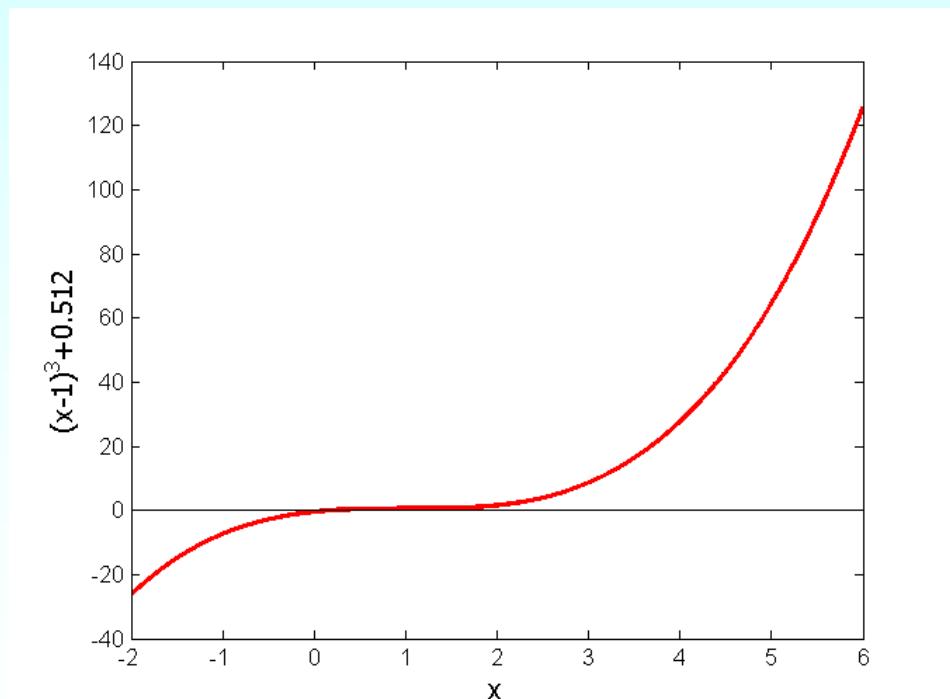


Figure 8 Divergence at inflection point for
 $f(x) = (x - 1)^3 + 0.512 = 0$

Drawbacks – Division by Zero

2. Division by zero

For the equation

$$f(x) = x^3 - 0.03x^2 + 2.4 \times 10^{-6} = 0$$

the Newton-Raphson method reduces to

$$x_{i+1} = x_i - \frac{x_i^3 - 0.03x_i^2 + 2.4 \times 10^{-6}}{3x_i^2 - 0.06x_i}$$

For $x_0 = 0$ or $x_0 = 0.02$, the denominator will equal zero.

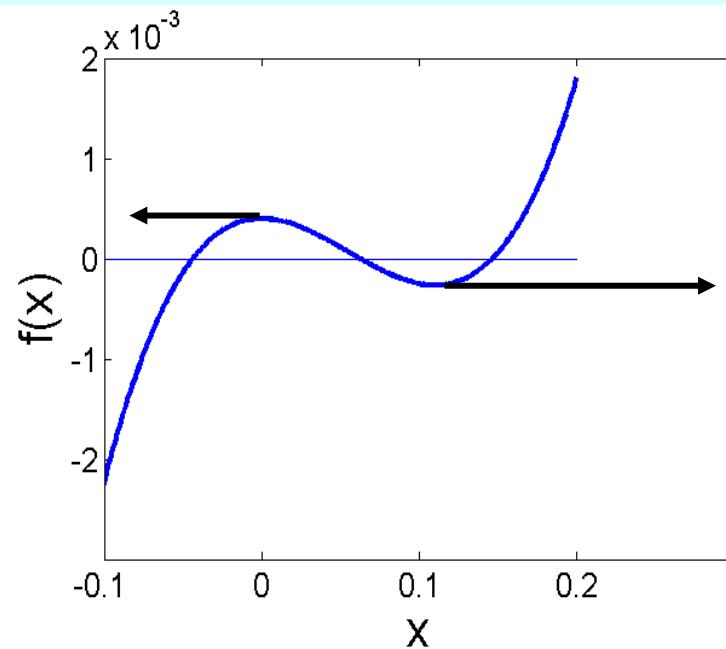


Figure 9 Pitfall of division by zero or near a zero number

Drawbacks – Oscillations near local maximum and minimum

3. Oscillations near local maximum and minimum

Results obtained from the Newton-Raphson method may oscillate about the local maximum or minimum without converging on a root but converging on the local maximum or minimum.

Eventually, it may lead to division by a number close to zero and may diverge.

For example for $f(x) = x^2 + 2 = 0$ the equation has no real roots.

Drawbacks – Oscillations near local maximum and minimum

Table 3 Oscillations near local maxima and minima in Newton-Raphson method.

Iteration Number	x_i	$f(x_i)$	$ \in_a \%$
0	-1.0000	3.00	
1	0.5	2.25	300.00
2	-1.75	5.063	128.571
3	-0.30357	2.092	476.47
4	3.1423	11.874	109.66
5	1.2529	3.570	150.80
6	-0.17166	2.029	829.88
7	5.7395	34.942	102.99
8	2.6955	9.266	112.93
9	0.97678	2.954	175.96

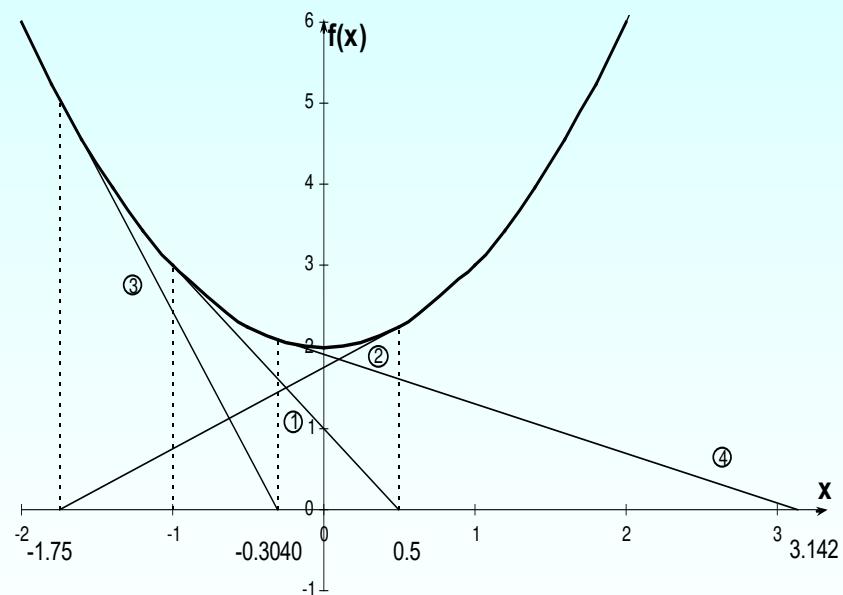


Figure 10 Oscillations around local minima for $f(x) = x^2 + 2$.

Drawbacks – Root Jumping

4. Root Jumping

In some cases where the function $f(x)$ is oscillating and has a number of roots, one may choose an initial guess close to a root. However, the guesses may jump and converge to some other root.

For example

$$f(x) = \sin x = 0$$

Choose

$$x_0 = 2.4\pi = 7.539822$$

It will converge to $x = 0$

instead of $x = 2\pi = 6.2831853$

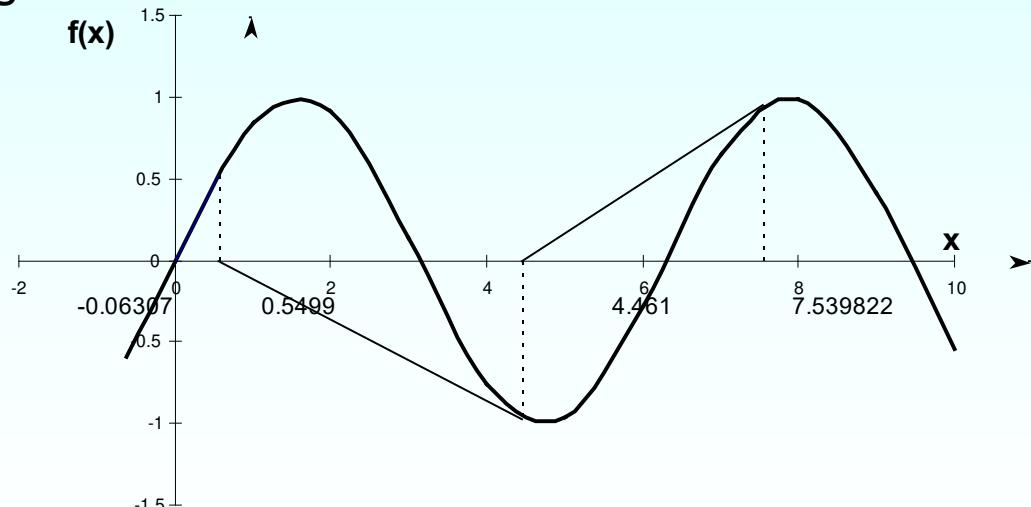


Figure 11 Root jumping from intended location of root for $f(x) = \sin x = 0$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/newton_ra_phson.html

THE END

<http://numericalmethods.eng.usf.edu>

Secant Method

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Secant Method

<http://numericalmethods.eng.usf.edu>

Secant Method – Derivation

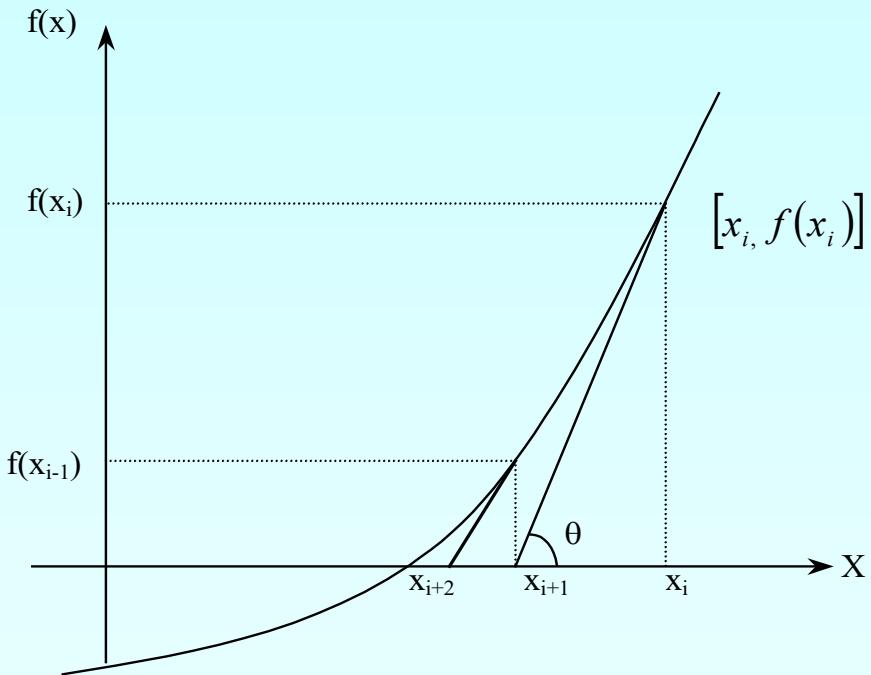


Figure 1 Geometrical illustration of the Newton-Raphson method.

Newton's Method

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (1)$$

Approximate the derivative

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \quad (2)$$

Substituting Equation (2) into Equation (1) gives the Secant method

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$

Secant Method – Derivation

The secant method can also be derived from geometry:

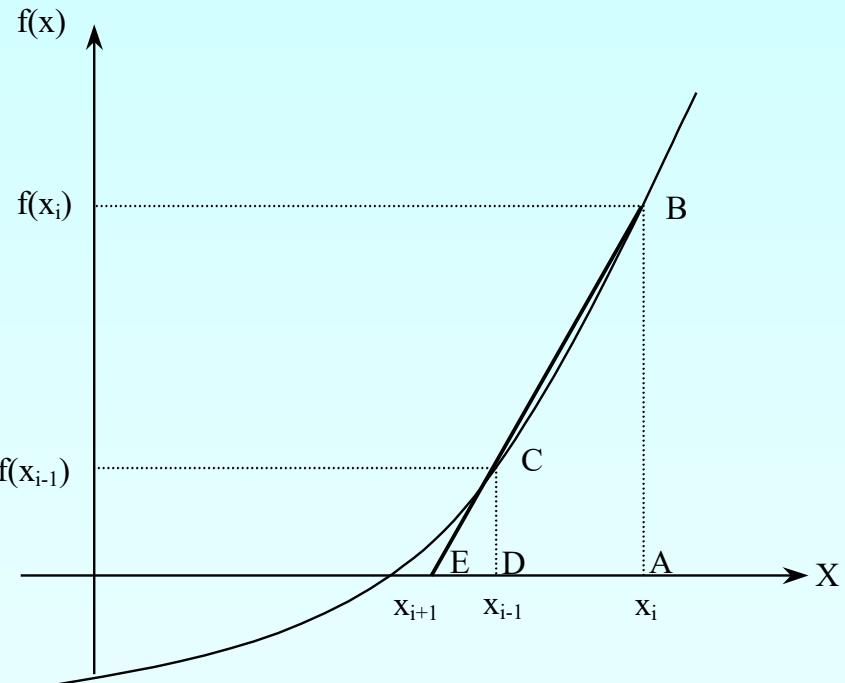


Figure 2 Geometrical representation of the Secant method.

The Geometric Similar Triangles

$$\frac{AB}{AE} = \frac{DC}{DE}$$

can be written as

$$\frac{f(x_i)}{x_i - x_{i+1}} = \frac{f(x_{i-1})}{x_{i-1} - x_{i+1}}$$

On rearranging, the secant method is given as

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$

Algorithm for Secant Method

Step 1

Calculate the next estimate of the root from two initial guesses

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$

Find the absolute relative approximate error

$$|\epsilon_a| = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| \times 100$$

Step 2

Find if the absolute relative approximate error is greater than the prespecified relative error tolerance.

If so, go back to step 1, else stop the algorithm.

Also check if the number of iterations has exceeded the maximum number of iterations.

Example 1

You are working for 'DOWN THE TOILET COMPANY' that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

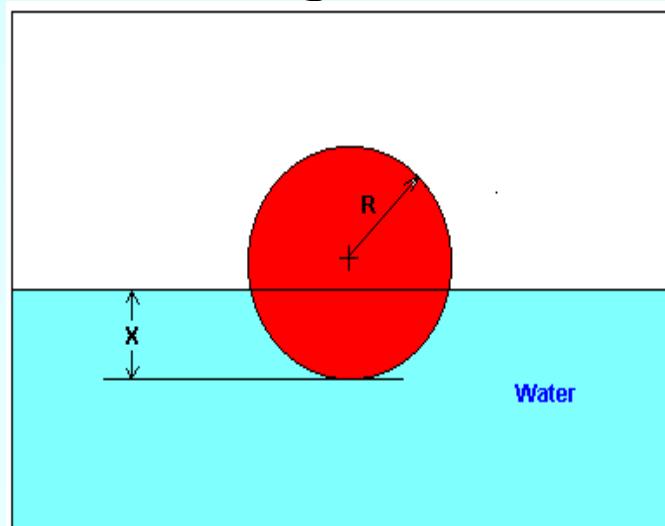


Figure 3 Floating Ball Problem.

Example 1 Cont.

The equation that gives the depth x to which the ball is submerged under water is given by

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

Use the Secant method of finding roots of equations to find the depth x to which the ball is submerged under water.

- Conduct three iterations to estimate the root of the above equation.
- Find the absolute relative approximate error and the number of significant digits at least correct at the end of each iteration.

Example 1 Cont.

Solution

To aid in the understanding of how this method works to find the root of an equation, the graph of $f(x)$ is shown to the right,

where

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

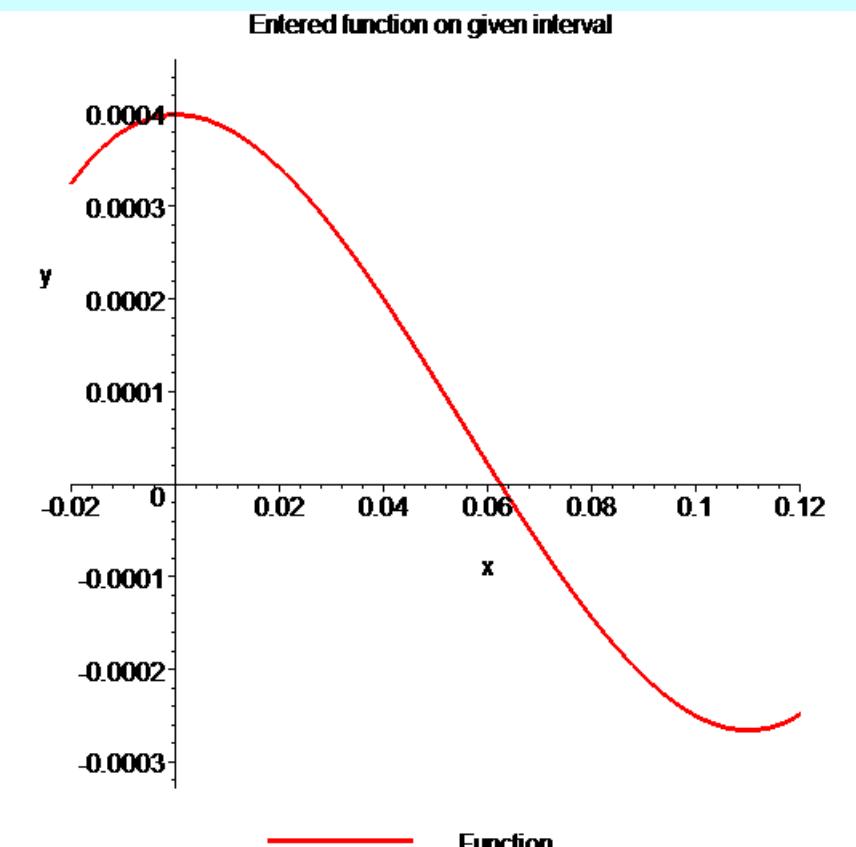


Figure 4 Graph of the function $f(x)$.

Example 1 Cont.

Let us assume the initial guesses of the root of $f(x)=0$ as $x_{-1} = 0.02$ and $x_0 = 0.05$.

Iteration 1

The estimate of the root is

$$\begin{aligned}x_1 &= x_0 - \frac{f(x_0)(x_0 - x_{-1})}{f(x_0) - f(x_{-1})} \\&= 0.05 - \frac{(0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4})(0.05 - 0.02)}{(0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}) - (0.02^3 - 0.165(0.02)^2 + 3.993 \times 10^{-4})} \\&= 0.06461\end{aligned}$$

Example 1 Cont.

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 1 is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_1 - x_0}{x_1} \right| \times 100 \\ &= \left| \frac{0.06461 - 0.05}{0.06461} \right| \times 100 \\ &= 22.62\% \end{aligned}$$

The number of significant digits at least correct is 0, as you need an absolute relative approximate error of 5% or less for one significant digits to be correct in your result.

Example 1 Cont.

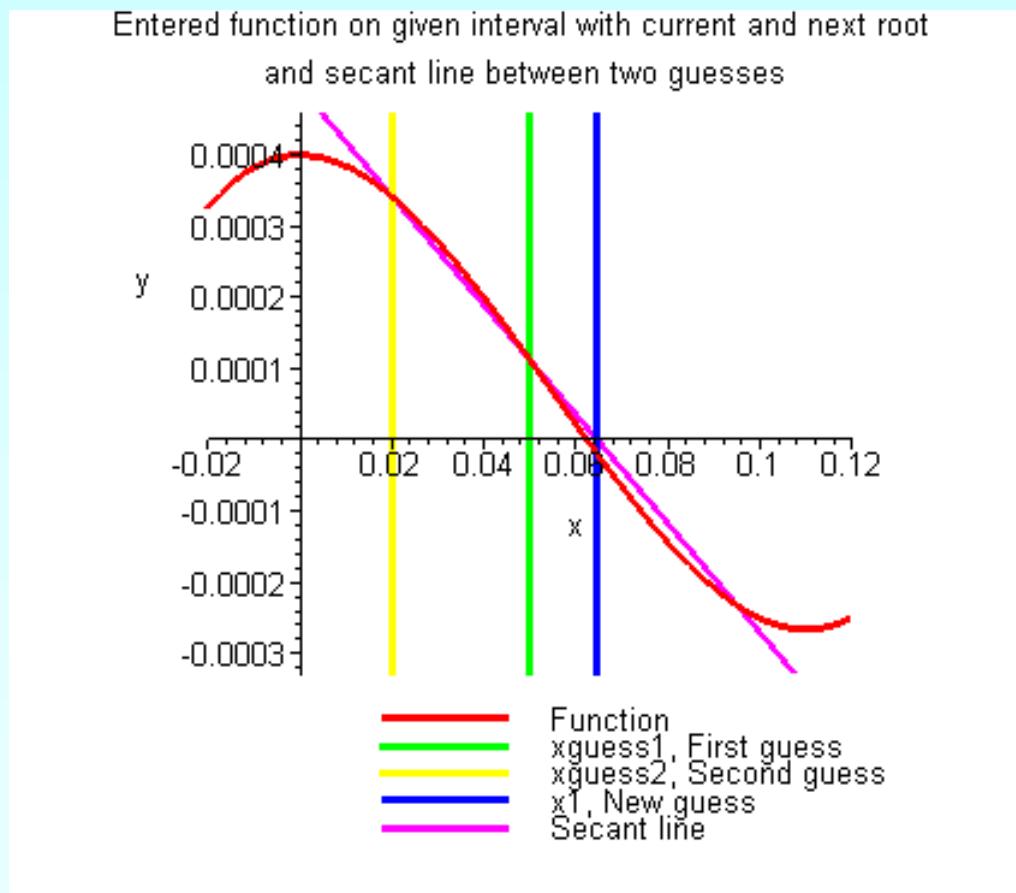


Figure 5 Graph of results of Iteration 1.

Example 1 Cont.

Iteration 2

The estimate of the root is

$$\begin{aligned}x_2 &= x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)} \\&= 0.06461 - \frac{(0.06461^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4})(0.06461 - 0.05)}{(0.06461^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4}) - (0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4})} \\&= 0.06241\end{aligned}$$

Example 1 Cont.

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 2 is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_2 - x_1}{x_2} \right| \times 100 \\ &= \left| \frac{0.06241 - 0.06461}{0.06241} \right| \times 100 \\ &= 3.525\% \end{aligned}$$

The number of significant digits at least correct is 1, as you need an absolute relative approximate error of 5% or less.

Example 1 Cont.

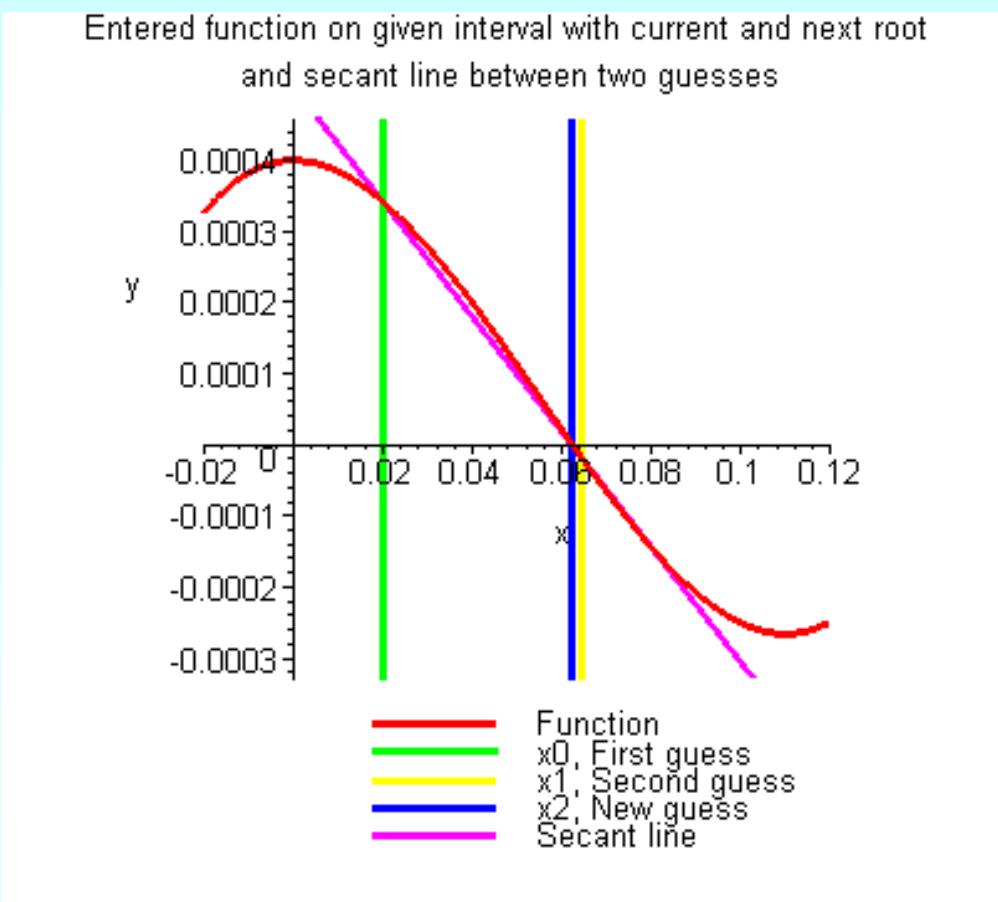


Figure 6 Graph of results of Iteration 2.

Example 1 Cont.

Iteration 3

The estimate of the root is

$$\begin{aligned}x_3 &= x_2 - \frac{f(x_2)(x_2 - x_1)}{f(x_2) - f(x_1)} \\&= 0.06241 - \frac{(0.06241^3 - 0.165(0.06241)^2 + 3.993 \times 10^{-4})(0.06241 - 0.06461)}{(0.06241^3 - 0.165(0.06241)^2 + 3.993 \times 10^{-4}) - (0.05^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4})} \\&= 0.06238\end{aligned}$$

Example 1 Cont.

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 3 is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_3 - x_2}{x_3} \right| \times 100 \\ &= \left| \frac{0.06238 - 0.06241}{0.06238} \right| \times 100 \\ &= 0.0595\% \end{aligned}$$

The number of significant digits at least correct is 5, as you need an absolute relative approximate error of 0.5% or less.

Iteration #3

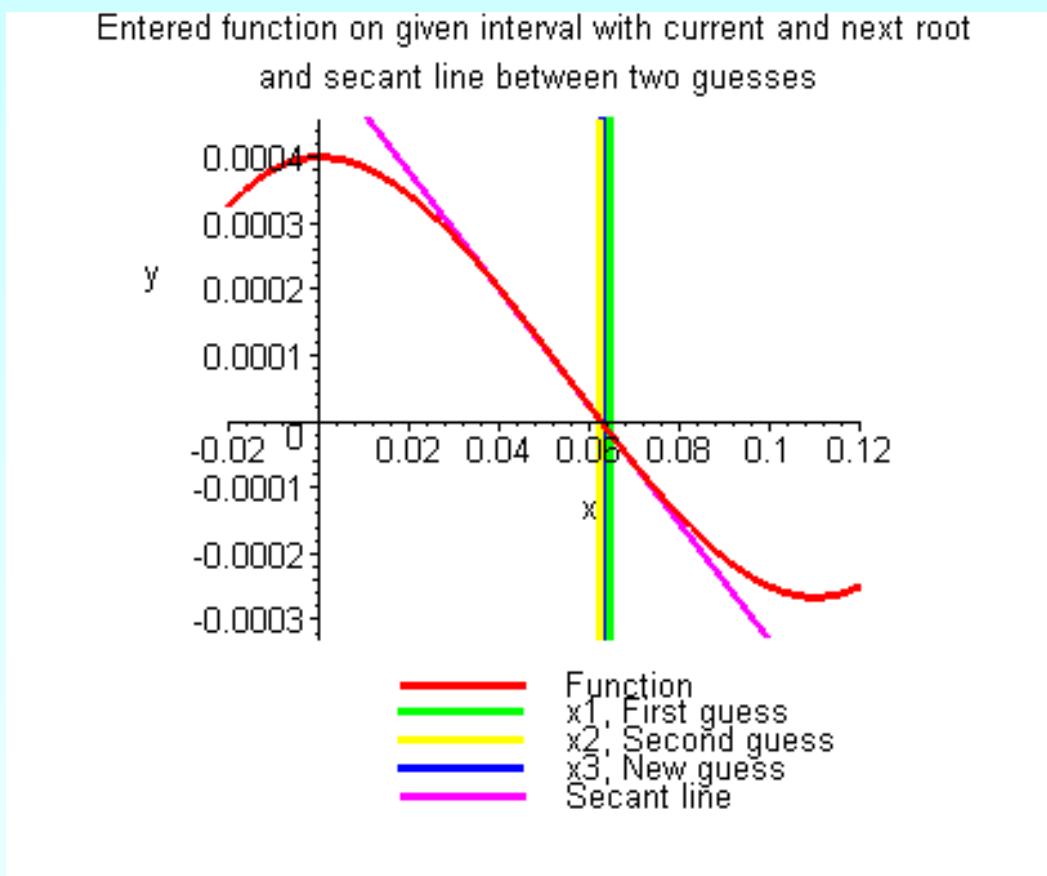
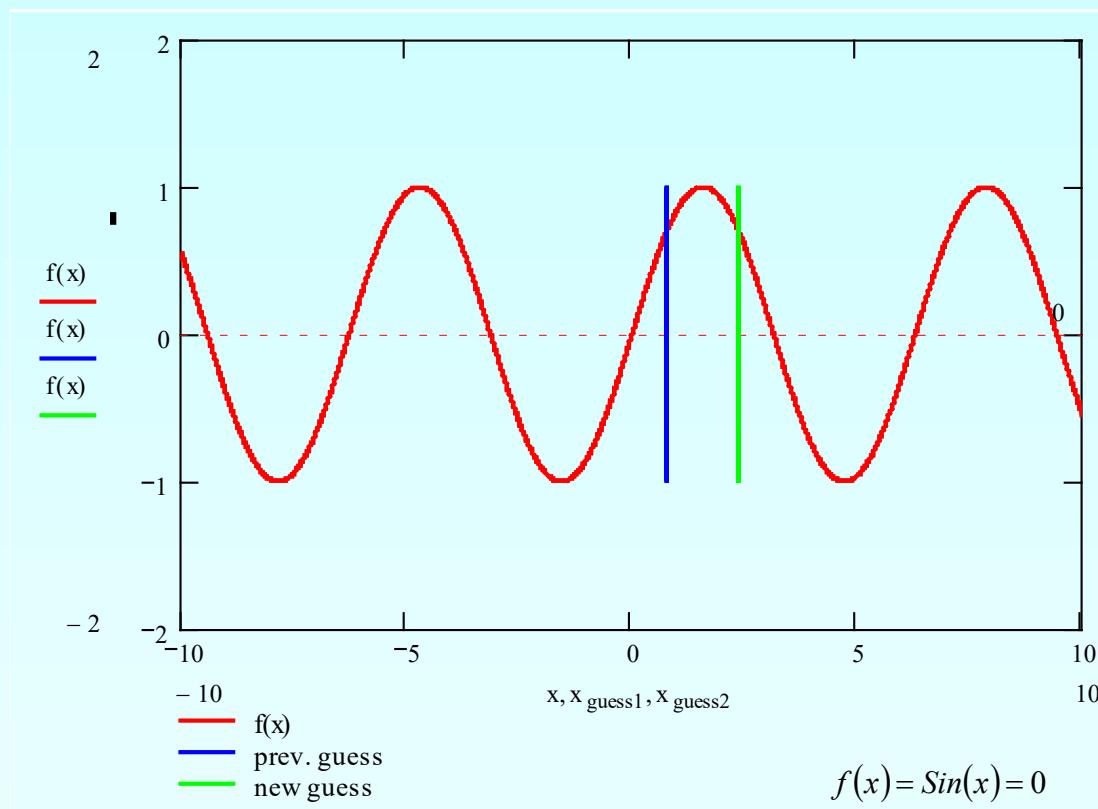


Figure 7 Graph of results of Iteration 3.

Advantages

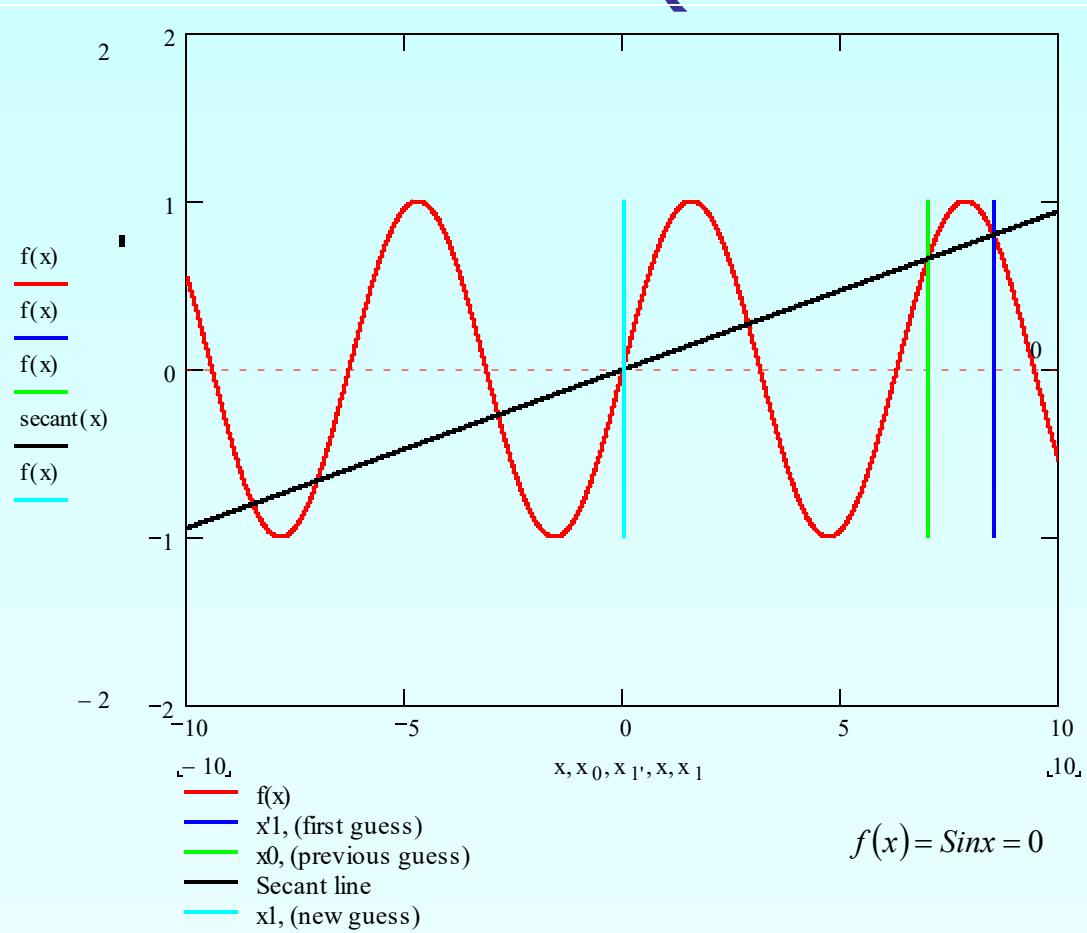
- Converges fast, if it converges
- Requires two guesses that do not need to bracket the root

Drawbacks



Division by zero

Drawbacks (continued)



Root Jumping

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/secant_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Gaussian Elimination

Major: All Engineering Majors

Author(s): Autar Kaw

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Naïve Gauss Elimination

<http://numericalmethods.eng.usf.edu>

Naïve Gaussian Elimination

A method to solve simultaneous linear equations of the form $[A][X]=[C]$

Two steps

1. Forward Elimination
2. Back Substitution

Forward Elimination

The goal of forward elimination is to transform the coefficient matrix into an upper triangular matrix

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$



$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

Forward Elimination

A set of n equations and n unknowns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = b_2$$

⋮
⋮
⋮
⋮

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n = b_n$$

($n-1$) steps of forward elimination

Forward Elimination

Step 1

For Equation 2, divide Equation 1 by a_{11} and multiply by a_{21} .

$$\left[\frac{a_{21}}{a_{11}} \right] (a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1)$$

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \dots + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}b_1$$

Forward Elimination

Subtract the result from Equation 2.

$$\begin{array}{rcl} a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n & = & b_2 \\ - \quad a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \dots + \frac{a_{21}}{a_{11}}a_{1n}x_n & = & \frac{a_{21}}{a_{11}}b_1 \\ \hline \left(a_{22} - \frac{a_{21}}{a_{11}}a_{12} \right)x_2 + \dots + \left(a_{2n} - \frac{a_{21}}{a_{11}}a_{1n} \right)x_n & = & b_2 - \frac{a_{21}}{a_{11}}b_1 \end{array}$$

or $\overset{'}{a_{22}}x_2 + \dots + \overset{'}{a_{2n}}x_n = \overset{'}{b_2}$

Forward Elimination

Repeat this procedure for the remaining equations to reduce the set of equations as

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$\dot{a_{22}}x_2 + \dot{a_{23}}x_3 + \dots + \dot{a_{2n}}x_n = \dot{b_2}$$

$$\dot{a_{32}}x_2 + \dot{a_{33}}x_3 + \dots + \dot{a_{3n}}x_n = \dot{b_3}$$

$$\begin{matrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{matrix}$$

$$\dot{a_{n2}}x_2 + \dot{a_{n3}}x_3 + \dots + \dot{a_{nn}}x_n = \dot{b_n}$$

End of Step 1

Forward Elimination

Step 2

Repeat the same procedure for the 3rd term of Equation 3.

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{22}'x_2 + a_{23}'x_3 + \dots + a_{2n}'x_n = b_2'$$

$$a_{33}''x_3 + \dots + a_{3n}''x_n = b_3''$$

$$\begin{matrix} \cdot \\ \cdot \\ \cdot \end{matrix}$$

$$a_{n3}''x_3 + \dots + a_{nn}''x_n = b_n''$$

End of Step 2

Forward Elimination

At the end of (n-1) Forward Elimination steps, the system of equations will look like

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{22}'x_2 + a_{23}'x_3 + \dots + a_{2n}'x_n = b_2'$$

$$a_{33}''x_3 + \dots + a_{3n}''x_n = b_3''$$

.

.

.

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}$$

End of Step (n-1)

Matrix Form at End of Forward Elimination

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a'_{22} & a'_{23} & \cdots & a'_{2n} \\ 0 & 0 & a''_{33} & \cdots & a''_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & a_{nn}^{(n-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b'_2 \\ b''_3 \\ \vdots \\ b_n^{(n-1)} \end{bmatrix}$$

Back Substitution

Solve each equation starting from the last equation

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

Example of a system of 3 equations

Back Substitution Starting Eqns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{22}'x_2 + a_{23}'x_3 + \dots + a_{2n}'x_n = b_2'$$

$$a_{33}''x_3 + \dots + a_n''x_n = b_3''$$

⋮ ⋮ ⋮

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}$$

Back Substitution

Start with the last equation because it has only one unknown

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

Back Substitution

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

$$x_i = \frac{b_i^{(i-1)} - a_{i,i+1}^{(i-1)}x_{i+1} - a_{i,i+2}^{(i-1)}x_{i+2} - \dots - a_{i,n}^{(i-1)}x_n}{a_{ii}^{(i-1)}} \text{ for } i = n-1, \dots, 1$$

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^n a_{ij}^{(i-1)}x_j}{a_{ii}^{(i-1)}} \text{ for } i = n-1, \dots, 1$$

THE END

<http://numericalmethods.eng.usf.edu>

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/gaussian_elimination.html

Naïve Gauss Elimination Example

<http://numericalmethods.eng.usf.edu>

Example 1

The upward velocity of a rocket is given at three different times

Table 1 Velocity vs. time data.

Time, t (s)	Velocity, v (m/s)
5	106.8
8	177.2
12	279.2



The velocity data is approximated by a polynomial as:

$$v(t) = a_1 t^2 + a_2 t + a_3 , \quad 5 \leq t \leq 12.$$

Find the velocity at $t=6$ seconds .

Example 1 Cont.

Assume

$$v(t) = a_1 t^2 + a_2 t + a_3, \quad 5 \leq t \leq 12.$$

Results in a matrix template of the form:

$$\begin{bmatrix} t_1^2 & t_1 & 1 \\ t_2^2 & t_2 & 1 \\ t_3^2 & t_3 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

Using data from Table 1, the matrix becomes:

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Example 1 Cont.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix} \Rightarrow \begin{bmatrix} 25 & 5 & 1 & : & 106.8 \\ 64 & 8 & 1 & : & 177.2 \\ 144 & 12 & 1 & : & 279.2 \end{bmatrix}$$

1. Forward Elimination
2. Back Substitution

Forward Elimination

Number of Steps of Forward Elimination

Number of steps of forward elimination is
 $(n-1)=(3-1)=2$

Forward Elimination: Step 1

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 64 & 8 & 1 & : 177.2 \\ 144 & 12 & 1 & : 279.2 \end{array} \right]$$

Divide Equation 1 by 25 and multiply it by 64, $\frac{64}{25} = 2.56$.

$$[25 \ 5 \ 1 \ : \ 106.8] \times 2.56 = [64 \ 12.8 \ 2.56 \ : \ 273.408]$$

Subtract the result from Equation 2

$$\begin{array}{r} [64 \ 8 \ 1 \ : \ 177.2] \\ - [64 \ 12.8 \ 2.56 \ : \ 273.408] \\ \hline [0 \ -4.8 \ -1.56 \ : \ -96.208] \end{array}$$

Substitute new equation for Equation 2

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 0 & -4.8 & -1.56 & : -96.208 \\ 144 & 12 & 1 & : 279.2 \end{array} \right]$$

Forward Elimination: Step 1 (cont.)

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 0 & -4.8 & -1.56 & : -96.208 \\ 144 & 12 & 1 & : 279.2 \end{array} \right] \quad \begin{array}{l} \text{Divide Equation 1 by 25 and} \\ \text{multiply it by 144, } \frac{144}{25} = 5.76. \end{array}$$

$$[25 \ 5 \ 1 \ : \ 106.8] \times 5.76 = [144 \ 28.8 \ 5.76 \ : \ 615.168]$$

Subtract the result from
Equation 3

$$\begin{array}{r} [144 \ 12 \ 1 \ : \ 279.2] \\ - [144 \ 28.8 \ 5.76 \ : \ 615.168] \\ \hline [0 \ -16.8 \ -4.76 \ : \ -335.968] \end{array}$$

Substitute new equation for
Equation 3

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 0 & -4.8 & -1.56 & : -96.208 \\ 0 & -16.8 & -4.76 & : -335.968 \end{array} \right]$$

Forward Elimination: Step 2

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : & 106.8 \\ 0 & -4.8 & -1.56 & : & -96.208 \\ 0 & -16.8 & -4.76 & : & -335.968 \end{array} \right]$$

Divide Equation 2 by -4.8
and multiply it by -16.8 ,
 $\frac{-16.8}{-4.8} = 3.5$.

$$[0 \quad -4.8 \quad -1.56 \quad : \quad -96.208] \times 3.5 = [0 \quad -16.8 \quad -5.46 \quad : \quad -336.728]$$

Subtract the result from
Equation 3

$$\begin{array}{r} [0 \quad -16.8 \quad -4.76 \quad : \quad 335.968] \\ - [0 \quad -16.8 \quad -5.46 \quad : \quad -336.728] \\ \hline [0 \quad 0 \quad 0.7 \quad : \quad 0.76] \end{array}$$

Substitute new equation for
Equation 3

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : & 106.8 \\ 0 & -4.8 & -1.56 & : & -96.208 \\ 0 & 0 & 0.7 & : & 0.76 \end{array} \right]$$

Back Substitution

Back Substitution

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 0 & -4.8 & -1.56 & : -96.2 \\ 0 & 0 & 0.7 & : 0.7 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 0 & -4.8 & -1.56 & : -96.2 \\ 0 & 0 & 0.7 & : 0.7 \end{array} \right] \left[\begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right] = \left[\begin{array}{c} 106.8 \\ -96.2 \\ 0.76 \end{array} \right]$$

Solving for a_3

$$0.7a_3 = 0.76$$

$$a_3 = \frac{0.76}{0.7}$$

$$a_3 = 1.08571$$

Back Substitution (cont.)

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.208 \\ 0.76 \end{bmatrix}$$

Solving for a_2

$$-4.8a_2 - 1.56a_3 = -96.208$$

$$a_2 = \frac{-96.208 + 1.56a_3}{-4.8}$$

$$a_2 = \frac{-96.208 + 1.56 \times 1.08571}{-4.8}$$

$$a_2 = 19.6905$$

Back Substitution (cont.)

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.2 \\ 0.76 \end{bmatrix}$$

Solving for a_1

$$25a_1 + 5a_2 + a_3 = 106.8$$

$$\begin{aligned} a_1 &= \frac{106.8 - 5a_2 - a_3}{25} \\ &= \frac{106.8 - 5 \times 19.6905 - 1.08571}{25} \\ &= 0.290472 \end{aligned}$$

Naïve Gaussian Elimination Solution

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.290472 \\ 19.6905 \\ 1.08571 \end{bmatrix}$$

Example 1 Cont.

Solution

The solution vector is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.290472 \\ 19.6905 \\ 1.08571 \end{bmatrix}$$

The polynomial that passes through the three data points is then:

$$\begin{aligned} v(t) &= a_1 t^2 + a_2 t + a_3 \\ &= 0.290472t^2 + 19.6905t + 1.08571, \quad 5 \leq t \leq 12 \end{aligned}$$

$$\begin{aligned} v(6) &= 0.290472(6)^2 + 19.6905(6) + 1.08571 \\ &= 129.686 \text{ m/s.} \end{aligned}$$

THE END

<http://numericalmethods.eng.usf.edu>

Naïve Gauss Elimination Pitfalls

<http://numericalmethods.eng.usf.edu>

Pitfall#1. Division by zero

$$10x_2 - 7x_3 = 3$$

$$6x_1 + 2x_2 + 3x_3 = 11$$

$$5x_1 - x_2 + 5x_3 = 9$$

$$\begin{bmatrix} 0 & 10 & -7 \\ 6 & 2 & 3 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 11 \\ 9 \end{bmatrix}$$

Is division by zero an issue here?

$$12x_1 + 10x_2 - 7x_3 = 15$$

$$6x_1 + 5x_2 + 3x_3 = 14$$

$$5x_1 - x_2 + 5x_3 = 9$$

$$\begin{bmatrix} 12 & 10 & -7 \\ 6 & 5 & 3 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15 \\ 14 \\ 9 \end{bmatrix}$$

Is division by zero an issue here?

YES

$$12x_1 + 10x_2 - 7x_3 = 15$$

$$6x_1 + 5x_2 + 3x_3 = 14$$

$$24x_1 - x_2 + 5x_3 = 28$$

$$\begin{bmatrix} 12 & 10 & -7 \\ 6 & 5 & 3 \\ 24 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15 \\ 14 \\ 28 \end{bmatrix} \rightarrow \begin{bmatrix} 12 & 10 & -7 \\ 0 & 0 & 6.5 \\ 12 & -21 & 19 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15 \\ 6.5 \\ -2 \end{bmatrix}$$

Division by zero is a possibility at any step
of forward elimination

Pitfall#2. Large Round-off Errors

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

Exact Solution

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Pitfall#2. Large Round-off Errors

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

Solve it on a computer using 6 significant digits with chopping

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.9625 \\ 1.05 \\ 0.999995 \end{bmatrix}$$

Pitfall#2. Large Round-off Errors

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

Solve it on a computer using **5** significant digits with chopping

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.625 \\ 1.5 \\ 0.99995 \end{bmatrix}$$

Is there a way to reduce the round off error?

Avoiding Pitfalls

Increase the number of significant digits

- Decreases round-off error
- Does not avoid division by zero

Avoiding Pitfalls

Gaussian Elimination with Partial Pivoting

- Avoids division by zero
- Reduces round off error

THE END

<http://numericalmethods.eng.usf.edu>

Gauss Elimination with Partial Pivoting

<http://numericalmethods.eng.usf.edu>

Pitfalls of Naïve Gauss Elimination

- Possible division by zero
- Large round-off errors

Avoiding Pitfalls

Increase the number of significant digits

- Decreases round-off error
- Does not avoid division by zero

Avoiding Pitfalls

Gaussian Elimination with Partial Pivoting

- Avoids division by zero
- Reduces round off error

What is Different About Partial Pivoting?

At the beginning of the k^{th} step of forward elimination,
find the maximum of

$$|a_{kk}|, |a_{k+1,k}|, \dots, |a_{nk}|$$

If the maximum of the values is $|a_{pk}|$
in the p^{th} row, $k \leq p \leq n$, then switch rows p and k .

Matrix Form at Beginning of 2nd Step of Forward Elimination

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a'_{22} & a'_{23} & \cdots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & \cdots & a'_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & a'_{n2} & a'_{n3} & a'_{n4} & a'_{nn} \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{array} \right] = \left[\begin{array}{c} b_1 \\ b'_2 \\ b'_3 \\ \vdots \\ b'_n \end{array} \right]$$

Example (2nd step of FE)

$$\begin{bmatrix} 6 & 14 & 5.1 & 3.7 & 6 \\ 0 & -7 & 6 & 1 & 2 \\ 0 & 4 & 12 & 1 & 11 \\ 0 & 9 & 23 & 6 & 8 \\ 0 & -17 & 12 & 11 & 43 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 5 \\ -6 \\ 8 \\ 9 \\ 3 \end{bmatrix}$$

Which two rows would you switch?

Example (2nd step of FE)

$$\begin{bmatrix} 6 & 14 & 5.1 & 3.7 & 6 \\ 0 & -17 & 12 & 11 & 43 \\ 0 & 4 & 12 & 1 & 11 \\ 0 & 9 & 23 & 6 & 8 \\ 0 & -7 & 6 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 8 \\ 9 \\ -6 \end{bmatrix}$$

Switched Rows

Gaussian Elimination with Partial Pivoting

A method to solve simultaneous linear equations of the form $[A][X]=[C]$

Two steps

1. Forward Elimination
2. Back Substitution

Forward Elimination

Same as naïve Gauss elimination method except that we switch rows before **each** of the $(n-1)$ steps of forward elimination.

Example: Matrix Form at Beginning of 2nd Step of Forward Elimination

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a'_{22} & a'_{23} & \cdots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & \cdots & a'_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & a'_{n2} & a'_{n3} & a'_{n4} & a'_{nn} \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{array} \right] = \left[\begin{array}{c} b_1 \\ b'_2 \\ b'_3 \\ \vdots \\ b'_n \end{array} \right]$$

Matrix Form at End of Forward Elimination

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a'_{22} & a'_{23} & \cdots & a'_{2n} \\ 0 & 0 & a''_{33} & \cdots & a''_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & a_{nn}^{(n-1)} \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{array} \right] = \left[\begin{array}{c} b_1 \\ b'_2 \\ b''_3 \\ \vdots \\ b_n^{(n-1)} \end{array} \right]$$

Back Substitution Starting Eqns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$\overset{'}{a_{22}}x_2 + \overset{'}{a_{23}}x_3 + \dots + \overset{'}{a_{2n}}x_n = \overset{'}{b_2}$$

$$\overset{''}{a_{33}}x_3 + \dots + \overset{''}{a_{nn}}x_n = \overset{''}{b_3}$$

⋮ ⋮ ⋮ ⋮

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}$$

Back Substitution

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^n a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}} \text{ for } i = n-1, \dots, 1$$

THE END

<http://numericalmethods.eng.usf.edu>

Gauss Elimination with Partial Pivoting Example

<http://numericalmethods.eng.usf.edu>

Example 2

Solve the following set of equations
by Gaussian elimination with partial
pivoting

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Example 2 Cont.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix} \Rightarrow \begin{bmatrix} 25 & 5 & 1 & : & 106.8 \\ 64 & 8 & 1 & : & 177.2 \\ 144 & 12 & 1 & : & 279.2 \end{bmatrix}$$

1. Forward Elimination
2. Back Substitution

Forward Elimination

Number of Steps of Forward Elimination

Number of steps of forward elimination is
 $(n-1)=(3-1)=2$

Forward Elimination: Step 1

- Examine absolute values of first column, first row and below.

$|25|, |64|, |144|$

- Largest absolute value is 144 and exists in row 3.
- Switch row 1 and row 3.

$$\left[\begin{array}{ccc|c} 25 & 5 & 1 & : 106.8 \\ 64 & 8 & 1 & : 177.2 \\ 144 & 12 & 1 & : 279.2 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 64 & 8 & 1 & : 177.2 \\ 25 & 5 & 1 & : 106.8 \end{array} \right]$$

Forward Elimination: Step 1 (cont.)

$$\begin{bmatrix} 144 & 12 & 1 & : & 279.2 \\ 64 & 8 & 1 & : & 177.2 \\ 25 & 5 & 1 & : & 106.8 \end{bmatrix}$$

Divide Equation 1 by 144 and multiply it by 64, $\frac{64}{144} = 0.4444$.

$$[144 \ 12 \ 1 \ : \ 279.2] \times 0.4444 = [63.99 \ 5.333 \ 0.4444 \ : \ 124.1]$$

Subtract the result from
Equation 2

$$\begin{array}{r} [64 \qquad \qquad \qquad 1 \ : \ 177.2] \\ - [63.99 \qquad 5.333 \qquad 0.4444 \ : \ 124.1] \\ \hline [0 \qquad 2.667 \qquad 0.5556 \ : \ 53.10] \end{array}$$

Substitute new equation for
Equation 2

$$\begin{bmatrix} 144 & 12 & 1 & : & 279.2 \\ 0 & 2.667 & 0.5556 & : & 53.10 \\ 25 & 5 & 1 & : & 106.8 \end{bmatrix}$$

Forward Elimination: Step 1 (cont.)

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.667 & 0.5556 & : 53.10 \\ 25 & 5 & 1 & : 106.8 \end{array} \right] \quad \text{Divide Equation 1 by 144 and multiply it by 25, } \frac{25}{144} = 0.1736.$$

$$[144 \ 12 \ 1 \ : \ 279.2] \times 0.1736 = [25.00 \ 2.083 \ 0.1736 \ : \ 48.47]$$

Subtract the result from Equation 3

$$\begin{array}{r} [25 \ 5 \ 1 \ : \ 106.8] \\ - [25 \ 2.083 \ 0.1736 \ : \ 48.47] \\ \hline [0 \ 2.917 \ 0.8264 \ : \ 58.33] \end{array}$$

Substitute new equation for Equation 3

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.667 & 0.5556 & : 53.10 \\ 0 & 2.917 & 0.8264 & : 58.33 \end{array} \right]$$

Forward Elimination: Step 2

- Examine absolute values of second column, second row and below.

$$|2.667|, |2.917|$$

- Largest absolute value is 2.917 and exists in row 3.
- Switch row 2 and row 3.

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.667 & 0.5556 & : 53.10 \\ 0 & 2.917 & 0.8264 & : 58.33 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.917 & 0.8264 & : 58.33 \\ 0 & 2.667 & 0.5556 & : 53.10 \end{array} \right]$$

Forward Elimination: Step 2 (cont.)

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.917 & 0.8264 & : 58.33 \\ 0 & 2.667 & 0.5556 & : 53.10 \end{array} \right]$$

Divide Equation 2 by 2.917 and multiply it by 2.667,
 $\frac{2.667}{2.917} = 0.9143$.

$$[0 \ 2.917 \ 0.8264 \ : \ 58.33] \times 0.9143 = [0 \ 2.667 \ 0.7556 \ : \ 53.33]$$

Subtract the result from Equation 3

$$\begin{array}{r} [0 \ 2.667 \ 0.5556 \ : \ 53.10] \\ - [0 \ 2.667 \ 0.7556 \ : \ 53.33] \\ \hline [0 \ 0 \ -0.2 \ : \ -0.23] \end{array}$$

Substitute new equation for Equation 3

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.917 & 0.8264 & : 58.33 \\ 0 & 0 & -0.2 & : -0.23 \end{array} \right]$$

Back Substitution

Back Substitution

$$\left[\begin{array}{ccc|c} 144 & 12 & 1 & : 279.2 \\ 0 & 2.917 & 0.8264 & : 58.33 \\ 0 & 0 & -0.2 & : -0.23 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 144 & 12 & 1 & a_1 \\ 0 & 2.917 & 0.8264 & a_2 \\ 0 & 0 & -0.2 & a_3 \end{array} \right] = \left[\begin{array}{c} 279.2 \\ 58.33 \\ -0.23 \end{array} \right]$$

Solving for a_3

$$-0.2a_3 = -0.23$$

$$a_3 = \frac{-0.23}{-0.2} = 1.15$$

Back Substitution (cont.)

$$\begin{bmatrix} 144 & 12 & 1 \\ 0 & 2.917 & 0.8264 \\ 0 & 0 & -0.2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 279.2 \\ 58.33 \\ -0.23 \end{bmatrix}$$

Solving for a_2

$$2.917a_2 + 0.8264a_3 = 58.33$$

$$\begin{aligned} a_2 &= \frac{58.33 - 0.8264a_3}{2.917} \\ &= \frac{58.33 - 0.8264 \times 1.15}{2.917} \\ &= 19.67 \end{aligned}$$

Back Substitution (cont.)

$$\begin{bmatrix} 144 & 12 & 1 \\ 0 & 2.917 & 0.8264 \\ 0 & 0 & -0.2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 279.2 \\ 58.33 \\ -0.23 \end{bmatrix}$$

Solving for a_1

$$144a_1 + 12a_2 + a_3 = 279.2$$

$$\begin{aligned} a_1 &= \frac{279.2 - 12a_2 - a_3}{144} \\ &= \frac{279.2 - 12 \times 19.67 - 1.15}{144} \\ &= 0.2917 \end{aligned}$$

Gaussian Elimination with Partial Pivoting Solution

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.2917 \\ 19.67 \\ 1.15 \end{bmatrix}$$

Gauss Elimination with Partial Pivoting Another Example

<http://numericalmethods.eng.usf.edu>

Partial Pivoting: Example

Consider the system of equations

$$10x_1 - 7x_2 = 7$$

$$-3x_1 + 2.099x_2 + 6x_3 = 3.901$$

$$5x_1 - x_2 + 5x_3 = 6$$

In matrix form

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix}$$

Solve using Gaussian Elimination with Partial Pivoting using five significant digits with chopping

Partial Pivoting: Example

Forward Elimination: Step 1

Examining the values of the first column

$|10|$, $|-3|$, and $|5|$ or 10, 3, and 5

The largest absolute value is 10, which means, to follow the rules of Partial Pivoting, we switch row1 with row1.

Performing Forward Elimination

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix} \implies \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix}$$

Partial Pivoting: Example

Forward Elimination: Step 2

Examining the values of the first column

$|-0.001|$ and $|2.5|$ or 0.0001 and 2.5

The largest absolute value is 2.5 , so row 2 is switched with row 3

Performing the row swap

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix} \implies \begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & -0.001 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.5 \\ 6.001 \end{bmatrix}$$

Partial Pivoting: Example

Forward Elimination: Step 2

Performing the Forward Elimination results in:

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.5 \\ 6.002 \end{bmatrix}$$

Partial Pivoting: Example

Back Substitution

Solving the equations through back substitution

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.5 \\ 6.002 \end{bmatrix}$$
$$x_3 = \frac{6.002}{6.002} = 1$$
$$x_2 = \frac{2.5 - 5x_3}{2.5} = -1$$
$$x_1 = \frac{7 + 7x_2 - 0x_3}{10} = 0$$

Partial Pivoting: Example

Compare the calculated and exact solution

The fact that they are equal is coincidence, but it does illustrate the advantage of Partial Pivoting

$$[X]_{calculated} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \quad [X]_{exact} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

THE END

<http://numericalmethods.eng.usf.edu>

Determinant of a Square Matrix Using Naïve Gauss Elimination

Example

<http://numericalmethods.eng.usf.edu>

Theorem of Determinants

If a multiple of one row of $[A]_{n \times n}$ is added or subtracted to another row of $[A]_{n \times n}$ to result in $[B]_{n \times n}$ then $\det(A) = \det(B)$

Theorem of Determinants

The determinant of an upper triangular matrix

$[A]_{n \times n}$ is given by

$$\det(A) = a_{11} \times a_{22} \times \dots \times a_{ii} \times \dots \times a_{nn}$$

$$= \prod_{i=1}^n a_{ii}$$

Forward Elimination of a Square Matrix

Using forward elimination to transform $[A]_{n \times n}$ to an upper triangular matrix, $[U]_{n \times n}$.

$$[A]_{n \times n} \rightarrow [U]_{n \times n}$$

$$\det(A) = \det(U)$$

Example

Using naïve Gaussian elimination find the determinant of the following square matrix.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

Forward Elimination

Forward Elimination: Step 1

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

Divide Equation 1 by 25 and multiply it by 64, $\frac{64}{25} = 2.56$.

$$[25 \ 5 \ 1] \times 2.56 = [64 \ 12.8 \ 2.56]$$

$$\begin{bmatrix} 64 & 8 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 64 & 12.8 & 2.56 \end{bmatrix}$$

$$\begin{bmatrix} 0 & -4.8 & -1.56 \end{bmatrix}$$

Subtract the result from Equation 2

Substitute new equation for Equation 2

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix}$$

Forward Elimination: Step 1 (cont.)

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix}$$

Divide Equation 1 by 25 and multiply it by 144, $\frac{144}{25} = 5.76$.

$$[25 \ 5 \ 1] \times 5.76 = [144 \ 28.8 \ 5.76]$$

$$\begin{array}{r} [144 \ 12 \ 1] \\ - [144 \ 28.8 \ 5.76] \\ \hline [0 \ -16.8 \ -4.76] \end{array}$$

Subtract the result from Equation 3

Substitute new equation for Equation 3

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

Forward Elimination: Step 2

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

Divide Equation 2 by -4.8
and multiply it by -16.8 ,
 $\frac{-16.8}{-4.8} = 3.5$.

$$([0 \quad -4.8 \quad -1.56]) \times 3.5 = [0 \quad -16.8 \quad -5.46]$$

Subtract the result from
Equation 3

$$\begin{array}{r} [0 \quad -16.8 \quad -4.76] \\ - [0 \quad -16.8 \quad -5.46] \\ \hline [0 \quad 0 \quad 0.7] \end{array}$$

Substitute new equation for
Equation 3

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

Finding the Determinant

After forward elimination

$$\left[\begin{array}{ccc} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{array} \right] \rightarrow \left[\begin{array}{ccc} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{array} \right]$$

$$\begin{aligned}\det(A) &= u_{11} \times u_{22} \times u_{33} \\ &= 25 \times (-4.8) \times 0.7 \\ &= -84.00\end{aligned}$$

Summary

- Forward Elimination
- Back Substitution
- Pitfalls
- Improvements
- Partial Pivoting
- Determinant of a Matrix

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/gaussian_elimination.html

THE END

<http://numericalmethods.eng.usf.edu>

LU Decomposition

Major: All Engineering Majors

Authors: Autar Kaw

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

LU Decomposition

<http://numericalmethods.eng.usf.edu>

LU Decomposition

LU Decomposition is another method to solve a set of simultaneous linear equations

Which is better, Gauss Elimination or LU Decomposition?

To answer this, a closer look at LU decomposition is needed.

LU Decomposition

Method

For most non-singular matrix $[A]$ that one could conduct Naïve Gauss Elimination forward elimination steps, one can always write it as

$$[A] = [L][U]$$

where

$[L]$ = lower triangular matrix

$[U]$ = upper triangular matrix

How does LU Decomposition work?

If solving a set of linear equations

If $[A] = [L][U]$ then

Multiply by

Which gives

Remember $[L]^{-1}[L] = [I]$ which leads to

Now, if $[I][U] = [U]$ then

Now, let

Which ends with

and

$$[A][X] = [C]$$

$$[L][U][X] = [C]$$

$$[L]^{-1}$$

$$[L]^{-1}[L][U][X] = [L]^{-1}[C]$$

$$[I][U][X] = [L]^{-1}[C]$$

$$[U][X] = [L]^{-1}[C]$$

$$[L]^{-1}[C] = [Z]$$

$$[L][Z] = [C] \quad (1)$$

$$[U][X] = [Z] \quad (2)$$

LU Decomposition

How can this be used?

Given $[A][X] = [C]$

1. Decompose $[A]$ into $[L]$ and $[U]$
2. Solve $[L][Z] = [C]$ for $[Z]$
3. Solve $[U][X] = [Z]$ for $[X]$

Is LU Decomposition better than Gaussian Elimination?

Solve $[A][X] = [B]$

T = clock cycle time and nxn = size of the matrix

Forward Elimination

$$CT|_{FE} = T \left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3} \right)$$

Back Substitution

$$CT|_{BS} = T(4n^2 + 12n)$$

Decomposition to LU

$$CT|_{DE} = T \left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3} \right)$$

Forward Substitution

$$CT|_{FS} = T(4n^2 - 4n)$$

Back Substitution

$$CT|_{BS} = T(4n^2 + 12n)$$

Is LU Decomposition better than Gaussian Elimination?

To solve $[A][X] = [B]$

Time taken by methods

Gaussian Elimination	LU Decomposition
$T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right)$	$T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right)$

T = clock cycle time and nxn = size of the matrix

So both methods are equally efficient.

To find inverse of [A]

Time taken by Gaussian Elimination

$$\begin{aligned} &= n(CT|_{FE} + CT|_{BS}) \\ &= T\left(\frac{8n^4}{3} + 12n^3 + \frac{4n^2}{3}\right) \end{aligned}$$

Time taken by LU Decomposition

$$\begin{aligned} &= CT|_{DE} + n \times CT|_{FS} + n \times CT|_{BS} \\ &= T\left(\frac{32n^3}{3} + 12n^2 - \frac{20n}{3}\right) \end{aligned}$$

To find inverse of [A]

Time taken by Gaussian Elimination

$$T\left(\frac{8n^4}{3} + 12n^3 + \frac{4n^2}{3}\right)$$

Time taken by LU Decomposition

$$T\left(\frac{32n^3}{3} + 12n^2 - \frac{20n}{3}\right)$$

Table 1 Comparing computational times of finding inverse of a matrix using LU decomposition and Gaussian elimination.

n	10	100	1000	10000
$CT _{\text{inverse GE}} / CT _{\text{inverse LU}}$	3.288	25.84	250.8	2501

For large n , $CT|_{\text{inverse GE}} / CT|_{\text{inverse LU}} \approx n/4$

Method: [A] Decomposes to [L] and [U]

$$[A] = [L][U] = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

[U] is the same as the coefficient matrix at the end of the forward elimination step.

[L] is obtained using the *multipliers* that were used in the forward elimination process

Finding the $[U]$ matrix

Using the Forward Elimination Procedure of Gauss Elimination

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

Step 1: $\frac{64}{25} = 2.56$; $Row2 - Row1(2.56) = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix}$

$\frac{144}{25} = 5.76$; $Row3 - Row1(5.76) = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$

Finding the [U] Matrix

Matrix after Step 1:

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

Step 2: $\frac{-16.8}{-4.8} = 3.5$; $Row3 - Row2(3.5) =$

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

$$[U] = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

Finding the [L] matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix}$$

Using the multipliers used during the Forward Elimination Procedure

From the first step
of forward
elimination

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

$$\ell_{21} = \frac{a_{21}}{a_{11}} = \frac{64}{25} = 2.56$$

$$\ell_{31} = \frac{a_{31}}{a_{11}} = \frac{144}{25} = 5.76$$

Finding the [L] Matrix

From the second
step of forward
elimination

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix} \quad \ell_{32} = \frac{a_{32}}{a_{22}} = \frac{-16.8}{-4.8} = 3.5$$

$$[L] = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix}$$

Does $[L][U] = [A]$?

$$[L][U] = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} = ?$$

Using LU Decomposition to solve SLEs

Solve the following set of linear equations using LU Decomposition

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Using the procedure for finding the $[L]$ and $[U]$ matrices

$$[A] = [L][U] = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 25 \\ 2.56 & 1 & 0 & 0 \\ 5.76 & 3.5 & 1 & 0 \end{array} \right] \left[\begin{array}{ccc} 5 & 1 & 1 \\ -4.8 & -1.56 & 0 \\ 0 & 0 & 0.7 \end{array} \right]$$

Example

Set $[L][Z] = [C]$

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Solve for $[Z]$

$$z_1 = 10$$

$$2.56z_1 + z_2 = 177.2$$

$$5.76z_1 + 3.5z_2 + z_3 = 279.2$$

Example

Complete the forward substitution to solve for [Z]

$$z_1 = 106.8$$

$$\begin{aligned} z_2 &= 177.2 - 2.56z_1 \\ &= 177.2 - 2.56(106.8) \end{aligned}$$

$$= -96.2$$

$$\begin{aligned} z_3 &= 279.2 - 5.76z_1 - 3.5z_2 \\ &= 279.2 - 5.76(106.8) - 3.5(-96.21) \\ &= 0.735 \end{aligned}$$

$$[Z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

Example

Set $[U][X] = [Z]$

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

Solve for $[X]$

The 3 equations become

$$25a_1 + 5a_2 + a_3 = 106.8$$

$$-4.8a_2 - 1.56a_3 = -96.21$$

$$0.7a_3 = 0.735$$

Example

From the 3rd equation

$$0.7a_3 = 0.735$$

$$a_3 = \frac{0.735}{0.7}$$

$$a_3 = 1.050$$

Substituting in a_3 and using the second equation

$$-4.8a_2 - 1.56a_3 = -96.21$$

$$a_2 = \frac{-96.21 + 1.56a_3}{-4.8}$$

$$a_2 = \frac{-96.21 + 1.56(1.050)}{-4.8}$$

$$a_2 = 19.70$$

Example

Substituting in a_3 and a_2 using
the first equation

$$25a_1 + 5a_2 + a_3 = 106.8$$

$$\begin{aligned}a_1 &= \frac{106.8 - 5a_2 - a_3}{25} \\&= \frac{106.8 - 5(19.70) - 1.050}{25} \\&= 0.2900\end{aligned}$$

Hence the Solution Vector is:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.2900 \\ 19.70 \\ 1.050 \end{bmatrix}$$

Finding the inverse of a square matrix

The inverse $[B]$ of a square matrix $[A]$ is defined as

$$[A][B] = [I] = [B][A]$$

Finding the inverse of a square matrix

How can LU Decomposition be used to find the inverse?

Assume the first column of $[B]$ to be $[b_{11} \ b_{12} \ \dots \ b_{n1}]^T$

Using this and the definition of matrix multiplication

First column of $[B]$

$$[A] \begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Second column of $[B]$

$$[A] \begin{bmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

The remaining columns in $[B]$ can be found in the same manner

Example: Inverse of a Matrix

Find the inverse of a square matrix $[A]$

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

Using the decomposition procedure, the $[L]$ and $[U]$ matrices are found to be

$$[A] = [L][U] = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

Example: Inverse of a Matrix

Solving for the each column of $[B]$ requires two steps

- 1) Solve $[L][Z] = [C]$ for $[Z]$
- 2) Solve $[U][X] = [Z]$ for $[X]$

$$\text{Step 1: } [L][Z] = [C] \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

This generates the equations:

$$z_1 = 1$$

$$2.56z_1 + z_2 = 0$$

$$5.76z_1 + 3.5z_2 + z_3 = 0$$

Example: Inverse of a Matrix

Solving for [Z]

$$z_1 = 1$$

$$\begin{aligned} z_2 &= 0 - 2.56z_1 \\ &= 0 - 2.56(1) \\ &= -2.56 \end{aligned}$$

$$\begin{aligned} z_3 &= 0 - 5.76z_1 - 3.5z_2 \\ &= 0 - 5.76(1) - 3.5(-2.56) \\ &= 3.2 \end{aligned}$$

$$[Z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix}$$

Example: Inverse of a Matrix

Solving $[U][X] = [Z]$ for $[X]$

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix}$$

$$25b_{11} + 5b_{21} + b_{31} = 1$$

$$-4.8b_{21} - 1.56b_{31} = -2.56$$

$$0.7b_{31} = 3.2$$

Example: Inverse of a Matrix

Using Backward Substitution

$$b_{31} = \frac{3.2}{0.7} = 4.571$$

$$\begin{aligned} b_{21} &= \frac{-2.56 + 1.560b_{31}}{-4.8} \\ &= \frac{-2.56 + 1.560(4.571)}{-4.8} = -0.9524 \end{aligned}$$

$$\begin{aligned} b_{11} &= \frac{1 - 5b_{21} - b_{31}}{25} \\ &= \frac{1 - 5(-0.9524) - 4.571}{25} = 0.04762 \end{aligned}$$

So the first column of
the inverse of $[A]$ is:

$$\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 0.04762 \\ -0.9524 \\ 4.571 \end{bmatrix}$$

Example: Inverse of a Matrix

Repeating for the second and third columns of the inverse

Second Column

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} -0.08333 \\ 1.417 \\ -5.000 \end{bmatrix}$$

Third Column

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0.03571 \\ -0.4643 \\ 1.429 \end{bmatrix}$$

Example: Inverse of a Matrix

The inverse of $[A]$ is

$$[A]^{-1} = \begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix}$$

To check your work do the following operation

$$[A][A]^{-1} = [I] = [A]^{-1}[A]$$

Additional Resources

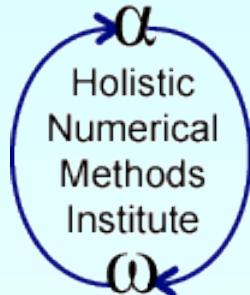
For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/lu_decomposition.html

THE END

<http://numericalmethods.eng.usf.edu>

Gauss-Siedel Method



Major: All Engineering Majors

Authors: Autar Kaw

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Gauss-Seidel Method

<http://numericalmethods.eng.usf.edu>

Gauss-Seidel Method

An iterative method.

Basic Procedure:

- Algebraically solve each linear equation for x_i
- Assume an initial guess solution array
- Solve for each x_i and repeat
- Use absolute relative approximate error after each iteration to check if error is within a pre-specified tolerance.

Gauss-Seidel Method

Why?

The Gauss-Seidel Method allows the user to control round-off error.

Elimination methods such as Gaussian Elimination and LU Decomposition are prone to round-off error.

Also: If the physics of the problem are understood, a close initial guess can be made, decreasing the number of iterations needed.

Gauss-Seidel Method

Algorithm

A set of n equations and n unknowns:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = b_2$$

⋮
⋮
⋮
⋮

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n = b_n$$

If: the diagonal elements are non-zero

Rewrite each equation solving for the corresponding unknown

ex:

First equation, solve for x_1

Second equation, solve for x_2

Gauss-Seidel Method

Algorithm

Rewriting each equation

$$x_1 = \frac{c_1 - a_{12}x_2 - a_{13}x_3 - \dots - a_{1n}x_n}{a_{11}} \quad \xleftarrow{\hspace{10em}} \text{From Equation 1}$$

$$x_2 = \frac{c_2 - a_{21}x_1 - a_{23}x_3 - \dots - a_{2n}x_n}{a_{22}} \quad \xleftarrow{\hspace{10em}} \text{From equation 2}$$
$$\vdots \quad \vdots \quad \vdots$$

$$x_{n-1} = \frac{c_{n-1} - a_{n-1,1}x_1 - a_{n-1,2}x_2 - \dots - a_{n-1,n-2}x_{n-2} - a_{n-1,n}x_n}{a_{n-1,n-1}} \quad \xleftarrow{\hspace{10em}} \text{From equation n-1}$$

$$x_n = \frac{c_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{n,n-1}x_{n-1}}{a_{nn}} \quad \xleftarrow{\hspace{10em}} \text{From equation n}$$

Gauss-Seidel Method

Algorithm

General Form of each equation

$$x_1 = \frac{c_1 - \sum_{\substack{j=1 \\ j \neq 1}}^n a_{1j} x_j}{a_{11}}$$
$$x_{n-1} = \frac{c_{n-1} - \sum_{\substack{j=1 \\ j \neq n-1}}^n a_{n-1,j} x_j}{a_{n-1,n-1}}$$
$$x_2 = \frac{c_2 - \sum_{\substack{j=1 \\ j \neq 2}}^n a_{2j} x_j}{a_{22}}$$
$$x_n = \frac{c_n - \sum_{\substack{j=1 \\ j \neq n}}^n a_{nj} x_j}{a_{nn}}$$

Gauss-Seidel Method

Algorithm

General Form for any row 'i'

$$x_i = \frac{c_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j}{a_{ii}}, i = 1, 2, \dots, n.$$

How or where can this equation be used?

Gauss-Seidel Method

Solve for the unknowns

Assume an initial guess for [X]

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}$$

Use rewritten equations to solve for each value of x_i .

Important: Remember to use the most recent value of x_i . Which means to apply values calculated to the calculations remaining in the **current** iteration.

Gauss-Seidel Method

Calculate the Absolute Relative Approximate Error

$$|\epsilon_a|_i = \left| \frac{x_i^{new} - x_i^{old}}{x_i^{new}} \right| \times 100$$

So when has the answer been found?

The iterations are stopped when the absolute relative approximate error is less than a prespecified tolerance for all unknowns.

Gauss-Seidel Method: Example 1

The upward velocity of a rocket is given at three different times

Table 1 Velocity vs. Time data.

Time, t (s)	Velocity v (m/s)
5	106.8
8	177.2
12	279.2



The velocity data is approximated by a polynomial as:

$$v(t) = a_1 t^2 + a_2 t + a_3, \quad 5 \leq t \leq 12.$$

Gauss-Seidel Method: Example 1

Using a Matrix template of the form

$$\begin{bmatrix} t_1^2 & t_1 & 1 \\ t_2^2 & t_2 & 1 \\ t_3^2 & t_3 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

The system of equations becomes

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Initial Guess: Assume an initial guess of

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

Gauss-Seidel Method: Example 1

Rewriting each equation

$$a_1 = \frac{106.8 - 5a_2 - a_3}{25}$$

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

$$a_2 = \frac{177.2 - 64a_1 - a_3}{8}$$

$$a_3 = \frac{279.2 - 144a_1 - 12a_2}{1}$$

Gauss-Seidel Method: Example 1

Applying the initial guess and solving for a_i

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

$$a_1 = \frac{106.8 - 5(2) - (5)}{25} = 3.6720$$

Initial Guess

$$a_2 = \frac{177.2 - 64(3.6720) - (5)}{8} = -7.8510$$

$$a_3 = \frac{279.2 - 144(3.6720) - 12(-7.8510)}{1} = -155.36$$

When solving for a_2 , how many of the initial guess values were used?

Gauss-Seidel Method: Example 1

Finding the absolute relative approximate error

$$|e_a|_i = \left| \frac{x_i^{new} - x_i^{old}}{x_i^{new}} \right| \times 100$$

$$|e_a|_1 = \left| \frac{3.6720 - 1.0000}{3.6720} \right| \times 100 = 72.76\%$$

$$|e_a|_2 = \left| \frac{-7.8510 - 2.0000}{-7.8510} \right| \times 100 = 125.47\%$$

$$|e_a|_3 = \left| \frac{-155.36 - 5.0000}{-155.36} \right| \times 100 = 103.22\%$$

At the end of the first iteration

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

The maximum absolute relative approximate error is 125.47%

Gauss-Seidel Method: Example 1

Using

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

from iteration #1

Iteration #2

the values of a_i are found:

$$a_1 = \frac{106.8 - 5(-7.8510) - 155.36}{25} = 12.056$$

$$a_2 = \frac{177.2 - 64(12.056) - 155.36}{8} = -54.882$$

$$a_3 = \frac{279.2 - 144(12.056) - 12(-54.882)}{1} = -798.34$$

Gauss-Seidel Method: Example 1

Finding the absolute relative approximate error

$$|\epsilon_a|_1 = \left| \frac{12.056 - 3.6720}{12.056} \right| \times 100 = 69.543\%$$

$$|\epsilon_a|_2 = \left| \frac{-54.882 - (-7.8510)}{-54.882} \right| \times 100 = 85.695\%$$

$$|\epsilon_a|_3 = \left| \frac{-798.34 - (-155.36)}{-798.34} \right| \times 100 = 80.540\%$$

At the end of the second iteration

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 12.056 \\ -54.882 \\ -798.54 \end{bmatrix}$$

The maximum absolute
relative approximate error is
85.695%

Gauss-Seidel Method: Example 1

Repeating more iterations, the following values are obtained

Iteration	a_1	$ e_a _1 \%$	a_2	$ e_a _2 \%$	a_3	$ e_a _3 \%$
1	3.6720	72.767	-7.8510	125.47	-155.36	103.22
2	12.056	69.543	-54.882	85.695	-798.34	80.540
3	47.182	74.447	-255.51	78.521	-3448.9	76.852
4	193.33	75.595	-1093.4	76.632	-14440	76.116
5	800.53	75.850	-4577.2	76.112	-60072	75.963
6	3322.6	75.906	-19049	75.972	-249580	75.931

Notice – The relative errors are not decreasing at any significant rate

Also, the solution is not converging to the true solution of

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.29048 \\ 19.690 \\ 1.0857 \end{bmatrix}$$

Gauss-Seidel Method: Pitfall

What went wrong?

Even though done correctly, the answer is not converging to the correct answer

This example illustrates a pitfall of the Gauss-Siedel method: not all systems of equations will converge.

Is there a fix?

One class of system of equations always converges: One with a *diagonally dominant* coefficient matrix.

Diagonally dominant: [A] in $[A][X] = [C]$ is diagonally dominant if:

$$\left|a_{ii}\right| \geq \sum_{\substack{j=1 \\ j \neq i}}^n \left|a_{ij}\right| \quad \text{for all 'i'} \quad \text{and} \quad \left|a_{ii}\right| > \sum_{\substack{j=1 \\ j \neq i}}^n \left|a_{ij}\right| \quad \text{for at least one 'i'}$$

Gauss-Seidel Method: Pitfall

Diagonally dominant: The coefficient on the diagonal must be at least equal to the sum of the other coefficients in that row and at least one row with a diagonal coefficient greater than the sum of the other coefficients in that row.

Which coefficient matrix is diagonally dominant?

$$[A] = \begin{bmatrix} 2 & 5.81 & 34 \\ 45 & 43 & 1 \\ 123 & 16 & 1 \end{bmatrix}$$

$$[B] = \begin{bmatrix} 124 & 34 & 56 \\ 23 & 53 & 5 \\ 96 & 34 & 129 \end{bmatrix}$$

Most physical systems do result in simultaneous linear equations that have diagonally dominant coefficient matrices.

Gauss-Seidel Method: Example 2

Given the system of equations

$$12x_1 + 3x_2 - 5x_3 = 1$$

$$x_1 + 5x_2 + 3x_3 = 28$$

$$3x_1 + 7x_2 + 13x_3 = 76$$

The coefficient matrix is:

$$[A] = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$

With an initial guess of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Will the solution converge using the Gauss-Siedel method?

Gauss-Seidel Method: Example 2

Checking if the coefficient matrix is diagonally dominant

$$[A] = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$
$$|a_{11}| = |12| = 12 \geq |a_{12}| + |a_{13}| = |3| + |-5| = 8$$
$$|a_{22}| = |5| = 5 \geq |a_{21}| + |a_{23}| = |1| + |3| = 4$$
$$|a_{33}| = |13| = 13 \geq |a_{31}| + |a_{32}| = |3| + |7| = 10$$

The inequalities are all true and at least one row is *strictly* greater than:

Therefore: The solution should converge using the Gauss-Siedel Method

Gauss-Seidel Method: Example 2

Rewriting each equation

$$\begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 28 \\ 76 \end{bmatrix}$$

$$x_1 = \frac{1 - 3x_2 + 5x_3}{12}$$

$$x_2 = \frac{28 - x_1 - 3x_3}{5}$$

$$x_3 = \frac{76 - 3x_1 - 7x_2}{13}$$

With an initial guess of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$x_1 = \frac{1 - 3(0) + 5(1)}{12} = 0.50000$$

$$x_2 = \frac{28 - (0.5) - 3(1)}{5} = 4.9000$$

$$x_3 = \frac{76 - 3(0.50000) - 7(4.9000)}{13} = 3.0923$$

Gauss-Seidel Method: Example 2

The absolute relative approximate error

$$|\epsilon_a|_1 = \left| \frac{0.50000 - 1.0000}{0.50000} \right| \times 100 = 100.00\%$$

$$|\epsilon_a|_2 = \left| \frac{4.9000 - 0}{4.9000} \right| \times 100 = 100.00\%$$

$$|\epsilon_a|_3 = \left| \frac{3.0923 - 1.0000}{3.0923} \right| \times 100 = 67.662\%$$

The maximum absolute relative error after the first iteration is 100%

Gauss-Seidel Method: Example 2

After Iteration #1

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.5000 \\ 4.9000 \\ 3.0923 \end{bmatrix}$$

Substituting the x values into the equations

$$x_1 = \frac{1 - 3(4.9000) + 5(3.0923)}{12} = 0.14679$$

$$x_2 = \frac{28 - (0.14679) - 3(3.0923)}{5} = 3.7153$$

$$x_3 = \frac{76 - 3(0.14679) - 7(4.900)}{13} = 3.8118$$

After Iteration #2

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.14679 \\ 3.7153 \\ 3.8118 \end{bmatrix}$$

Gauss-Seidel Method: Example 2

Iteration #2 absolute relative approximate error

$$|\epsilon_a|_1 = \left| \frac{0.14679 - 0.50000}{0.14679} \right| \times 100 = 240.61\%$$

$$|\epsilon_a|_2 = \left| \frac{3.7153 - 4.9000}{3.7153} \right| \times 100 = 31.889\%$$

$$|\epsilon_a|_3 = \left| \frac{3.8118 - 3.0923}{3.8118} \right| \times 100 = 18.874\%$$

The maximum absolute relative error after the first iteration is 240.61%

This is much larger than the maximum absolute relative error obtained in iteration #1. Is this a problem?

Gauss-Seidel Method: Example 2

Repeating more iterations, the following values are obtained

Iteration	a_1	$ e_a _1 \%$	a_2	$ e_a _2 \%$	a_3	$ e_a _3 \%$
1	0.50000	100.00	4.9000	100.00	3.0923	67.662
2	0.14679	240.61	3.7153	31.889	3.8118	18.876
3	0.74275	80.236	3.1644	17.408	3.9708	4.0042
4	0.94675	21.546	3.0281	4.4996	3.9971	0.65772
5	0.99177	4.5391	3.0034	0.82499	4.0001	0.074383
6	0.99919	0.74307	3.0001	0.10856	4.0001	0.00101

The solution obtained $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.99919 \\ 3.0001 \\ 4.0001 \end{bmatrix}$ is close to the exact solution of $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$.

Gauss-Seidel Method: Example 3

Given the system of equations

$$3x_1 + 7x_2 + 13x_3 = 76$$

$$x_1 + 5x_2 + 3x_3 = 28$$

$$12x_1 + 3x_2 - 5x_3 = 1$$

With an initial guess of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Rewriting the equations

$$x_1 = \frac{76 - 7x_2 - 13x_3}{3}$$

$$x_2 = \frac{28 - x_1 - 3x_3}{5}$$

$$x_3 = \frac{1 - 12x_1 - 3x_2}{-5}$$

Gauss-Seidel Method: Example 3

Conducting six iterations, the following values are obtained

Iteration	a_1	$ e_a _1 \%$	A_2	$ e_a _2 \%$	a_3	$ e_a _3 \%$
1	21.000	95.238	0.80000	100.00	50.680	98.027
2	-196.15	110.71	14.421	94.453	-462.30	110.96
3	-1995.0	109.83	-116.02	112.43	4718.1	109.80
4	-20149	109.90	1204.6	109.63	-47636	109.90
5	2.0364×10^5	109.89	-12140	109.92	4.8144×10^5	109.89
6	-2.0579×10^5	109.89	1.2272×10^5	109.89	-4.8653×10^6	109.89

The values are not converging.

Does this mean that the Gauss-Seidel method cannot be used?

Gauss-Seidel Method

The Gauss-Seidel Method can still be used

The coefficient matrix is not diagonally dominant

$$[A] = \begin{bmatrix} 3 & 7 & 13 \\ 1 & 5 & 3 \\ 12 & 3 & -5 \end{bmatrix}$$

But this is the same set of equations used in example #2, which did converge.

$$[A] = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$

If a system of linear equations is not diagonally dominant, check to see if rearranging the equations can form a diagonally dominant matrix.

Gauss-Seidel Method

Not every system of equations can be rearranged to have a diagonally dominant coefficient matrix.

Observe the set of equations

$$x_1 + x_2 + x_3 = 3$$

$$2x_1 + 3x_2 + 4x_3 = 9$$

$$x_1 + 7x_2 + x_3 = 9$$

Which equation(s) prevents this set of equation from having a diagonally dominant coefficient matrix?

Gauss-Seidel Method

Summary

- Advantages of the Gauss-Seidel Method
- Algorithm for the Gauss-Seidel Method
- Pitfalls of the Gauss-Seidel Method

Gauss-Seidel Method

Questions?

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/gauss_seidel.html

THE END

<http://numericalmethods.eng.usf.edu>

Direct Method of Interpolation

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Direct Method of Interpolation

<http://numericalmethods.eng.usf.edu>

What is Interpolation ?

Given $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, find the value of 'y' at a value of 'x' that is not given.

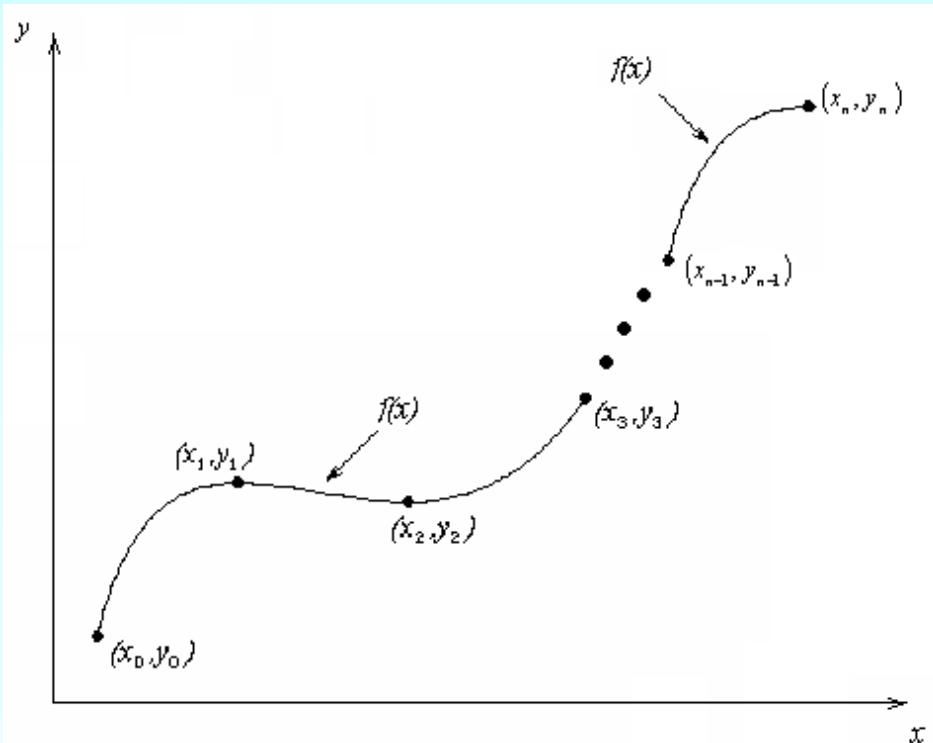


Figure 1 Interpolation of discrete.

Interpolants

Polynomials are the most common choice of interpolants because they are easy to:

- Evaluate
- Differentiate, and
- Integrate

Direct Method

Given 'n+1' data points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, pass a polynomial of order 'n' through the data as given below:

$$y = a_0 + a_1 x + \dots + a_n x^n.$$

where a_0, a_1, \dots, a_n are real constants.

- Set up 'n+1' equations to find 'n+1' constants.
- To find the value 'y' at a given value of 'x', simply substitute the value of 'x' in the above polynomial.



Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

Find the velocity at $t=16$ seconds using the direct method for linear interpolation.

Table 1 Velocity as a function of time.

$t, (\text{s})$	$v(t), (\text{m/s})$
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

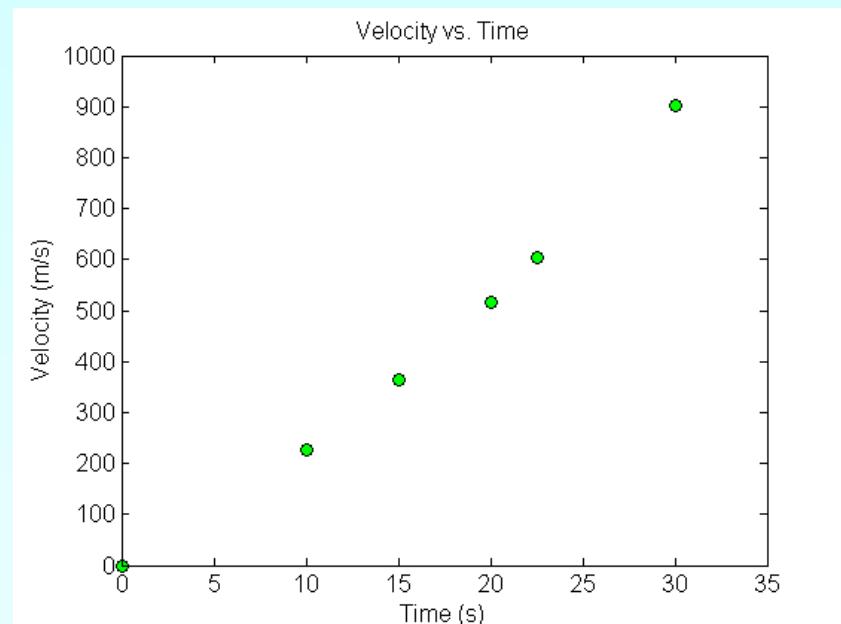


Figure 2 Velocity vs. time data for the rocket example

Linear Interpolation

$$v(t) = a_0 + a_1 t$$

$$v(15) = a_0 + a_1(15) = 362.78$$

$$v(20) = a_0 + a_1(20) = 517.35$$

Solving the above two equations gives,

$$a_0 = -100.93 \quad a_1 = 30.914$$

Hence

$$v(t) = -100.93 + 30.914t, \quad 15 \leq t \leq 20.$$

$$v(16) = -100.93 + 30.914(16) = 393.7 \text{ m/s}$$

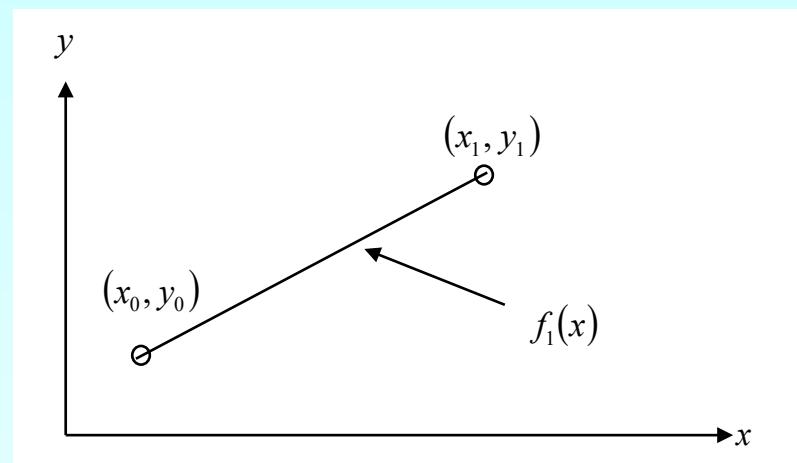


Figure 3 Linear interpolation.



Example 2

The upward velocity of a rocket is given as a function of time in Table 2.

Find the velocity at $t=16$ seconds using the direct method for quadratic interpolation.

Table 2 Velocity as a function of time.

$t, (\text{s})$	$v(t), (\text{m/s})$
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

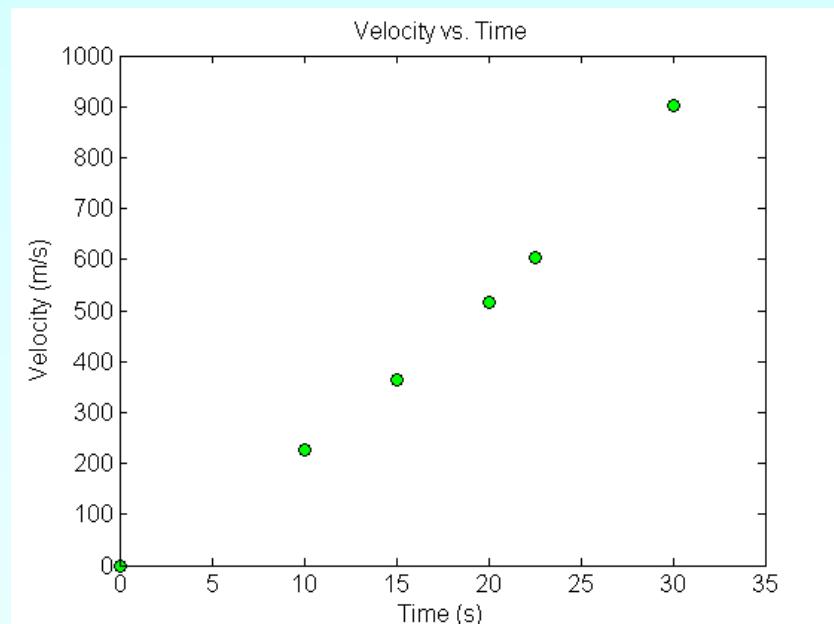


Figure 5 Velocity vs. time data for the rocket example

Quadratic Interpolation

$$v(t) = a_0 + a_1 t + a_2 t^2$$

$$v(10) = a_0 + a_1(10) + a_2(10)^2 = 227.04$$

$$v(15) = a_0 + a_1(15) + a_2(15)^2 = 362.78$$

$$v(20) = a_0 + a_1(20) + a_2(20)^2 = 517.35$$

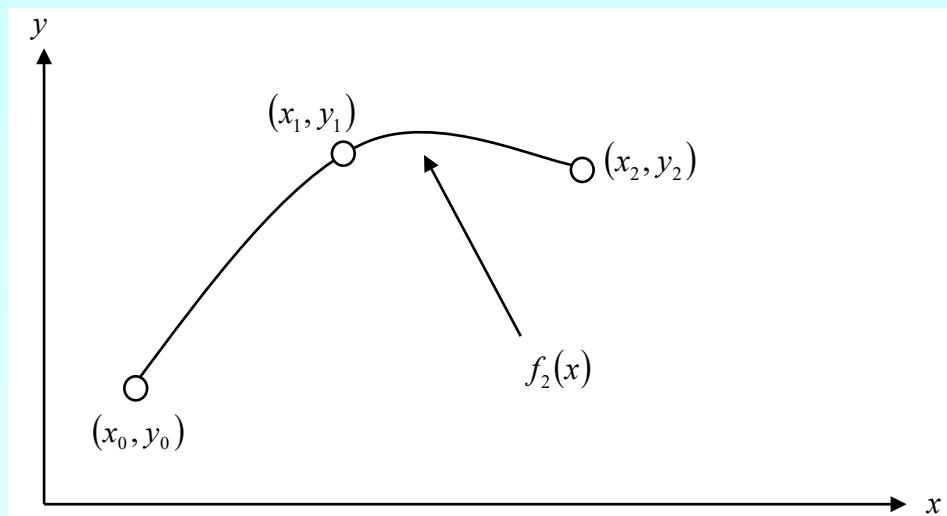


Figure 6 Quadratic interpolation.

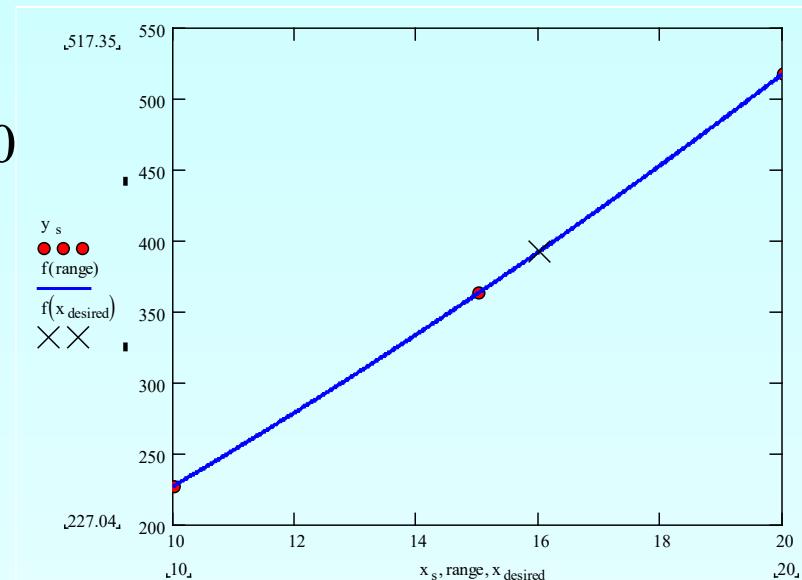
Solving the above three equations gives

$$a_0 = 12.05 \quad a_1 = 17.733 \quad a_2 = 0.3766$$

Quadratic Interpolation (cont.)

$$v(t) = 12.05 + 17.733t + 0.3766t^2, \quad 10 \leq t \leq 20$$

$$\begin{aligned} v(16) &= 12.05 + 17.733(16) + 0.3766(16)^2 \\ &= 392.19 \text{ m/s} \end{aligned}$$



The absolute relative approximate error $|e_a|$ obtained between the results from the first and second order polynomial is

$$\begin{aligned} |e_a| &= \left| \frac{392.19 - 393.70}{392.19} \right| \times 100 \\ &= 0.38410\% \end{aligned}$$



Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

Find the velocity at $t=16$ seconds using the direct method for cubic interpolation.

Table 3 Velocity as a function of time.

$t, (\text{s})$	$v(t), (\text{m/s})$
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

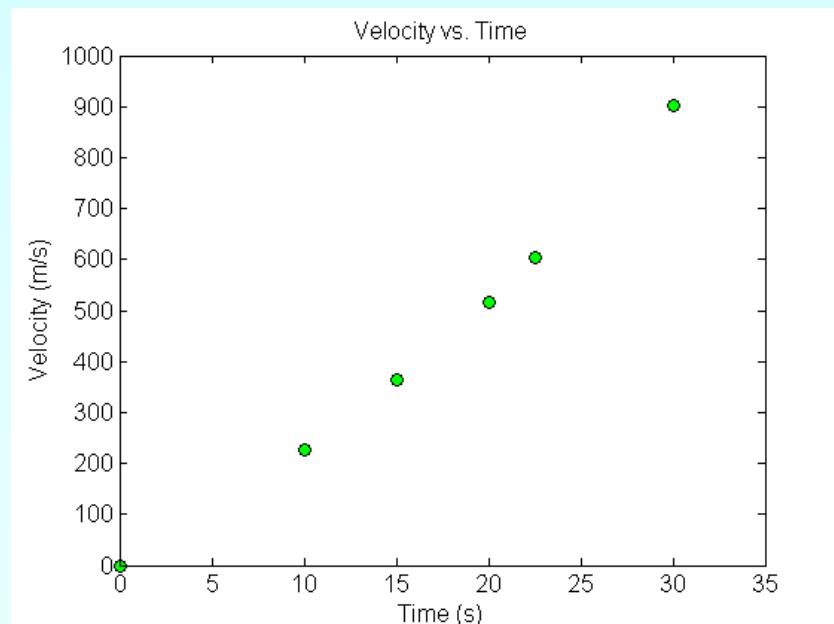


Figure 6 Velocity vs. time data for the rocket example

Cubic Interpolation

$$v(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

$$v(10) = 227.04 = a_0 + a_1(10) + a_2(10)^2 + a_3(10)^3$$

$$v(15) = 362.78 = a_0 + a_1(15) + a_2(15)^2 + a_3(15)^3$$

$$v(20) = 517.35 = a_0 + a_1(20) + a_2(20)^2 + a_3(20)^3$$

$$v(22.5) = 602.97 = a_0 + a_1(22.5) + a_2(22.5)^2 + a_3(22.5)^3$$

$$a_0 = -4.2540 \quad a_1 = 21.266 \quad a_2 = 0.13204 \quad a_3 = 0.0054347$$

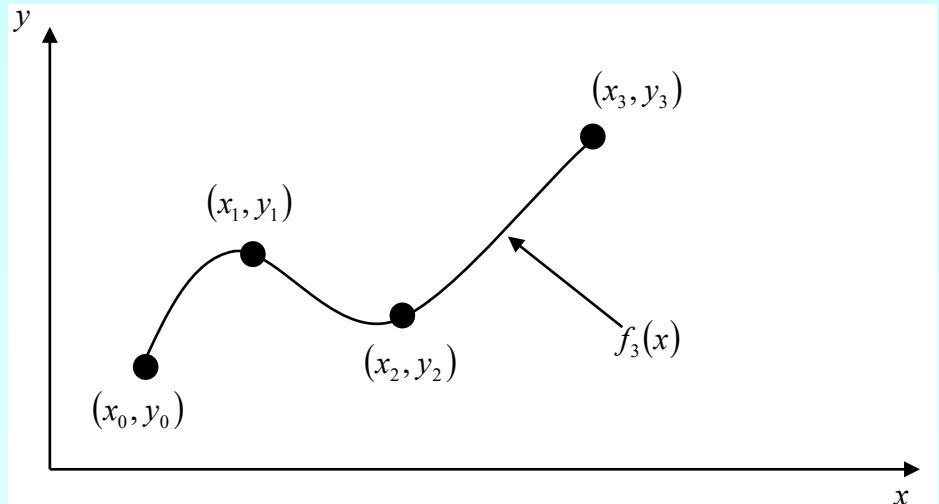
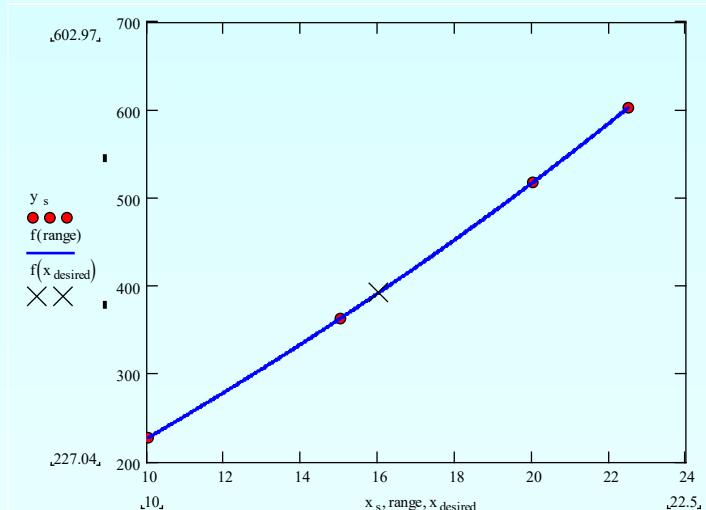


Figure 7 Cubic interpolation.

Cubic Interpolation (contd)

$$v(t) = -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \leq t \leq 22.5$$

$$\begin{aligned} v(16) &= -4.2540 + 21.266(16) + 0.13204(16)^2 + 0.0054347(16)^3 \\ &= 392.06 \text{ m/s} \end{aligned}$$



The absolute percentage relative approximate error $|e_a|$ between second and third order polynomial is

$$\begin{aligned} |e_a| &= \left| \frac{392.06 - 392.19}{392.06} \right| \times 100 \\ &= 0.033269\% \end{aligned}$$

Comparison Table

Table 4 Comparison of different orders of the polynomial.

Order of Polynomial	1	2	3
$v(t = 16)$ m/s	393.7	392.19	392.06
Absolute Relative Approximate Error	-----	0.38410 %	0.033269 %

Distance from Velocity Profile

Find the distance covered by the rocket from $t=11\text{s}$ to $t=16\text{s}$?

$$v(t) = -4.3810 + 21.289t + 0.13064t^2 + 0.0054606t^3, \quad 10 \leq t \leq 22.5$$

$$\begin{aligned} s(16) - s(11) &= \int_{11}^{16} v(t) dt \\ &= \int_{11}^{16} \left(-4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3 \right) dt \\ &= \left[-4.2540t + 21.266 \frac{t^2}{2} + 0.13204 \frac{t^3}{3} + 0.0054347 \frac{t^4}{4} \right]_{11}^{16} \\ &= 1605 \text{ m} \end{aligned}$$

Acceleration from Velocity Profile

Find the acceleration of the rocket at $t=16s$ given that
 $v(t) = -4.2540 + 21.266t + 0.13204^2 + 0.0054347t^3, 10 \leq t \leq 22.5$

$$\begin{aligned}a(t) &= \frac{d}{dt} v(t) \\&= \frac{d}{dt} (-4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3) \\&= 21.289 + 0.26130t + 0.016382t^2, \quad 10 \leq t \leq 22.5\end{aligned}$$

$$\begin{aligned}a(16) &= 21.266 + 0.26408(16) + 0.016304(16)^2 \\&= 29.665 \text{ m/s}^2\end{aligned}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/direct_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Newton's Divided Difference Polynomial Method of Interpolation

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

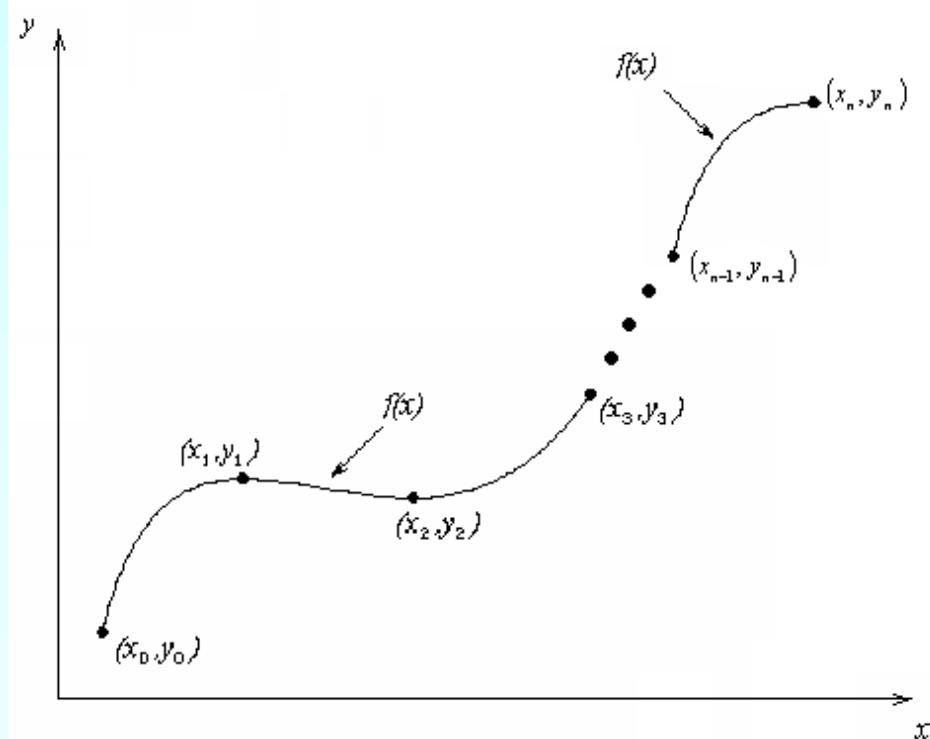
Transforming Numerical Methods Education for STEM
Undergraduates

Newton's Divided Difference Method of Interpolation

<http://numericalmethods.eng.usf.edu>

What is Interpolation ?

Given $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, find the value of 'y' at a value of 'x' that is not given.



Interpolants

Polynomials are the most common choice of interpolants because they are easy to:

- Evaluate
- Differentiate, and
- Integrate.

Newton's Divided Difference Method

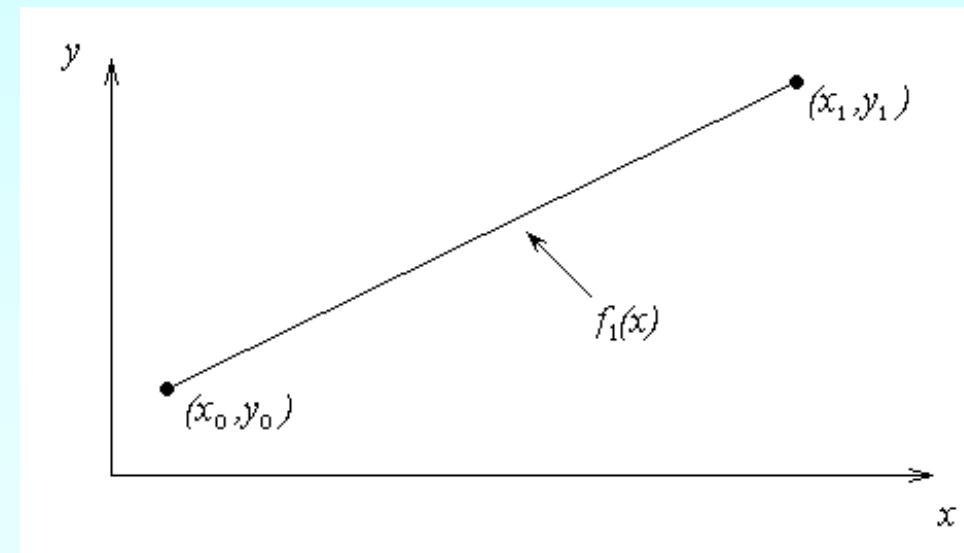
Linear interpolation: Given $(x_0, y_0), (x_1, y_1)$, pass a linear interpolant through the data

$$f_1(x) = b_0 + b_1(x - x_0)$$

where

$$b_0 = f(x_0)$$

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$



Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Newton Divided Difference method for linear interpolation.

Table. Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

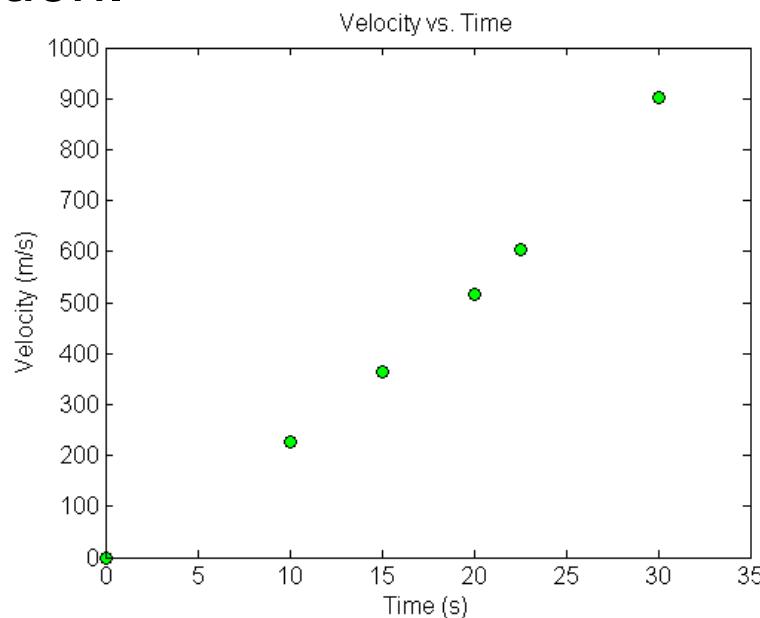


Figure. Velocity vs. time data
for the rocket example

Linear Interpolation

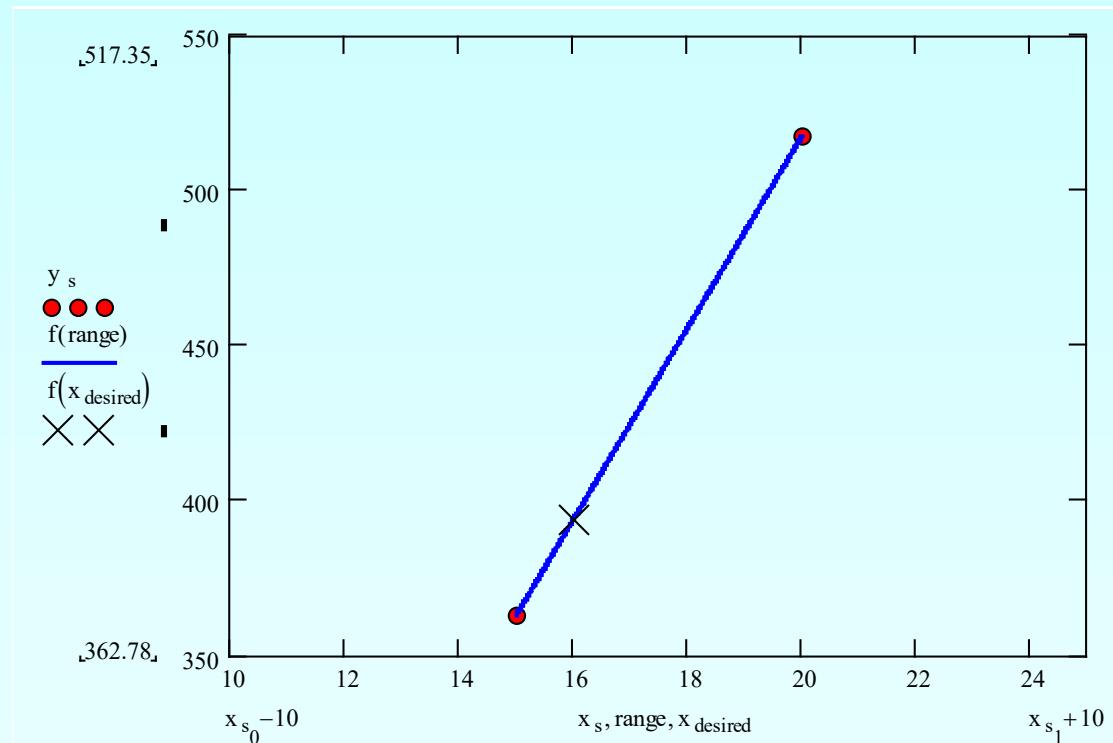
$$v(t) = b_0 + b_1(t - t_0)$$

$$t_0 = 15, v(t_0) = 362.78$$

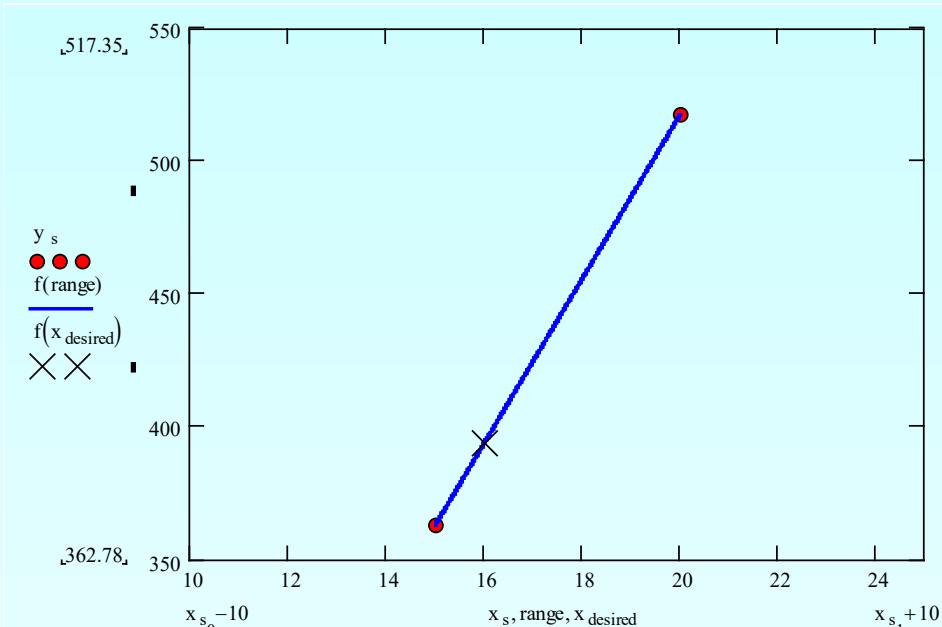
$$t_1 = 20, v(t_1) = 517.35$$

$$b_0 = v(t_0) = 362.78$$

$$b_1 = \frac{v(t_1) - v(t_0)}{t_1 - t_0} = 30.914$$



Linear Interpolation (contd)



$$v(t) = b_0 + b_1(t - t_0)$$

$$= 362.78 + 30.914(t - 15), \quad 15 \leq t \leq 20$$

At $t = 16$

$$v(16) = 362.78 + 30.914(16 - 15)$$

$$= 393.69 \text{ m/s}$$

Quadratic Interpolation

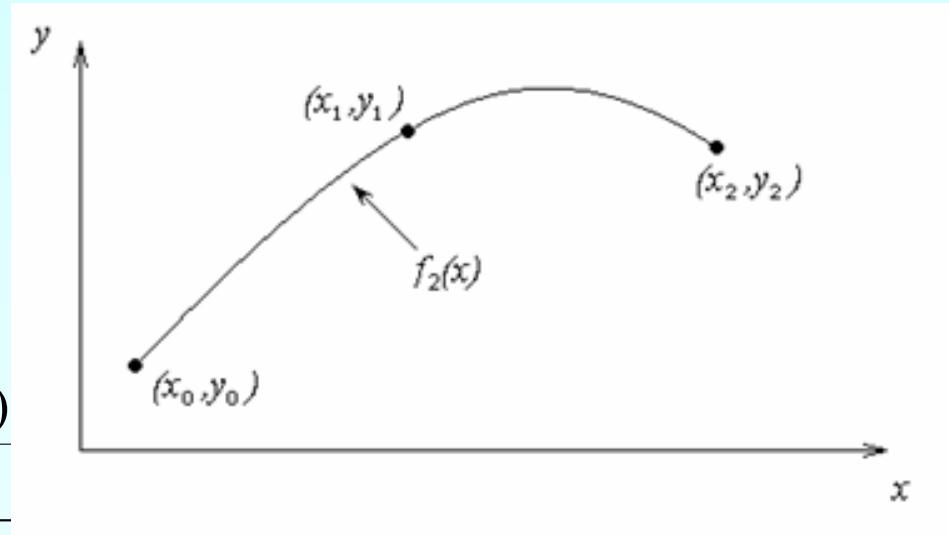
Given (x_0, y_0) , (x_1, y_1) , and (x_2, y_2) , fit a quadratic interpolant through the data.

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

$$b_0 = f(x_0)$$

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$b_2 = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$



Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Newton Divided Difference method for quadratic interpolation.

Table. Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

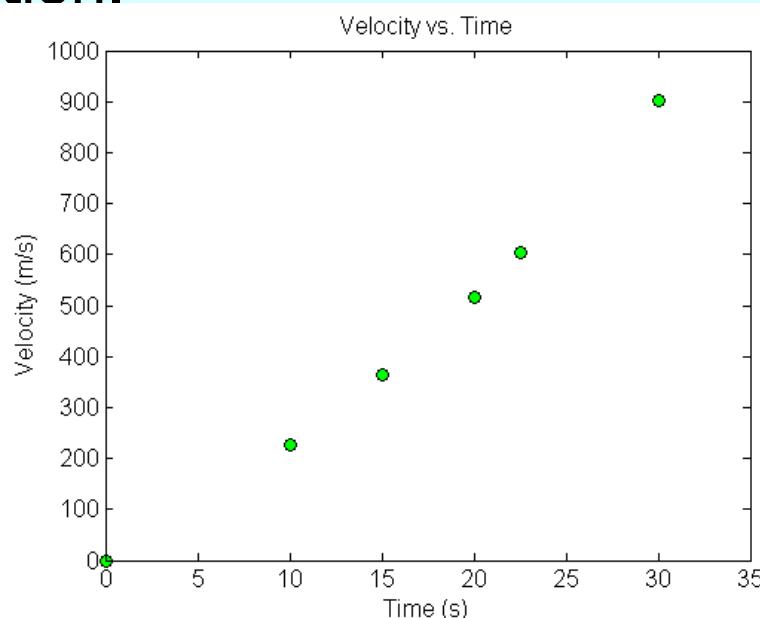
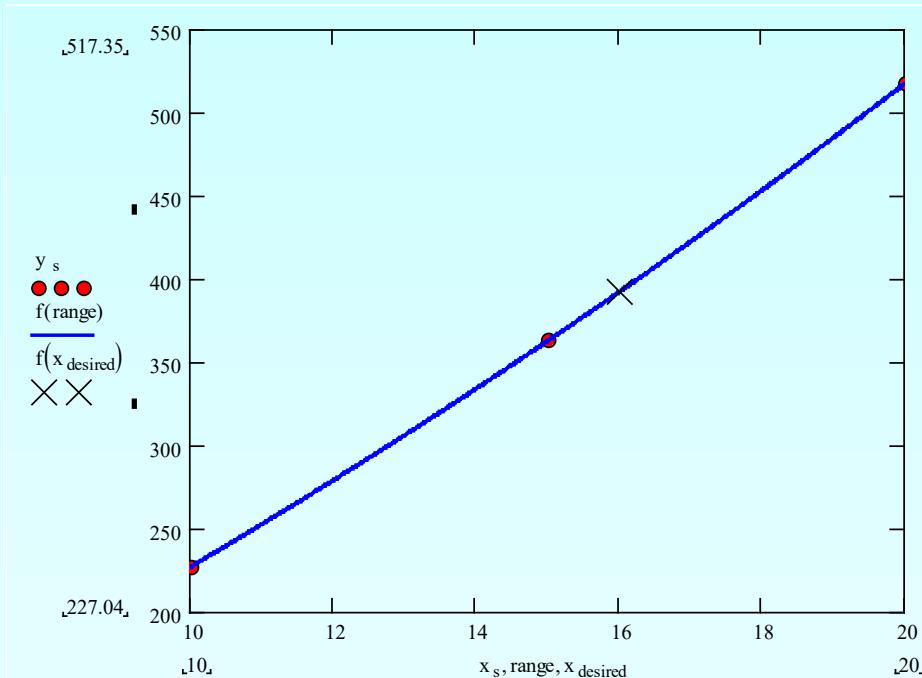


Figure. Velocity vs. time data
for the rocket example

Quadratic Interpolation (contd)



$$t_0 = 10, v(t_0) = 227.04$$

$$t_1 = 15, v(t_1) = 362.78$$

$$t_2 = 20, v(t_2) = 517.35$$

Quadratic Interpolation (contd)

$$b_0 = v(t_0)$$

$$= 227.04$$

$$b_1 = \frac{v(t_1) - v(t_0)}{t_1 - t_0} = \frac{362.78 - 227.04}{15 - 10}$$

$$= 27.148$$

$$b_2 = \frac{\frac{v(t_2) - v(t_1)}{t_2 - t_1} - \frac{v(t_1) - v(t_0)}{t_1 - t_0}}{t_2 - t_0} = \frac{\frac{517.35 - 362.78}{20 - 15} - \frac{362.78 - 227.04}{15 - 10}}{20 - 10}$$
$$= \frac{30.914 - 27.148}{10}$$
$$= 0.37660$$

Quadratic Interpolation (contd)

$$\begin{aligned}v(t) &= b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) \\&= 227.04 + 27.148(t - 10) + 0.37660(t - 10)(t - 15), \quad 10 \leq t \leq 20\end{aligned}$$

At $t = 16$,

$$v(16) = 227.04 + 27.148(16 - 10) + 0.37660(16 - 10)(16 - 15) = 392.19 \text{ m/s}$$

The absolute relative approximate error $|e_a|$ obtained between the results from the first order and second order polynomial is

$$\begin{aligned}|e_a| &= \left| \frac{392.19 - 393.69}{392.19} \right| \times 100 \\&= 0.38502 \%\end{aligned}$$

General Form

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

where

$$b_0 = f[x_0] = f(x_0)$$

$$b_1 = f[x_1, x_0] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$b_2 = f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

Rewriting

$$f_2(x) = f[x_0] + f[x_1, x_0](x - x_0) + f[x_2, x_1, x_0](x - x_0)(x - x_1)$$

General Form

Given $(n + 1)$ data points, $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$ as

$$f_n(x) = b_0 + b_1(x - x_0) + \dots + b_n(x - x_0)(x - x_1)\dots(x - x_{n-1})$$

where

$$b_0 = f[x_0]$$

$$b_1 = f[x_1, x_0]$$

$$b_2 = f[x_2, x_1, x_0]$$

⋮

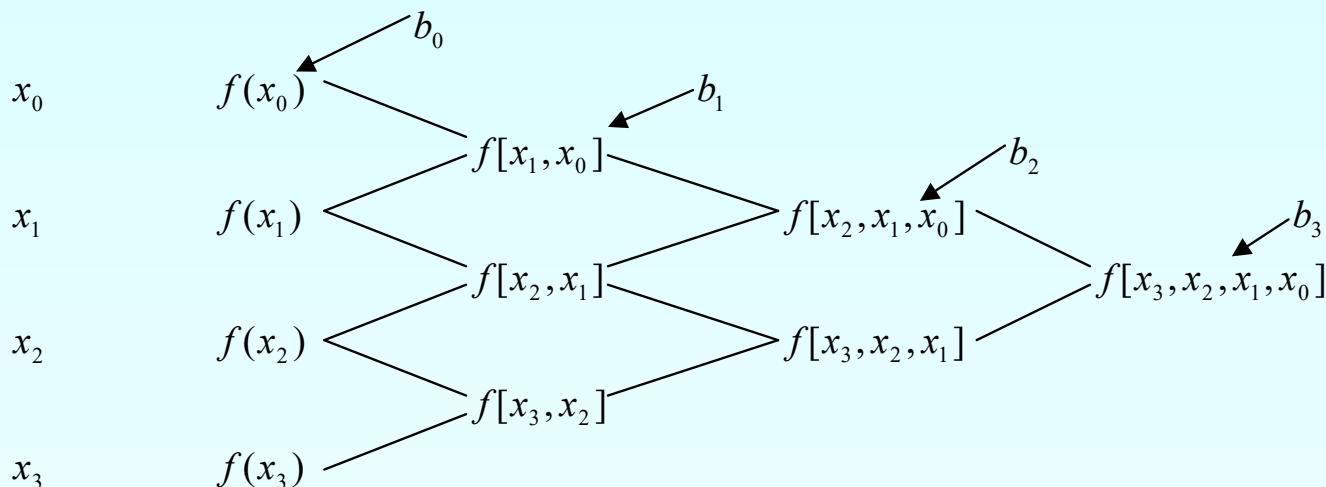
$$b_{n-1} = f[x_{n-1}, x_{n-2}, \dots, x_0]$$

$$b_n = f[x_n, x_{n-1}, \dots, x_0]$$

General form

The third order polynomial, given $(x_0, y_0), (x_1, y_1), (x_2, y_2)$, and (x_3, y_3) , is

$$f_3(x) = f[x_0] + f[x_1, x_0](x - x_0) + f[x_2, x_1, x_0](x - x_0)(x - x_1) \\ + f[x_3, x_2, x_1, x_0](x - x_0)(x - x_1)(x - x_2)$$



Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Newton Divided Difference method for cubic interpolation.

Table. Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

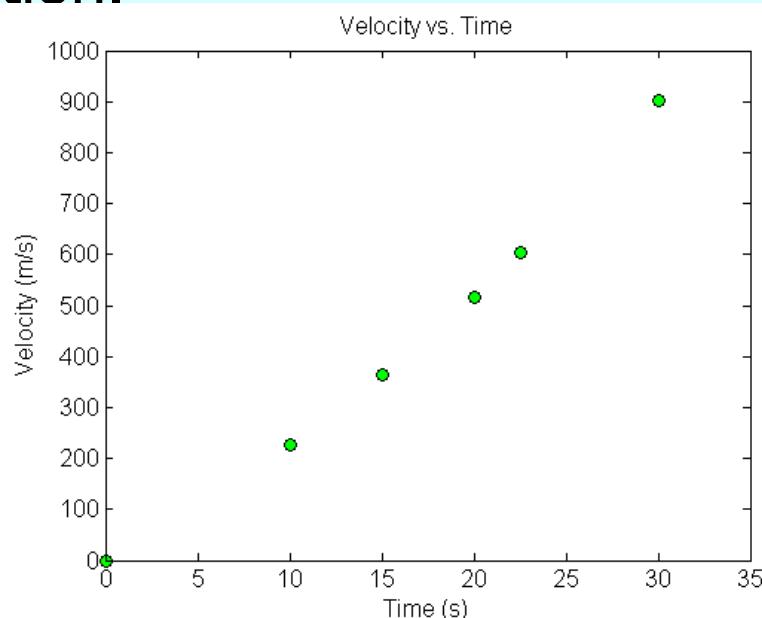


Figure. Velocity vs. time data
for the rocket example

Example

The velocity profile is chosen as

$$v(t) = b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) + b_3(t - t_0)(t - t_1)(t - t_2)$$

we need to choose four data points that are closest to $t = 16$

$$t_0 = 10, \quad v(t_0) = 227.04$$

$$t_1 = 15, \quad v(t_1) = 362.78$$

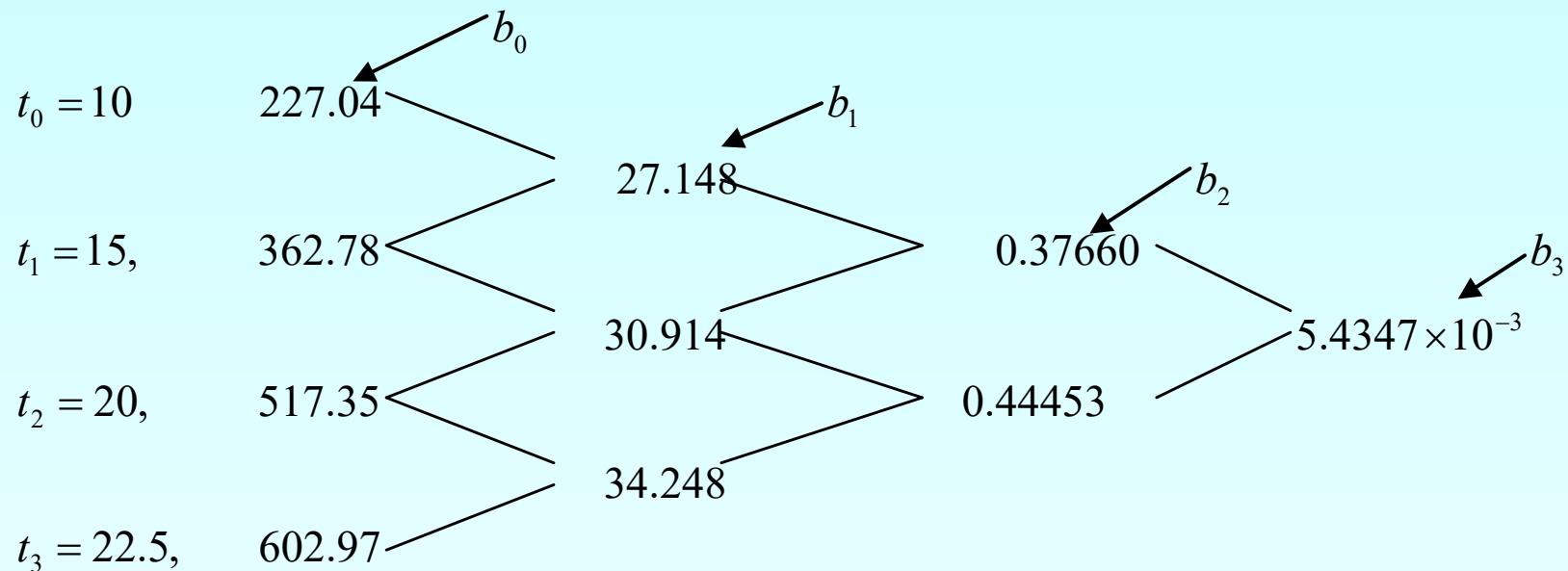
$$t_2 = 20, \quad v(t_2) = 517.35$$

$$t_3 = 22.5, \quad v(t_3) = 602.97$$

The values of the constants are found as:

$$b_0 = 227.04; \quad b_1 = 27.148; \quad b_2 = 0.37660; \quad b_3 = 5.4347 \times 10^{-3}$$

Example



$$b_0 = 227.04; b_1 = 27.148; b_2 = 0.37660; b_3 = 5.4347 \times 10^{-3}$$

Example

Hence

$$\begin{aligned}v(t) &= b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) + b_3(t - t_0)(t - t_1)(t - t_2) \\&= 227.04 + 27.148(t - 10) + 0.37660(t - 10)(t - 15) \\&\quad + 5.4347 * 10^{-3} (t - 10)(t - 15)(t - 20)\end{aligned}$$

At $t = 16$,

$$\begin{aligned}v(16) &= 227.04 + 27.148(16 - 10) + 0.37660(16 - 10)(16 - 15) \\&\quad + 5.4347 * 10^{-3} (16 - 10)(16 - 15)(16 - 20) \\&= 392.06 \text{ m/s}\end{aligned}$$

The absolute relative approximate error $|e_a|$ obtained is

$$|e_a| = \left| \frac{392.06 - 392.19}{392.06} \right| \times 100$$

$$= 0.033427 \%$$

Comparison Table

Order of Polynomial	1	2	3
$v(t=16)$ m/s	393.69	392.19	392.06
Absolute Relative Approximate Error	-----	0.38502 %	0.033427 %

Distance from Velocity Profile

Find the distance covered by the rocket from $t=11\text{s}$ to $t=16\text{s}$?

$$\begin{aligned}v(t) &= 227.04 + 27.148(t - 10) + 0.37660(t - 10)(t - 15) && 10 \leq t \leq 22.5 \\&\quad + 5.4347 * 10^{-3} (t - 10)(t - 15)(t - 20) \\&= -4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3 && 10 \leq t \leq 22.5\end{aligned}$$

So

$$\begin{aligned}s(16) - s(11) &= \int_{11}^{16} v(t) dt \\&= \int_{11}^{16} (-4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3) dt \\&= \left[-4.2541t + 21.265 \frac{t^2}{2} + 0.13204 \frac{t^3}{3} + 0.0054347 \frac{t^4}{4} \right]_{11}^{16} \\&= 1605 \text{ m}\end{aligned}$$

Acceleration from Velocity Profile

Find the acceleration of the rocket at t=16s given that

$$v(t) = -4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3$$

$$a(t) = \frac{d}{dt} v(t) = \frac{d}{dt} (-4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3)$$

$$= 21.265 + 0.26408t + 0.016304t^2$$

$$a(16) = 21.265 + 0.26408(16) + 0.016304(16)^2$$

$$= 29.664 \text{ m/s}^2$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/newton_divided_difference_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Lagrangian Interpolation

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

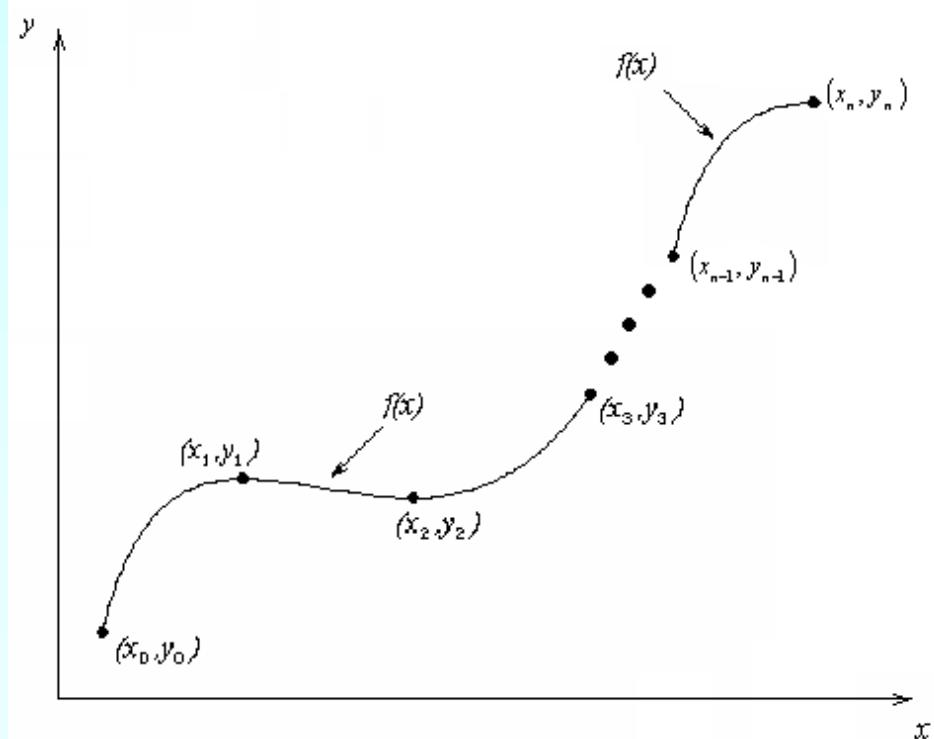
Transforming Numerical Methods Education for STEM
Undergraduates

Lagrange Method of Interpolation

<http://numericalmethods.eng.usf.edu>

What is Interpolation ?

Given $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, find the value of 'y' at a value of 'x' that is not given.



Interpolants

Polynomials are the most common choice of interpolants because they are easy to:

- Evaluate
- Differentiate, and
- Integrate.

Lagrangian Interpolation

Lagrangian interpolating polynomial is given by

$$f_n(x) = \sum_{i=0}^n L_i(x)f(x_i)$$

where ‘ n ’ in $f_n(x)$ stands for the n^{th} order polynomial that approximates the function $y = f(x)$ given at $(n + 1)$ data points as $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, and

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

$L_i(x)$ is a weighting function that includes a product of $(n - 1)$ terms with terms of $j = i$ omitted.

Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Lagrangian method for linear interpolation.

Table Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

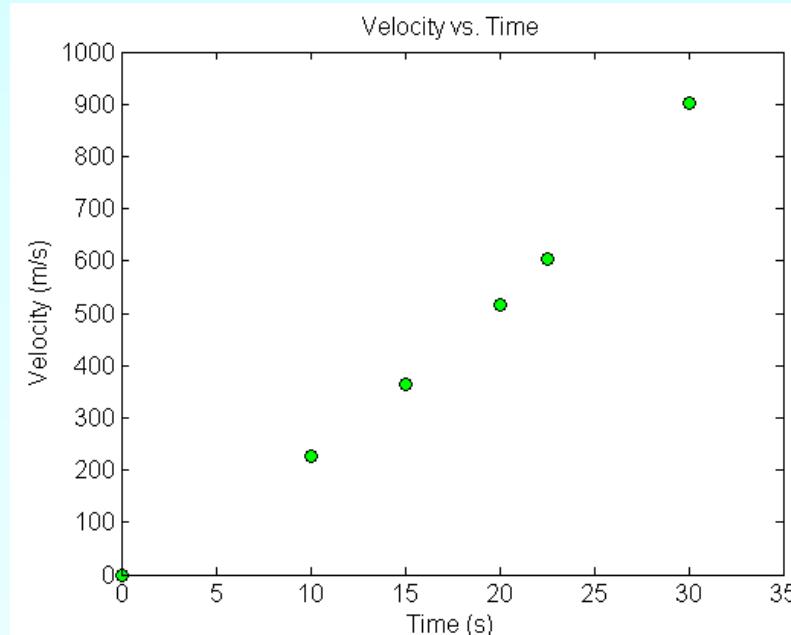


Figure. Velocity vs. time data for the rocket example

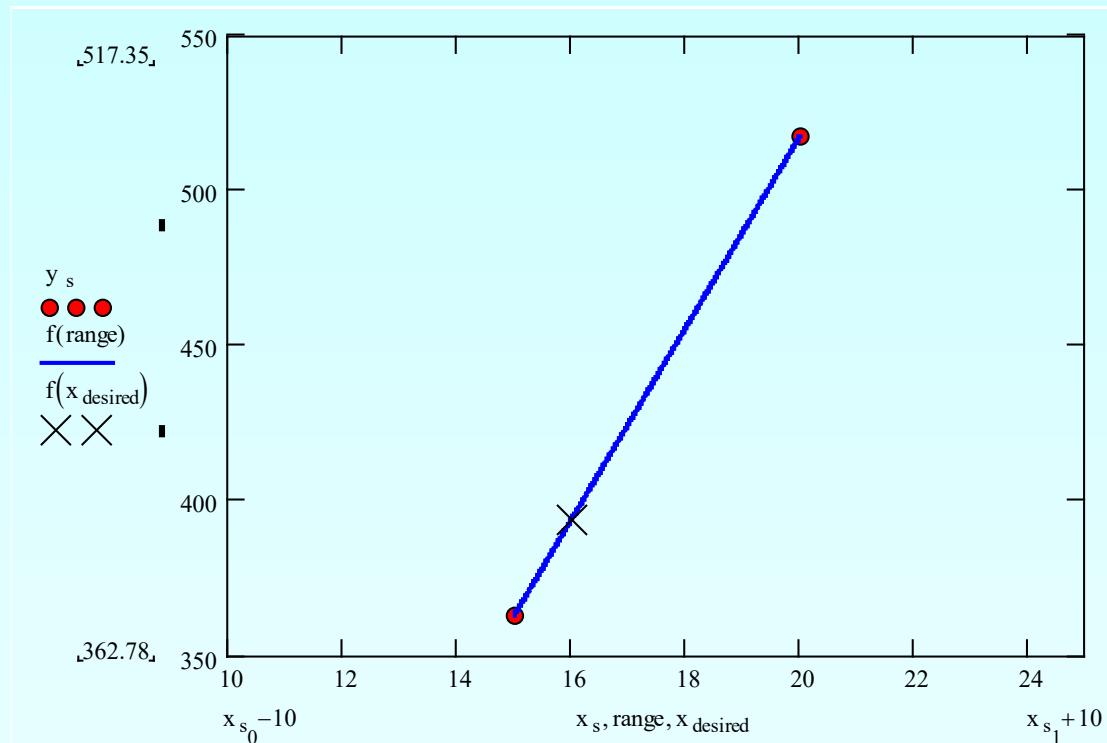
Linear Interpolation

$$v(t) = \sum_{i=0}^1 L_i(t)v(t_i)$$

$$= L_0(t)v(t_0) + L_1(t)v(t_1)$$

$$t_0 = 15, v(t_0) = 362.78$$

$$t_1 = 20, v(t_1) = 517.35$$



Linear Interpolation (contd)

$$L_0(t) = \prod_{\substack{j=0 \\ j \neq 0}}^1 \frac{t - t_j}{t_0 - t_j} = \frac{t - t_1}{t_0 - t_1}$$

$$L_1(t) = \prod_{\substack{j=0 \\ j \neq 1}}^1 \frac{t - t_j}{t_1 - t_j} = \frac{t - t_0}{t_1 - t_0}$$

$$v(t) = \frac{t - t_1}{t_0 - t_1} v(t_0) + \frac{t - t_0}{t_1 - t_0} v(t_1) = \frac{t - 20}{15 - 20} (362.78) + \frac{t - 15}{20 - 15} (517.35)$$

$$v(16) = \frac{16 - 20}{15 - 20} (362.78) + \frac{16 - 15}{20 - 15} (517.35)$$

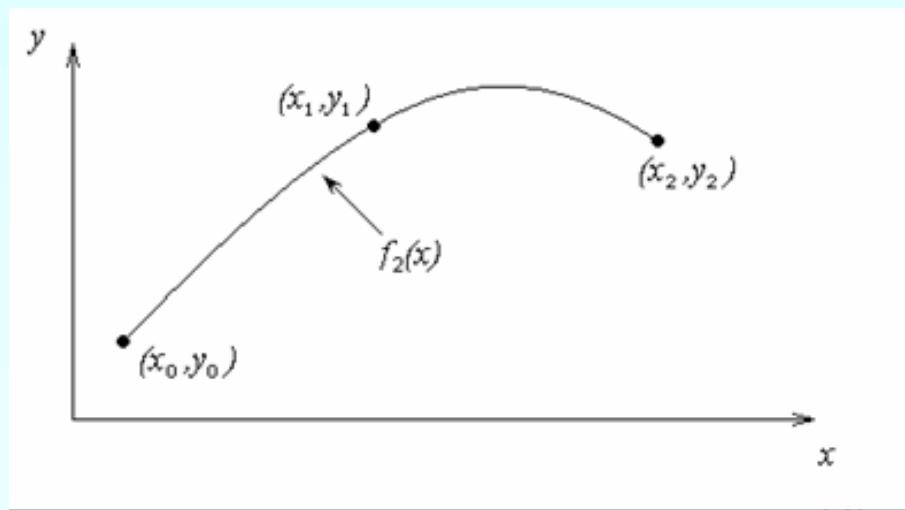
$$= 0.8(362.78) + 0.2(517.35)$$

$$= 393.7 \text{ m/s.}$$

Quadratic Interpolation

For the second order polynomial interpolation (also called quadratic interpolation), we choose the velocity given by

$$\begin{aligned}v(t) &= \sum_{i=0}^2 L_i(t)v(t_i) \\&= L_0(t)v(t_0) + L_1(t)v(t_1) + L_2(t)v(t_2)\end{aligned}$$



Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Lagrangian method for quadratic interpolation.

Table Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

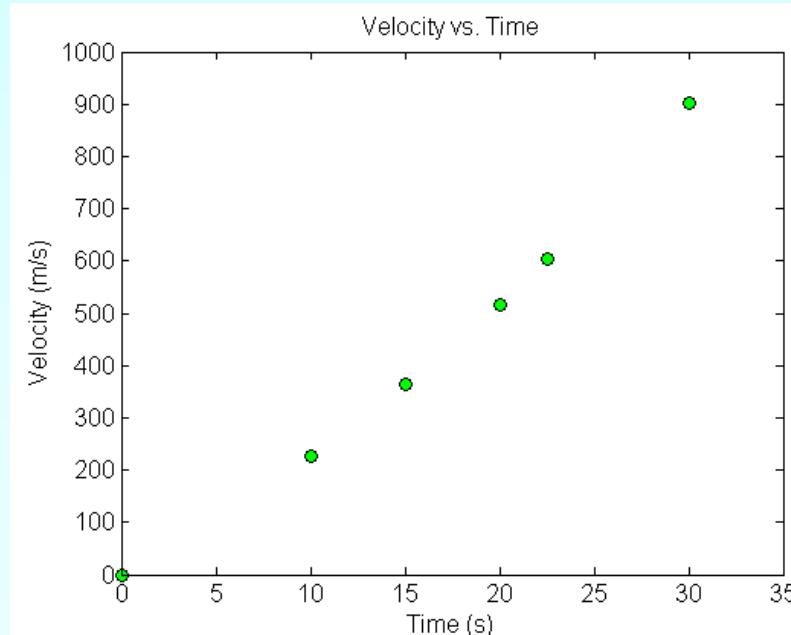


Figure. Velocity vs. time data for the rocket example

Quadratic Interpolation (contd)

$$t_0 = 10, v(t_0) = 227.04$$

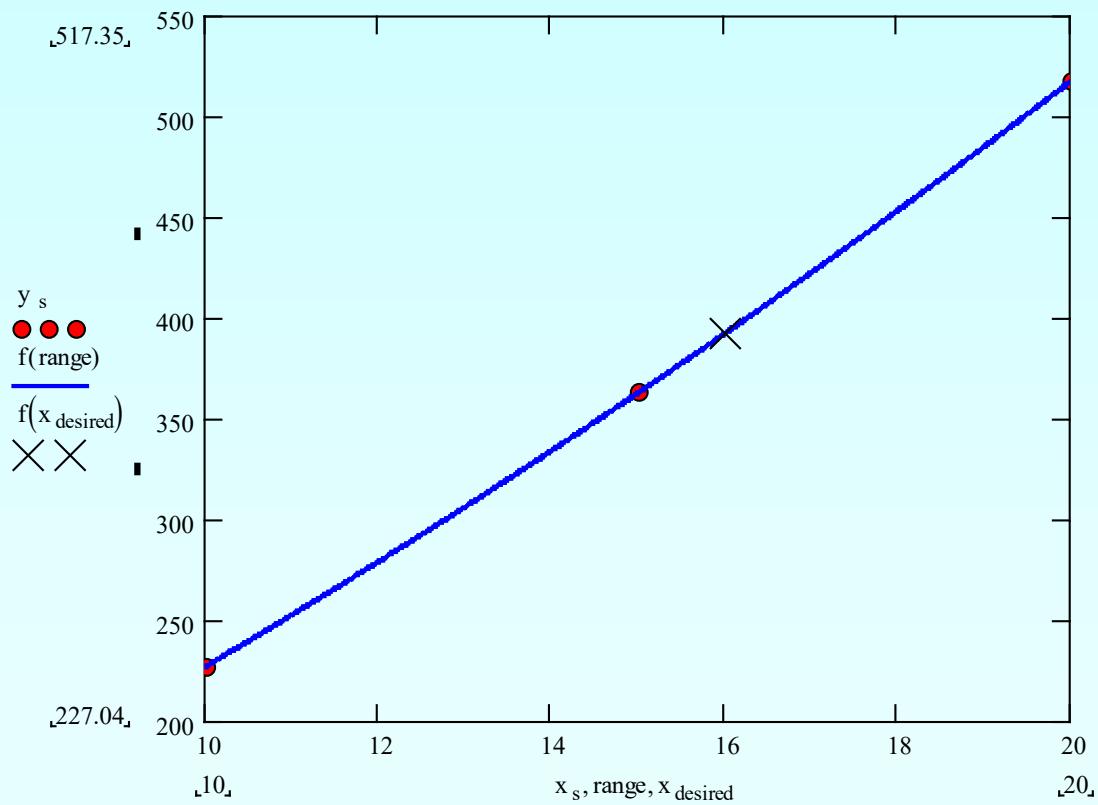
$$t_1 = 15, v(t_1) = 362.78$$

$$t_2 = 20, v(t_2) = 517.35$$

$$L_0(t) = \prod_{\substack{j=0 \\ j \neq 0}}^2 \frac{t - t_j}{t_0 - t_j} = \left(\frac{t - t_1}{t_0 - t_1} \right) \left(\frac{t - t_2}{t_0 - t_2} \right)$$

$$L_1(t) = \prod_{\substack{j=0 \\ j \neq 1}}^2 \frac{t - t_j}{t_1 - t_j} = \left(\frac{t - t_0}{t_1 - t_0} \right) \left(\frac{t - t_2}{t_1 - t_2} \right)$$

$$L_2(t) = \prod_{\substack{j=0 \\ j \neq 2}}^2 \frac{t - t_j}{t_2 - t_j} = \left(\frac{t - t_0}{t_2 - t_0} \right) \left(\frac{t - t_1}{t_2 - t_1} \right)$$



Quadratic Interpolation (contd)

$$v(t) = \left(\frac{t-t_1}{t_0-t_1} \right) \left(\frac{t-t_2}{t_0-t_2} \right) v(t_0) + \left(\frac{t-t_0}{t_1-t_0} \right) \left(\frac{t-t_2}{t_1-t_2} \right) v(t_1) + \left(\frac{t-t_0}{t_2-t_0} \right) \left(\frac{t-t_1}{t_2-t_1} \right) v(t_2)$$
$$v(16) = \left(\frac{16-15}{10-15} \right) \left(\frac{16-20}{10-20} \right) (227.04) + \left(\frac{16-10}{15-10} \right) \left(\frac{16-20}{15-20} \right) (362.78) + \left(\frac{16-10}{20-10} \right) \left(\frac{16-15}{20-15} \right) (517.35)$$
$$= (-0.08)(227.04) + (0.96)(362.78) + (0.12)(527.35)$$
$$= 392.19 \text{ m/s}$$

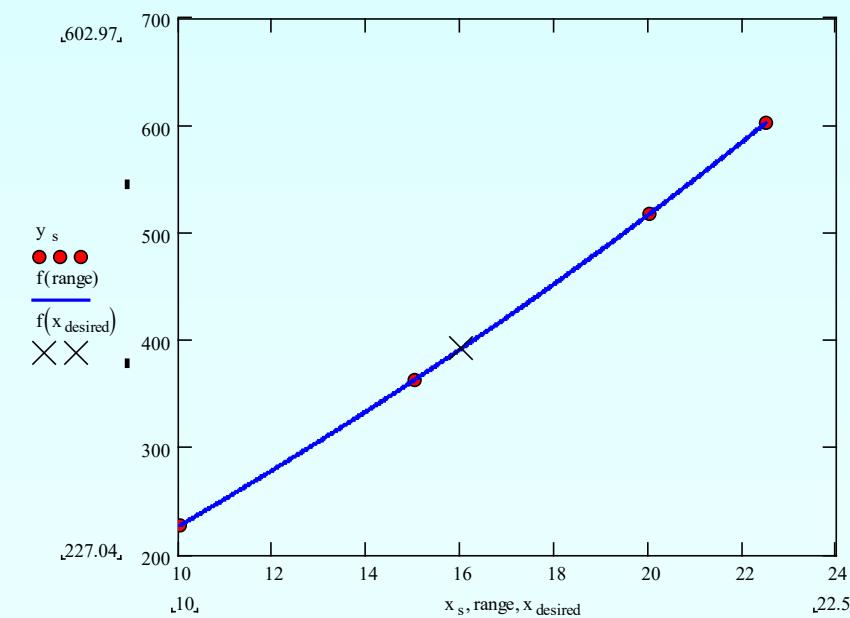
The absolute relative approximate error $|e_a|$ obtained between the results from the first and second order polynomial is

$$|e_a| = \left| \frac{392.19 - 393.70}{392.19} \right| \times 100$$
$$= 0.38410\%$$

Cubic Interpolation

For the third order polynomial (also called cubic interpolation), we choose the velocity given by

$$\begin{aligned}v(t) &= \sum_{i=0}^3 L_i(t)v(t_i) \\&= L_0(t)v(t_0) + L_1(t)v(t_1) + L_2(t)v(t_2) + L_3(t)v(t_3)\end{aligned}$$



Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using the Lagrangian method for cubic interpolation.

Table Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

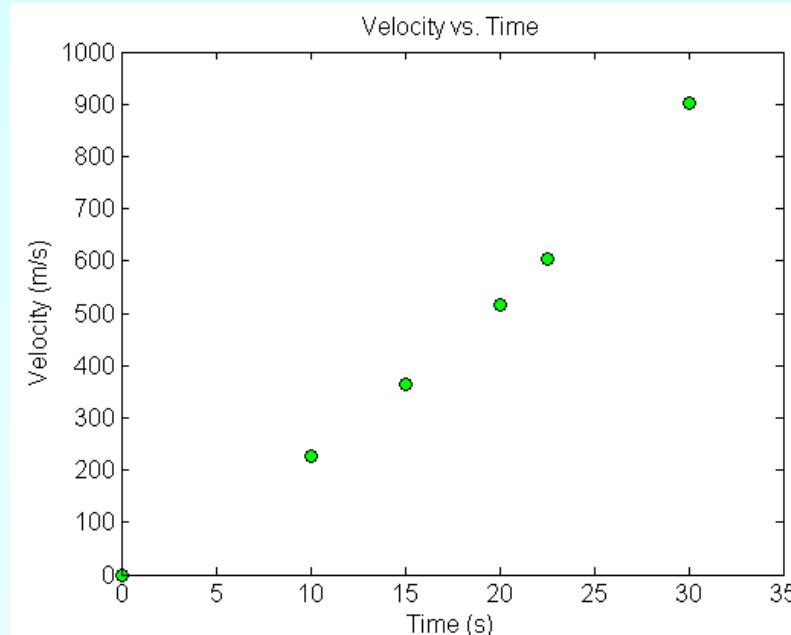


Figure. Velocity vs. time data for the rocket example

Cubic Interpolation (contd)

$$t_o = 10, \quad v(t_o) = 227.04 \quad t_1 = 15, \quad v(t_1) = 362.78$$

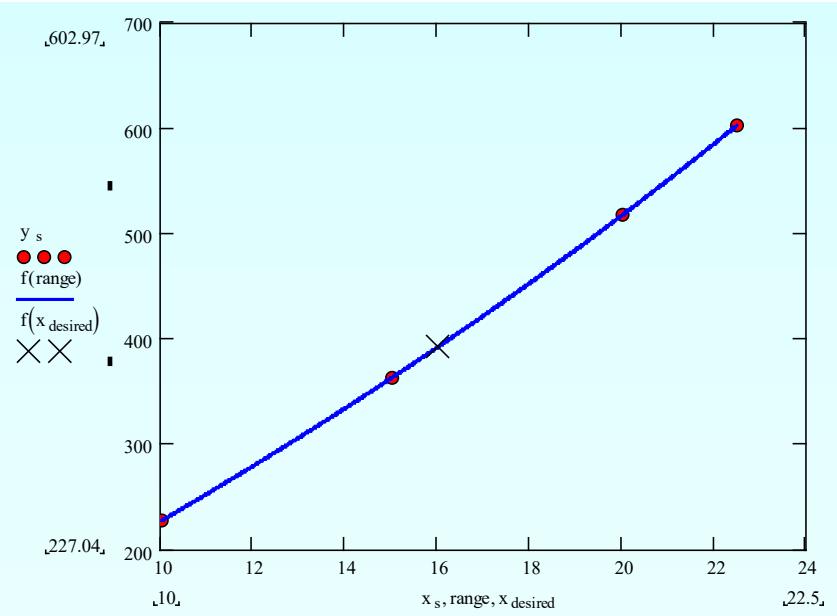
$$t_2 = 20, \quad v(t_2) = 517.35 \quad t_3 = 22.5, \quad v(t_3) = 602.97$$

$$L_0(t) = \prod_{\substack{j=0 \\ j \neq 0}}^3 \frac{t - t_j}{t_0 - t_j} = \left(\frac{t - t_1}{t_0 - t_1} \right) \left(\frac{t - t_2}{t_0 - t_2} \right) \left(\frac{t - t_3}{t_0 - t_3} \right);$$

$$L_1(t) = \prod_{\substack{j=0 \\ j \neq 1}}^3 \frac{t - t_j}{t_1 - t_j} = \left(\frac{t - t_0}{t_1 - t_0} \right) \left(\frac{t - t_2}{t_1 - t_2} \right) \left(\frac{t - t_3}{t_1 - t_3} \right)$$

$$L_2(t) = \prod_{\substack{j=0 \\ j \neq 2}}^3 \frac{t - t_j}{t_2 - t_j} = \left(\frac{t - t_0}{t_2 - t_0} \right) \left(\frac{t - t_1}{t_2 - t_1} \right) \left(\frac{t - t_3}{t_2 - t_3} \right);$$

$$L_3(t) = \prod_{\substack{j=0 \\ j \neq 3}}^3 \frac{t - t_j}{t_3 - t_j} = \left(\frac{t - t_0}{t_3 - t_0} \right) \left(\frac{t - t_1}{t_3 - t_1} \right) \left(\frac{t - t_2}{t_3 - t_2} \right)$$



Cubic Interpolation (contd)

$$v(t) = \left(\frac{t-t_1}{t_0-t_1} \right) \left(\frac{t-t_2}{t_0-t_2} \right) \left(\frac{t-t_3}{t_0-t_3} \right) v(t_1) + \left(\frac{t-t_0}{t_1-t_0} \right) \left(\frac{t-t_2}{t_1-t_2} \right) \left(\frac{t-t_3}{t_1-t_3} \right) v(t_2)$$
$$+ \left(\frac{t-t_0}{t_2-t_0} \right) \left(\frac{t-t_1}{t_2-t_1} \right) \left(\frac{t-t_3}{t_2-t_3} \right) v(t_2) + \left(\frac{t-t_1}{t_3-t_1} \right) \left(\frac{t-t_2}{t_3-t_2} \right) \left(\frac{t-t_3}{t_3-t_2} \right) v(t_3)$$
$$v(16) = \left(\frac{16-15}{10-15} \right) \left(\frac{16-20}{10-20} \right) \left(\frac{16-22.5}{10-22.5} \right) (227.04) + \left(\frac{16-10}{15-10} \right) \left(\frac{16-20}{15-20} \right) \left(\frac{16-22.5}{15-22.5} \right) (362.78)$$
$$+ \left(\frac{16-10}{20-10} \right) \left(\frac{16-15}{20-15} \right) \left(\frac{16-22.5}{20-22.5} \right) (517.35) + \left(\frac{16-10}{22.5-10} \right) \left(\frac{16-15}{22.5-15} \right) \left(\frac{16-20}{22.5-20} \right) (602.97)$$
$$= (-0.0416)(227.04) + (0.832)(362.78) + (0.312)(517.35) + (-0.1024)(602.97)$$
$$= 392.06 \text{ m/s}$$

The absolute relative approximate error $|e_a|$ obtained between the results from the first and second order polynomial is

$$|e_a| = \left| \frac{392.06 - 392.19}{392.06} \right| \times 100$$
$$= 0.033269\%$$

Comparison Table

Order of Polynomial	1	2	3
v(t=16) m/s	393.69	392.19	392.06
Absolute Relative Approximate Error	-----	0.38410%	0.033269%

Distance from Velocity Profile

Find the distance covered by the rocket from $t=11\text{s}$ to $t=16\text{s}$?

$$v(t) = (t^3 - 57.5t^2 + 1087.5t - 6750)(-0.36326) + (t^3 - 52.5t^2 + 875t - 4500)(1.9348) \\ + (t^3 - 47.5t^2 + 712.5t - 3375)(-4.1388) + (t^3 - 45t^2 + 650t - 3000)(2.5727)$$

$$v(t) = -4.245 + 21.265t + 0.13195t^2 + 0.00544t^3, \quad 10 \leq t \leq 22.5$$

$$s(16) - s(11) = \int_{11}^{16} v(t) dt \\ \approx \int_{11}^{16} (-4.245 + 21.265t + 0.13195t^2 + 0.00544t^3) dt \\ = \left[-4.245t + 21.265 \frac{t^2}{2} + 0.13195 \frac{t^3}{3} + 0.00544 \frac{t^4}{4} \right]_{11}^{16} \\ = 1605 \text{ m}$$

Acceleration from Velocity Profile

Find the acceleration of the rocket at t=16s given that

$$v(t) = -4.245 + 21.265t + 0.13195t^2 + 0.00544t^3, \quad 10 \leq t \leq 22.5$$

$$\begin{aligned} a(t) &= \frac{d}{dt}v(t) = \frac{d}{dt}(-4.245 + 21.265t + 0.13195t^2 + 0.00544t^3) \\ &= 21.265 + 0.26390t + 0.01632t^2 \end{aligned}$$

$$\begin{aligned} a(16) &= 21.265 + 0.26390(16) + 0.01632(16)^2 \\ &= 29.665 \text{ m/s}^2 \end{aligned}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/lagrange_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Spline Interpolation Method

Major: All Engineering Majors

Authors: Autar Kaw, Jai Paul

<http://numericalmethods.eng.usf.edu>

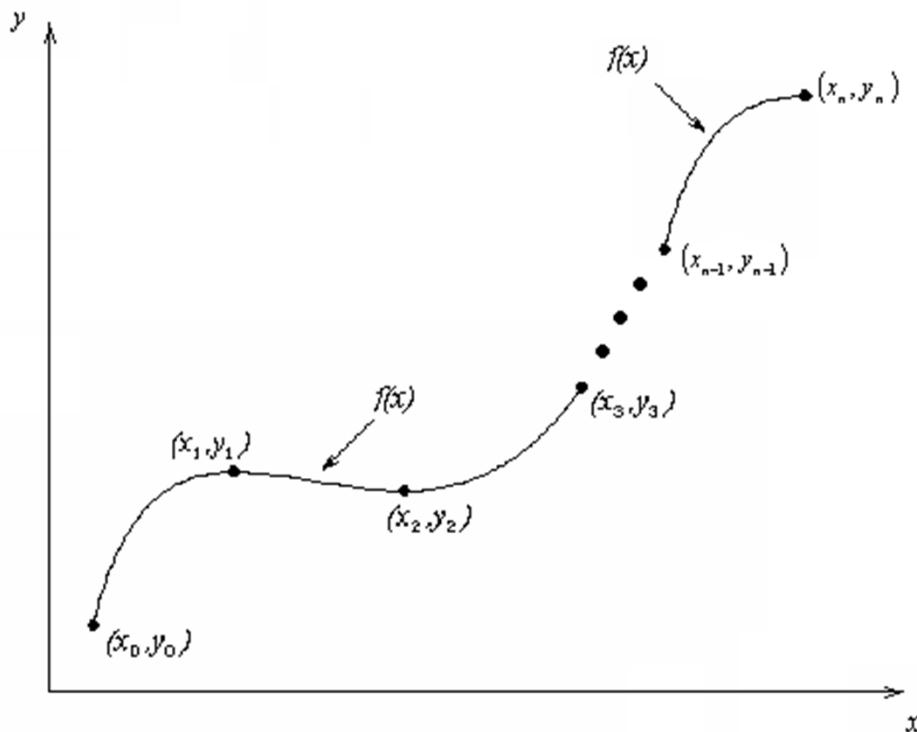
Transforming Numerical Methods Education for STEM
Undergraduates

Spline Method of Interpolation

<http://numericalmethods.eng.usf.edu>

What is Interpolation ?

Given $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, find the value of 'y' at a value of 'x' that is not given.



Interpolants

Polynomials are the most common choice of interpolants because they are easy to:

- Evaluate
- Differentiate, and
- Integrate.

Rocket Example Results

t (s)	v (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Polynomial Order	Velocity at t=16 in m/s	Absolute Relative Approxima- te Error	Least Number of Significant Digits Correct
1	393.69	-----	
2	392.19	0.38%	2
3	392.05	0.036%	3
4	392.07	0.0051%	3
5	392.06	0.0026%	4

Why Splines ?

$$f(x) = \frac{1}{1 + 25x^2}$$

Table : Six equidistantly spaced points in [-1, 1]

x	$y = \frac{1}{1 + 25x^2}$
-1.0	0.038461
-0.6	0.1
-0.2	0.5
0.2	0.5
0.6	0.1
1.0	0.038461

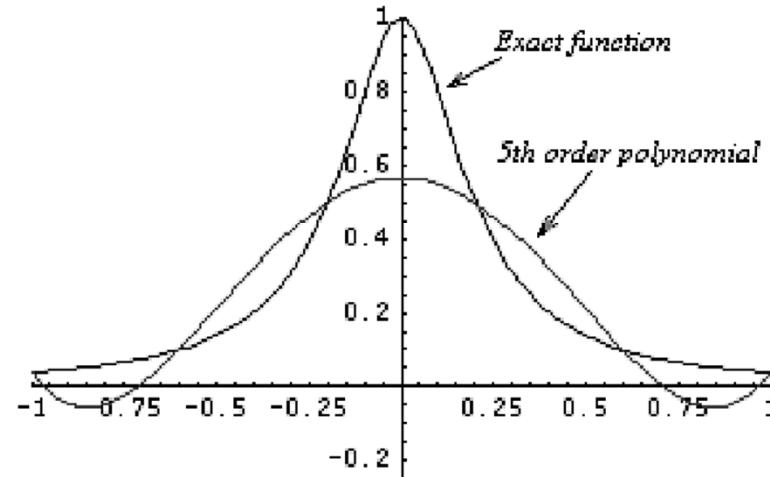


Figure : 5th order polynomial vs. exact function

Why Splines ?

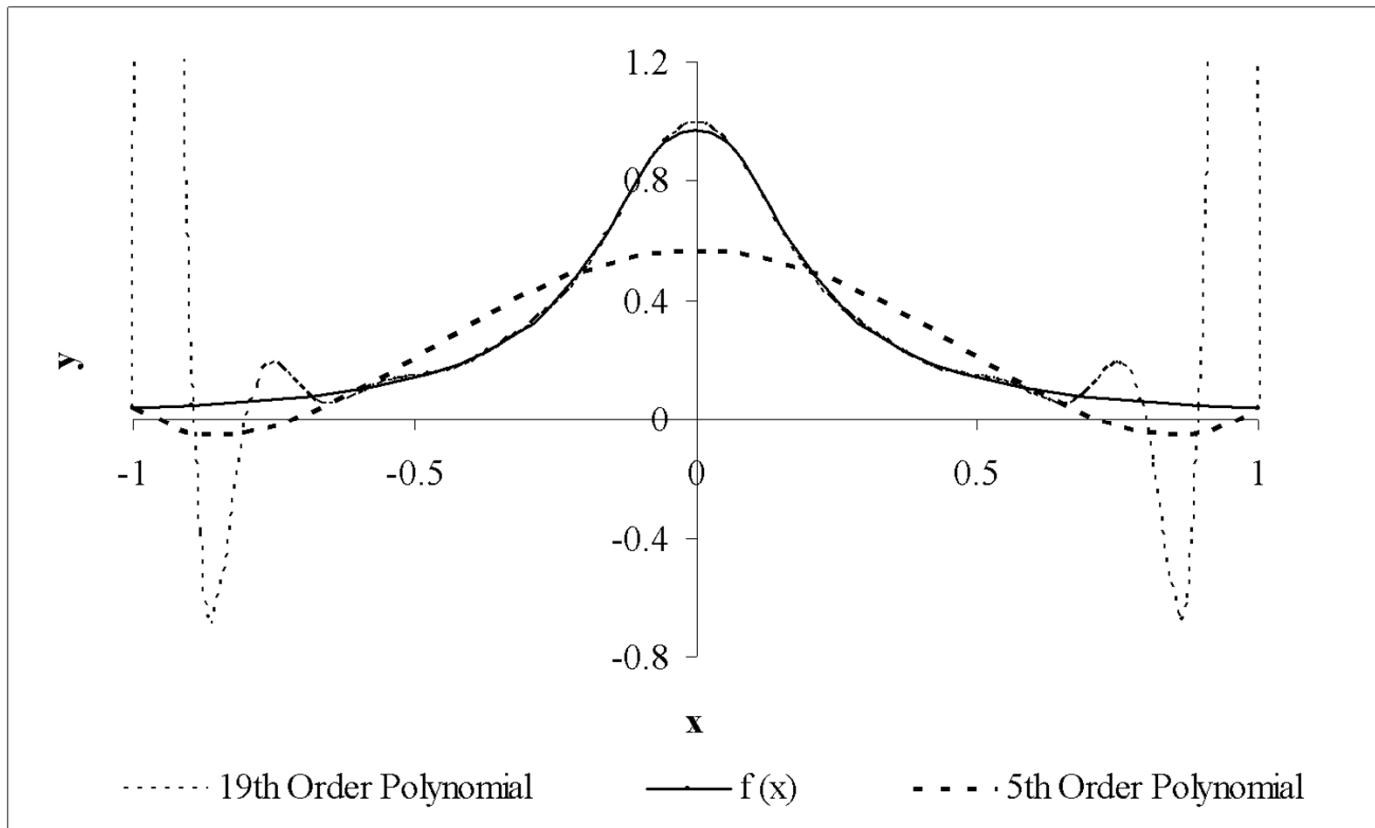
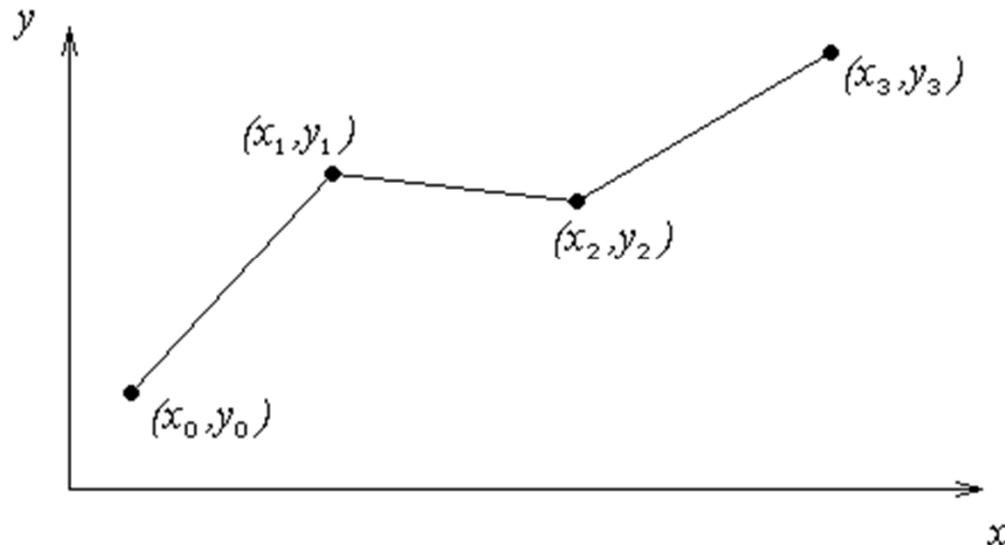


Figure : Higher order polynomial interpolation is a bad idea

Linear Interpolation

Given $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, fit linear splines to the data. This simply involves forming the consecutive data through straight lines. So if the above data is given in an ascending order, the linear splines are given by $(y_i = f(x_i))$

Figure : Linear splines



Linear Interpolation (contd)

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0), \quad x_0 \leq x \leq x_1$$

$$= f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1), \quad x_1 \leq x \leq x_2$$

.

.

.

$$= f(x_{n-1}) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_{n-1}), \quad x_{n-1} \leq x \leq x_n$$

Note the terms of

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

in the above function are simply slopes between x_{i-1} and x_i .

Example

The upward velocity of a rocket is given as a function of time in Table 1. Find the velocity at $t=16$ seconds using linear splines.

Table Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

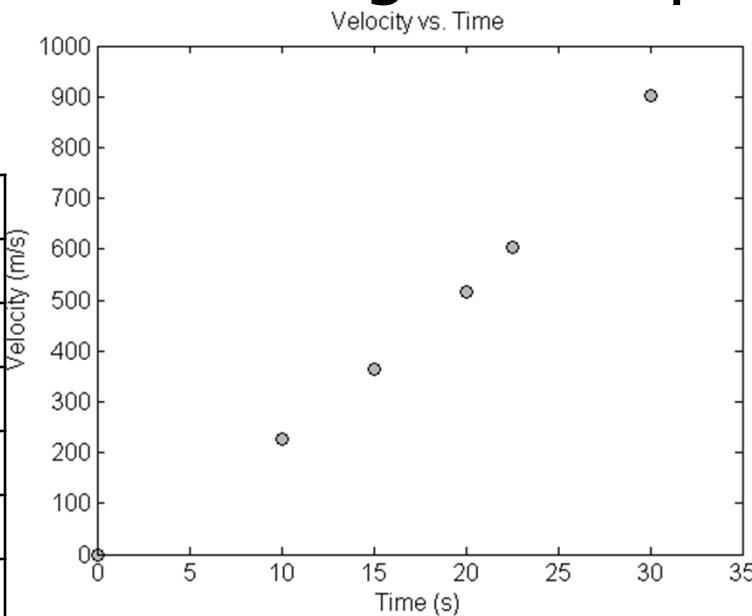


Figure. Velocity vs. time data for the rocket example



Linear Interpolation

$$t_0 = 15, \quad v(t_0) = 362.78$$

$$t_1 = 20, \quad v(t_1) = 517.35$$

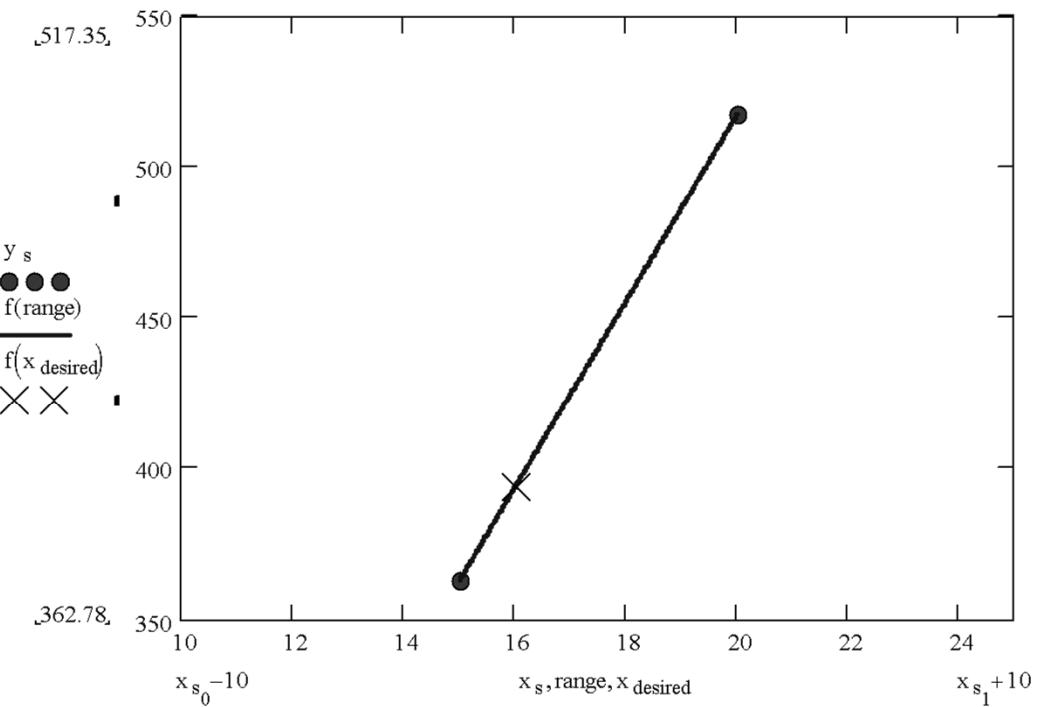
$$\begin{aligned}v(t) &= v(t_0) + \frac{v(t_1) - v(t_0)}{t_1 - t_0}(t - t_0) \\&= 362.78 + \frac{517.35 - 362.78}{20 - 15}(t - 15)\end{aligned}$$

$$v(t) = 362.78 + 30.913(t - 15)$$

At $t = 16$,

$$v(16) = 362.78 + 30.913(16 - 15)$$

$$= 393.7 \text{ m/s}$$



Quadratic Interpolation

Given $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, fit quadratic splines through the data. The splines are given by

$$f(x) = a_1 x^2 + b_1 x + c_1, \quad x_0 \leq x \leq x_1$$

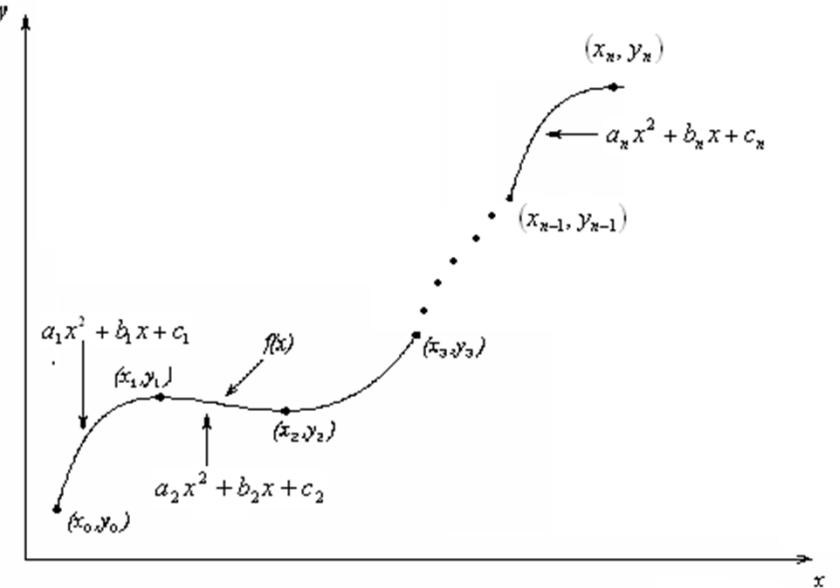
$$= a_2 x^2 + b_2 x + c_2, \quad x_1 \leq x \leq x_2$$

.

.

.

$$= a_n x^2 + b_n x + c_n, \quad x_{n-1} \leq x \leq x_n$$



Find $a_i, b_i, c_i, i = 1, 2, \dots, n$

Quadratic Interpolation (contd)

Each quadratic spline goes through two consecutive data points

$$a_1 x_0^2 + b_1 x_0 + c_1 = f(x_0)$$

$$a_1 x_1^2 + b_1 x_1 + c_1 = f(x_1)$$

.

.

$$a_i x_{i-1}^2 + b_i x_{i-1} + c_i = f(x_{i-1})$$

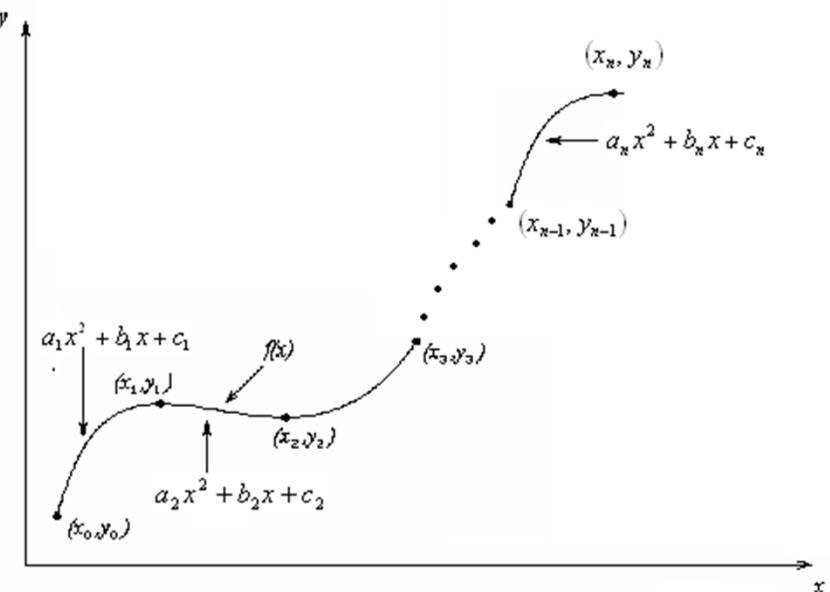
$$a_i x_i^2 + b_i x_i + c_i = f(x_i)$$

.

.

$$a_n x_{n-1}^2 + b_n x_{n-1} + c_n = f(x_{n-1})$$

$$a_n x_n^2 + b_n x_n + c_n = f(x_n)$$



This condition gives $2n$ equations

Quadratic Splines (contd)

The first derivatives of two quadratic splines are continuous at the interior points.

For example, the derivative of the first spline

$$a_1 x^2 + b_1 x + c_1 \text{ is } 2a_1 x + b_1$$

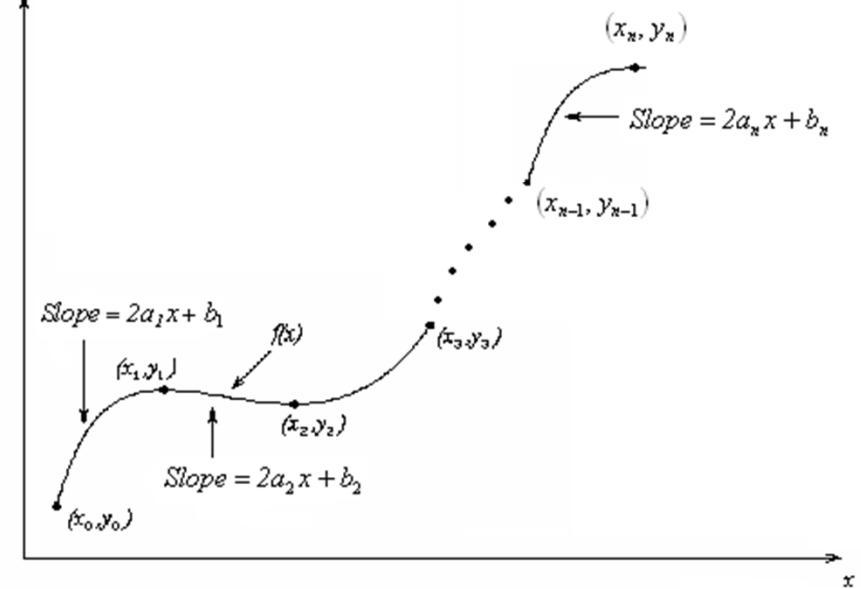
The derivative of the second spline

$$a_2 x^2 + b_2 x + c_2 \text{ is } 2a_2 x + b_2$$

and the two are equal at $x = x_1$ giving

$$2a_1 x_1 + b_1 = 2a_2 x_1 + b_2$$

$$2a_1 x_1 + b_1 - 2a_2 x_1 - b_2 = 0$$



Quadratic Splines (contd)

Similarly at the other interior points,

$$2a_2x_2 + b_2 - 2a_3x_2 - b_3 = 0$$

.

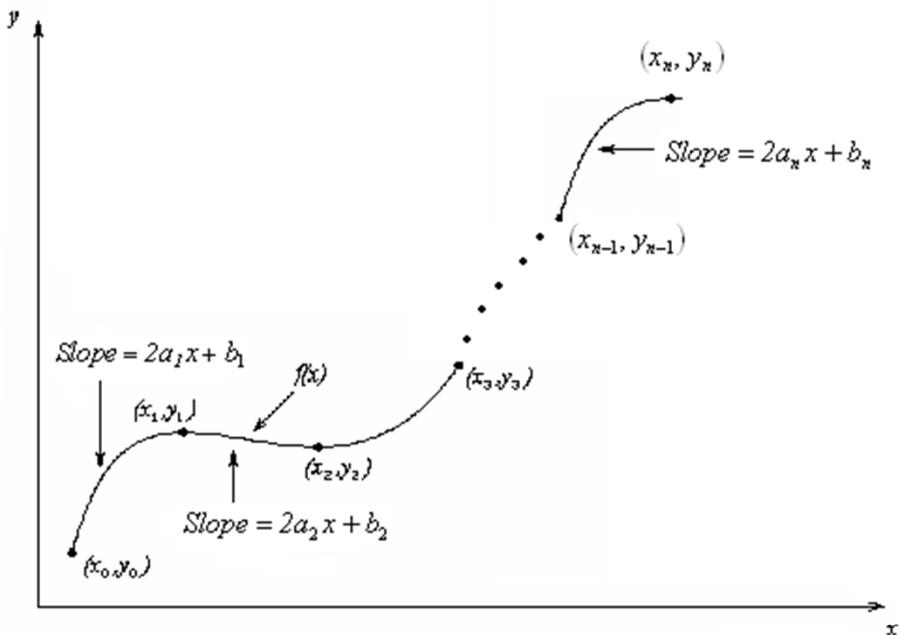
$$2a_ix_i + b_i - 2a_{i+1}x_i - b_{i+1} = 0$$

.

$$2a_{n-1}x_{n-1} + b_{n-1} - 2a_nx_{n-1} - b_n = 0$$

We have $(n-1)$ such equations. The total number of equations is $(2n) + (n-1) = (3n-1)$.

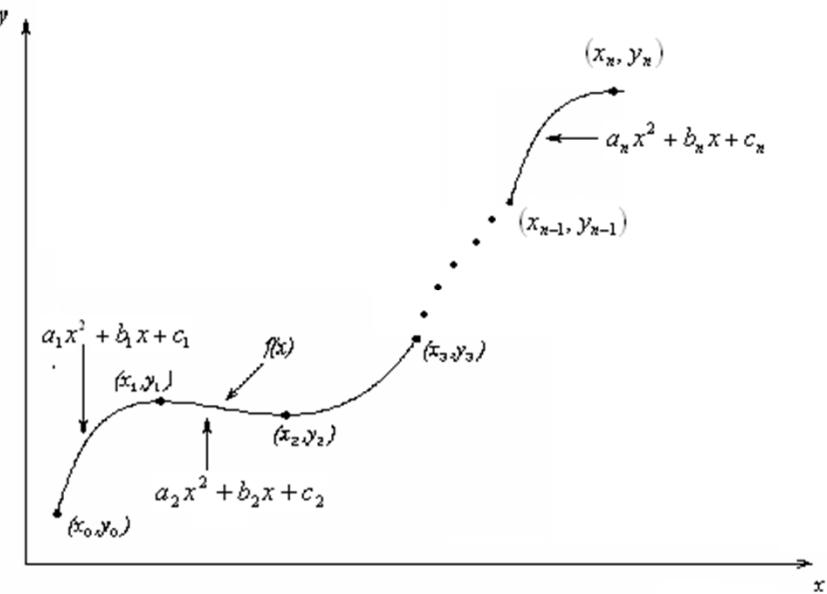
We can assume that the first spline is linear, that is $a_1 = 0$



Quadratic Splines (contd)

This gives us ‘ $3n$ ’ equations and ‘ $3n$ ’ unknowns. Once we find the ‘ $3n$ ’ constants, we can find the function at any value of ‘ x ’ using the splines,

$$\begin{aligned}f(x) &= a_1x^2 + b_1x + c_1, & x_0 \leq x \leq x_1 \\&= a_2x^2 + b_2x + c_2, & x_1 \leq x \leq x_2 \\&\vdots \\&= a_nx^2 + b_nx + c_n, & x_{n-1} \leq x \leq x_n\end{aligned}$$



Quadratic Spline Example

The upward velocity of a rocket is given as a function of time. Using quadratic splines

- a) Find the velocity at $t=16$ seconds
- b) Find the acceleration at $t=16$ seconds
- c) Find the distance covered between $t=11$ and $t=16$ seconds

Table Velocity as a function of time

t (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

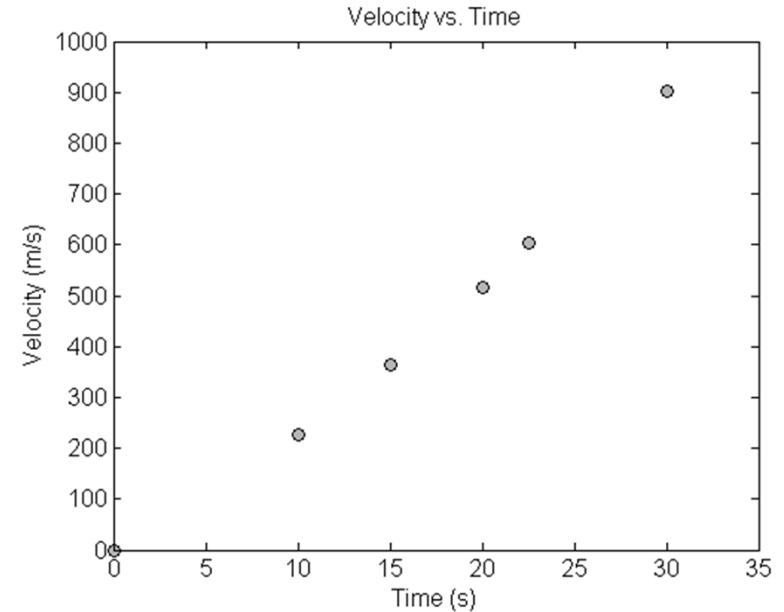


Figure. Velocity vs. time data for the rocket example

Solution

$$\begin{aligned}v(t) &= a_1 t^2 + b_1 t + c_1, \quad 0 \leq t \leq 10 \\&= a_2 t^2 + b_2 t + c_2, \quad 10 \leq t \leq 15 \\&= a_3 t^2 + b_3 t + c_3, \quad 15 \leq t \leq 20 \\&= a_4 t^2 + b_4 t + c_4, \quad 20 \leq t \leq 22.5 \\&= a_5 t^2 + b_5 t + c_5, \quad 22.5 \leq t \leq 30\end{aligned}$$

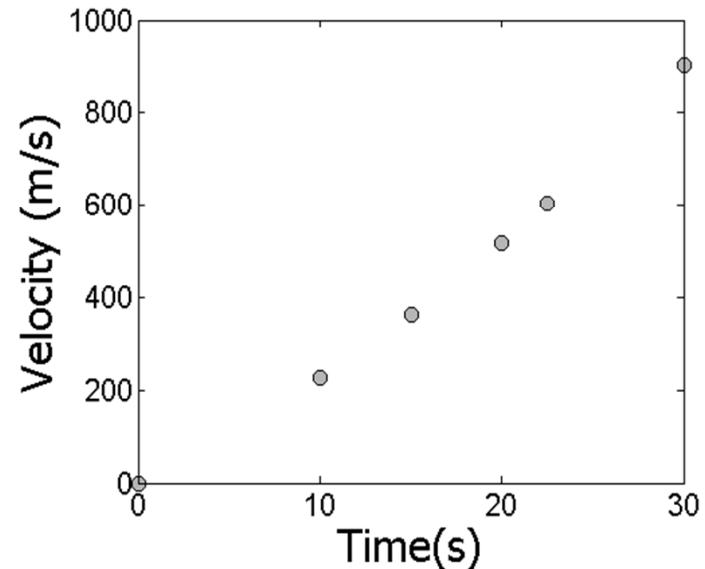
Let us set up the equations

Each Spline Goes Through Two Consecutive Data Points

$$v(t) = a_1 t^2 + b_1 t + c_1, \quad 0 \leq t \leq 10$$

$$a_1(0)^2 + b_1(0) + c_1 = 0$$

$$a_1(10)^2 + b_1(10) + c_1 = 227.04$$



Each Spline Goes Through Two Consecutive Data Points

t s	v(t) m/s
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

$$a_2(10)^2 + b_2(10) + c_2 = 227.04$$

$$a_2(15)^2 + b_2(15) + c_2 = 362.78$$

$$a_3(15)^2 + b_3(15) + c_3 = 362.78$$

$$a_3(20)^2 + b_3(20) + c_3 = 517.35$$

$$a_4(20)^2 + b_4(20) + c_4 = 517.35$$

$$a_4(22.5)^2 + b_4(22.5) + c_4 = 602.97$$

$$a_5(22.5)^2 + b_5(22.5) + c_5 = 602.97$$

$$a_5(30)^2 + b_5(30) + c_5 = 901.67$$

Derivatives are Continuous at Interior Data Points

$$v(t) = a_1 t^2 + b_1 t + c_1, \quad 0 \leq t \leq 10 \\ = a_2 t^2 + b_2 t + c_2, \quad 10 \leq t \leq 15$$

$$\left. \frac{d}{dt} (a_1 t^2 + b_1 t + c_1) \right|_{t=10} = \left. \frac{d}{dt} (a_2 t^2 + b_2 t + c_2) \right|_{t=10}$$

$$(2a_1 t + b_1) \Big|_{t=10} = (2a_2 t + b_2) \Big|_{t=10}$$

$$2a_1(10) + b_1 = 2a_2(10) + b_2$$

$$20a_1 + b_1 - 20a_2 - b_2 = 0$$

Derivatives are continuous at Interior Data Points

At t=10

$$2a_1(10) + b_1 - 2a_2(10) - b_2 = 0$$

At t=15

$$2a_2(15) + b_2 - 2a_3(15) - b_3 = 0$$

At t=20

$$2a_3(20) + b_3 - 2a_4(20) - b_4 = 0$$

At t=22.5

$$2a_4(22.5) + b_4 - 2a_5(22.5) - b_5 = 0$$

Last Equation

$$a_1 = 0$$

Final Set of Equations

$$\begin{bmatrix}
 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 100 & 10 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 100 & 10 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 225 & 15 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 225 & 15 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 400 & 20 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 400 & 20 & 1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 506.25 & 22.5 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 506.25 & 22.5 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 900 & 30 & 1 & 0 \\
 20 & 1 & 0 & -20 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 30 & 1 & 0 & -30 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 40 & 1 & 0 & -40 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 45 & 1 & 0 & -45 & -1 & 0 & 0 \\
 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix} = \begin{bmatrix}
 a_1 & 0 \\
 b_1 & 227.04 \\
 c_1 & 227.04 \\
 a_2 & 362.78 \\
 b_2 & 362.78 \\
 c_2 & 517.35 \\
 a_3 & 517.35 \\
 b_3 & 602.97 \\
 c_3 & 602.97 \\
 a_4 & 901.67 \\
 b_4 & 0 \\
 c_4 & 0 \\
 a_5 & 0 \\
 b_5 & 0 \\
 c_5 & 0
 \end{bmatrix}$$

Coefficients of Spline

i	a_i	b_i	c_i
1	0	22.704	0
2	0.8888	4.928	88.88
3	-0.1356	35.66	-141.61
4	1.6048	-33.956	554.55
5	0.20889	28.86	-152.13

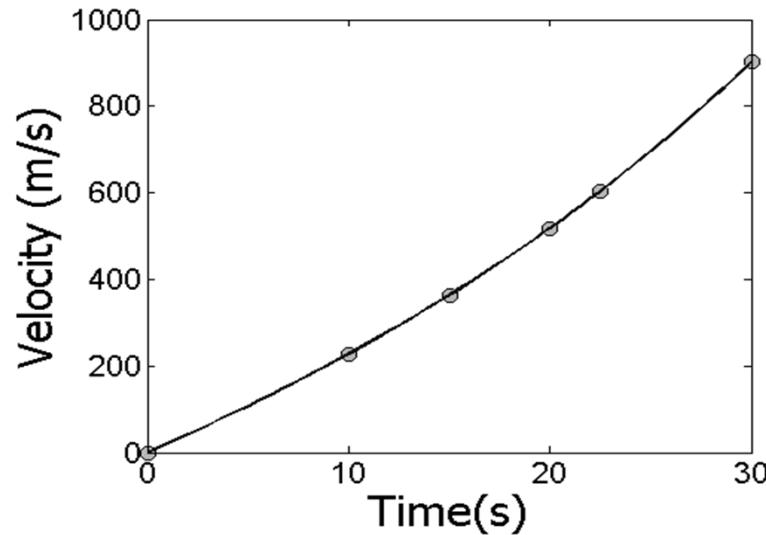
Quadratic Spline Interpolation

Part 2 of 2

<http://numericalmethods.eng.usf.edu>

Final Solution

$$\begin{aligned}v(t) &= 22.704t, & 0 \leq t \leq 10 \\&= 0.8888t^2 + 4.928t + 88.88, & 10 \leq t \leq 15 \\&= -0.1356t^2 + 35.66t - 141.61, & 15 \leq t \leq 20 \\&= 1.6048t^2 - 33.956t + 554.55, & 20 \leq t \leq 22.5 \\&= 0.20889t^2 + 28.86t - 152.13, & 22.5 \leq t \leq 30\end{aligned}$$



Velocity at a Particular Point

a) Velocity at t=16

$$\begin{aligned}v(t) &= 22.704t, & 0 \leq t \leq 10 \\&= 0.8888t^2 + 4.928t + 88.88, & 10 \leq t \leq 15 \\&= -0.1356t^2 + 35.66t - 141.61, & 15 \leq t \leq 20 \\&= 1.6048t^2 - 33.956t + 554.55, & 20 \leq t \leq 22.5 \\&= 0.20889t^2 + 28.86t - 152.13, & 22.5 \leq t \leq 30\end{aligned}$$

$$\begin{aligned}v(16) &= -0.1356(16)^2 + 35.66(16) - 141.61 \\&= 394.24 \text{ m/s}\end{aligned}$$

Acceleration from Velocity Profile

b) The quadratic spline valid at $t=16$ is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16}$$

$$v(t) = -0.1356 t^2 + 35.66t - 141.61, \quad 15 \leq t \leq 20$$

$$\begin{aligned} a(t) &= \frac{d}{dt} (-0.1356t^2 + 35.66t - 141.61) \\ &= -0.2712t + 35.66, \quad 15 \leq t \leq 20 \end{aligned}$$

$$a(16) = -0.2712(16) + 35.66 = 31.321 \text{m/s}^2$$

Distance from Velocity Profile

c) Find the distance covered by the rocket from $t=11\text{s}$ to $t=16\text{s}$.

$$S(16) - S(11) = \int_{11}^{16} v(t) dt$$

$$v(t) = 0.8888t^2 + 4.928t + 88.88, 10 \leq t \leq 15$$

$$= -0.1356t^2 + 35.66t - 141.61, 15 \leq t \leq 20$$

$$\begin{aligned} S(16) - S(11) &= \int_{11}^{16} v(t) dt = \int_{11}^{15} v(t) dt + \int_{15}^{16} v(t) dt \\ &= \int_{11}^{15} (0.8888t^2 + 4.928t + 88.88) dt + \int_{15}^{16} (-0.1356t^2 + 35.66t - 141.61) dt \\ &= 1595.9 \text{ m} \end{aligned}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/spline_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Linear Regression

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Linear Regression

<http://numericalmethods.eng.usf.edu>

What is Regression?

What is regression? Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
best fit $y = f(x)$ to the data.

Residual at each point E_i is $y_i - f(x_i)$

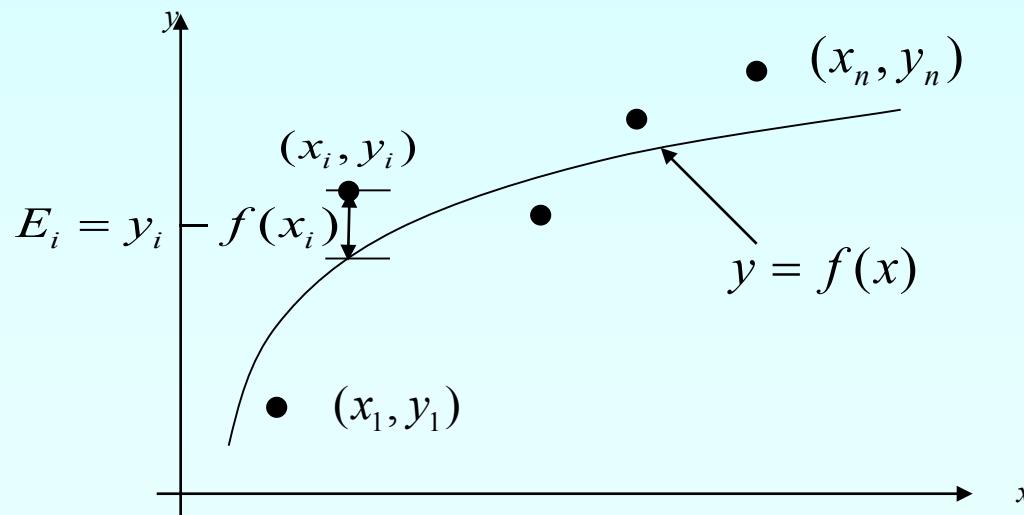


Figure. Basic model for regression

Linear Regression-Criterion#1

Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = a_0 + a_1 x$ to the data.

Does minimizing $\sum_{i=1}^n E_i$ work as a criterion?

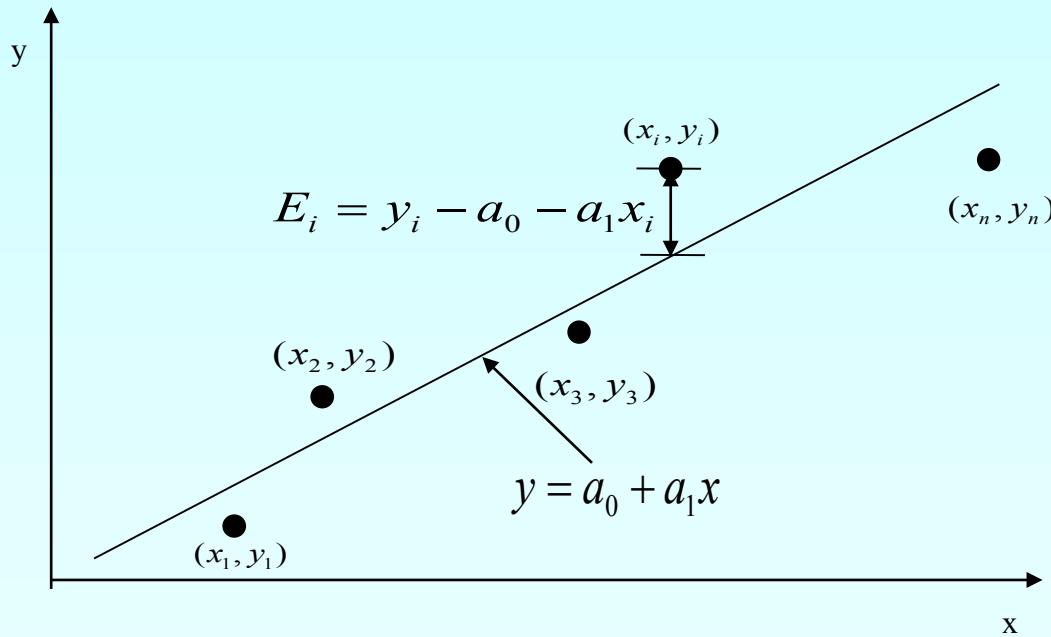


Figure. Linear regression of y vs x data showing residuals at a typical point, x_i .

Example for Criterion#1

Example: Given the data points (2,4), (3,6), (2,6) and (3,8), best fit the data to a straight line using Criterion#1

$$\text{Minimize} \sum_{i=1}^n E_i$$

Table. Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

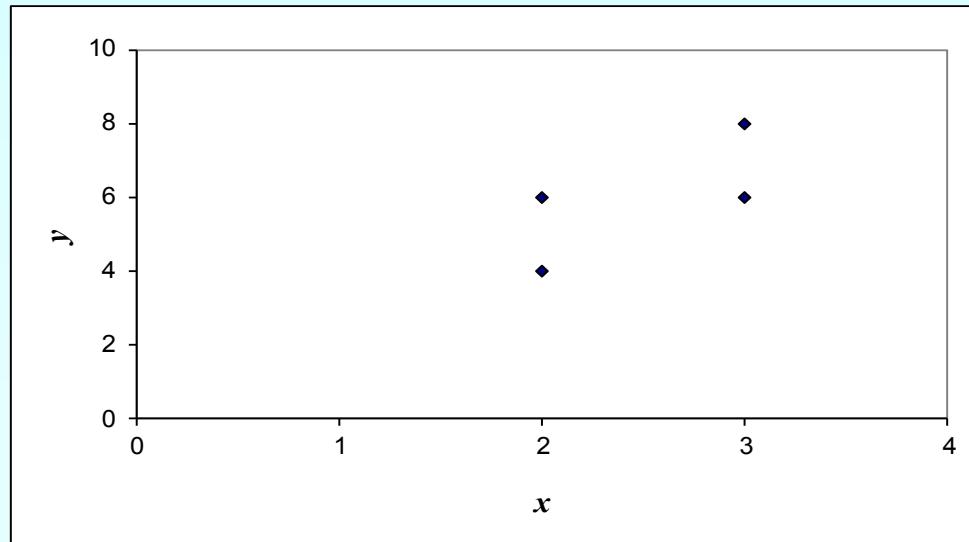


Figure. Data points for y vs x data.

Linear Regression-Criteria#1

Using $y=4x - 4$ as the regression curve

Table. Residuals at each point
for regression model $y=4x - 4$

x	y	$y_{predicted}$	$E = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
		$\sum_{i=1}^4 E_i = 0$	

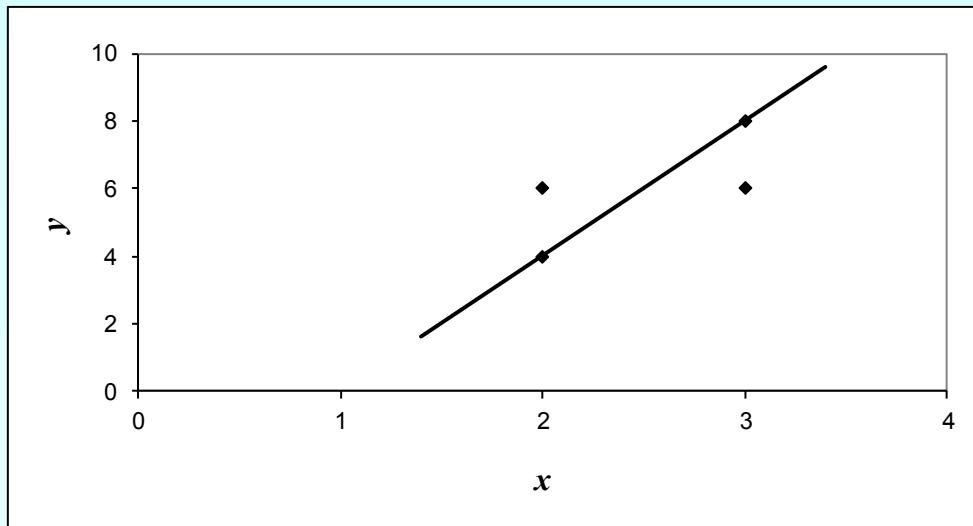


Figure. Regression curve $y=4x - 4$ and y vs x data

Linear Regression-Criterion#1

Using $y=6$ as a regression curve

Table. Residuals at each point
for regression model $y=6$

x	y	$y_{predicted}$	$E = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
		$\sum_{i=1}^4 E_i = 0$	

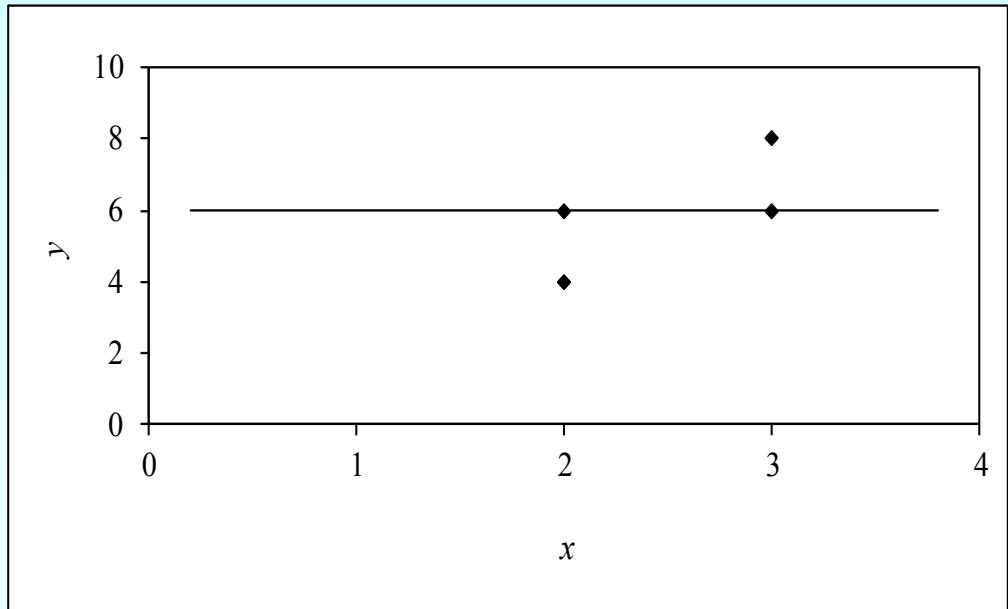


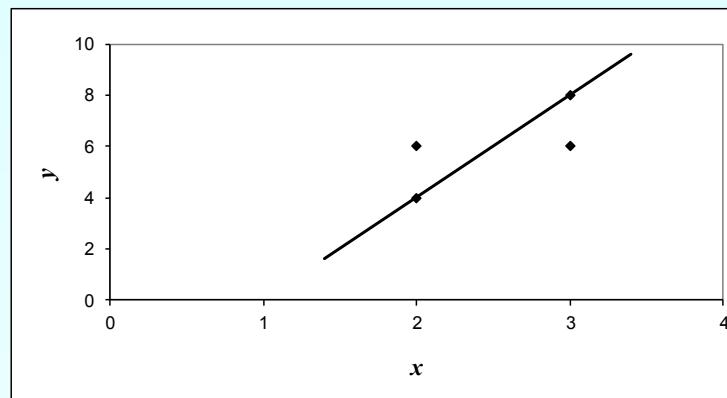
Figure. Regression curve $y=6$ and y vs x data

Linear Regression – Criterion #1

$$\sum_{i=1}^4 E_i = 0 \quad \text{for both regression models of } y=4x-4 \text{ and } y=6$$

The sum of the residuals is minimized, in this case it is zero, but the regression model is not unique.

Hence the criterion of minimizing the sum of the residuals is a bad criterion.



Linear Regression-Criterion#1

Using $y=4x - 4$ as the regression curve

Table. Residuals at each point
for regression model $y=4x - 4$

x	y	$y_{predicted}$	$E = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
		$\sum_{i=1}^4 E_i = 0$	

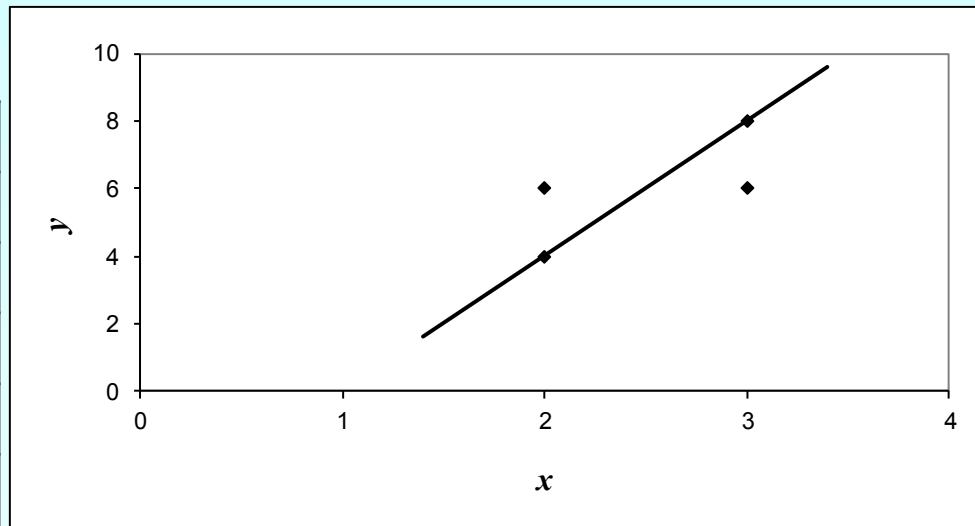


Figure. Regression curve $y=4x-4$ and y vs x data

Linear Regression-Criterion#2

Will minimizing $\sum_{i=1}^n |E_i|$ work any better?

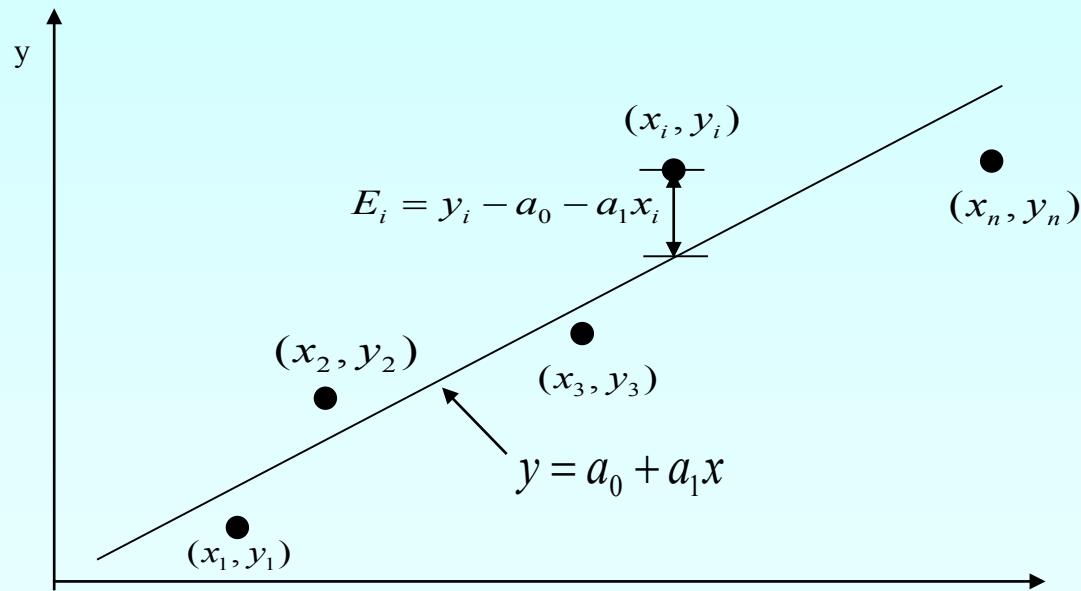


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i .

Example for Criterion#2

Example: Given the data points (2,4), (3,6), (2,6) and (3,8), best fit the data to a straight line using Criterion#2

$$\text{Minimize } \sum_{i=1}^n |E_i|$$

Table. Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

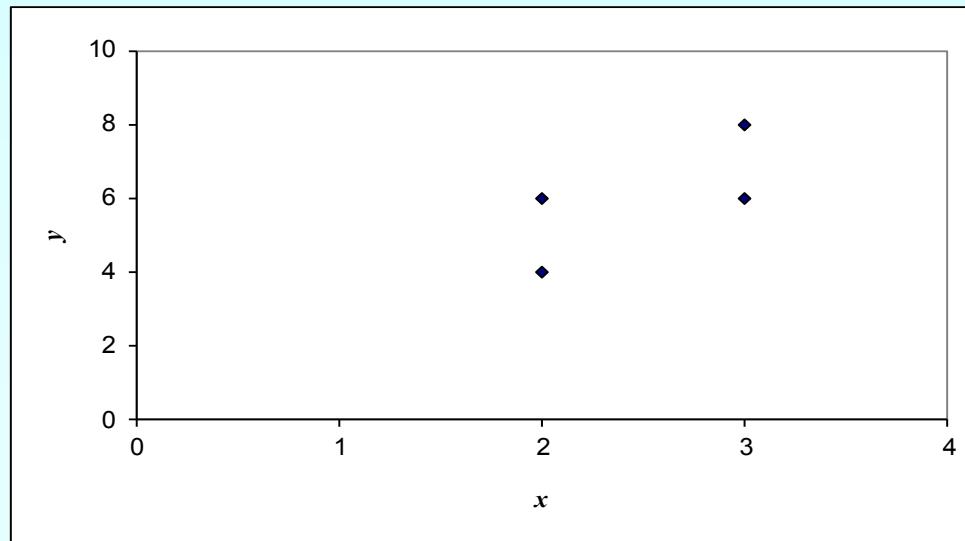


Figure. Data points for y vs. x data.

Linear Regression-Criterion#2

Using $y=4x - 4$ as the regression curve

Table. Residuals at each point

for regression model $y=4x - 4$

x	y	$y_{predicted}$	$E = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
		$\sum_{i=1}^4 E_i = 4$	

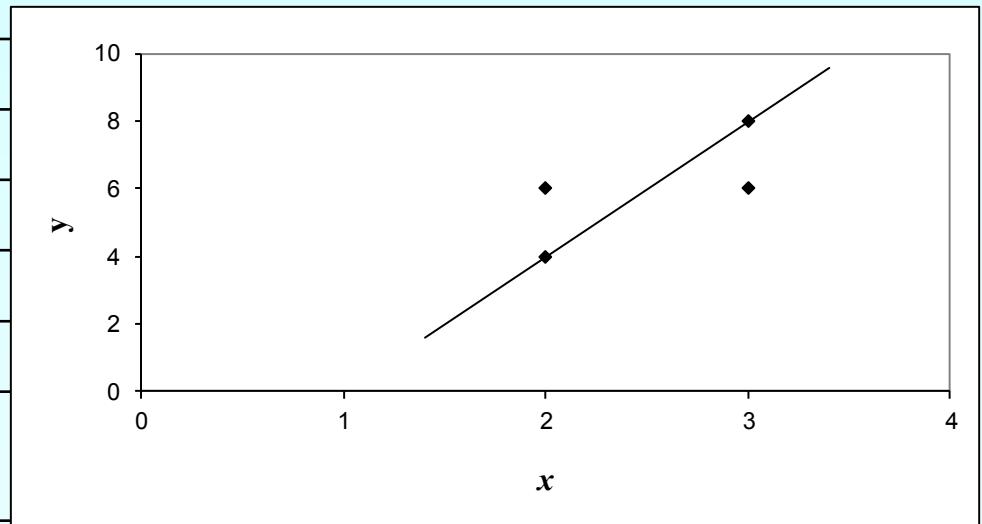


Figure. Regression curve $y=4x - 4$ and y vs. x data

Linear Regression-Criterion#2

Using $y=6$ as a regression curve

Table. Residuals at each point
for regression model $y=6$

x	y	$y_{predicted}$	$E = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
		$\sum_{i=1}^4 E_i = 4$	

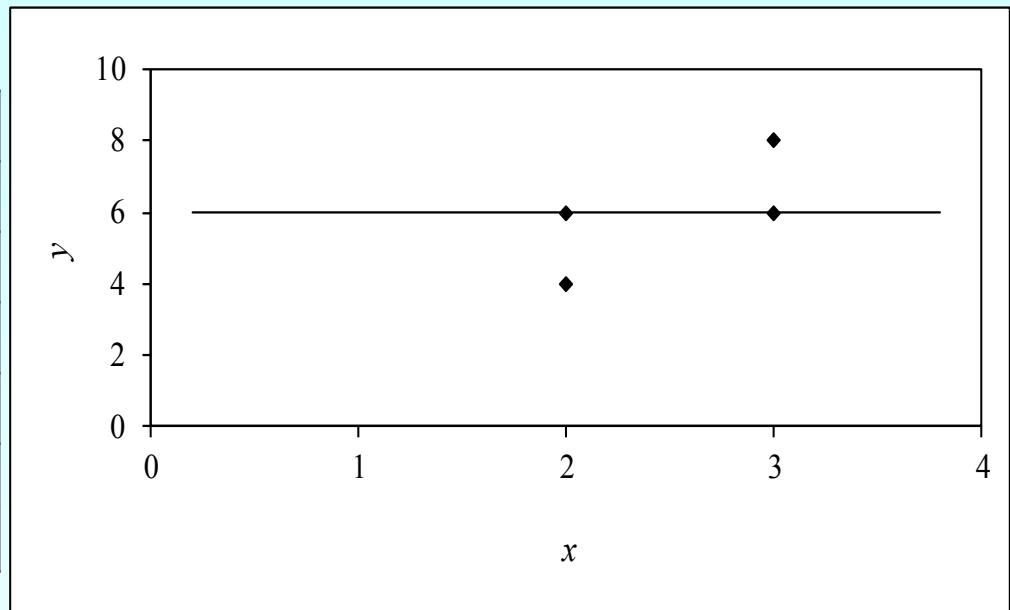


Figure. Regression curve $y=6$ and y vs x data

Linear Regression-Criterion#2

$$\sum_{i=1}^4 |E_i| = 4 \quad \text{for both regression models of } y=4x - 4 \text{ and } y=6.$$

The sum of the absolute residuals has been made as small as possible, that is 4, but the regression model is not unique.

Hence the criterion of minimizing the sum of the absolute value of the residuals is also a bad criterion.

Least Squares Criterion

The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.

$$S_r = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$$

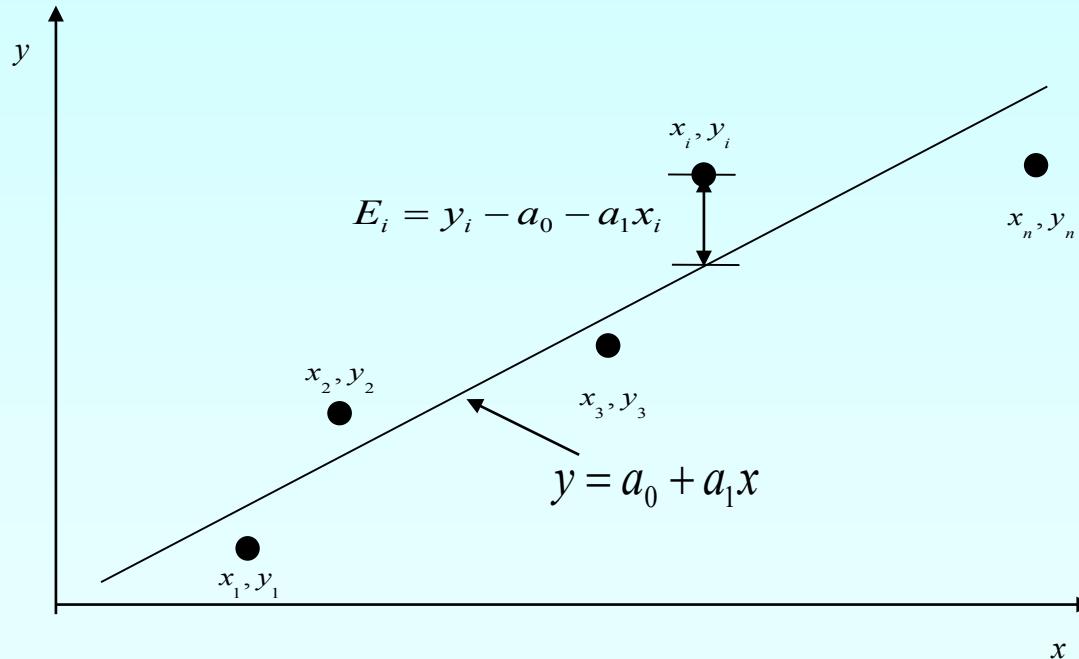


Figure. Linear regression of y vs x data showing residuals at a typical point, x_i .

Finding Constants of Linear Model

Minimize the sum of the square of the residuals: $S_r = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$

To find a_0 and a_1 we minimize S_r with respect to a_1 and a_0 .

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-x_i) = 0$$

giving

$$\sum_{i=1}^n a_0 + \sum_{i=1}^n a_1 x_i = \sum_{i=1}^n y_i$$

$$\sum_{i=1}^n a_0 x_i + \sum_{i=1}^n a_1 x_i^2 = \sum_{i=1}^n y_i x_i$$

Finding Constants of Linear Model

Solving for a_0 and a_1 directly yields,

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

and

$$a_0 = \bar{y} - a_1 \bar{x}$$
$$a_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

Example 1

The torque, T needed to turn the torsion spring of a mousetrap through an angle, θ , is given below. Find the constants for the model given by

$$T = k_1 + k_2\theta$$

Table: Torque vs Angle for a torsional spring

Angle, θ	Torque, T
Radians	N-m
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

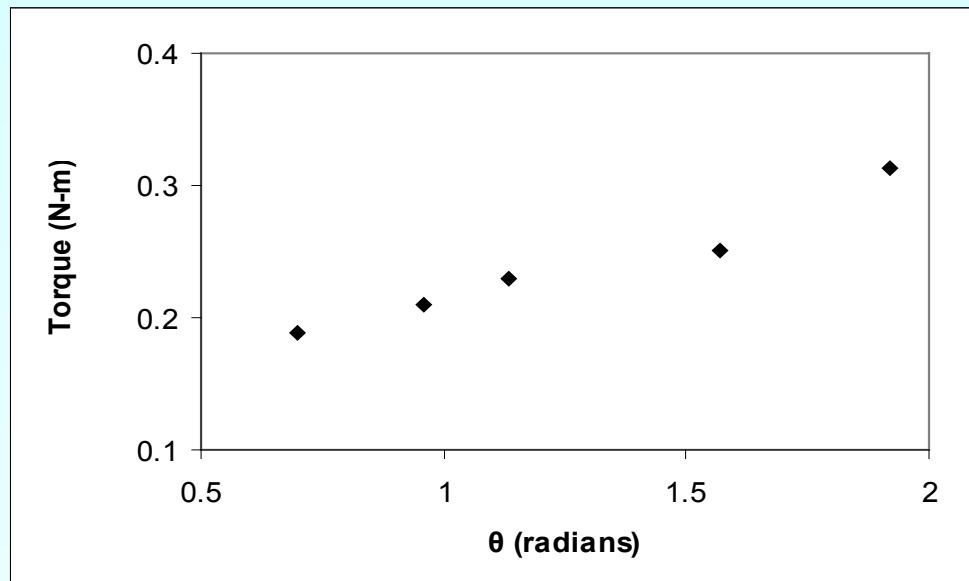


Figure. Data points for Torque vs Angle data

Example 1 cont.

The following table shows the summations needed for the calculations of the constants in the regression model.

Table. Tabulation of data for calculation of important summations

θ	T	θ^2	$T\theta$
Radians	N-m	Radians ²	N-m-Radians
0.698132	0.188224	0.487388	0.131405
0.959931	0.209138	0.921468	0.200758
1.134464	0.230052	1.2870	0.260986
1.570796	0.250965	2.4674	0.394215
1.919862	0.313707	3.6859	0.602274
$\sum_{i=1}^5 =$	6.2831	8.8491	1.5896

Using equations described for a_0 and a_1 with $n = 5$

$$k_2 = \frac{n \sum_{i=1}^5 \theta_i T_i - \sum_{i=1}^5 \theta_i \sum_{i=1}^5 T_i}{n \sum_{i=1}^5 \theta_i^2 - \left(\sum_{i=1}^5 \theta_i \right)^2}$$
$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$
$$= 9.6091 \times 10^{-2} \text{ N-m/rad}$$

Example 1 cont.

Use the average torque and average angle to calculate k_1

$$\bar{T} = \frac{\sum_{i=1}^5 T_i}{n}$$

$$= \frac{1.1921}{5}$$

$$= 2.3842 \times 10^{-1}$$

$$\bar{\theta} = \frac{\sum_{i=1}^5 \theta_i}{n}$$

$$= \frac{6.2831}{5}$$

$$= 1.2566$$

Using,

$$k_1 = \bar{T} - k_2 \bar{\theta}$$

$$= 2.3842 \times 10^{-1} - (9.6091 \times 10^{-2})(1.2566)$$

$$= 1.1767 \times 10^{-1} \text{ N-m}$$

Example 1 Results

Using linear regression, a trend line is found from the data

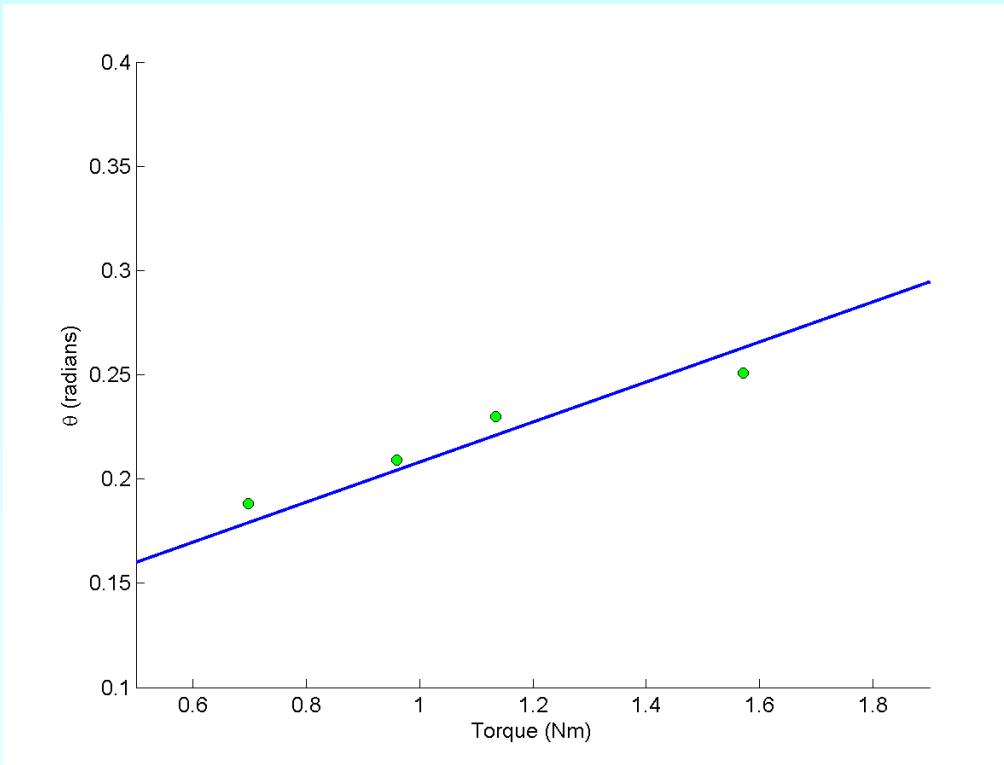


Figure. Linear regression of Torque versus Angle data

Can you find the energy in the spring if it is twisted from 0 to 180 degrees?

Linear Regression (special case)

Given

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

best fit

$$y = a_1 x$$

to the data.

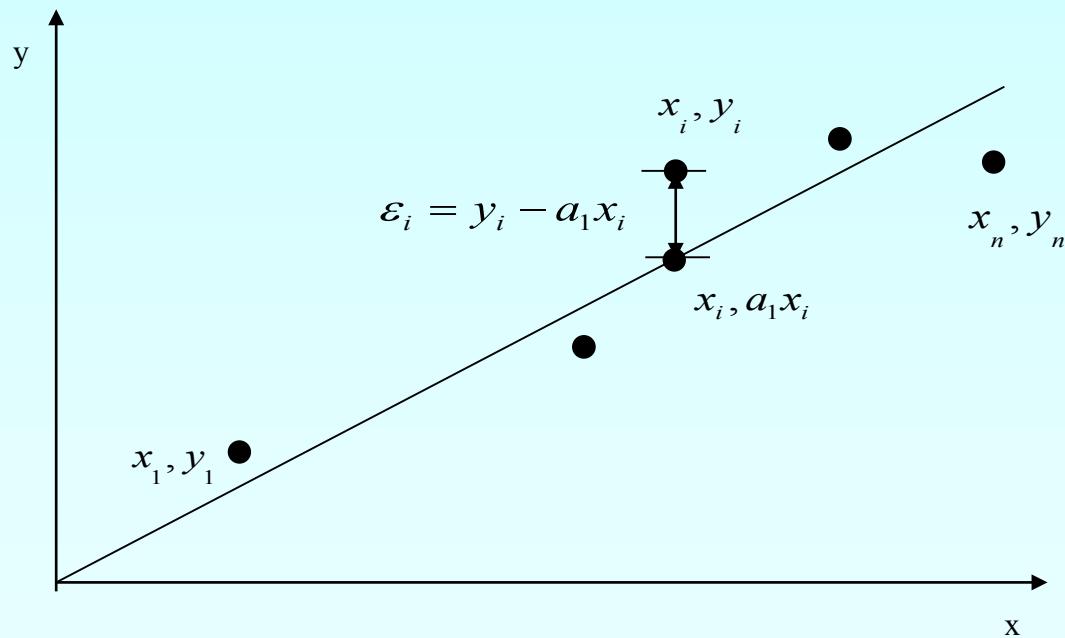
Linear Regression (special case cont.)

$$y = a_1 x$$

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

Is this correct?

Linear Regression (special case cont.)



Linear Regression (special case cont.)

Residual at each data point

$$\varepsilon_i = y_i - a_1 x_i$$

Sum of square of residuals

$$\begin{aligned} S_r &= \sum_{i=1}^n \varepsilon_i^2 \\ &= \sum_{i=1}^n (y_i - a_1 x_i)^2 \end{aligned}$$

Linear Regression (special case cont.)

Differentiate with respect to a_1

$$\frac{dS_r}{da_1} = \sum_{i=1}^n 2(y_i - a_1 x_i)(-x_i)$$

$$= \sum_{i=1}^n (-2y_i x_i + 2a_1 x_i^2)$$

$$\frac{dS_r}{da_1} = 0$$

gives

$$a_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Linear Regression (special case cont.)

Does this value of a_1 correspond to a local minima or local maxima?

$$a_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

$$\frac{dS_r}{da_1} = \sum_{i=1}^n (-2y_i x_i + 2a_1 x_i^2)$$

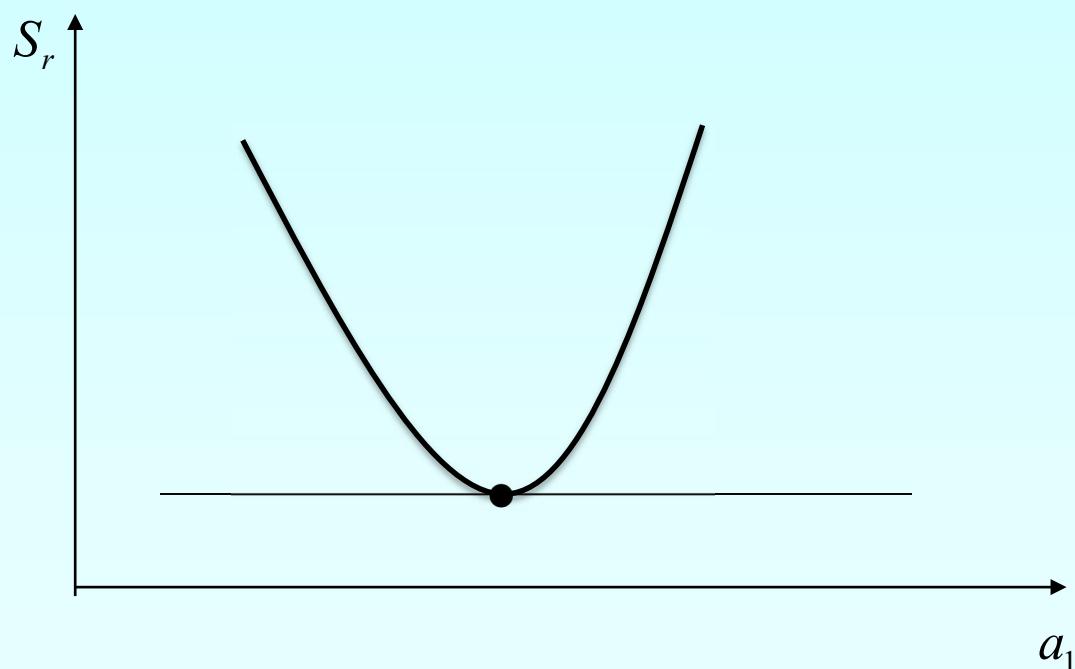
$$\frac{d^2 S_r}{da_1^2} = \sum_{i=1}^n 2x_i^2 > 0$$

Yes, it corresponds to a local minima.

$$a_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Linear Regression (special case cont.)

Is this local minima of S_r an absolute minimum of S_r ?



Example 2

To find the longitudinal modulus of composite, the following data is collected. Find the longitudinal modulus, E using the regression model

Table. Stress vs. Strain data

Strain (%)	Stress (MPa)
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.56	2896

$\sigma = E\varepsilon$ and the sum of the square of the residuals.

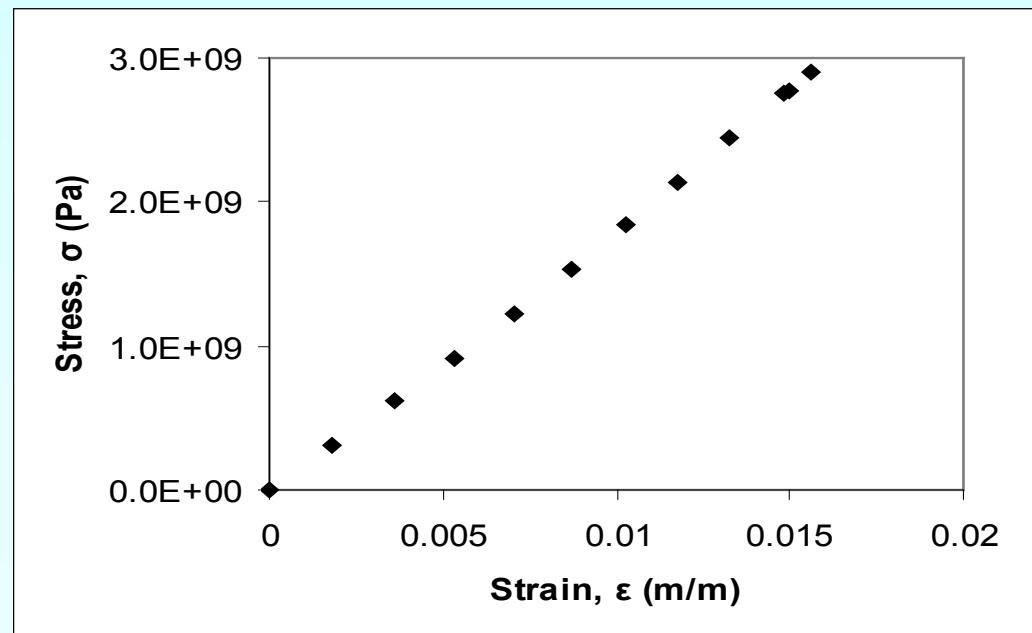


Figure. Data points for Stress vs. Strain data

Example 2 cont.

Table. Summation data for regression model

i	ε	σ	ε^2	$\varepsilon\sigma$
1	0.0000	0.0000	0.0000	0.0000
2	1.8300×10^{-3}	3.0600×10^8	3.3489×10^{-6}	5.5998×10^5
3	3.6000×10^{-3}	6.1200×10^8	1.2960×10^{-5}	2.2032×10^6
4	5.3240×10^{-3}	9.1700×10^8	2.8345×10^{-5}	4.8821×10^6
5	7.0200×10^{-3}	1.2230×10^9	4.9280×10^{-5}	8.5855×10^6
6	8.6700×10^{-3}	1.5290×10^9	7.5169×10^{-5}	1.3256×10^7
7	1.0244×10^{-2}	1.8350×10^9	1.0494×10^{-4}	1.8798×10^7
8	1.1774×10^{-2}	2.1400×10^9	1.3863×10^{-4}	2.5196×10^7
9	1.3290×10^{-2}	2.4460×10^9	1.7662×10^{-4}	3.2507×10^7
10	1.4790×10^{-2}	2.7520×10^9	2.1874×10^{-4}	4.0702×10^7
11	1.5000×10^{-2}	2.7670×10^9	2.2500×10^{-4}	4.1505×10^7
12	1.5600×10^{-2}	2.8960×10^9	2.4336×10^{-4}	4.5178×10^7
$\sum_{i=1}^{12}$			1.2764×10^{-3}	2.3337×10^8

$$E = \frac{\sum_{i=1}^n \sigma_i \varepsilon_i}{\sum_{i=1}^n \varepsilon_i^2}$$

$$\sum_{i=1}^{12} \varepsilon_i^2 = 1.2764 \times 10^{-3}$$

$$\sum_{i=1}^{12} \sigma_i \varepsilon_i = 2.3337 \times 10^8$$

$$E = \frac{\sum_{i=1}^{12} \sigma_i \varepsilon_i}{\sum_{i=1}^{12} \varepsilon_i^2}$$

$$= \frac{2.3337 \times 10^8}{1.2764 \times 10^{-3}} \\ = 182.84 \text{ GPa}$$

Example 2 Results

The equation $\sigma = 182.84 \times 10^9 \varepsilon$ describes the data.

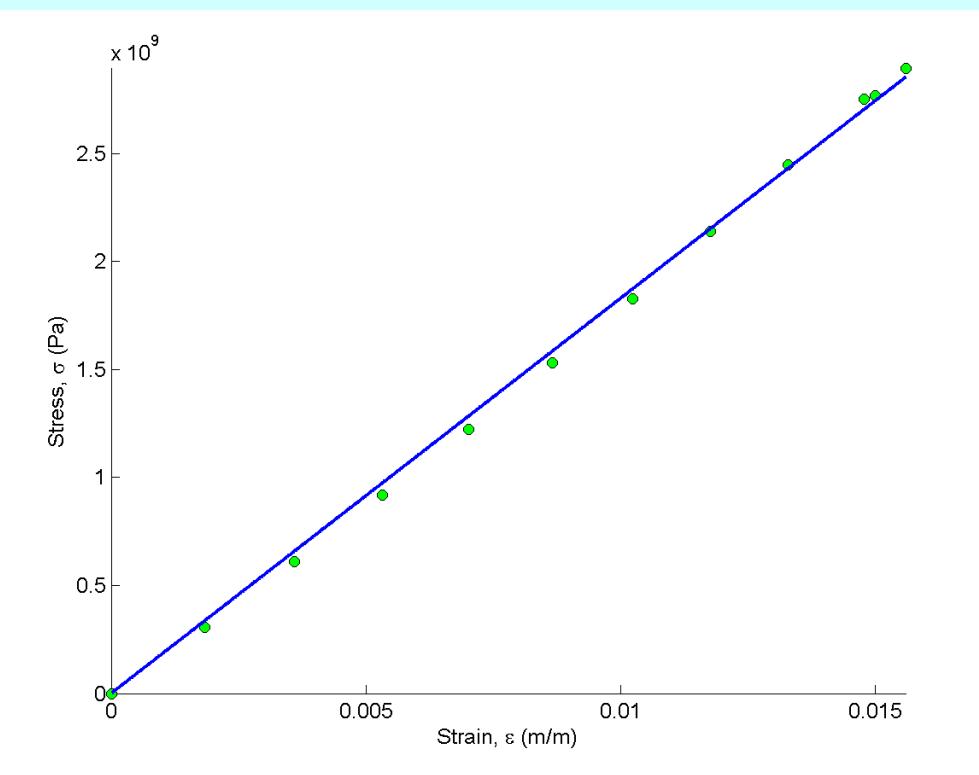


Figure. Linear regression for stress vs. strain data

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/linear_regression.html

THE END

<http://numericalmethods.eng.usf.edu>

Nonlinear Regression

Major: All Engineering Majors

Authors: Autar Kaw, Luke Snyder

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Nonlinear Regression

<http://numericalmethods.eng.usf.edu>

Nonlinear Regression

Some popular nonlinear regression models:

1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^b)$
3. Saturation growth model: $\left(y = \frac{ax}{b+x} \right)$
4. Polynomial model: $(y = a_0 + a_1x + \dots + a_mx^m)$

Nonlinear Regression

Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = f(x)$ to the data, where $f(x)$ is a nonlinear function of x .

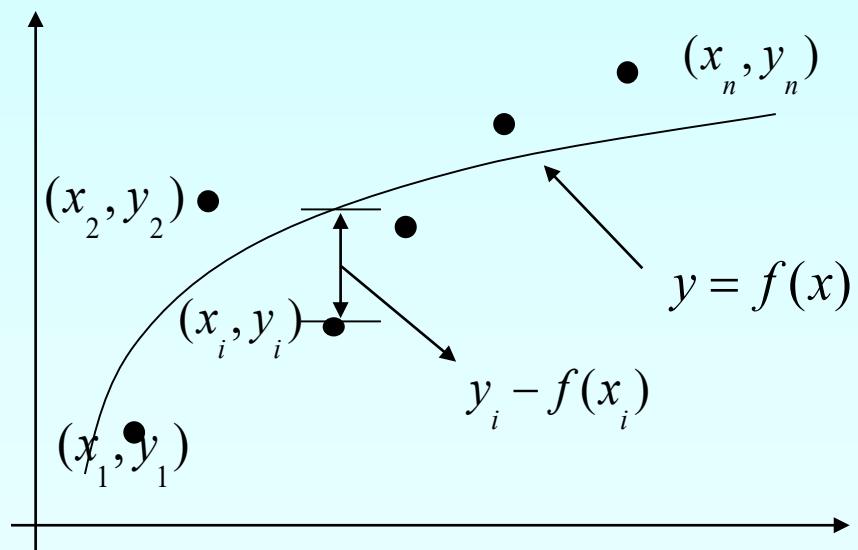


Figure. Nonlinear regression model for discrete y vs. x data

Regression Exponential Model

Exponential Model

Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = ae^{bx}$ to the data.

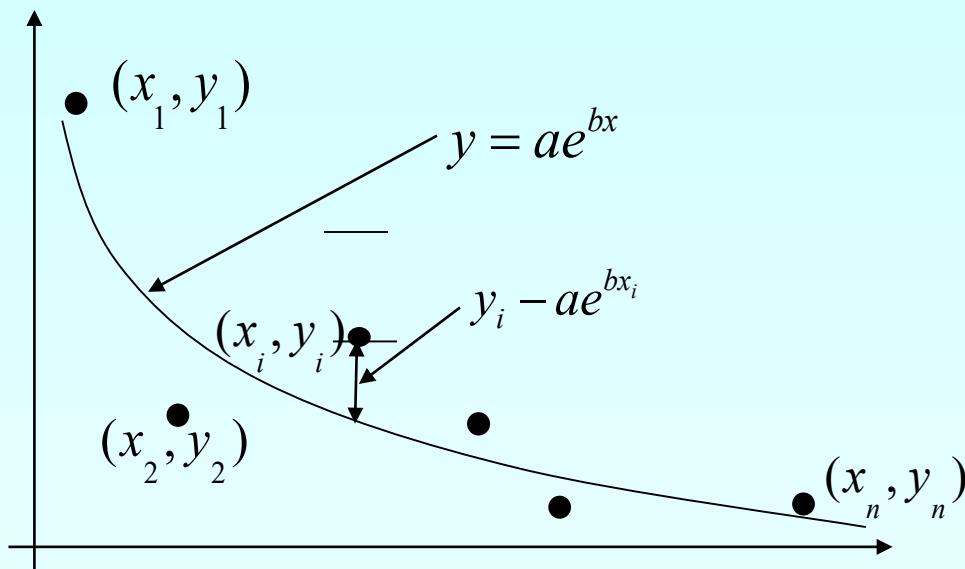


Figure. Exponential model of nonlinear regression for y vs. x data

Finding Constants of Exponential Model

The sum of the square of the residuals is defined as

$$S_r = \sum_{i=1}^n \left(y_i - ae^{bx_i} \right)^2$$

Differentiate with respect to a and b

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^n 2 \left(y_i - ae^{bx_i} \right) \left(-e^{bx_i} \right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^n 2 \left(y_i - ae^{bx_i} \right) \left(-ax_i e^{bx_i} \right) = 0$$

Finding Constants of Exponential Model

Rewriting the equations, we obtain

$$-\sum_{i=1}^n y_i e^{bx_i} + a \sum_{i=1}^n e^{2bx_i} = 0$$

$$\sum_{i=1}^n y_i x_i e^{bx_i} - a \sum_{i=1}^n x_i e^{2bx_i} = 0$$

Finding constants of Exponential Model

Solving the first equation for a yields

$$a = \frac{\sum_{i=1}^n y_i e^{bx_i}}{\sum_{i=1}^n e^{2bx_i}}$$

Substituting a back into the previous equation

$$\sum_{i=1}^n y_i x_i e^{bx_i} - \frac{\sum_{i=1}^n y_i e^{bx_i}}{\sum_{i=1}^n e^{2bx_i}} \sum_{i=1}^n x_i e^{2bx_i} = 0$$

The constant b can be found through numerical methods such as bisection method.

Example 1-Exponential Model

Many patients get concerned when a test involves injection of a radioactive material. For example for scanning a gallbladder, a few drops of Technetium-99m isotope is used. Half of the Technetium-99m would be gone in about 6 hours. It, however, takes about 24 hours for the radiation levels to reach what we are exposed to in day-to-day activities. Below is given the relative intensity of radiation as a function of time.

Table. Relative intensity of radiation as a function of time.

t(hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

Example 1-Exponential Model cont.

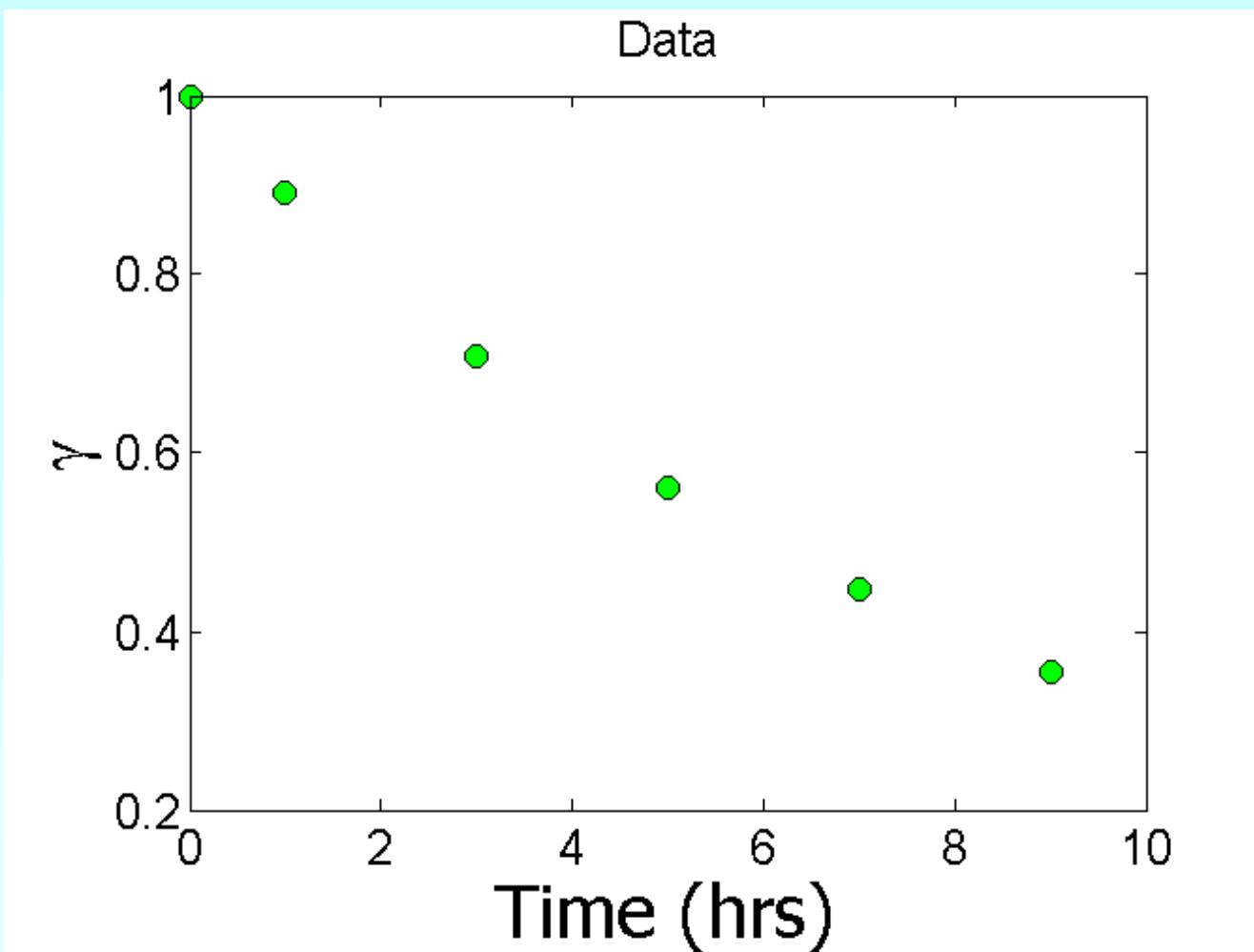
The relative intensity is related to time by the equation

$$\gamma = Ae^{\lambda t}$$

Find:

- a) The value of the regression constants A and λ
- b) The half-life of Technetium-99m
- c) Radiation intensity after 24 hours

Plot of data



Constants of the Model

$$\gamma = Ae^{\lambda t}$$

The value of λ is found by solving the nonlinear equation

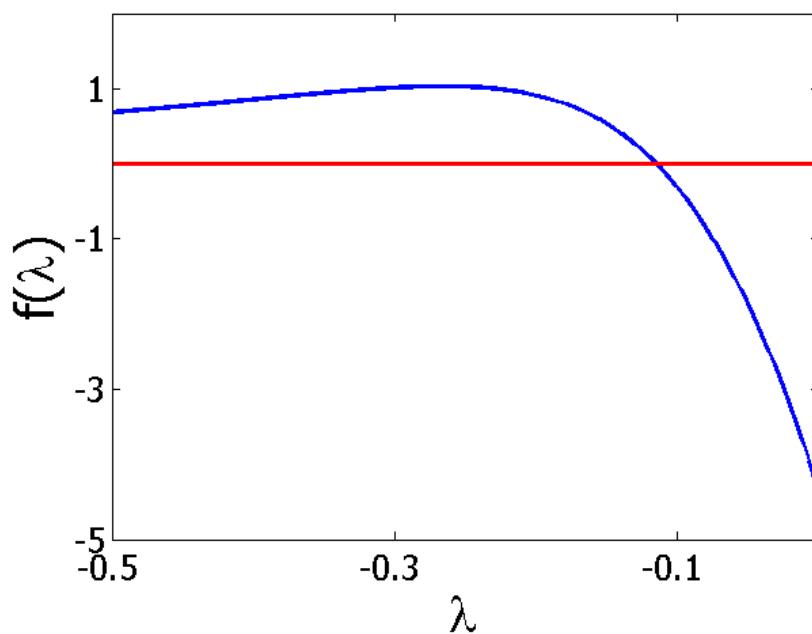
$$f(\lambda) = \sum_{i=1}^n \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}} \sum_{i=1}^n t_i e^{2\lambda t_i} = 0$$

$$A = \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}}$$

Setting up the Equation in MATLAB

$$f(\lambda) = \sum_{i=1}^n \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}} \sum_{i=1}^n t_i e^{2\lambda t_i} = 0$$

$f(\lambda)$ vs λ



t (hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

Setting up the Equation in MATLAB

$$f(\lambda) = \sum_{i=1}^n \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}} \sum_{i=1}^n t_i e^{2\lambda t_i} = 0$$

$$\lambda = -0.1151$$

```
t=[0 1 3 5 7 9]
gamma=[1 0.891 0.708 0.562 0.447 0.355]
syms lamda
sum1=sum(gamma.*t.*exp(lamda*t));
sum2=sum(gamma.*exp(lamda*t));
sum3=sum(exp(2*lamda*t));
sum4=sum(t.*exp(2*lamda*t));
f=sum1-sum2/sum3*sum4;
```

Calculating the Other Constant

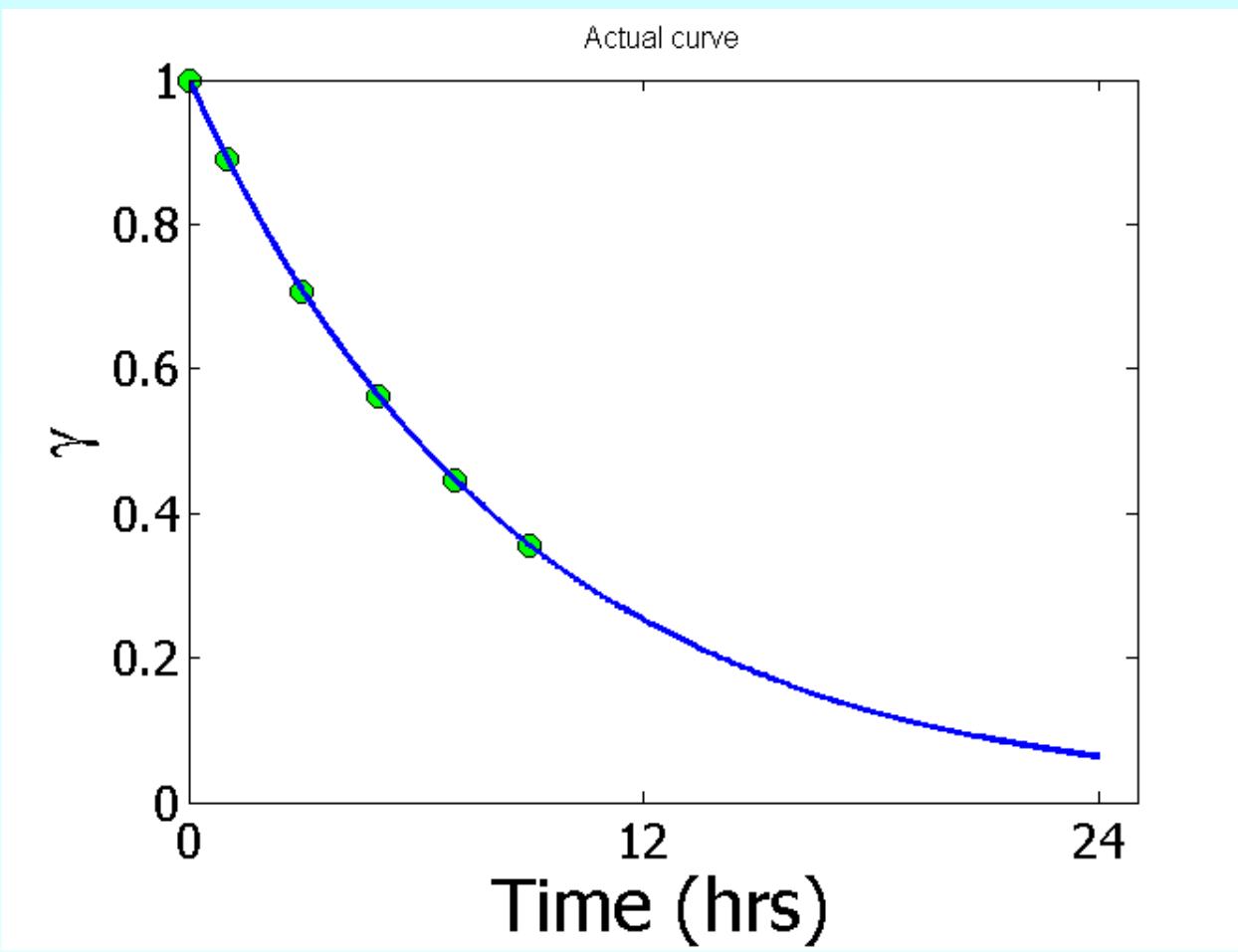
The value of A can now be calculated

$$A = \frac{\sum_{i=1}^6 \gamma_i e^{\lambda t_i}}{\sum_{i=1}^6 e^{2\lambda t_i}} = 0.9998$$

The exponential regression model then is

$$\gamma = 0.9998 e^{-0.1151t}$$

Plot of data and regression curve



Relative Intensity After 24 hrs

The relative intensity of radiation after 24 hours

$$\begin{aligned}\gamma &= 0.9998 \times e^{-0.1151(24)} \\ &= 6.3160 \times 10^{-2}\end{aligned}$$

This result implies that only

$$\frac{6.316 \times 10^{-2}}{0.9998} \times 100 = 6.317\%$$

radioactive intensity is left after 24 hours.

Homework

- What is the half-life of Technetium-99m isotope?
- Write a program in the language of your choice to find the constants of the model.
- Compare the constants of this regression model with the one where the data is transformed.
- What if the model was $\gamma = e^{\lambda t}$?

THE END

<http://numericalmethods.eng.usf.edu>

Polynomial Model

Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = a_0 + a_1 x + \dots + a_m x^m$ ($m \leq n - 2$) to a given data set.

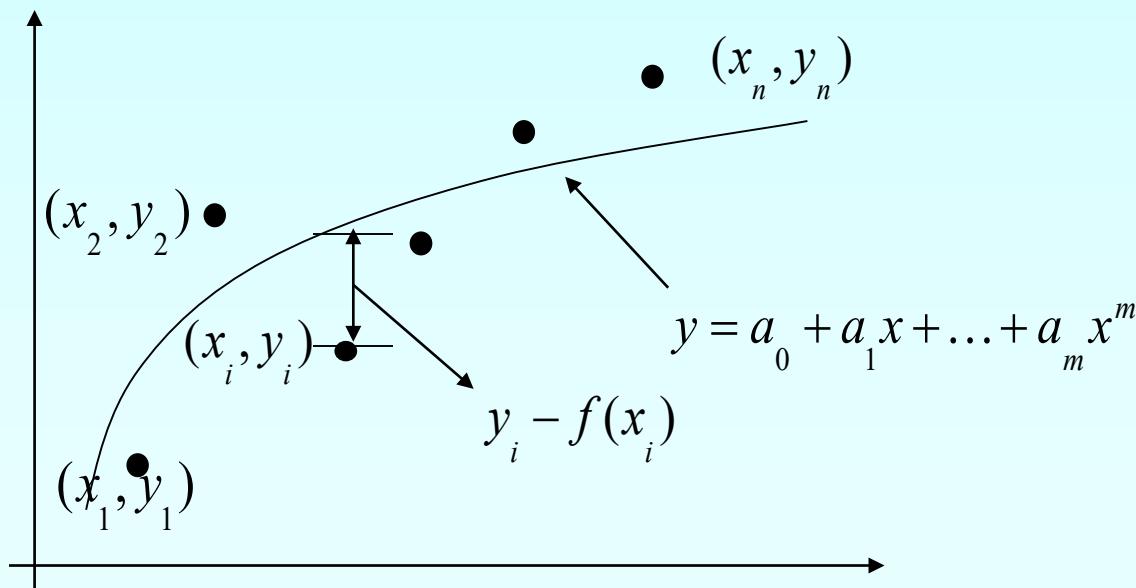


Figure. Polynomial model for nonlinear regression of y vs. x data

Polynomial Model cont.

The residual at each data point is given by

$$E_i = y_i - a_0 - a_1 x_i - \dots - a_m x_i^m$$

The sum of the square of the residuals then is

$$\begin{aligned} S_r &= \sum_{i=1}^n E_i^2 \\ &= \sum_{i=1}^n (y_i - a_0 - a_1 x_i - \dots - a_m x_i^m)^2 \end{aligned}$$

Polynomial Model cont.

To find the constants of the polynomial model, we set the derivatives with respect to a_i where $i = 1, \dots, m$, equal to zero.

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^n 2.(y_i - a_0 - a_1 x_i - \dots - a_m x_i^m)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^n 2.(y_i - a_0 - a_1 x_i - \dots - a_m x_i^m)(-x_i) = 0$$

$$\vdots \quad \vdots \quad \vdots$$

$$\frac{\partial S_r}{\partial a_m} = \sum_{i=1}^n 2.(y_i - a_0 - a_1 x_i - \dots - a_m x_i^m)(-x_i^m) = 0$$

Polynomial Model cont.

These equations in matrix form are given by

$$\begin{bmatrix} n & \left(\sum_{i=1}^n x_i\right) & \dots & \left(\sum_{i=1}^n x_i^m\right) \\ \left(\sum_{i=1}^n x_i\right) & \left(\sum_{i=1}^n x_i^2\right) & \dots & \left(\sum_{i=1}^n x_i^{m+1}\right) \\ \dots & \dots & \dots & \dots \\ \left(\sum_{i=1}^n x_i^m\right) & \left(\sum_{i=1}^n x_i^{m+1}\right) & \dots & \left(\sum_{i=1}^n x_i^{2m}\right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \dots \\ \sum_{i=1}^n x_i^m y_i \end{bmatrix}$$

The above equations are then solved for a_0, a_1, \dots, a_m

Example 2-Polynomial Model

Regress the thermal expansion coefficient vs. temperature data to a second order polynomial.

Table. Data points for temperature vs α

Temperature, T (°F)	Coefficient of thermal expansion, α (in/in/°F)
80	6.47×10^{-6}
40	6.24×10^{-6}
-40	5.72×10^{-6}
-120	5.09×10^{-6}
-200	4.30×10^{-6}
-280	3.33×10^{-6}
-340	2.45×10^{-6}

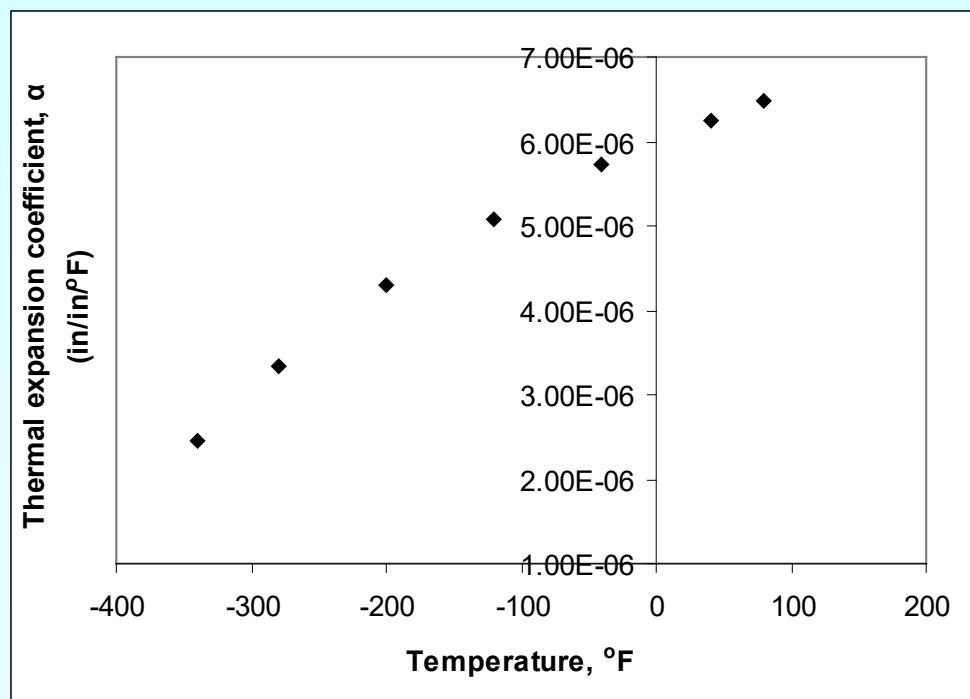


Figure. Data points for thermal expansion coefficient vs temperature.

Example 2-Polynomial Model cont.

We are to fit the data to the polynomial regression model

$$\alpha = a_0 + a_1 T + a_2 T^2$$

The coefficients a_0, a_1, a_2 are found by differentiating the sum of the square of the residuals with respect to each variable and setting the values equal to zero to obtain

$$\begin{bmatrix} n & \left(\sum_{i=1}^n T_i \right) & \left(\sum_{i=1}^n T_i^2 \right) \\ \left(\sum_{i=1}^n T_i \right) & \left(\sum_{i=1}^n T_i^2 \right) & \left(\sum_{i=1}^n T_i^3 \right) \\ \left(\sum_{i=1}^n T_i^2 \right) & \left(\sum_{i=1}^n T_i^3 \right) & \left(\sum_{i=1}^n T_i^4 \right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \alpha_i \\ \sum_{i=1}^n T_i \alpha_i \\ \sum_{i=1}^n T_i^2 \alpha_i \end{bmatrix}$$

Example 2-Polynomial Model cont.

The necessary summations are as follows

Table. Data points for temperature vs. α

Temperature, T (°F)	Coefficient of thermal expansion, α (in/in/°F)
80	6.47×10^{-6}
40	6.24×10^{-6}
-40	5.72×10^{-6}
-120	5.09×10^{-6}
-200	4.30×10^{-6}
-280	3.33×10^{-6}
-340	2.45×10^{-6}

$$\sum_{i=1}^7 T_i^2 = 2.5580 \times 10^5$$

$$\sum_{i=1}^7 T_i^3 = -7.0472 \times 10^7$$

$$\sum_{i=1}^7 T_i^4 = 2.1363 \times 10^{10}$$

$$\sum_{i=1}^7 \alpha_i = 3.3600 \times 10^{-5}$$

$$\sum_{i=1}^7 T_i \alpha_i = -2.6978 \times 10^{-3}$$

$$\sum_{i=1}^7 T_i^2 \alpha_i = 8.5013 \times 10^{-1}$$

Example 2-Polynomial Model cont.

Using these summations, we can now calculate a_0, a_1, a_2

$$\begin{bmatrix} 7.0000 & -8.6000 \times 10^2 & 2.5800 \times 10^5 \\ -8.600 \times 10^2 & 2.5800 \times 10^5 & -7.0472 \times 10^7 \\ 2.5800 \times 10^5 & -7.0472 \times 10^7 & 2.1363 \times 10^{10} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 3.3600 \times 10^{-5} \\ -2.6978 \times 10^{-3} \\ 8.5013 \times 10^{-1} \end{bmatrix}$$

Solving the above system of simultaneous linear equations we have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.0217 \times 10^{-6} \\ 6.2782 \times 10^{-9} \\ -1.2218 \times 10^{-11} \end{bmatrix}$$

The polynomial regression model is then

$$\begin{aligned} a &= a_0 + a_1 T + a_2 T^2 \\ &= 6.0217 \times 10^{-6} + 6.2782 \times 10^{-9} T - 1.2218 \times 10^{-11} T^2 \end{aligned}$$

Transformation of Data

To find the constants of many nonlinear models, it results in solving simultaneous nonlinear equations. For mathematical convenience, some of the data for such models can be transformed. For example, the data for an exponential model can be transformed.

As shown in the previous example, many chemical and physical processes are governed by the equation,

$$y = ae^{bx}$$

Taking the natural log of both sides yields,

$$\ln y = \ln a + bx$$

Let $z = \ln y$ and $a_0 = \ln a$

We now have a linear regression model where $z = a_0 + a_1 x$

(implying) $a = e^{a_0}$ with $a_1 = b$

Transformation of data cont.

Using linear model regression methods,

$$a_1 = \frac{n \sum_{i=1}^n x_i z_i - \sum_{i=1}^n x_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$a_0 = \bar{z} - a_1 \bar{x}$$

Once a_0, a_1 are found, the original constants of the model are found as

$$b = a_1$$

$$a = e^{a_0}$$

THE END

<http://numericalmethods.eng.usf.edu>

Example 3-Transformation of data

Many patients get concerned when a test involves injection of a radioactive material. For example for scanning a gallbladder, a few drops of Technetium-99m isotope is used. Half of the Technetium-99m would be gone in about 6 hours. It, however, takes about 24 hours for the radiation levels to reach what we are exposed to in day-to-day activities. Below is given the relative intensity of radiation as a function of time.

Table. Relative intensity of radiation as a function of time

t(hr)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

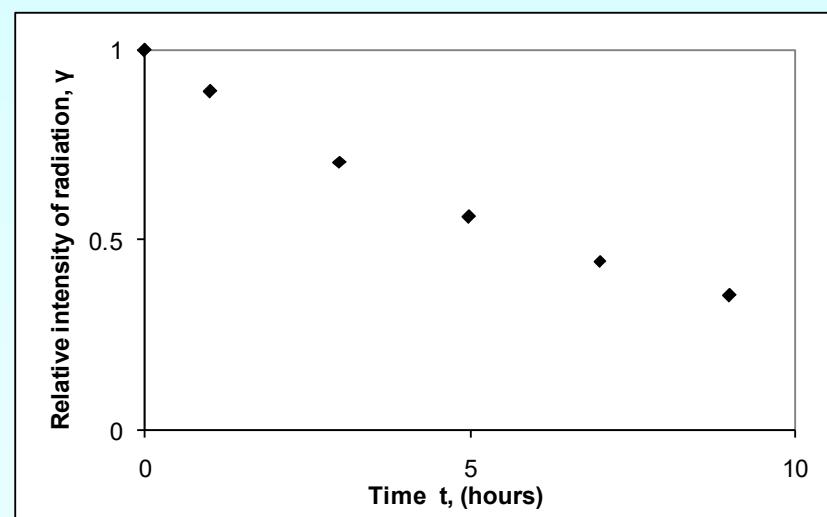


Figure. Data points of relative radiation intensity vs. time

Example 3-Transformation of data cont.

Find:

- a) The value of the regression constants A and λ
- b) The half-life of Technetium-99m
- c) Radiation intensity after 24 hours

The relative intensity is related to time by the equation

$$\gamma = Ae^{\lambda t}$$

Example 3-Transformation of data cont.

Exponential model given as,

$$\gamma = Ae^{\lambda t}$$

$$\ln(\gamma) = \ln(A) + \lambda t$$

Assuming $z = \ln \gamma$, $a_0 = \ln(A)$ and $a_1 = \lambda$ we obtain

$$z = a_0 + a_1 t$$

This is a linear relationship between z and t

Example 3-Transformation of data cont.

Using this linear relationship, we can calculate a_0, a_1 where

$$a_1 = \frac{n \sum_{i=1}^n t_i z_i - \sum_{i=1}^n t_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n t_i^2 - \left(\sum_{i=1}^n t_i \right)^2}$$

and

$$a_0 = \bar{z} - a_1 \bar{t}$$

$$\lambda = a_1$$

$$A = e^{a_0}$$

Example 3-Transformation of Data cont.

Summations for data transformation are as follows

Table. Summation data for Transformation of data

i	t_i	γ_i	model $z_i = \ln \gamma_i$	$t_i z_i$	t_i^2
1	0	1	0.00000	0.0000	0.0000
2	1	0.891	-0.11541	-0.11541	1.0000
3	3	0.708	-0.34531	-1.0359	9.0000
4	5	0.562	-0.57625	-2.8813	25.000
5	7	0.447	-0.80520	-5.6364	49.000
6	9	0.355	-1.0356	-9.3207	81.000
Σ	25.000		-2.8778	-18.990	165.00

With $n = 6$

$$\sum_{i=1}^6 t_i = 25.000$$

$$\sum_{i=1}^6 z_i = -2.8778$$

$$\sum_{i=1}^6 t_i z_i = -18.990$$

$$\sum_{i=1}^6 t_i^2 = 165.00$$

Example 3-Transformation of Data cont.

Calculating a_0, a_1

$$a_1 = \frac{6(-18.990) - (25)(-2.8778)}{6(165.00) - (25)^2} = -0.11505$$

$$a_0 = \frac{-2.8778}{6} - (-0.11505)\frac{25}{6} = -2.6150 \times 10^{-4}$$

Since

$$a_0 = \ln(A)$$

$$A = e^{a_0}$$

$$= e^{-2.6150 \times 10^{-4}} = 0.99974$$

also

$$\lambda = a_1 = -0.11505$$

Example 3-Transformation of Data cont.

Resulting model is $\gamma = 0.99974 \times e^{-0.11505t}$

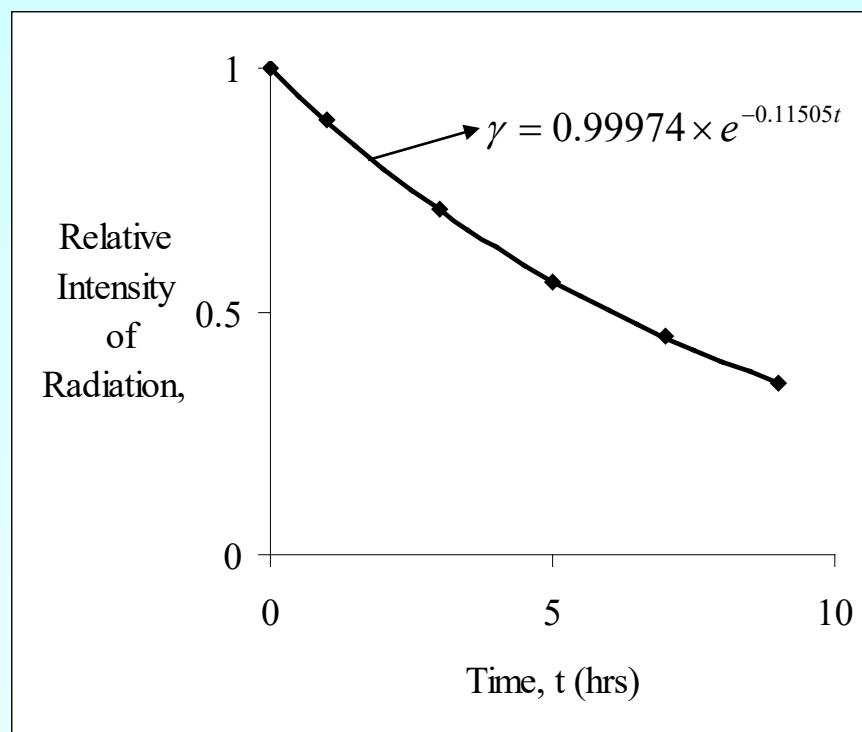


Figure. Relative intensity of radiation as a function of temperature using transformation of data model.

Example 3-Transformation of Data cont.

The regression formula is then

$$\gamma = 0.99974 \times e^{-0.11505t}$$

b) Half life of Technetium-99m is when $\gamma = \frac{1}{2} \gamma \Big|_{t=0}$

$$0.99974 \times e^{-0.11505t} = \frac{1}{2}(0.99974)e^{-0.11505(0)}$$

$$e^{-0.11508t} = 0.5$$

$$-0.11505t = \ln(0.5)$$

$$t = 6.0248 \text{ hours}$$

Example 3-Transformation of Data cont.

- c) The relative intensity of radiation after 24 hours is then

$$\begin{aligned}\gamma &= 0.99974e^{-0.11505(24)} \\ &= 0.063200\end{aligned}$$

This implies that only $\frac{6.3200 \times 10^{-2}}{0.99983} \times 100 = 6.3216\%$ of the radioactive material is left after 24 hours.

Comparison

Comparison of exponential model with and without data Transformation:

Table. Comparison for exponential model with and without data Transformation.

	With data Transformation (Example 3)	Without data Transformation (Example 1)
A	0.99974	0.99983
λ	-0.11505	-0.11508
Half-Life (hrs)	6.0248	6.0232
Relative intensity after 24 hrs.	6.3200×10^{-2}	6.3160×10^{-2}

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/nonlinear_regression.html

THE END

<http://numericalmethods.eng.usf.edu>

Trapezoidal Rule of Integration

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Trapezoidal Rule of Integration

<http://numericalmethods.eng.usf.edu>

What is Integration

Integration:

The process of measuring the area under a function plotted on a graph.

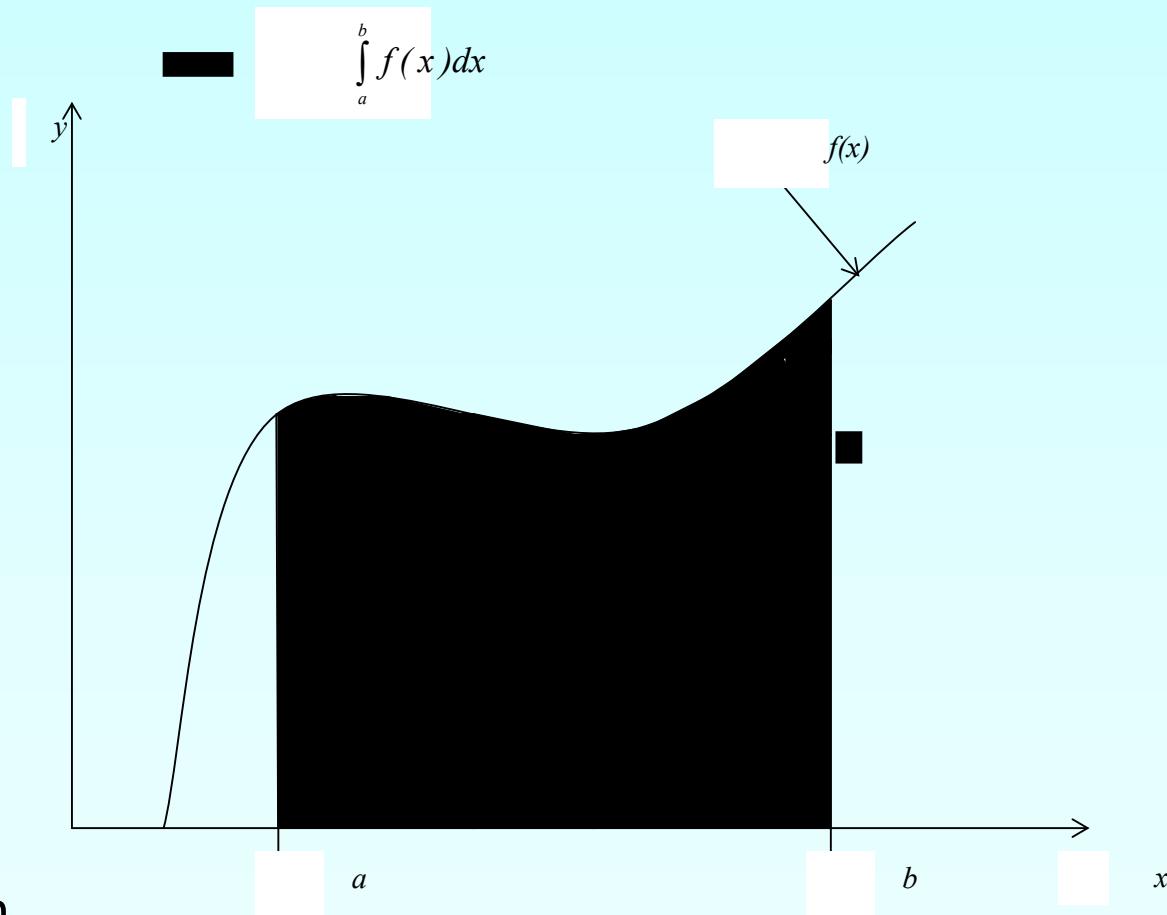
$$I = \int_a^b f(x)dx$$

Where:

$f(x)$ is the integrand

a= lower limit of integration

b= upper limit of integration



Basis of Trapezoidal Rule

Trapezoidal Rule is based on the Newton-Cotes Formula that states if one can approximate the integrand as an n^{th} order polynomial...

$$I = \int_a^b f(x)dx \quad \text{where} \quad f(x) \approx f_n(x)$$

and $f_n(x) = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n$

Basis of Trapezoidal Rule

Then the integral of that function is approximated by the integral of that n^{th} order polynomial.

$$\int_a^b f(x) \approx \int_a^b f_n(x)$$

Trapezoidal Rule assumes $n=1$, that is, the area under the linear polynomial,

$$\int_a^b f(x) dx = (b-a) \left[\frac{f(a)+f(b)}{2} \right]$$

Derivation of the Trapezoidal Rule

Method Derived From Geometry

The area under the curve is a trapezoid.
The integral

$$\begin{aligned} \int_a^b f(x)dx &\approx \text{Area of trapezoid} \\ &= \frac{1}{2}(\text{Sum of parallel sides})(\text{height}) \\ &= \frac{1}{2}(f(b) + f(a))(b - a) \\ &= (b - a) \left[\frac{f(a) + f(b)}{2} \right] \end{aligned}$$

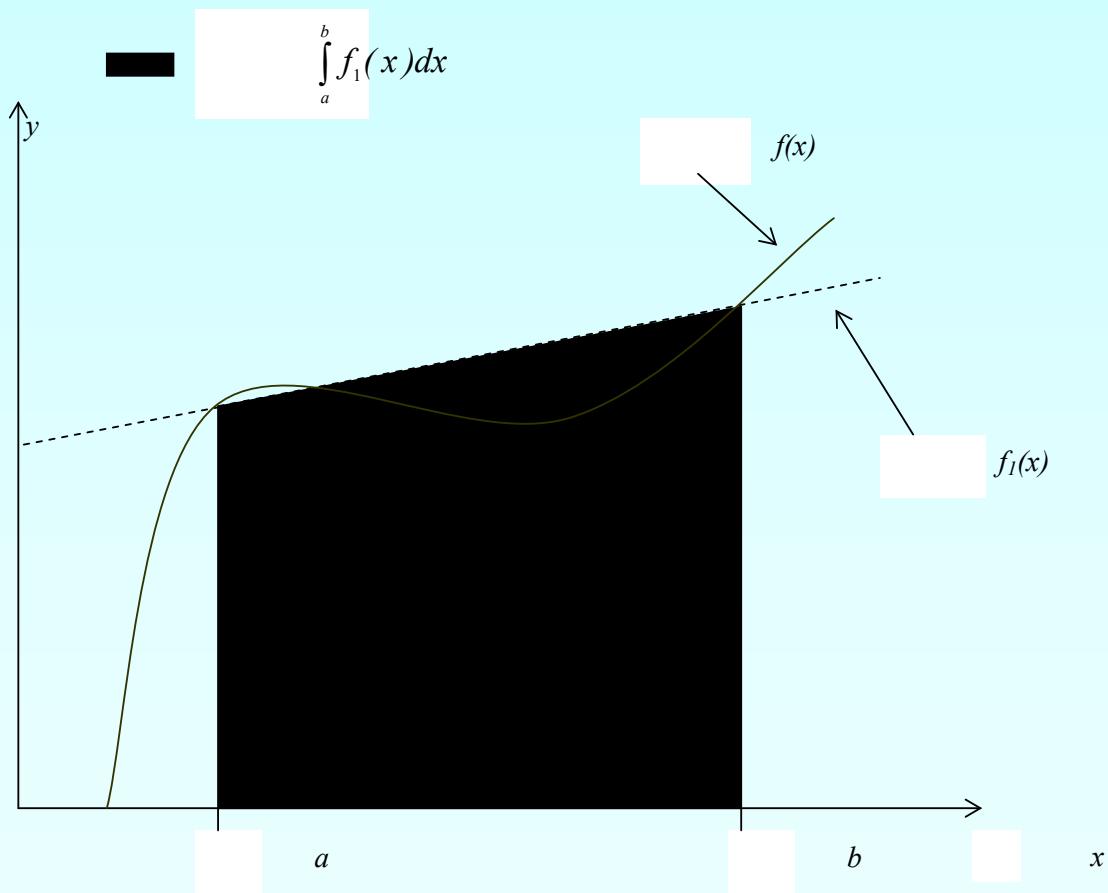


Figure 2: Geometric Representation

Example 1

The vertical distance covered by a rocket from $t=8$ to $t=30$ seconds is given by:

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use single segment Trapezoidal rule to find the distance covered.
- b) Find the true error, E_t for part (a).
- c) Find the absolute relative true error, $|\epsilon_a|$ for part (a).

Solution

a) $I \approx (b - a) \left[\frac{f(a) + f(b)}{2} \right]$

$$a = 8 \quad b = 30$$

$$f(t) = 2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$f(8) = 2000 \ln \left[\frac{140000}{140000 - 2100(8)} \right] - 9.8(8) = 177.27 \text{ m/s}$$

$$f(30) = 2000 \ln \left[\frac{140000}{140000 - 2100(30)} \right] - 9.8(30) = 901.67 \text{ m/s}$$

Solution (cont)

a)
$$I = (30 - 8) \left[\frac{177.27 + 901.67}{2} \right]$$

$$= 11868 \text{ m}$$

b) The exact value of the above integral is

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt = 11061 \text{ m}$$

Solution (cont)

b) $E_t = \text{True Value} - \text{Approximate Value}$

$$= 11061 - 11868$$
$$= -807 \text{ m}$$

c) The absolute relative true error, $|\epsilon_t|$, would be

$$|\epsilon_t| = \left| \frac{11061 - 11868}{11061} \right| \times 100 = 7.2959\%$$

Multiple Segment Trapezoidal Rule

In Example 1, the true error using single segment trapezoidal rule was large. We can divide the interval [8,30] into [8,19] and [19,30] intervals and apply Trapezoidal rule over each segment.

$$f(t) = 2000 \ln\left(\frac{140000}{140000 - 2100t}\right) - 9.8t$$

$$\int_8^{30} f(t) dt = \int_8^{19} f(t) dt + \int_{19}^{30} f(t) dt$$

$$= (19 - 8) \left[\frac{f(8) + f(19)}{2} \right] + (30 - 19) \left[\frac{f(19) + f(30)}{2} \right]$$

Multiple Segment Trapezoidal Rule

With

$$f(8) = 177.27 \text{ m/s}$$

$$f(30) = 901.67 \text{ m/s}$$

$$f(19) = 484.75 \text{ m/s}$$

Hence:

$$\int_8^{30} f(t) dt = (19 - 8) \left[\frac{177.27 + 484.75}{2} \right] + (30 - 19) \left[\frac{484.75 + 901.67}{2} \right]$$

$$= 11266 \text{ m}$$

Multiple Segment Trapezoidal Rule

The true error is:

$$\begin{aligned} E_t &= 11061 - 11266 \\ &= -205 \text{ m} \end{aligned}$$

The true error now is reduced from -807 m to -205 m.

Extending this procedure to divide the interval into equal segments to apply the Trapezoidal rule; the sum of the results obtained for each segment is the approximate value of the integral.

Multiple Segment Trapezoidal Rule

Divide into equal segments as shown in Figure 4. Then the width of each segment is:

$$h = \frac{b - a}{n}$$

The integral I is:

$$I = \int_a^b f(x)dx$$

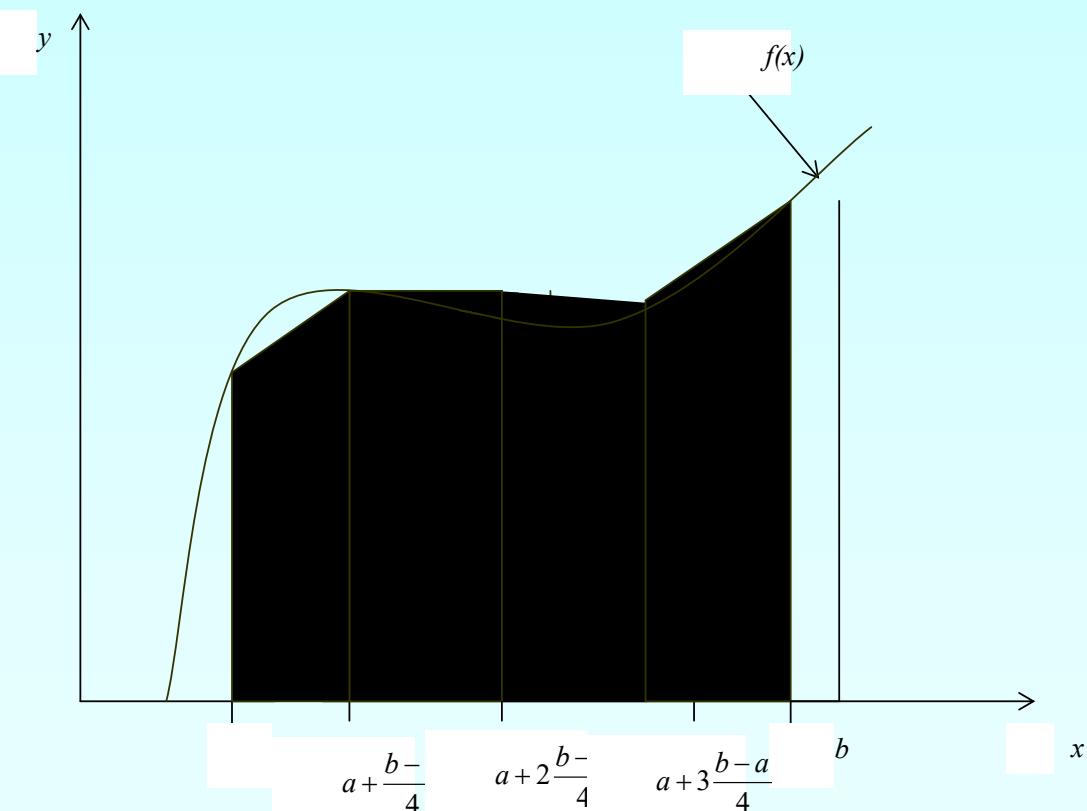


Figure 4: Multiple (n=4) Segment Trapezoidal Rule

Multiple Segment Trapezoidal Rule

The integral I can be broken into h integrals as:

$$\int_a^b f(x)dx = \int_a^{a+h} f(x)dx + \int_{a+h}^{a+2h} f(x)dx + \dots + \int_{a+(n-2)h}^{a+(n-1)h} f(x)dx + \int_{a+(n-1)h}^b f(x)dx$$

Applying Trapezoidal rule on each segment gives:

$$\int_a^b f(x)dx = \frac{b-a}{2n} \left[f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a + ih) \right\} + f(b) \right]$$

Example 2

The vertical distance covered by a rocket from $t = 8$ to $t = 30$ seconds is given by:

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- Use two-segment Trapezoidal rule to find the distance covered.
- Find the true error, E_t for part (a).
- Find the absolute relative true error, $|\epsilon_a|$ for part (a).

Solution

a) The solution using 2-segment Trapezoidal rule is

$$I = \frac{b-a}{2n} \left[f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a + ih) \right\} + f(b) \right]$$

$$n = 2 \quad a = 8 \quad b = 30$$

$$h = \frac{b-a}{n} = \frac{30-8}{2} = 11$$

Solution (cont)

Then:

$$\begin{aligned} I &= \frac{30 - 8}{2(2)} \left[f(8) + 2 \left\{ \sum_{i=1}^{2-1} f(a + ih) \right\} + f(30) \right] \\ &= \frac{22}{4} [f(8) + 2f(19) + f(30)] \\ &= \frac{22}{4} [177.27 + 2(484.75) + 901.67] \\ &= 11266 \text{ m} \end{aligned}$$

Solution (cont)

b) The exact value of the above integral is

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt = 11061 \text{ m}$$

so the true error is

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$= 11061 - 11266$$

Solution (cont)

The absolute relative true error, $|\epsilon_t|$, would be

$$|\epsilon_t| = \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100$$

$$= \left| \frac{11061 - 11266}{11061} \right| \times 100$$

$$= 1.8534\%$$

Solution (cont)

Table 1 gives the values obtained using multiple segment Trapezoidal rule for:

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

n	Value	E_t	E _t %	E _a %
1	11868	-807	7.296	---
2	11266	-205	1.853	5.343
3	11153	-91.4	0.8265	1.019
4	11113	-51.5	0.4655	0.3594
5	11094	-33.0	0.2981	0.1669
6	11084	-22.9	0.2070	0.09082
7	11078	-16.8	0.1521	0.05482
8	11074	-12.9	0.1165	0.03560

Table 1: Multiple Segment Trapezoidal Rule Values

Example 3

Use Multiple Segment Trapezoidal Rule to find
the area under the curve

$$f(x) = \frac{300x}{1 + e^x} \quad \text{from } x = 0 \quad \text{to } x = 10$$

Using two segments, we get $h = \frac{10 - 0}{2} = 5$ and

$$f(0) = \frac{300(0)}{1 + e^0} = 0 \quad f(5) = \frac{300(5)}{1 + e^5} = 10.039 \quad f(10) = \frac{300(10)}{1 + e^{10}} = 0.136$$

Solution

Then:

$$\begin{aligned} I &= \frac{b-a}{2n} \left[f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a + ih) \right\} + f(b) \right] \\ &= \frac{10-0}{2(2)} \left[f(0) + 2 \left\{ \sum_{i=1}^{2-1} f(0 + 5) \right\} + f(10) \right] \\ &= \frac{10}{4} [f(0) + 2f(5) + f(10)] = \frac{10}{4} [0 + 2(10.039) + 0.136] \\ &= 50.535 \end{aligned}$$

Solution (cont)

So what is the true value of this integral?

$$\int_0^{10} \frac{300x}{1 + e^x} dx = 246.59$$

Making the absolute relative true error:

$$|\epsilon_t| = \left| \frac{246.59 - 50.535}{246.59} \right| \times 100\%$$

$$= 79.506\%$$

Solution (cont)

Table 2: Values obtained using Multiple Segment

Trapezoidal Rule for:

$$\int_0^{10} \frac{300x}{1 + e^x} dx$$

n	Approximate Value	E_t	$ e_t $
1	0.681	245.91	99.724%
2	50.535	196.05	79.505%
4	170.61	75.978	30.812%
8	227.04	19.546	7.927%
16	241.70	4.887	1.982%
32	245.37	1.222	0.495%
64	246.28	0.305	0.124%

Error in Multiple Segment Trapezoidal Rule

The true error for a single segment Trapezoidal rule is given by:

$$E_t = \frac{(b-a)^3}{12} f''(\zeta), \quad a < \zeta < b \quad \text{where } \zeta \text{ is some point in } [a,b]$$

What is the error, then in the multiple segment Trapezoidal rule? It will be simply the sum of the errors from each segment, where the error in each segment is that of the single segment Trapezoidal rule.

The error in each segment is

$$E_1 = \frac{[(a+h)-a]^3}{12} f''(\zeta_1), \quad a < \zeta_1 < a+h$$

$$= \frac{h^3}{12} f''(\zeta_1)$$

Error in Multiple Segment Trapezoidal Rule

Similarly:

$$\begin{aligned} E_i &= \frac{[(a + ih) - (a + (i-1)h)]^3}{12} f''(\zeta_i), \quad a + (i-1)h < \zeta_i < a + ih \\ &= \frac{h^3}{12} f''(\zeta_i) \end{aligned}$$

It then follows that:

$$\begin{aligned} E_n &= \frac{[b - \{a + (n-1)h\}]^3}{12} f''(\zeta_n), \quad a + (n-1)h < \zeta_n < b \\ &= \frac{h^3}{12} f''(\zeta_n) \end{aligned}$$

Error in Multiple Segment Trapezoidal Rule

Hence the total error in multiple segment Trapezoidal rule is

$$E_t = \sum_{i=1}^n E_i = \frac{h^3}{12} \sum_{i=1}^n f''(\zeta_i) = \frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^n f''(\zeta_i)}{n}$$

The term $\frac{\sum_{i=1}^n f''(\zeta_i)}{n}$ is an approximate average value of the $f''(x)$, $a < x < b$

Hence:

$$E_t = \frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^n f''(\zeta_i)}{n}$$

Error in Multiple Segment Trapezoidal Rule

Below is the table for the integral

$$\int_{8}^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

as a function of the number of segments. You can visualize that as the number of segments are doubled, the true error gets approximately quartered.

n	Value	E_t	$ e_t \%$	$ e_a \%$
2	11266	-205	1.854	5.343
4	11113	-51.5	0.4655	0.3594
8	11074	-12.9	0.1165	0.03560
16	11065	-3.22	0.02913	0.00401

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/trapezoidal_rule.html

THE END

<http://numericalmethods.eng.usf.edu>

Simpson's 1/3rd Rule of Integration

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Simpson's 1/3rd Rule of Integration

<http://numericalmethods.eng.usf.edu>

What is Integration?

Integration

The process of measuring the area under a curve.

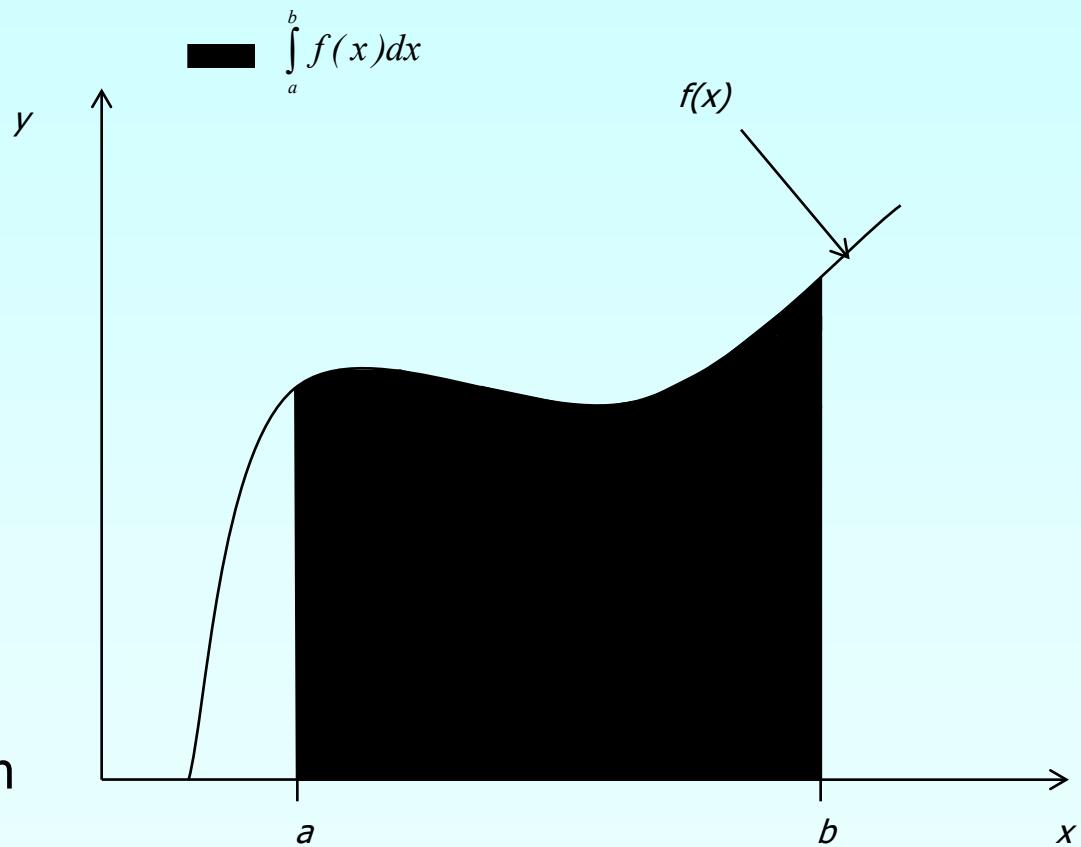
$$I = \int_a^b f(x)dx$$

Where:

$f(x)$ is the integrand

a= lower limit of integration

b= upper limit of integration



Simpson's 1/3rd Rule

Basis of Simpson's 1/3rd Rule

Trapezoidal rule was based on approximating the integrand by a first order polynomial, and then integrating the polynomial in the interval of integration. Simpson's 1/3rd rule is an extension of Trapezoidal rule where the integrand is approximated by a second order polynomial.

Hence

$$I = \int_a^b f(x)dx \approx \int_a^b f_2(x)dx$$

Where $f_2(x)$ is a second order polynomial.

$$f_2(x) = a_0 + a_1x + a_2x^2$$

Basis of Simpson's 1/3rd Rule

Choose

$$(a, f(a)), \left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right) \right), \text{ and } (b, f(b))$$

as the three points of the function to evaluate a_0 , a_1 and a_2 .

$$f(a) = f_2(a) = a_0 + a_1 a + a_2 a^2$$

$$f\left(\frac{a+b}{2}\right) = f_2\left(\frac{a+b}{2}\right) = a_0 + a_1\left(\frac{a+b}{2}\right) + a_2\left(\frac{a+b}{2}\right)^2$$

$$f(b) = f_2(b) = a_0 + a_1 b + a_2 b^2$$

Basis of Simpson's 1/3rd Rule

Solving the previous equations for a_0 , a_1 and a_2 give

$$a_0 = \frac{a^2 f(b) + abf(b) - 4abf\left(\frac{a+b}{2}\right) + abf(a) + b^2 f(a)}{a^2 - 2ab + b^2}$$
$$a_1 = -\frac{af(a) - 4af\left(\frac{a+b}{2}\right) + 3af(b) + 3bf(a) - 4bf\left(\frac{a+b}{2}\right) + bf(b)}{a^2 - 2ab + b^2}$$
$$a_2 = \frac{2\left(f(a) - 2f\left(\frac{a+b}{2}\right) + f(b)\right)}{a^2 - 2ab + b^2}$$

Basis of Simpson's 1/3rd Rule

Then

$$\begin{aligned} I &\approx \int_a^b f_2(x) dx \\ &= \int_a^b (a_0 + a_1 x + a_2 x^2) dx \\ &= \left[a_0 x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3} \right]_a^b \\ &= a_0(b-a) + a_1 \frac{b^2 - a^2}{2} + a_2 \frac{b^3 - a^3}{3} \end{aligned}$$

Basis of Simpson's 1/3rd Rule

Substituting values of a_0, a_1, a_2 give

$$\int_a^b f_2(x)dx = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

Since for Simpson's 1/3rd Rule, the interval $[a, b]$ is broken into 2 segments, the segment width

$$h = \frac{b-a}{2}$$

Basis of Simpson's 1/3rd Rule

Hence

$$\int_a^b f_2(x)dx = \frac{h}{3} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

Because the above form has 1/3 in its formula, it is called Simpson's 1/3rd Rule.

Example 1

The distance covered by a rocket from $t=8$ to $t=30$ is given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use Simpson's 1/3rd Rule to find the approximate value of x
- b) Find the true error, E_t
- c) Find the absolute relative true error, $|\epsilon_t|$

Solution

a)

$$x = \int_8^{30} f(t)dt$$

$$x = \left(\frac{b-a}{6} \right) \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

$$= \left(\frac{30-8}{6} \right) [f(8) + 4f(19) + f(30)]$$

$$= \left(\frac{22}{6} \right) [177.2667 + 4(484.7455) + 901.6740]$$

$$= 11065.72 \text{ m}$$

Solution (cont)

b) The exact value of the above integral is

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

$$= 11061.34 \text{ m}$$

True Error

$$E_t = 11061.34 - 11065.72$$

$$= -4.38 \text{ m}$$

Solution (cont)

a)c) Absolute relative true error,

$$|\epsilon_t| = \left| \frac{11061.34 - 11065.72}{11061.34} \right| \times 100\% \\ = 0.0396\%$$

Multiple Segment Simpson's 1/3rd Rule

Multiple Segment Simpson's 1/3rd Rule

Just like in multiple segment Trapezoidal Rule, one can subdivide the interval $[a, b]$ into n segments and apply Simpson's 1/3rd Rule repeatedly over every two segments. Note that n needs to be even. Divide interval $[a, b]$ into equal segments, hence the segment width

$$h = \frac{b - a}{n} \quad \int_a^b f(x) dx = \int_{x_0}^{x_n} f(x) dx$$

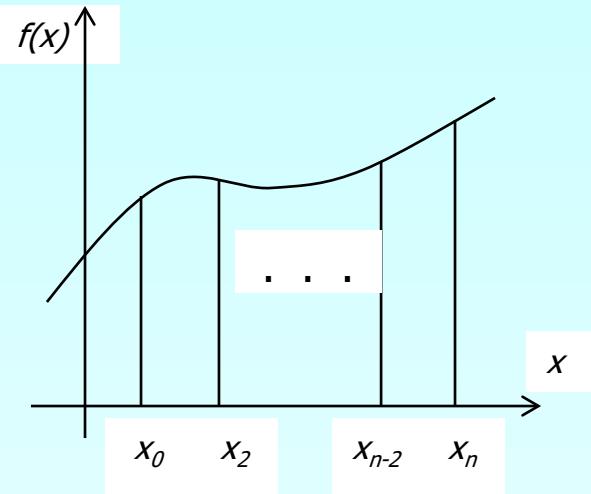
where

$$x_0 = a \quad x_n = b$$

Multiple Segment Simpson's 1/3rd Rule

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \dots$$

$$\dots + \int_{x_{n-4}}^{x_{n-2}} f(x)dx + \int_{x_{n-2}}^{x_n} f(x)dx$$



Apply Simpson's 1/3rd Rule over each interval,

$$\int_a^b f(x)dx = (x_2 - x_0) \left[\frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \right] + \dots$$

$$+ (x_4 - x_2) \left[\frac{f(x_2) + 4f(x_3) + f(x_4)}{6} \right] + \dots$$

Multiple Segment Simpson's 1/3rd Rule

$$\dots + (x_{n-2} - x_{n-4}) \left[\frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6} \right] + \dots$$

$$+ (x_n - x_{n-2}) \left[\frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6} \right]$$

Since

$$x_i - x_{i-2} = 2h \quad i = 2, 4, \dots, n$$

Multiple Segment Simpson's 1/3rd Rule

Then

$$\int_a^b f(x)dx = 2h \left[\frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \right] + \dots$$
$$+ 2h \left[\frac{f(x_2) + 4f(x_3) + f(x_4)}{6} \right] + \dots$$
$$+ 2h \left[\frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6} \right] + \dots$$
$$+ 2h \left[\frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6} \right]$$

Multiple Segment Simpson's 1/3rd Rule

$$\int_a^b f(x)dx = \frac{h}{3} [f(x_0) + 4\{f(x_1) + f(x_3) + \dots + f(x_{n-1})\} + \dots]$$

$$\dots + 2\{f(x_2) + f(x_4) + \dots + f(x_{n-2})\} + f(x_n)\}]$$

$$= \frac{h}{3} \left[f(x_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n) \right]$$

$$= \frac{b-a}{3n} \left[f(x_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n) \right]$$

Example 2

Use 4-segment Simpson's 1/3rd Rule to approximate the distance covered by a rocket from $t = 8$ to $t = 30$ as given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use four segment Simpson's 1/3rd Rule to find the approximate value of x .
- b) Find the true error, E_t for part (a).
- c) Find the absolute relative true error, $|e_a|$ for part (a).

Solution

a) Using n segment Simpson's 1/3rd Rule,

$$h = \frac{30 - 8}{4} = 5.5$$

So

$$f(t_0) = f(8)$$

$$f(t_1) = f(8 + 5.5) = f(13.5)$$

$$f(t_2) = f(13.5 + 5.5) = f(19)$$

$$f(t_3) = f(19 + 5.5) = f(24.5)$$

$$f(t_4) = f(30)$$

Solution (cont.)

$$x = \frac{b-a}{3n} \left[f(t_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(t_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(t_i) + f(t_n) \right]$$

$$= \frac{30-8}{3(4)} \left[f(8) + 4 \sum_{\substack{i=1 \\ i=odd}}^3 f(t_i) + 2 \sum_{\substack{i=2 \\ i=even}}^2 f(t_i) + f(30) \right]$$

$$= \frac{22}{12} [f(8) + 4f(t_1) + 4f(t_3) + 2f(t_2) + f(30)]$$

Solution (cont.)

cont.

$$= \frac{11}{6} [f(8) + 4f(13.5) + 4f(24.5) + 2f(19) + f(30)]$$

$$= \frac{11}{6} [177.2667 + 4(320.2469) + 4(676.0501) + 2(484.7455) + 901.6740]$$

$$= 11061.64 \text{ m}$$

Solution (cont.)

- b) In this case, the true error is

$$E_t = 11061.34 - 11061.64 = -0.30 \text{ m}$$

- c) The absolute relative true error

$$|\epsilon_t| = \left| \frac{11061.34 - 11061.64}{11061.34} \right| \times 100\%$$

$$= 0.0027\%$$

Solution (cont.)

Table 1: Values of Simpson's 1/3rd Rule for Example 2 with multiple segments

n	Approximate Value	E_t	$ E_t $
2	11065.72	4.38	0.0396%
4	11061.64	0.30	0.0027%
6	11061.40	0.06	0.0005%
8	11061.35	0.01	0.0001%
10	11061.34	0.00	0.0000%

Error in the Multiple Segment Simpson's 1/3rd Rule

The true error in a single application of Simpson's 1/3rd Rule is given as

$$E_t = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta), \quad a < \zeta < b$$

In Multiple Segment Simpson's 1/3rd Rule, the error is the sum of the errors in each application of Simpson's 1/3rd Rule. The error in n segment Simpson's 1/3rd Rule is given by

$$E_1 = -\frac{(x_2 - x_0)^5}{2880} f^{(4)}(\zeta_1) = -\frac{h^5}{90} f^{(4)}(\zeta_1), \quad x_0 < \zeta_1 < x_2$$

$$E_2 = -\frac{(x_4 - x_2)^5}{2880} f^{(4)}(\zeta_2) = -\frac{h^5}{90} f^{(4)}(\zeta_2), \quad x_2 < \zeta_2 < x_4$$

Error in the Multiple Segment Simpson's 1/3rd Rule

$$E_i = -\frac{(x_{2i} - x_{2(i-1)})^5}{2880} f^{(4)}(\zeta_i) = -\frac{h^5}{90} f^{(4)}(\zeta_i), \quad x_{2(i-1)} < \zeta_i < x_{2i}$$

.

.

.

$$E_{\frac{n}{2}-1} = -\frac{(x_{n-2} - x_{n-4})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right) = -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right), \quad x_{n-4} < \zeta_{\frac{n}{2}-1} < x_{n-2}$$

$$E_{\frac{n}{2}} = -\frac{(x_n - x_{n-2})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}}\right) = -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}}\right), \quad x_{n-2} < \zeta_{\frac{n}{2}} < x_n$$

Error in the Multiple Segment Simpson's 1/3rd Rule

Hence, the total error in Multiple Segment Simpson's 1/3rd Rule is

$$E_t = \sum_{i=1}^{\frac{n}{2}} E_i = -\frac{h^5}{90} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) = -\frac{(b-a)^5}{90n^5} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)$$

$$= -\frac{(b-a)^5}{90n^4} \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{n}$$

Error in the Multiple Segment Simpson's 1/3rd Rule

The term $\frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{n}$ is an approximate average value of $f^{(4)}(x)$, $a < x < b$

Hence

$$E_t = -\frac{(b-a)^5}{90n^4} \bar{f}^{(4)}$$

where

$$\bar{f}^{(4)} = \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{n}$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/simpsons_13rd_rule.html

THE END

<http://numericalmethods.eng.usf.edu>

Romberg Rule of Integration

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Romberg Rule of Integration

<http://numericalmethods.eng.usf.edu>

Basis of Romberg Rule

Integration

The process of measuring the area under a curve.

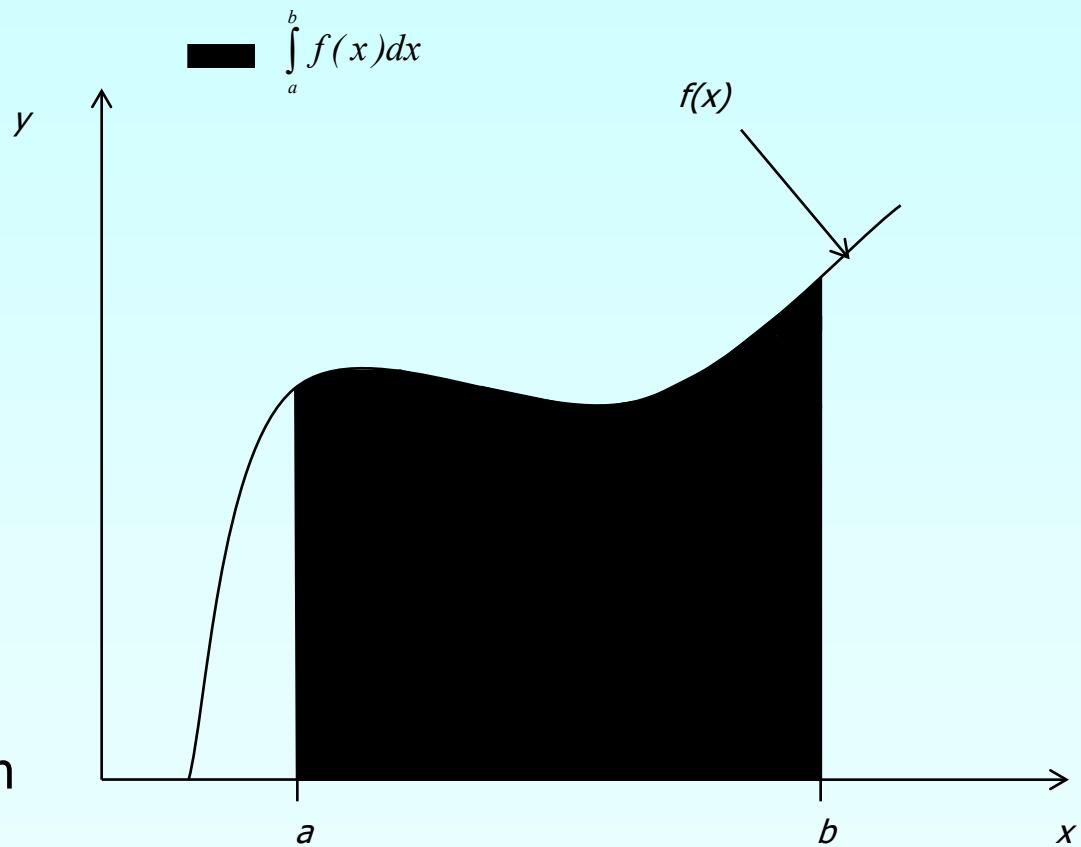
$$I = \int_a^b f(x)dx$$

Where:

$f(x)$ is the integrand

a= lower limit of integration

b= upper limit of integration



What is The Romberg Rule?

Romberg Integration is an extrapolation formula of the Trapezoidal Rule for integration. It provides a better approximation of the integral by reducing the True Error.

Error in Multiple Segment Trapezoidal Rule

The true error in a multiple segment Trapezoidal Rule with n segments for an integral

$$I = \int_a^b f(x)dx$$

Is given by

$$E_t = \frac{(b-a)^3}{12n^2} \sum_{i=1}^n f''(\xi_i)$$

where for each i , ξ_i is a point somewhere in the domain, $[a + (i-1)h, a + ih]$.

Error in Multiple Segment Trapezoidal Rule

The term $\frac{\sum_{i=1}^n f''(\xi_i)}{n}$ can be viewed as an approximate average value of $f''(x)$ in $[a,b]$.

This leads us to say that the true error, E_t previously defined can be approximated as

$$E_t \approx \alpha \frac{1}{n^2}$$

Error in Multiple Segment Trapezoidal Rule

Table 1 shows the results obtained for the integral using multiple segment Trapezoidal rule for

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

n	Value	E _t	e _t %	e _a %
1	11868	807	7.296	---
2	11266	205	1.854	5.343
3	11153	91.4	0.8265	1.019
4	11113	51.5	0.4655	0.3594
5	11094	33.0	0.2981	0.1669
6	11084	22.9	0.2070	0.09082
7	11078	16.8	0.1521	0.05482
8	11074	12.9	0.1165	0.03560

Table 1: Multiple Segment Trapezoidal Rule Values

Error in Multiple Segment Trapezoidal Rule

The true error gets approximately quartered as the number of segments is doubled. This information is used to get a better approximation of the integral, and is the basis of Richardson's extrapolation.

Richardson's Extrapolation for Trapezoidal Rule

The true error, E_t in the n -segment Trapezoidal rule is estimated as

$$E_t \approx \frac{C}{n^2}$$

where C is an *approximate constant* of proportionality. Since

$$E_t = TV - I_n$$

Where TV = true value and I_n = approx. value

Richardson's Extrapolation for Trapezoidal Rule

From the previous development, it can be shown that

$$\frac{C}{(2n)^2} \approx TV - I_{2n}$$

when the segment size is doubled and that

$$TV \approx I_{2n} + \frac{I_{2n} - I_n}{3}$$

which is Richardson's Extrapolation.

Example 1

The vertical distance covered by a rocket from 8 to 30 seconds is given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use Richardson's rule to find the distance covered.
Use the 2-segment and 4-segment Trapezoidal rule results given in Table 1.
- b) Find the true error, E_t for part (a).
- c) Find the absolute relative true error, $|\epsilon_a|$ for part (a).

Solution

a) $I_2 = 11266m$ $I_4 = 11113m$

Using Richardson's extrapolation formula
for Trapezoidal rule

$$TV \approx I_{2n} + \frac{I_{2n} - I_n}{3} \quad \text{and choosing } n=2,$$

$$TV \approx I_4 + \frac{I_4 - I_2}{3} = 11113 + \frac{11113 - 11266}{3}$$

$$= 11062m$$

Solution (cont.)

b) The exact value of the above integral is

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

$$= 11061 \text{ m}$$

Hence

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$= 11061 - 11062$$

$$= -1 \text{ m}$$

Solution (cont.)

c) The absolute relative true error $|\epsilon_t|$ would then be

$$|\epsilon_t| = \left| \frac{11061 - 11062}{11061} \right| \times 100$$

$$= 0.00904\%$$

Table 2 shows the Richardson's extrapolation results using 1, 2, 4, 8 segments. Results are compared with those of Trapezoidal rule.

Solution (cont.)

Table 2: The values obtained using Richardson's extrapolation formula for Trapezoidal rule for

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

n	Trapezoidal Rule	$ \in_t $ for Trapezoidal Rule	Richardson's Extrapolation	$ \in_t $ for Richardson's Extrapolation
1	11868	7.296	--	--
2	11266	1.854	11065	0.03616
4	11113	0.4655	11062	0.009041
8	11074	0.1165	11061	0.0000

Table 2: Richardson's Extrapolation Values

Romberg Integration

Romberg integration is same as Richardson's extrapolation formula as given previously. However, Romberg used a recursive algorithm for the extrapolation. Recall

$$TV \approx I_{2n} + \frac{I_{2n} - I_n}{3}$$

This can alternately be written as

$$(I_{2n})_R = I_{2n} + \frac{I_{2n} - I_n}{3} = I_{2n} + \frac{I_{2n} - I_n}{4^{2-1} - 1}$$

Romberg Integration

Note that the variable T is replaced by $(I_{2n})_R$ as the value obtained using Richardson's extrapolation formula. Note also that the sign \approx is replaced by = sign. Hence the estimate of the true value now is

$$TV \approx (I_{2n})_R + Ch^4$$

Where Ch^4 is an approximation of the true error.

Romberg Integration

Determine another integral value with further halving the step size (doubling the number of segments),

$$(I_{4n})_R = I_{4n} + \frac{I_{4n} - I_{2n}}{3}$$

It follows from the two previous expressions that the true value TV can be written as

$$TV \approx (I_{4n})_R + \frac{(I_{4n})_R - (I_{2n})_R}{15}$$

$$= I_{4n} + \frac{(I_{4n})_R - (I_{2n})_R}{4^{3-1} - 1}$$

Romberg Integration

A general expression for Romberg integration can be written as

$$I_{k,j} = I_{k-1,j+1} + \frac{I_{k-1,j+1} - I_{k-1,j}}{4^{k-1} - 1}, k \geq 2$$

The index k represents the order of extrapolation. $k=1$ represents the values obtained from the regular Trapezoidal rule, $k=2$ represents values obtained using the true estimate as $O(h^2)$. The index j represents the more and less accurate estimate of the integral.

Example 2

The vertical distance covered by a rocket from $t = 8$ to $t = 30$ seconds is given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

Use Romberg's rule to find the distance covered. Use the 1, 2, 4, and 8-segment Trapezoidal rule results as given in the Table 1.

Solution

From Table 1, the needed values from original Trapezoidal rule are

$$I_{1,1} = 11868$$

$$I_{1,2} = 11266$$

$$I_{1,3} = 11113$$

$$I_{1,4} = 11074$$

where the above four values correspond to using 1, 2, 4 and 8 segment Trapezoidal rule, respectively.

Solution (cont.)

To get the first order extrapolation values,

$$\begin{aligned} I_{2,1} &= I_{1,2} + \frac{I_{1,2} - I_{1,1}}{3} \\ &= 11266 + \frac{11266 - 11868}{3} \\ &= 11065 \end{aligned}$$

Similarly,

$$\begin{aligned} I_{2,2} &= I_{1,3} + \frac{I_{1,3} - I_{1,2}}{3} \\ &= 11113 + \frac{11113 - 11266}{3} \\ &= 11062 \end{aligned}$$

$$\begin{aligned} I_{2,3} &= I_{1,4} + \frac{I_{1,4} - I_{1,3}}{3} \\ &= 11074 + \frac{11074 - 11113}{3} \\ &= 11061 \end{aligned}$$

Solution (cont.)

For the second order extrapolation values,

$$\begin{aligned} I_{3,1} &= I_{2,2} + \frac{I_{2,2} - I_{2,1}}{15} \\ &= 11062 + \frac{11062 - 11065}{15} \\ &= 11062 \end{aligned}$$

Similarly,

$$\begin{aligned} I_{3,2} &= I_{2,3} + \frac{I_{2,3} - I_{2,2}}{15} \\ &= 11061 + \frac{11061 - 11062}{15} \\ &= 11061 \end{aligned}$$

Solution (cont.)

For the third order extrapolation values,

$$\begin{aligned} I_{4,1} &= I_{3,2} + \frac{I_{3,2} - I_{3,1}}{63} \\ &= 11061 + \frac{11061 - 11062}{63} \\ &= 11061m \end{aligned}$$

Table 3 shows these increased correct values in a tree graph.

Solution (cont.)

Table 3: Improved estimates of the integral value using Romberg Integration

		<i>First Order</i>	<i>Second Order</i>	<i>Third Order</i>
<i>1-segment</i>	11868			
<i>2-segment</i>	1126	11065		
<i>4-segment</i>	11113	11062	11062	11061
<i>8-segment</i>	11074	11061	11061	11061

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/romberg_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Gauss Quadrature Rule of Integration

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Gauss Quadrature Rule of Integration

<http://numericalmethods.eng.usf.edu>

What is Integration?

Integration

The process of measuring the area under a curve.

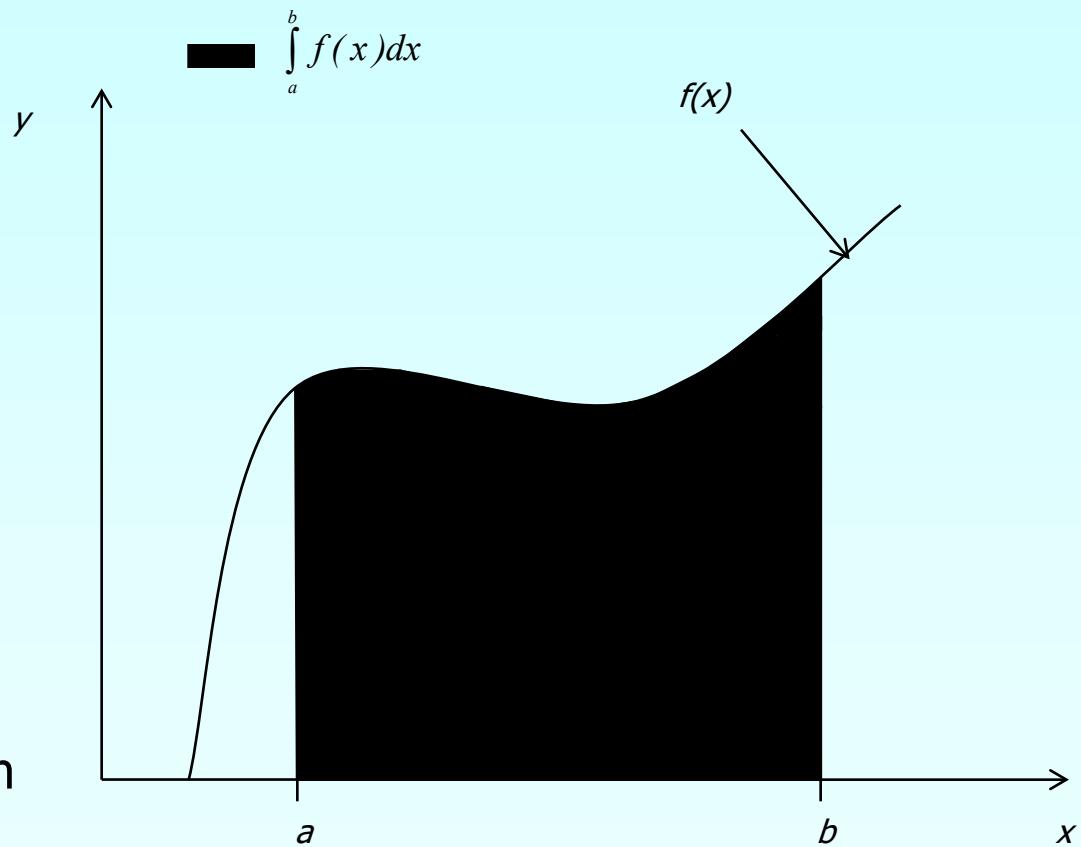
$$I = \int_a^b f(x)dx$$

Where:

$f(x)$ is the integrand

a= lower limit of integration

b= upper limit of integration



Two-Point Gaussian Quadrature Rule

Basis of the Gaussian Quadrature Rule

Previously, the Trapezoidal Rule was developed by the method of undetermined coefficients. The result of that development is summarized below.

$$\int_a^b f(x)dx \approx c_1f(a) + c_2f(b)$$
$$= \frac{b-a}{2}f(a) + \frac{b-a}{2}f(b)$$

Basis of the Gaussian Quadrature Rule

The two-point Gauss Quadrature Rule is an extension of the Trapezoidal Rule approximation where the arguments of the function are not predetermined as a and b but as unknowns x_1 and x_2 . In the two-point Gauss Quadrature Rule, the integral is approximated as

$$I = \int_a^b f(x)dx \approx c_1 f(x_1) + c_2 f(x_2)$$

Basis of the Gaussian Quadrature Rule

The four unknowns x_1 , x_2 , c_1 and c_2 are found by assuming that the formula gives exact results for integrating a general third order polynomial, $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$.

Hence

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b \left(a_0 + a_1x + a_2x^2 + a_3x^3\right)dx \\ &= \left[a_0x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3} + a_3 \frac{x^4}{4} \right]_a^b \\ &= a_0(b-a) + a_1\left(\frac{b^2 - a^2}{2}\right) + a_2\left(\frac{b^3 - a^3}{3}\right) + a_3\left(\frac{b^4 - a^4}{4}\right)\end{aligned}$$

Basis of the Gaussian Quadrature Rule

It follows that

$$\int_a^b f(x)dx = c_1(a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3) + c_2(a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3)$$

Equating Equations the two previous two expressions yield

$$\begin{aligned} & a_0(b-a) + a_1\left(\frac{b^2 - a^2}{2}\right) + a_2\left(\frac{b^3 - a^3}{3}\right) + a_3\left(\frac{b^4 - a^4}{4}\right) \\ &= c_1(a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3) + c_2(a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3) \\ &= a_0(c_1 + c_2) + a_1(c_1x_1 + c_2x_2) + a_2(c_1x_1^2 + c_2x_2^2) + a_3(c_1x_1^3 + c_2x_2^3) \end{aligned}$$

Basis of the Gaussian Quadrature Rule

Since the constants a_0, a_1, a_2, a_3 are arbitrary

$$b - a = c_1 + c_2$$

$$\frac{b^2 - a^2}{2} = c_1 x_1 + c_2 x_2$$

$$\frac{b^3 - a^3}{3} = c_1 {x_1}^2 + c_2 {x_2}^2$$

$$\frac{b^4 - a^4}{4} = c_1 {x_1}^3 + c_2 {x_2}^3$$

Basis of Gauss Quadrature

The previous four simultaneous nonlinear Equations have only one acceptable solution,

$$x_1 = \left(\frac{b-a}{2} \right) \left(-\frac{1}{\sqrt{3}} \right) + \frac{b+a}{2}$$

$$x_2 = \left(\frac{b-a}{2} \right) \left(\frac{1}{\sqrt{3}} \right) + \frac{b+a}{2}$$

$$c_1 = \frac{b-a}{2}$$

$$c_2 = \frac{b-a}{2}$$

Basis of Gauss Quadrature

Hence Two-Point Gaussian Quadrature Rule

$$\int_a^b f(x)dx \approx c_1 f(x_1) + c_2 f(x_2)$$
$$= \frac{b-a}{2} f\left(\frac{b-a}{2}\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) + \frac{b-a}{2} f\left(\frac{b-a}{2}\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right)$$

Higher Point Gaussian Quadrature Formulas

Higher Point Gaussian Quadrature Formulas

$$\int_a^b f(x)dx \approx c_1f(x_1) + c_2f(x_2) + c_3f(x_3)$$

is called the three-point Gauss Quadrature Rule.

The coefficients c_1 , c_2 , and c_3 , and the functional arguments x_1 , x_2 , and x_3 are calculated by assuming the formula gives exact expressions for integrating a fifth order polynomial

$$\int_a^b (a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5) dx$$

General n-point rules would approximate the integral

$$\int_a^b f(x)dx \approx c_1f(x_1) + c_2f(x_2) + \dots + c_nf(x_n)$$

Arguments and Weighing Factors for n-point Gauss Quadrature Formulas

In handbooks, coefficients and arguments given for n-point Gauss Quadrature Rule are given for integrals

$$\int_{-1}^1 g(x)dx \approx \sum_{i=1}^n c_i g(x_i)$$

as shown in Table 1.

Table 1: Weighting factors c and function arguments x used in Gauss Quadrature Formulas.

Points	Weighting Factors	Function Arguments
2	$c_1 = 1.000000000$ $c_2 = 1.000000000$	$x_1 = -0.577350269$ $x_2 = 0.577350269$
3	$c_1 = 0.555555556$ $c_2 = 0.888888889$ $c_3 = 0.555555556$	$x_1 = -0.774596669$ $x_2 = 0.000000000$ $x_3 = 0.774596669$
4	$c_1 = 0.347854845$ $c_2 = 0.652145155$ $c_3 = 0.652145155$ $c_4 = 0.347854845$	$x_1 = -0.861136312$ $x_2 = -0.339981044$ $x_3 = 0.339981044$ $x_4 = 0.861136312$

Arguments and Weighing Factors for n-point Gauss Quadrature Formulas

Table 1 (cont.) : Weighting factors c and function arguments x used in Gauss Quadrature Formulas.

Points	Weighting Factors	Function Arguments
5	$c_1 = 0.236926885$ $c_2 = 0.478628670$ $c_3 = 0.568888889$ $c_4 = 0.478628670$ $c_5 = 0.236926885$	$x_1 = -0.906179846$ $x_2 = -0.538469310$ $x_3 = 0.000000000$ $x_4 = 0.538469310$ $x_5 = 0.906179846$
6	$c_1 = 0.171324492$ $c_2 = 0.360761573$ $c_3 = 0.467913935$ $c_4 = 0.467913935$ $c_5 = 0.360761573$ $c_6 = 0.171324492$	$x_1 = -0.932469514$ $x_2 = -0.661209386$ $x_3 = -0.2386191860$ $x_4 = 0.2386191860$ $x_5 = 0.661209386$ $x_6 = 0.932469514$

Arguments and Weighing Factors for n-point Gauss Quadrature Formulas

So if the table is given for $\int_a^b g(x)dx$ integrals, how does one solve $\int_a^b f(x)dx$? The answer lies in that any integral with limits of $[a, b]$ can be converted into an integral with limits $[-1, 1]$. Let

$$x = mt + c$$

If $x = a$, then $t = -1$

Such that:

If $x = b$, then $t = 1$

$$m = \frac{b-a}{2}$$

Arguments and Weighing Factors for n-point Gauss Quadrature Formulas

Then

$$c = \frac{b+a}{2} \quad \text{Hence}$$

$$x = \frac{b-a}{2}t + \frac{b+a}{2} \quad dx = \frac{b-a}{2}dt$$

Substituting our values of x , and dx into the integral gives us

$$\int_a^b f(x)dx = \int_{-1}^1 f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right) \frac{b-a}{2} dt$$

Example 1

For an integral $\int_a^b f(x)dx$, derive the one-point Gaussian Quadrature Rule.

Solution

The one-point Gaussian Quadrature Rule is

$$\int_a^b f(x)dx \approx c_1 f(x_1)$$

Solution

The two unknowns x_1 , and c_1 are found by assuming that the formula gives exact results for integrating a general first order polynomial,

$$f(x) = a_0 + a_1 x.$$

$$\begin{aligned}\int_a^b f(x) dx &= \int_a^b (a_0 + a_1 x) dx \\ &= \left[a_0 x + a_1 \frac{x^2}{2} \right]_a^b \\ &= a_0(b-a) + a_1 \left(\frac{b^2 - a^2}{2} \right)\end{aligned}$$

Solution

It follows that

$$\int_a^b f(x)dx = c_1(a_0 + a_1x_1)$$

Equating Equations, the two previous two expressions yield

$$a_0(b-a) + a_1\left(\frac{b^2 - a^2}{2}\right) = c_1(a_0 + a_1x_1) = a_0(c_1) + a_1(c_1x_1)$$

Basis of the Gaussian Quadrature Rule

Since the constants a_0 , and a_1 are arbitrary

$$b - a = c_1$$

$$\frac{b^2 - a^2}{2} = c_1 x_1$$

giving

$$c_1 = b - a$$

$$x_1 = \frac{b + a}{2}$$

Solution

Hence One-Point Gaussian Quadrature Rule

$$\int_a^b f(x)dx \approx c_1 f(x_1) = (b-a) f\left(\frac{b+a}{2}\right)$$

Example 2

- a) Use two-point Gauss Quadrature Rule to approximate the distance covered by a rocket from $t=8$ to $t=30$ as given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- b) Find the true error, E_t for part (a).
- c) Also, find the absolute relative true error, $|E_a|$ for part (a).

Solution

First, change the limits of integration from [8,30] to [-1,1]
by previous relations as follows

$$\begin{aligned}\int_8^{30} f(t) dt &= \frac{30-8}{2} \int_{-1}^1 f\left(\frac{30-8}{2}x + \frac{30+8}{2}\right) dx \\ &= 11 \int_{-1}^1 f(11x + 19) dx\end{aligned}$$

Solution (cont)

Next, get weighting factors and function argument values from Table 1 for the two point rule,

$$c_1 = 1.000000000$$

$$x_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$x_2 = 0.577350269$$

Solution (cont.)

Now we can use the Gauss Quadrature formula

$$\begin{aligned} 11 \int_{-1}^1 f(11x + 19) dx &\approx 11c_1 f(11x_1 + 19) + 11c_2 f(11x_2 + 19) \\ &= 11f(11(-0.5773503) + 19) + 11f(11(0.5773503) + 19) \\ &= 11f(12.64915) + 11f(25.35085) \\ &= 11(296.8317) + 11(708.4811) \\ &= 11058.44 \text{ m} \end{aligned}$$

Solution (cont)

since

$$f(12.64915) = 2000 \ln \left[\frac{140000}{140000 - 2100(12.64915)} \right] - 9.8(12.64915)$$
$$= 296.8317$$

$$f(25.35085) = 2000 \ln \left[\frac{140000}{140000 - 2100(25.35085)} \right] - 9.8(25.35085)$$
$$= 708.4811$$

Solution (cont)

b) The true error, E_t , is

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$= 11061.34 - 11058.44$$

$$= 2.9000 \text{ m}$$

c) The absolute relative true error, $|\epsilon_t|$, is (Exact value = 11061.34m)

$$|\epsilon_t| = \left| \frac{11061.34 - 11058.44}{11061.34} \right| \times 100\%$$

$$= 0.0262\%$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/gauss_quadrature.html

THE END

<http://numericalmethods.eng.usf.edu>

Euler Method

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Euler Method

<http://numericalmethods.eng.usf.edu>

Euler's Method

$$\frac{dy}{dx} = f(x, y), y(0) = y_0$$

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}}$$

$$= \frac{y_1 - y_0}{x_1 - x_0}$$

$$= f(x_0, y_0)$$

$$\begin{aligned} y_1 &= y_0 + f(x_0, y_0)(x_1 - x_0) \\ &= y_0 + f(x_0, y_0)h \end{aligned}$$

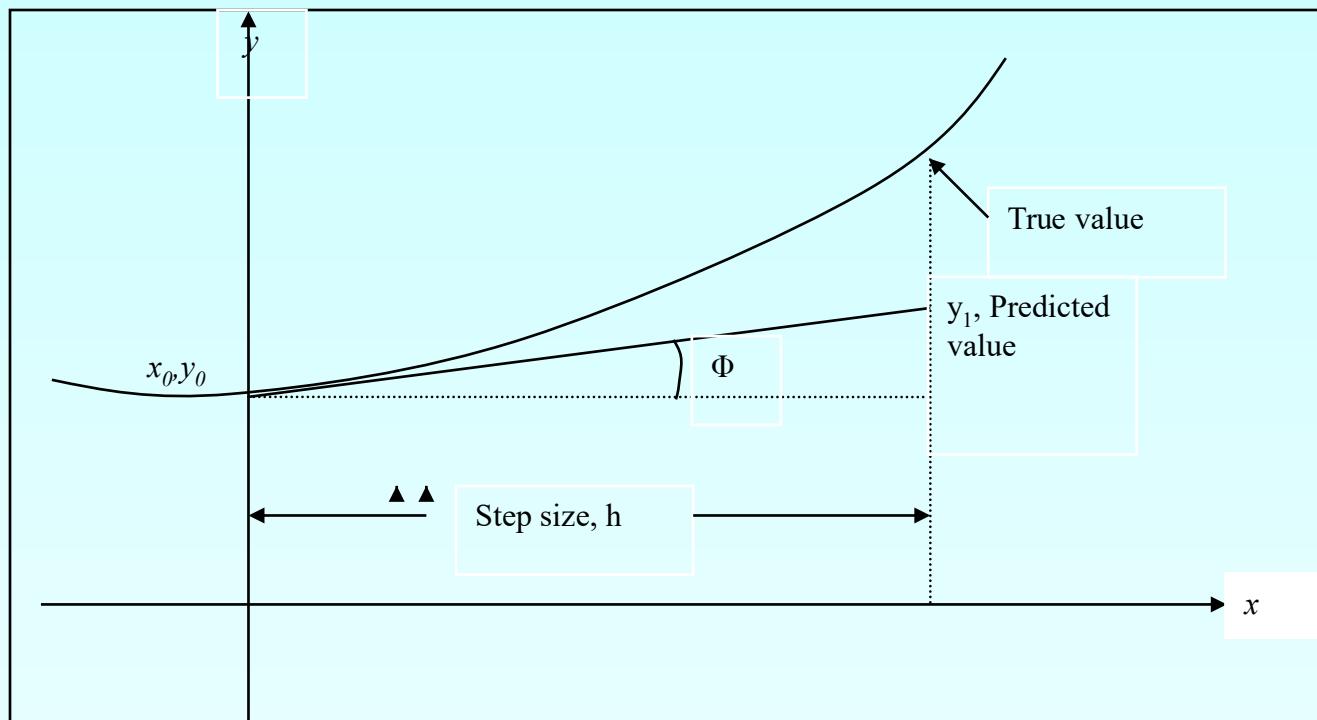


Figure 1 Graphical interpretation of the first step of Euler's method

Euler's Method

$$y_{i+1} = y_i + f(x_i, y_i)h$$

$$h = x_{i+1} - x_i$$

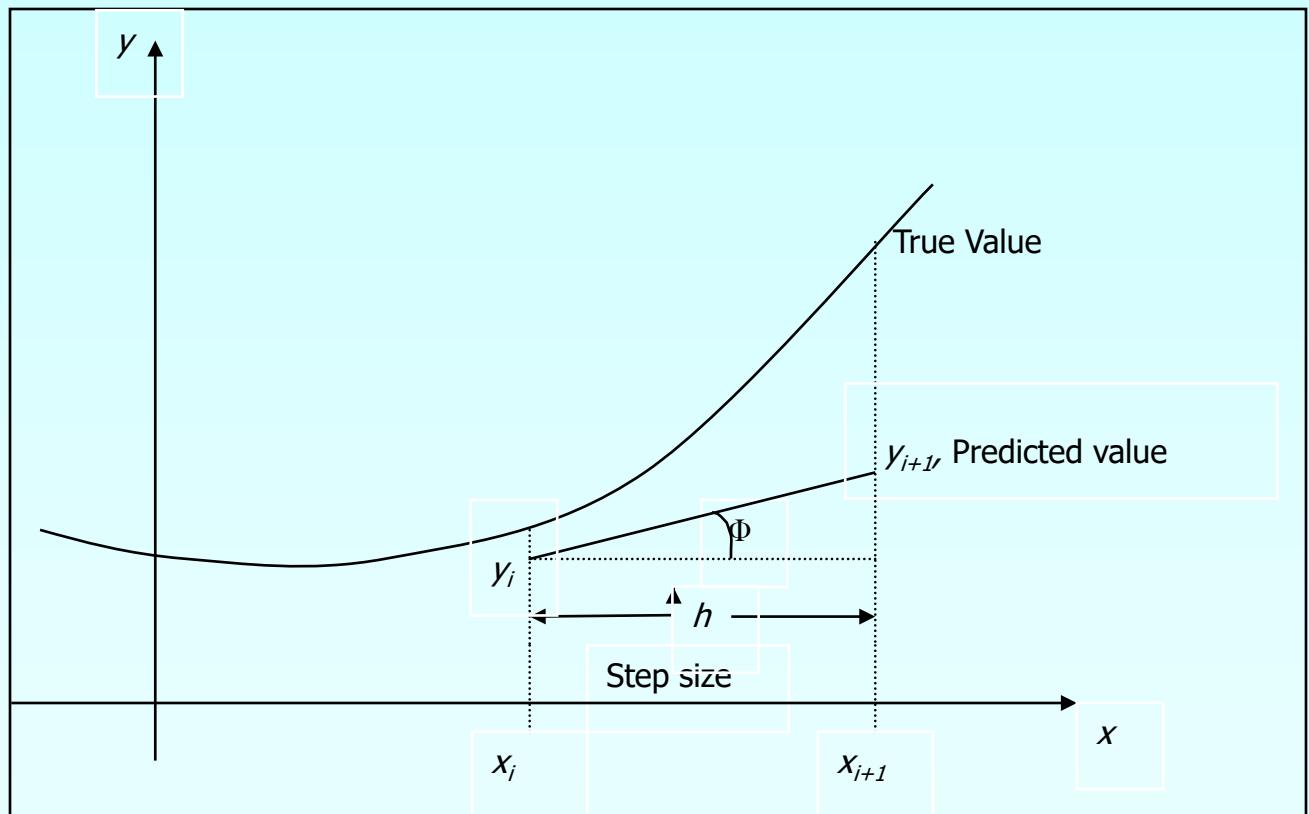


Figure 2. General graphical interpretation of Euler's method

How to write Ordinary Differential Equation

How does one write a first order differential equation in the form of

$$\frac{dy}{dx} = f(x, y)$$

Example

$$\frac{dy}{dx} + 2y = 1.3e^{-x}, y(0) = 5$$

is rewritten as

$$\frac{dy}{dx} = 1.3e^{-x} - 2y, y(0) = 5$$

In this case

$$f(x, y) = 1.3e^{-x} - 2y$$

Example

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8), \theta(0) = 1200K$$

Find the temperature at $t = 480$ seconds using Euler's method. Assume a step size of $h = 240$ seconds.

Solution

Step 1:

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

$$f(t, \theta) = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

$$\theta_{i+1} = \theta_i + f(t_i, \theta_i)h$$

$$\theta_1 = \theta_0 + f(t_0, \theta_0)h$$

$$= 1200 + f(0, 1200)240$$

$$= 1200 + (-2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8))240$$

$$= 1200 + (-4.5579)240$$

$$= 106.09K$$

θ_1 is the approximate temperature at $t = t_1 = t_0 + h = 0 + 240 = 240$

$$\theta(240) \approx \theta_1 = 106.09K$$

Solution Cont

Step 2: For $i = 1, t_1 = 240, \theta_1 = 106.09$

$$\begin{aligned}\theta_2 &= \theta_1 + f(t_1, \theta_1)h \\&= 106.09 + f(240, 106.09)240 \\&= 106.09 + \left(-2.2067 \times 10^{-12} (106.09^4 - 81 \times 10^8)\right)240 \\&= 106.09 + (0.017595)240 \\&= 110.32K\end{aligned}$$

θ_2 is the approximate temperature at $t = t_2 = t_1 + h = 240 + 240 = 480$

$$\theta(480) \approx \theta_2 = 110.32K$$

Solution Cont

The exact solution of the ordinary differential equation is given by the solution of a non-linear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}(0.00333\theta) = -0.22067 \times 10^{-3}t - 2.9282$$

The solution to this nonlinear equation at t=480 seconds is

$$\theta(480) = 647.57K$$

Comparison of Exact and Numerical Solutions

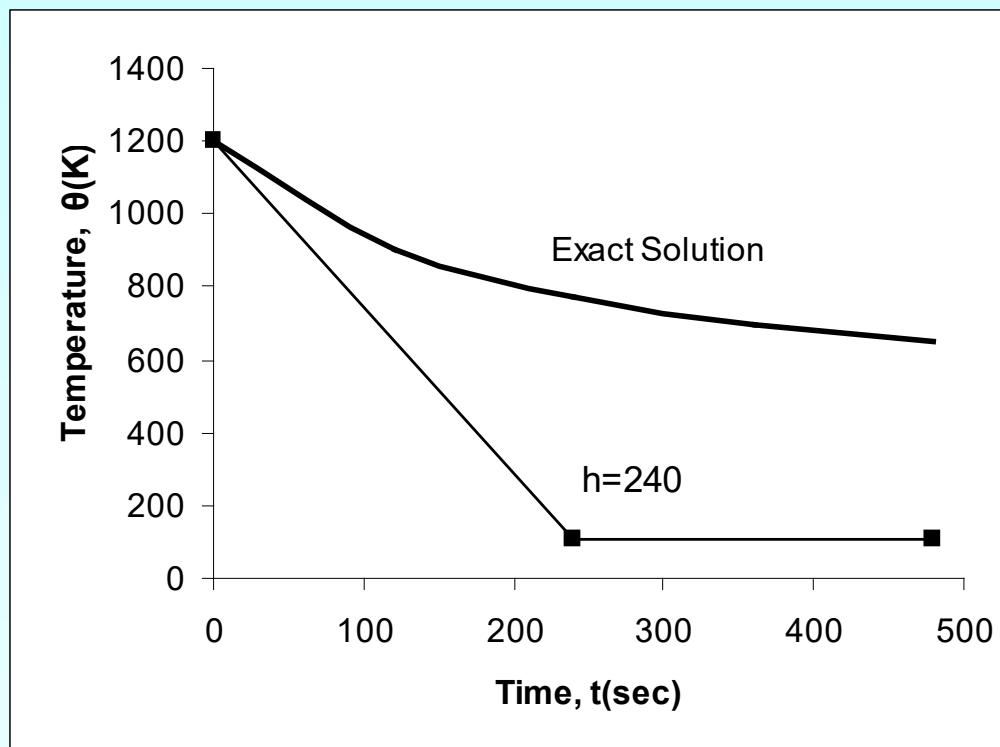


Figure 3. Comparing exact and Euler's method

Effect of step size

Table 1. Temperature at 480 seconds as a function of step size, h

Step, h	$\theta(480)$	E_t	$ \epsilon_t \%$
480	-987.81	1635.4	252.54
240	110.32	537.26	82.964
120	546.77	100.80	15.566
60	614.97	32.607	5.0352
30	632.77	14.806	2.2864

$$\theta(480) = 647.57K \quad (\text{exact})$$

Comparison with exact results

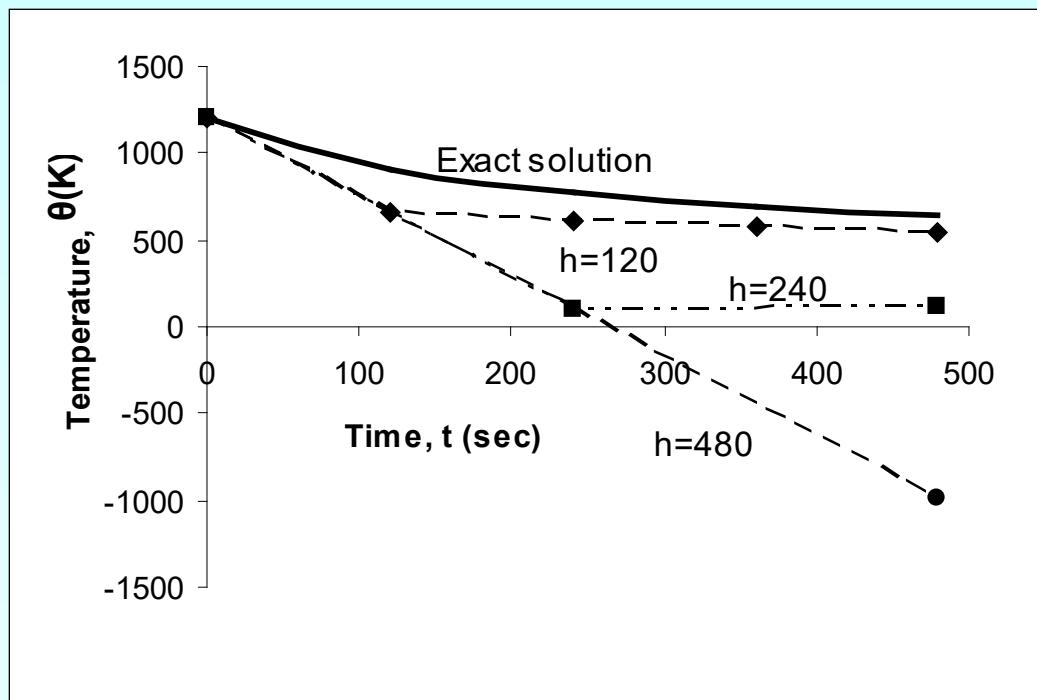


Figure 4. Comparison of Euler's method with exact solution for different step sizes

Effects of step size on Euler's Method

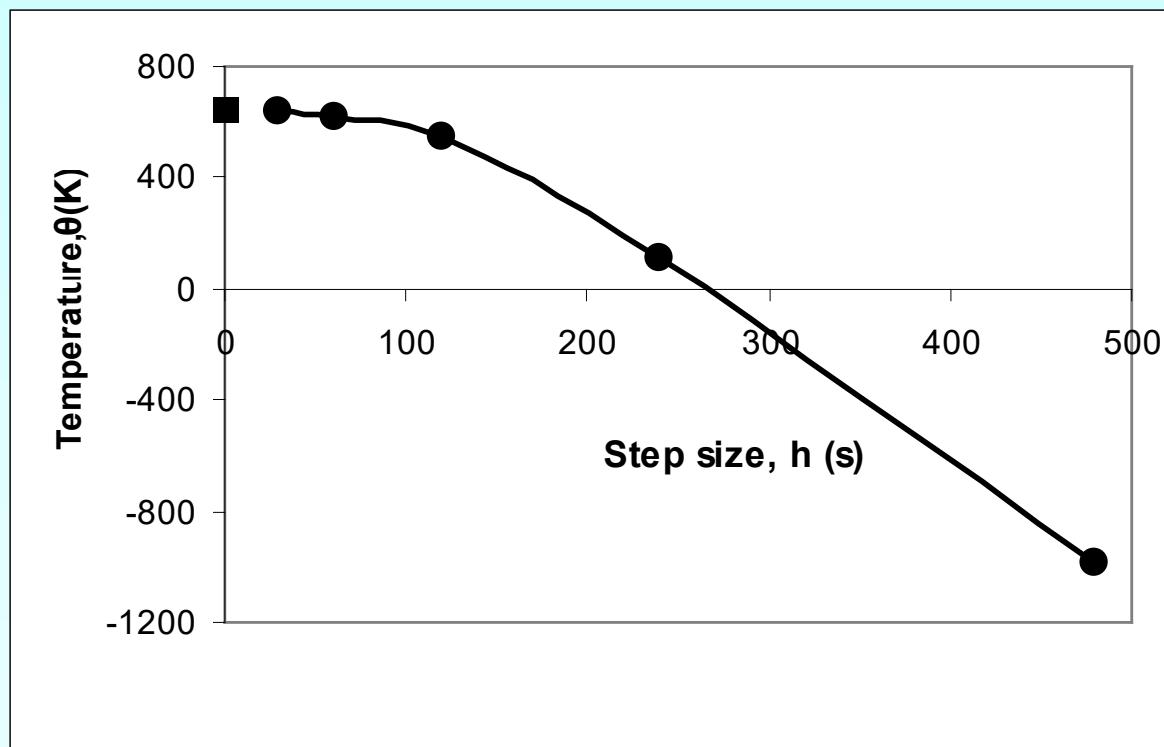


Figure 5. Effect of step size in Euler's method.

Errors in Euler's Method

It can be seen that Euler's method has large errors. This can be illustrated using Taylor series.

$$y_{i+1} = y_i + \left. \frac{dy}{dx} \right|_{x_i, y_i} (x_{i+1} - x_i) + \frac{1}{2!} \left. \frac{d^2 y}{dx^2} \right|_{x_i, y_i} (x_{i+1} - x_i)^2 + \frac{1}{3!} \left. \frac{d^3 y}{dx^3} \right|_{x_i, y_i} (x_{i+1} - x_i)^3 + \dots$$

$$y_{i+1} = y_i + f(x_i, y_i)(x_{i+1} - x_i) + \frac{1}{2!} f'(x_i, y_i)(x_{i+1} - x_i)^2 + \frac{1}{3!} f''(x_i, y_i)(x_{i+1} - x_i)^3 + \dots$$

As you can see the first two terms of the Taylor series

$$y_{i+1} = y_i + f(x_i, y_i)h \quad \text{are the Euler's method.}$$

The true error in the approximation is given by

$$E_t = \frac{f'(x_i, y_i)}{2!} h^2 + \frac{f''(x_i, y_i)}{3!} h^3 + \dots \quad E_t \propto h^2$$

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/euler_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Runge 2nd Order Method

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Runge-Kutta 2nd Order Method

<http://numericalmethods.eng.usf.edu>

Runge-Kutta 2nd Order Method

For $\frac{dy}{dx} = f(x, y), y(0) = y_0$

Runge Kutta 2nd order method is given by

$$y_{i+1} = y_i + (a_1 k_1 + a_2 k_2)h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + p_1 h, y_i + q_{11} k_1 h)$$

Heun's Method

Heun's method

Here $a_2=1/2$ is chosen

$$a_1 = \frac{1}{2}$$

$$p_1 = 1$$

$$q_{11} = 1$$

resulting in

$$y_{i+1} = y_i + \left(\frac{1}{2}k_1 + \frac{1}{2}k_2 \right)h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + h, y_i + k_1 h)$$

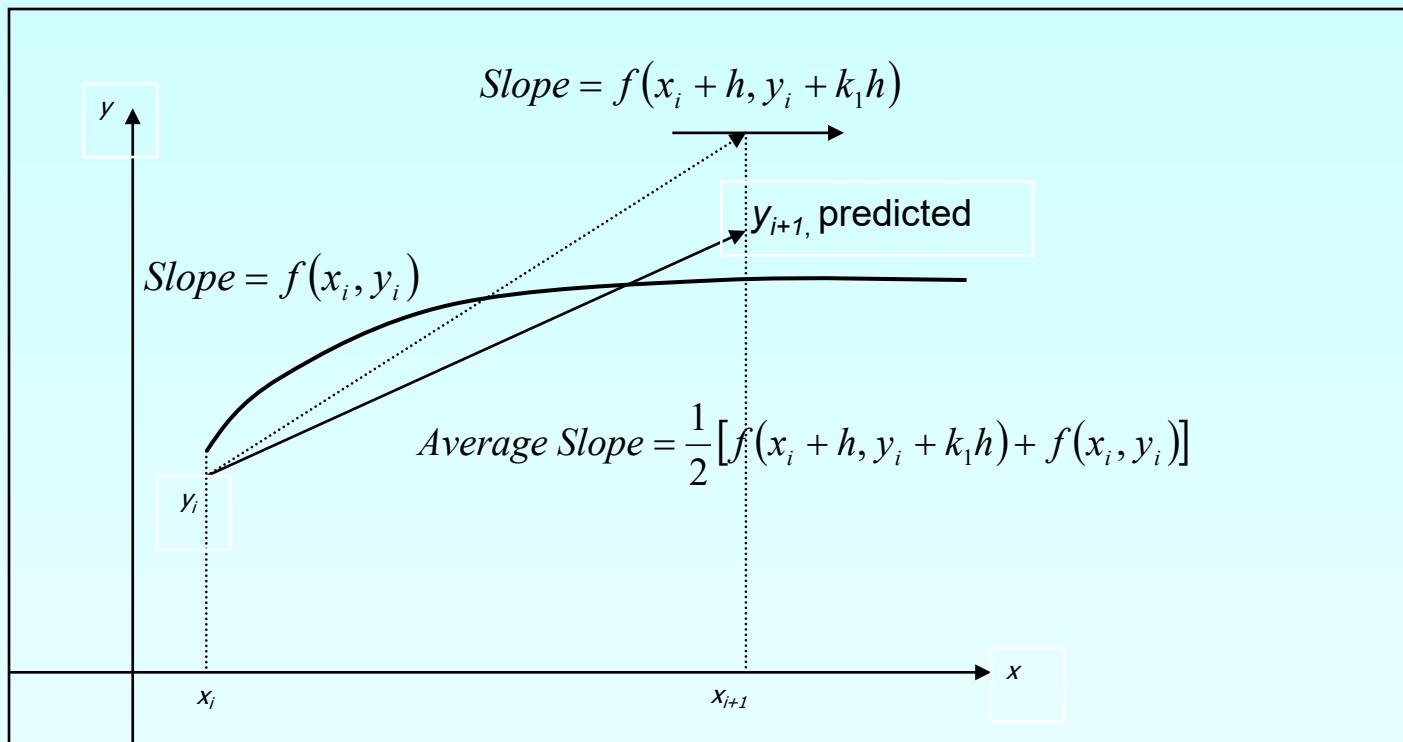


Figure 1 Runge-Kutta 2nd order method (Heun's method)

Midpoint Method

Here $a_2 = 1$ is chosen, giving

$$a_1 = 0$$

$$p_1 = \frac{1}{2}$$

$$q_{11} = \frac{1}{2}$$

resulting in

$$y_{i+1} = y_i + k_2 h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right)$$

Ralston's Method

Here $a_2 = \frac{2}{3}$ is chosen, giving

$$a_1 = \frac{1}{3}$$

$$p_1 = \frac{3}{4}$$

$$q_{11} = \frac{3}{4}$$

resulting in

$$y_{i+1} = y_i + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2 \right)h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f\left(x_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1 h\right)$$

How to write Ordinary Differential Equation

How does one write a first order differential equation in the form of

$$\frac{dy}{dx} = f(x, y)$$

Example

$$\frac{dy}{dx} + 2y = 1.3e^{-x}, y(0) = 5$$

is rewritten as

$$\frac{dy}{dx} = 1.3e^{-x} - 2y, y(0) = 5$$

In this case

$$f(x, y) = 1.3e^{-x} - 2y$$

Example

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8), \theta(0) = 1200K$$

Find the temperature at $t = 480$ seconds using Heun's method. Assume a step size of $h = 240$ seconds.

$$\begin{aligned}\frac{d\theta}{dt} &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8) \\ f(t, \theta) &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8) \\ \theta_{i+1} &= \theta_i + \left(\frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h\end{aligned}$$

Solution

Step 1: $i = 0, t_0 = 0, \theta_0 = \theta(0) = 1200K$

$$\begin{aligned} k_1 &= f(t_0, \theta_0) & k_2 &= f(t_0 + h, \theta_0 + k_1 h) \\ &= f(0, 1200) & &= f(0 + 240, 1200 + (-4.5579)240) \\ &= -2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8) & &= f(240, 106.09) \\ &= -4.5579 & &= -2.2067 \times 10^{-12} (106.09^4 - 81 \times 10^8) \\ & & &= 0.017595 \end{aligned}$$

$$\begin{aligned} \theta_1 &= \theta_0 + \left(\frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h \\ &= 1200 + \left(\frac{1}{2} (-4.5579) + \frac{1}{2} (0.017595) \right) 240 \\ &= 1200 + (-2.2702) 240 \\ &= 655.16K \end{aligned}$$

Solution Cont

Step 2: $i = 1, t_1 = t_0 + h = 0 + 240 = 240, \theta_1 = 655.16K$

$$\begin{aligned}k_1 &= f(t_1, \theta_1) \\&= f(240, 655.16) \\&= -2.2067 \times 10^{-12} (655.16^4 - 81 \times 10^8) \\&= -0.38869\end{aligned}$$

$$\begin{aligned}k_2 &= f(t_1 + h, \theta_1 + k_1 h) \\&= f(240 + 240, 655.16 + (-0.38869)240) \\&= f(480, 561.87) \\&= -2.2067 \times 10^{-12} (561.87^4 - 81 \times 10^8) \\&= -0.20206\end{aligned}$$

$$\begin{aligned}\theta_2 &= \theta_1 + \left(\frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h \\&= 655.16 + \left(\frac{1}{2}(-0.38869) + \frac{1}{2}(-0.20206) \right) 240 \\&= 655.16 + (-0.29538)240 \\&= 584.27K\end{aligned}$$

Solution Cont

The exact solution of the ordinary differential equation is given by the solution of a non-linear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}(0.0033333\theta) = -0.22067 \times 10^{-3}t - 2.9282$$

The solution to this nonlinear equation at t=480 seconds is

$$\theta(480) = 647.57K$$

Comparison with exact results

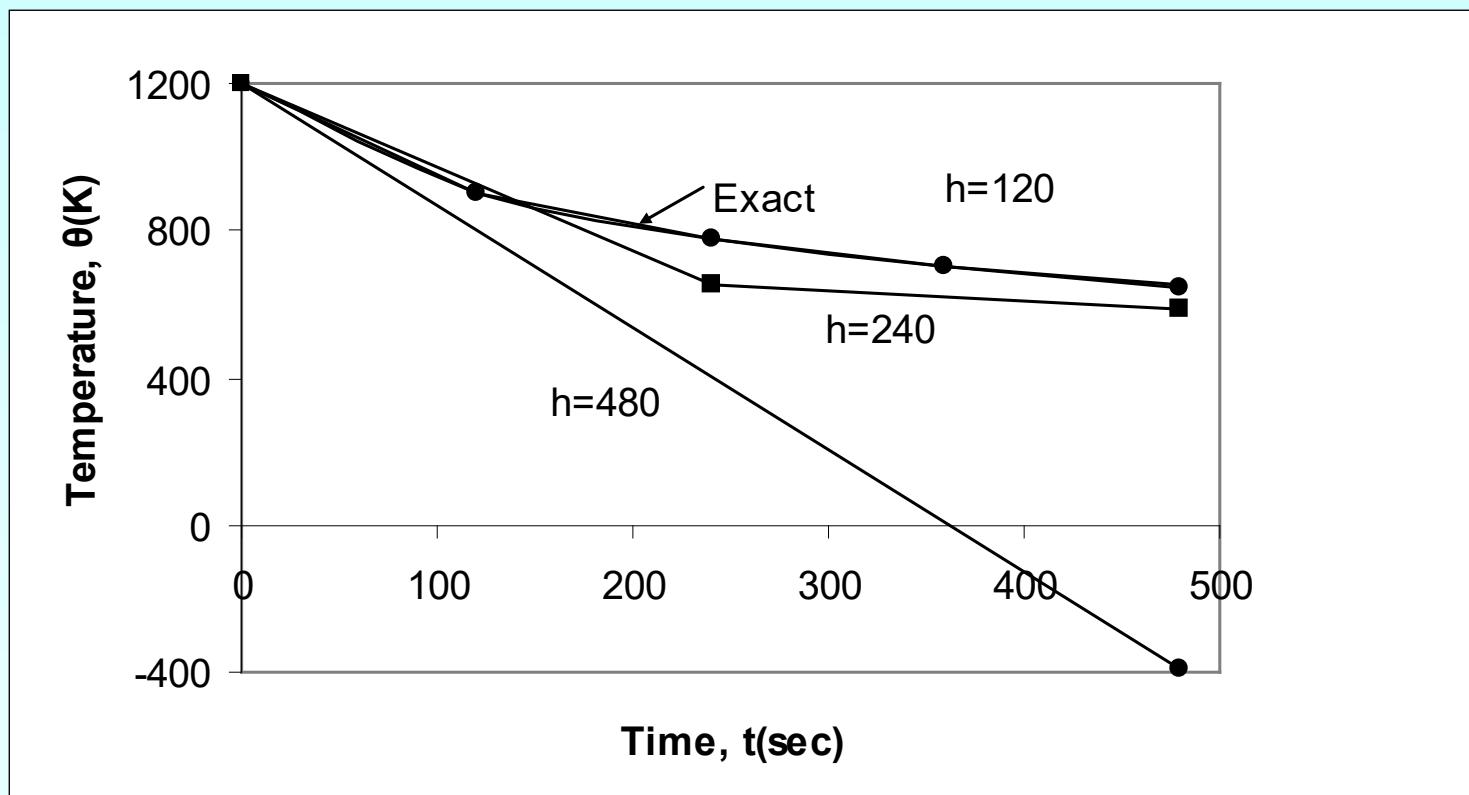


Figure 2. Heun's method results for different step sizes

Effect of step size

Table 1. Temperature at 480 seconds as a function of step size, h

Step size, h	$\theta(480)$	E_t	$ \epsilon_t \%$
480	-393.87	1041.4	160.82
240	584.27	63.304	9.7756
120	651.35	-3.7762	0.58313
60	649.91	-2.3406	0.36145
30	648.21	-0.63219	0.097625

$$\theta(480) = 647.57K \quad (\text{exact})$$

Effects of step size on Heun's Method

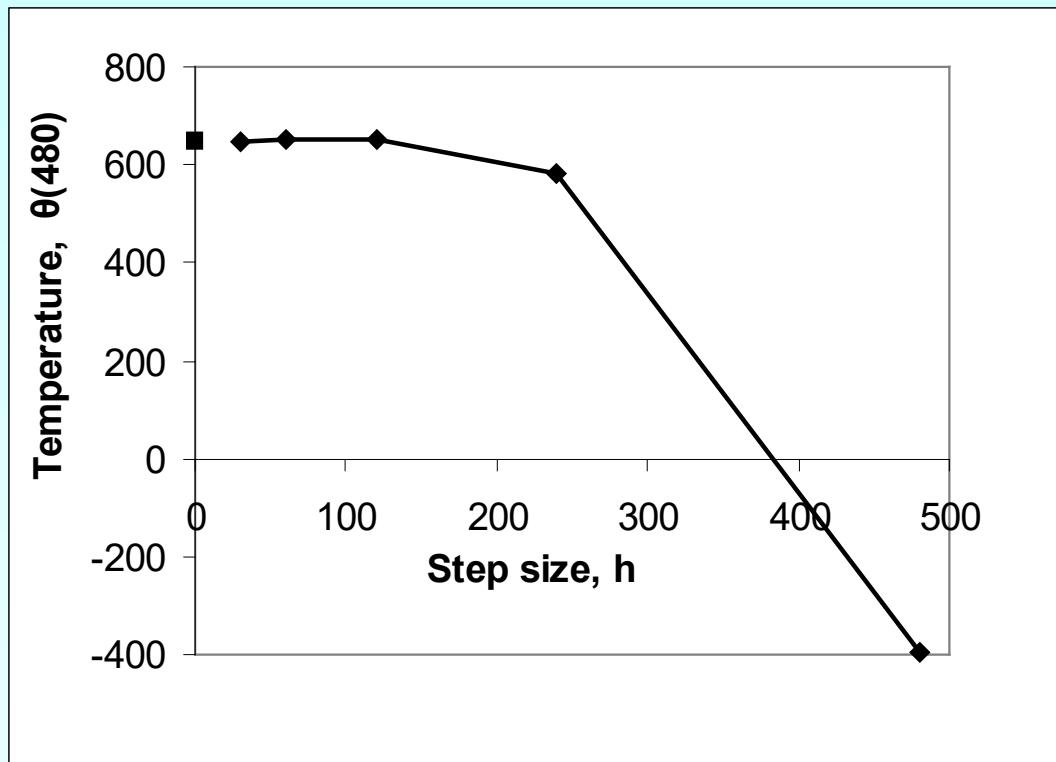


Figure 3. Effect of step size in Heun's method

Comparison of Euler and Runge-Kutta 2nd Order Methods

Table 2. Comparison of Euler and the Runge-Kutta methods

Step size, h	$\theta(480)$			
	Euler	Heun	Midpoint	Ralston
480	-987.84	-393.87	1208.4	449.78
240	110.32	584.27	976.87	690.01
120	546.77	651.35	690.20	667.71
60	614.97	649.91	654.85	652.25
30	632.77	648.21	649.02	648.61

$$\theta(480) = 647.57K \text{ (exact)}$$

Comparison of Euler and Runge-Kutta 2nd Order Methods

Table 2. Comparison of Euler and the Runge-Kutta methods

Step size, h	$ \epsilon_t \%$			
	Euler	Heun	Midpoint	Ralston
480	252.54	160.82	86.612	30.544
240	82.964	9.7756	50.851	6.5537
120	15.566	0.58313	6.5823	3.1092
60	5.0352	0.36145	1.1239	0.72299
30	2.2864	0.097625	0.22353	0.15940

$$\theta(480) = 647.57K \quad (\text{exact})$$

Comparison of Euler and Runge-Kutta 2nd Order Methods

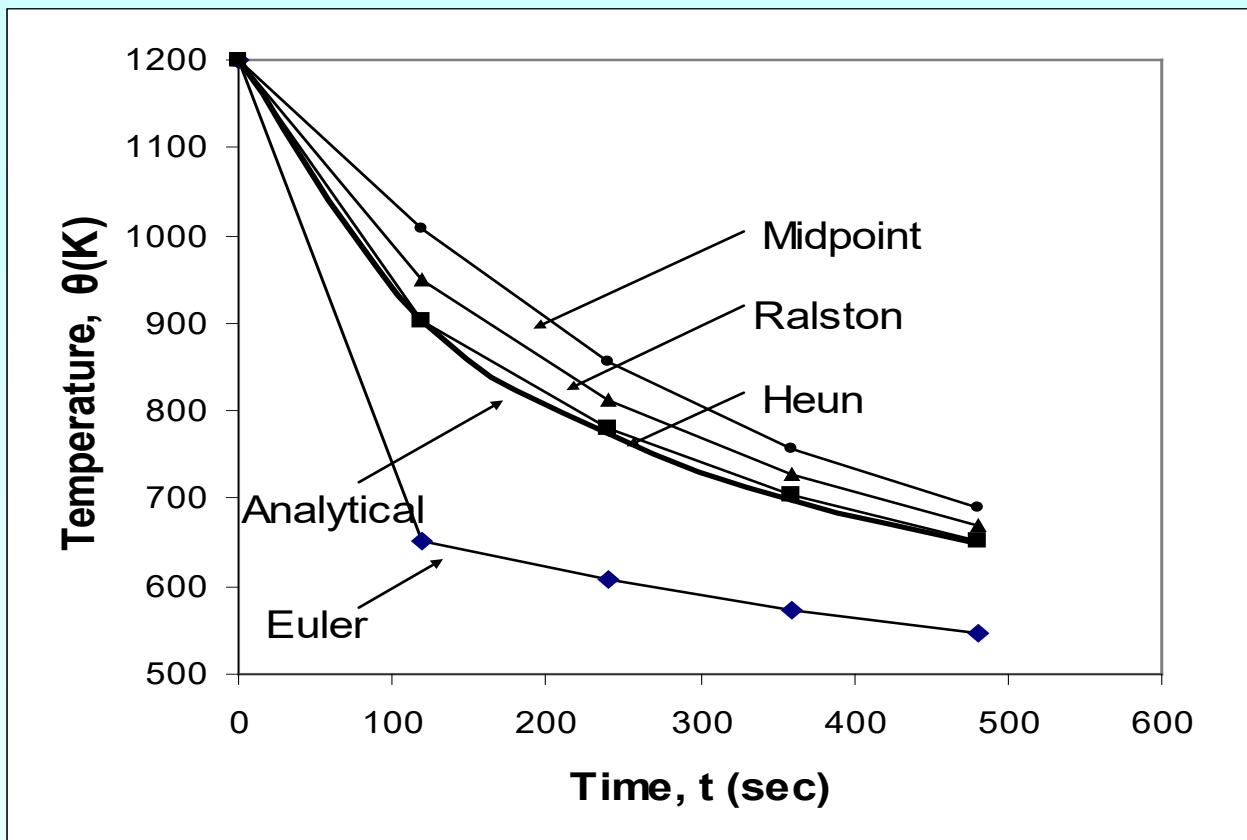


Figure 4. Comparison of Euler and Runge Kutta 2nd order methods with exact results.

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/runge_kutta_2nd_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Runge 4th Order Method

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Runge-Kutta 4th Order Method

<http://numericalmethods.eng.usf.edu>

Runge-Kutta 4th Order Method

For $\frac{dy}{dx} = f(x, y), y(0) = y_0$

Runge Kutta 4th order method is given by

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right)$$

$$k_3 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right)$$

$$k_4 = f(x_i + h, y_i + k_3 h)$$

How to write Ordinary Differential Equation

How does one write a first order differential equation in the form of

$$\frac{dy}{dx} = f(x, y)$$

Example

$$\frac{dy}{dx} + 2y = 1.3e^{-x}, y(0) = 5$$

is rewritten as

$$\frac{dy}{dx} = 1.3e^{-x} - 2y, y(0) = 5$$

In this case

$$f(x, y) = 1.3e^{-x} - 2y$$

Example

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8), \theta(0) = 1200K$$

Find the temperature at $t = 480$ seconds using Runge-Kutta 4th order method.

Assume a step size of $h = 240$ seconds.

$$\begin{aligned}\frac{d\theta}{dt} &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8) \\ f(t, \theta) &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)\end{aligned}$$

$$\theta_{i+1} = \theta_i + \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4)h$$

Solution

Step 1: $i = 0, t_0 = 0, \theta_0 = \theta(0) = 1200$

$$k_1 = f(t_0, \theta_o) = f(0, 1200) = -2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8) = -4.5579$$

$$\begin{aligned} k_2 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_1h\right) = f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-4.5579)240\right) \\ &= f(120, 653.05) = -2.2067 \times 10^{-12} (653.05^4 - 81 \times 10^8) = -0.38347 \end{aligned}$$

$$\begin{aligned} k_3 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_2h\right) = f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-0.38347)240\right) \\ &= f(120, 1154.0) = 2.2067 \times 10^{-12} (1154.0^4 - 81 \times 10^8) = -3.8954 \end{aligned}$$

$$\begin{aligned} k_4 &= f(t_0 + h, \theta_0 + k_3h) = f(0 + (240), 1200 + (-3.984)240) \\ &= f(240, 265.10) = 2.2067 \times 10^{-12} (265.10^4 - 81 \times 10^8) = 0.0069750 \end{aligned}$$

Solution Cont

$$\begin{aligned}\theta_1 &= \theta_0 + \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4)h \\&= 1200 + \frac{1}{6} (-4.5579 + 2(-0.38347) + 2(-3.8954) + (0.069750))240 \\&= 1200 + \frac{1}{6} (-2.1848)240 \\&= 675.65K\end{aligned}$$

θ_1 is the approximate temperature at

$$t = t_1 = t_0 + h = 0 + 240 = 240$$

$$\theta(240) \approx \theta_1 = 675.65K$$

Solution Cont

Step 2: $i = 1, t_1 = 240, \theta_1 = 675.65K$

$$k_1 = f(t_1, \theta_1) = f(240, 675.65) = -2.2067 \times 10^{-12} (675.65^4 - 81 \times 10^8) = -0.44199$$

$$\begin{aligned} k_2 &= f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_1h\right) = f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.44199)240\right) \\ &= f(360, 622.61) = -2.2067 \times 10^{-12} (622.61^4 - 81 \times 10^8) = -0.31372 \end{aligned}$$

$$\begin{aligned} k_3 &= f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_2h\right) = f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.31372)240\right) \\ &= f(360, 638.00) = 2.2067 \times 10^{-12} (638.00^4 - 81 \times 10^8) = -0.34775 \end{aligned}$$

$$\begin{aligned} k_4 &= f(t_1 + h, \theta_1 + k_3h) = f(240 + (240), 675.65 + (-0.34775)240) \\ &= f(480, 592.19) = 2.2067 \times 10^{-12} (592.19^4 - 81 \times 10^8) = -0.25351 \end{aligned}$$

Solution Cont

$$\begin{aligned}\theta_2 &= \theta_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h \\&= 675.65 + \frac{1}{6}(-0.44199 + 2(-0.31372) + 2(-0.34775) + (-0.25351))240 \\&= 675.65 + \frac{1}{6}(-2.0184)240 \\&= 594.91K\end{aligned}$$

θ_2 is the approximate temperature at

$$t_2 = t_1 + h = 240 + 240 = 480$$

$$\theta(480) \approx \theta_2 = 594.91K$$

Solution Cont

The exact solution of the ordinary differential equation is given by the solution of a non-linear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}(0.00333\theta) = -0.22067 \times 10^{-3} t - 2.9282$$

The solution to this nonlinear equation at t=480 seconds is

$$\theta(480) = 647.57K$$

Comparison with exact results

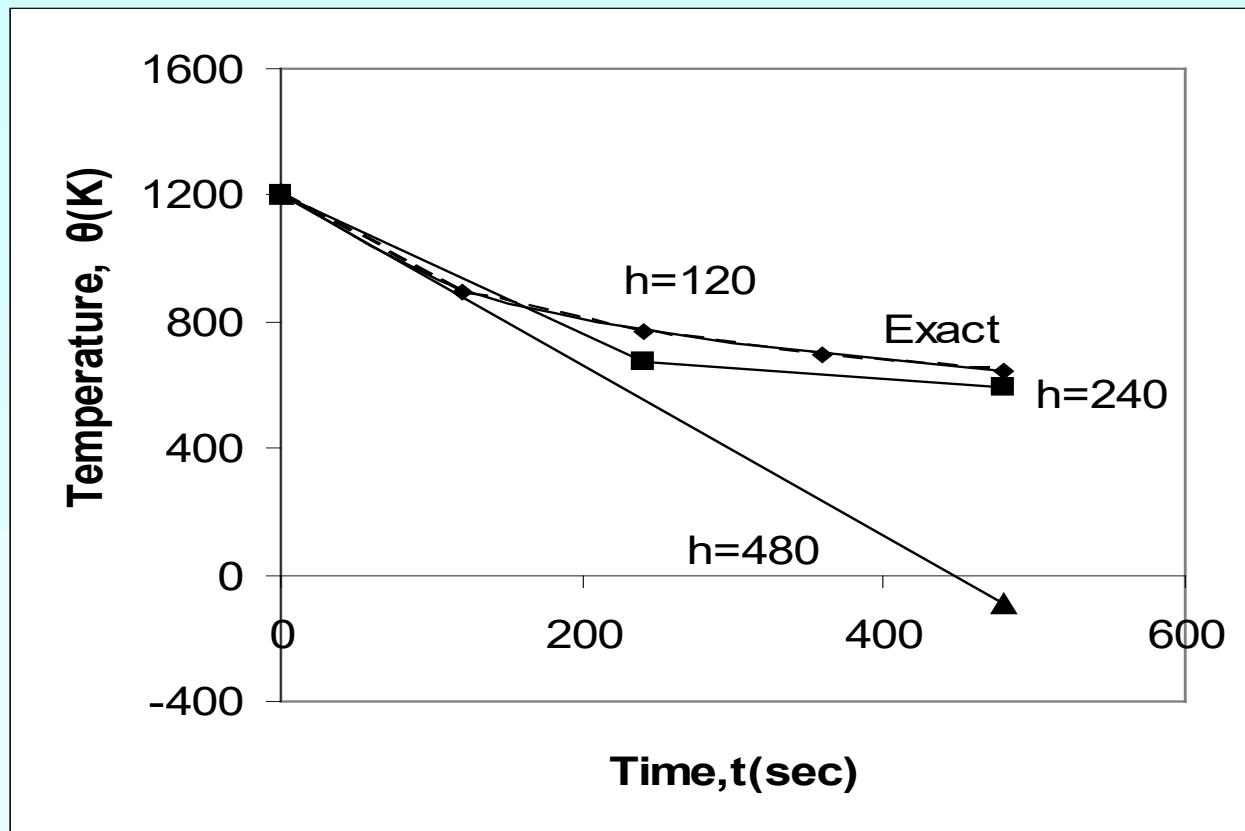


Figure 1. Comparison of Runge-Kutta 4th order method with exact solution

Effect of step size

Table 1. Temperature at 480 seconds as a function of step size, h

Step size, h	θ (480)	E_t	$ \epsilon_t \%$
480	-90.278	737.85	113.94
240	594.91	52.660	8.1319
120	646.16	1.4122	0.21807
60	647.54	0.033626	0.0051926
30	647.57	0.00086900	0.00013419

$$\theta(480) = 647.57K \quad (\text{exact})$$

Effects of step size on Runge-Kutta 4th Order Method

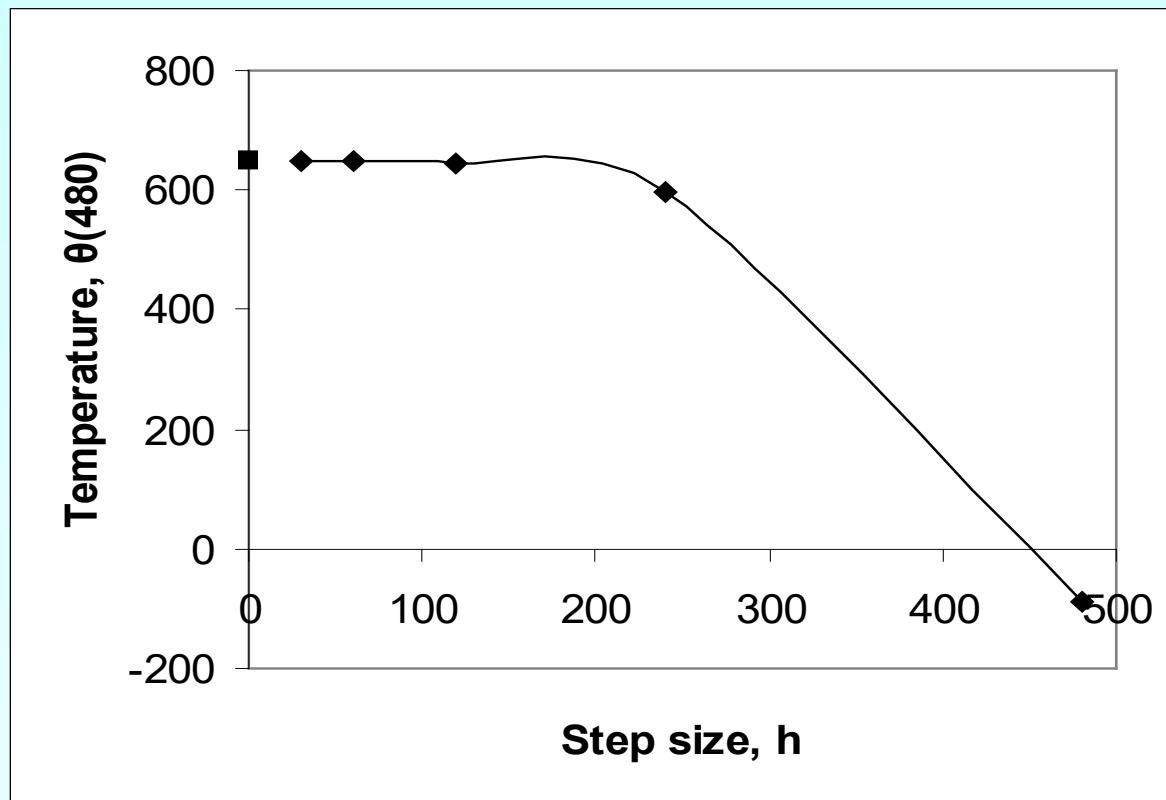


Figure 2. Effect of step size in Runge-Kutta 4th order method

Comparison of Euler and Runge-Kutta Methods

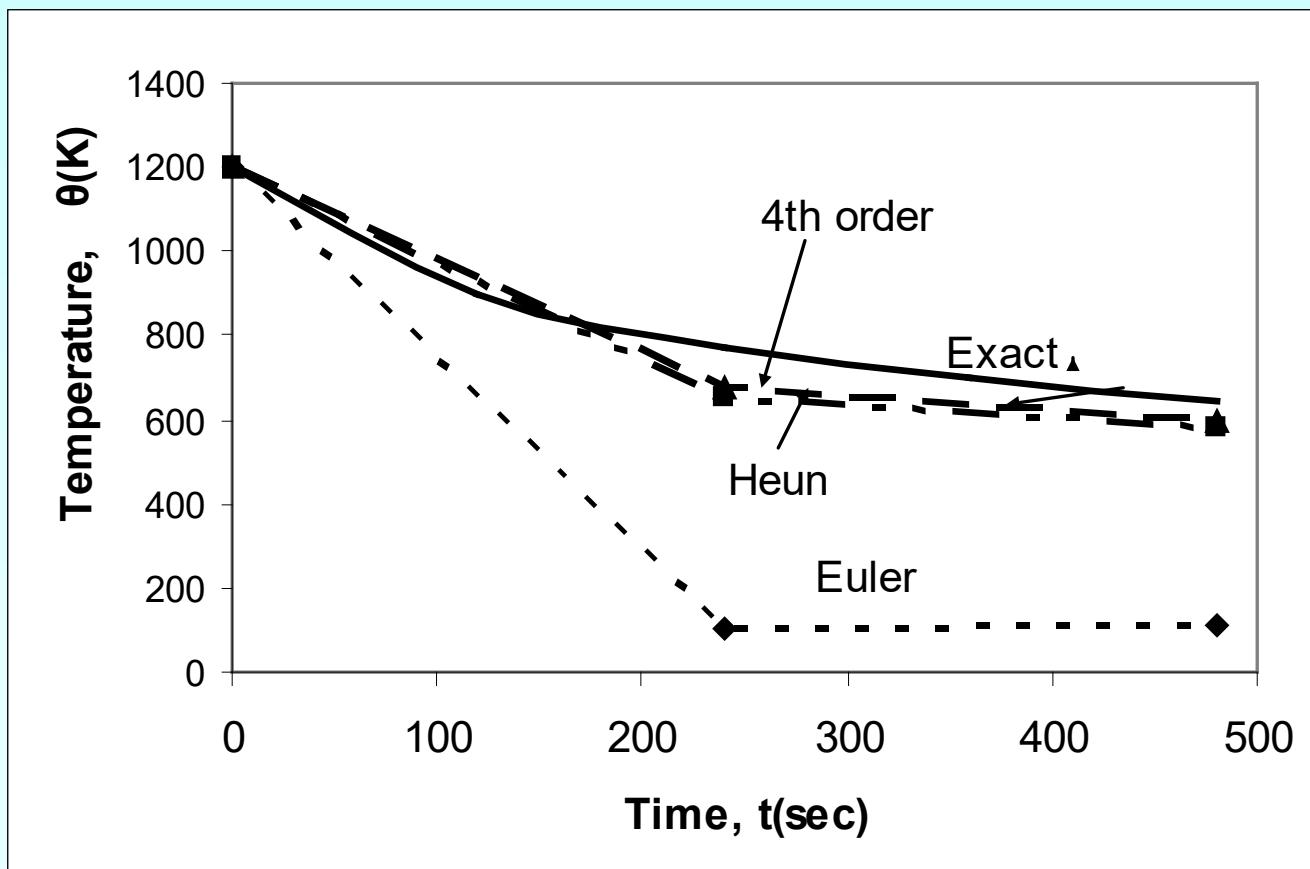


Figure 3. Comparison of Runge-Kutta methods of 1st, 2nd, and 4th order.

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/runge_kutta_4th_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Shooting Method

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Shooting Method

<http://numericalmethods.eng.usf.edu>

Shooting Method

The shooting method uses the methods used in solving initial value problems. This is done by assuming initial values that would have been given if the ordinary differential equation were a initial value problem. The boundary value obtained is compared with the actual boundary value. Using trial and error or some scientific approach, one tries to get as close to the boundary value as possible.

Example

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0,$$

$$u(5) = 0.0038731,$$

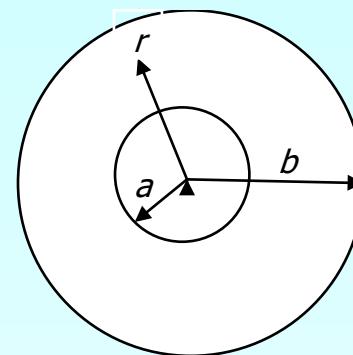
$$u(8) = 0.0030770$$

Let

$$\frac{du}{dr} = w$$

Then

$$\frac{dw}{dr} + \frac{1}{r} w - \frac{u}{r^2} = 0$$



Where $a = 5$
and $b = 8$

Solution

Two first order differential equations are given as

$$\frac{du}{dr} = w, \quad u(5) = 0.0038371$$

$$\frac{dw}{dr} = -\frac{w}{r} + \frac{u}{r^2}, \quad w(5) = \text{not known}$$

Let us assume

$$w(5) = \frac{du}{dr}(5) \approx \frac{u(8) - u(5)}{8 - 5} = -0.00026538$$

To set up initial value problem

$$\frac{du}{dr} = w = f_1(r, u, w), \quad u(5) = 0.0038371$$

$$\frac{dw}{dr} = -\frac{w}{r} + \frac{u}{r^2} = f_2(r, u, w), \quad w(5) = -0.00026538$$

Solution Cont

Using Euler's method,

$$u_{i+1} = u_i + f_1(r_i, u_i, w_i)h$$

$$w_{i+1} = w_i + f_2(r_i, u_i, w_i)h$$

Let us consider 4 segments between the two boundaries, $r = 5$ and $r = 8$ then,

$$h = \frac{8 - 5}{4} = 0.75$$

Solution Cont

For $i = 0, r_0 = 5, u_0 = 0.0038371, w_0 = -0.00026538$

$$\begin{aligned}u_1 &= u_0 + f_1(r_0, u_0, w_0)h \\&= 0.0038371 + f_1(5, 0.0038371, -0.00026538)(0.75) \\&= 0.0038371 + (-0.00026538)(0.75) \\&= 0.0036741\end{aligned}$$

$$\begin{aligned}w_1 &= w_0 + f_2(r_0, u_0, w_0)h \\&= -0.00026538 + f_2(5, 0.0038371, -0.00026538)(0.75) \\&= -0.00026538 + \left(-\frac{-0.00026538}{5} + \frac{0.0038371}{5^2} \right)(0.75) \\&= -0.00010938\end{aligned}$$

Solution Cont

For $i = 1, r_1 = r_0 + h = 5 + 0.75 = 5.75, u_1 = 0.0036741, w_1 = -0.00010940$

$$\begin{aligned}u_2 &= u_1 + f_1(r_1, u_1, w_1)h \\&= 0.0036741 + f_1(5.75, 0.0036741, -0.00010938)(0.75) \\&= 0.0036741 + (-0.00010938)(0.75) \\&= 0.0035920\end{aligned}$$

$$\begin{aligned}w_2 &= w_1 + f_2(r_1, u_1, w_1)h \\&= -0.00010938 + f_2(5.75, 0.0036741, -0.00010938)(0.75) \\&= -0.00010938 + (0.00013015)(0.75) \\&= -0.000011769\end{aligned}$$

Solution Cont

For $i = 2, r_2 = r_1 + h = 5.75 + 0.75 = 6.5 \quad u_2 = 0.0035920, w_2 = -0.000011785$

$$\begin{aligned} u_3 &= u_2 + f_1(r_2, u_2, w_2)h \\ &= 0.0035920 + f_1(6.5, 0.0035920, -0.000011769)(0.75) \\ &= 0.0035920 + (-0.000011769)(0.75) \\ &= 0.0035832 \end{aligned}$$

$$\begin{aligned} w_3 &= w_2 + f_2(r_2, u_2, w_2)h \\ &= -0.000011769 + f_2(6.5, 0.0035920, -0.000011769)(0.75) \\ &= -0.000011769 + (0.000086829)(0.75) \\ &= 0.000053352 \end{aligned}$$

Solution Cont

For $i = 3, r_3 = r_2 + h = 6.50 + 0.75 = 7.25 \quad u_3 = 0.0035832, w_3 = 0.000053332$

$$\begin{aligned} u_4 &= u_3 + f_1(r_3, u_3, w_3)h \\ &= 0.0035832 + f_1(7.25, 0.0035832, 0.000053352)(0.75) \\ &= 0.0035832 + (0.000053352)(0.75) \\ &= 0.0036232 \end{aligned}$$

$$\begin{aligned} w_4 &= w_3 + f_2(r_3, u_3, w_3)h \\ &= -0.000011785 + f_2(5.75, 0.0035832, -0.000053352)(0.75) \\ &= 0.000053352 + (0.000060811)(0.75) \\ &= 0.000098961 \end{aligned}$$

So at $r = r_4 = r_3 + h = 7.25 + 0.75 = 8$

$$u(8) \approx u_4 = 0.0036232$$

Solution Cont

Let us assume a new value for $\frac{du}{dr}(5)$

$$w(5) = 2 \frac{du}{dr}(5) \approx 2 \frac{u(8) - u(5)}{8 - 5} = 2(-0.00026538) = -0.00053076$$

Using $h = 0.75$ and Euler's method, we get

$$u(8) \approx u_4 = 0.0029665"$$

While the given value of this boundary condition is

$$u(8) \approx u_4 = 0.0030770$$

Solution Cont

Using linear interpolation on the obtained data for the two assumed values of

$\frac{du}{dr}(5)$ we get

$$u(8) = 0.00030770$$

$$\begin{aligned}\frac{du}{dr}(5) &\approx \frac{-0.00053076 - (-0.00026538)}{0.0029645 - 0.0036232} (0.0030770 - 0.0036232) + (-0.00026538) \\ &= -0.00048611\end{aligned}$$

Using $h = 0.75$ and repeating the Euler's method with $w(5) = -0.00048611$

$$u(8) \approx u_4 = 0.0030769$$

Solution Cont

Using linear interpolation to refine the value of u_4

till one gets close to the actual value of $u(8)$ which gives you,

$$u_1 = u(5) = 0.0038731$$

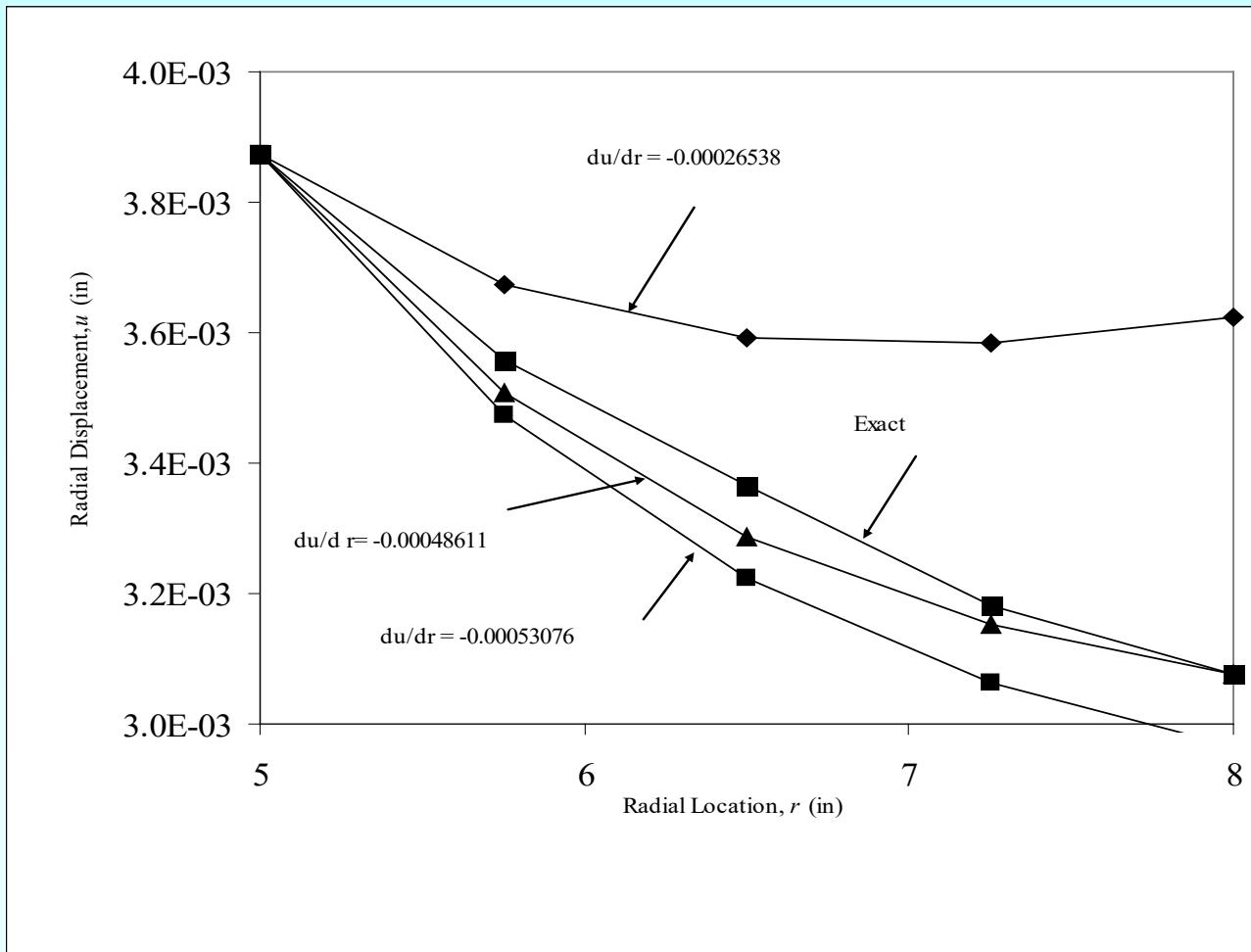
$$u(5.75) \approx u_2 = 0.0035085$$

$$u(6.50) \approx u_3 = 0.0032858$$

$$u(7.25) \approx u_4 = 0.0031518$$

$$u(8.00) \approx u_5 = 0.0030770$$

Comparisons of different initial guesses



Comparison of Euler and Runge-Kutta Results with exact results

Table 1 Comparison of Euler and Runge-Kutta results with exact results.

r (in)	Exact (in)	Euler (in)	$ e_t \%$	Runge-Kutta (in)	$ e_t \%$
5	3.8731×10^{-3}	3.8731×10^{-3}	0.0000	3.8731×10^{-3}	0.0000
5.75	3.5567×10^{-3}	3.5085×10^{-3}	1.3731	3.5554×10^{-3}	3.5824×10^{-2}
6.5	3.3366×10^{-3}	3.2858×10^{-3}	1.5482	3.3341×10^{-3}	7.4037×10^{-2}
7.25	3.1829×10^{-3}	3.1518×10^{-3}	9.8967×10^{-1}	3.1792×10^{-3}	1.1612×10^{-1}
8	3.0770×10^{-3}	3.0770×10^{-3}	1.9500	3.0723×10^{-3}	1.5168×10^{-1}

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/shooting_method.html

THE END

<http://numericalmethods.eng.usf.edu>

Finite Difference Method

Major: All Engineering Majors

Authors: Autar Kaw, Charlie Barker

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM
Undergraduates

Finite Difference Method

<http://numericalmethods.eng.usf.edu>

Finite Difference Method

An example of a boundary value ordinary differential equation is

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0, \quad u(5) = 0.008731", \quad u(8) = 0.0030769"$$

The derivatives in such ordinary differential equation are substituted by finite divided differences approximations, such as

$$\frac{dy}{dx} \approx \frac{y_{i+1} - y_i}{\Delta x}$$

$$\frac{d^2y}{dx^2} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2}$$

Example

Take the case of a pressure vessel that is being tested in the laboratory to check its ability to withstand pressure. For a thick pressure vessel of inner radius a and outer radius b , the differential equation for the radial displacement u of a point along the thickness is given by

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0$$

The pressure vessel can be modeled as,

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2}$$

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_i}{\Delta r}$$

Substituting these approximations gives you,

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i} \frac{u_{i+1} - u_i}{\Delta r} - \frac{u_i}{r_i^2} = 0$$

$$\left(\frac{1}{(\Delta r)^2} + \frac{1}{r_i \Delta r} \right) u_{i+1} + \left(-\frac{2}{(\Delta r)^2} - \frac{1}{r_i \Delta r} - \frac{1}{r_i^2} \right) u_i + \frac{1}{(\Delta r)^2} u_{i-1} = 0$$

Solution

Step 1 At node $i = 0$, $r_0 = a = 5"$ $u_0 = 0.0038731"$

Step 2 At node $i = 1$, $r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6"$

$$\frac{1}{(0.6)^2}u_0 + \left(-\frac{2}{(0.6)^2} - \frac{1}{(5.6)(0.6)} - \frac{1}{(5.6)^2}\right)u_1 + \left(\frac{1}{0.6^2} + \frac{1}{(5.6)(0.6)}\right)u_2 = 0$$

$$2.7778u_0 - 5.8851u_1 + 3.0754u_2 = 0$$

Step 3 At node $i = 2$, $r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2"$

$$\frac{1}{0.6^2}u_1 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.2)(0.6)} - \frac{1}{6.2^2}\right)u_2 + \left(\frac{1}{0.6^2} + \frac{1}{(6.2)(0.6)}\right)u_3 = 0$$

$$2.7778u_1 - 5.8504u_2 + 3.0466u_3 = 0$$

Solution Cont

Step 4 At node $i = 3$, $r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8"$

$$\frac{1}{0.6^2}u_2 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.8)(0.6)} - \frac{1}{6.8^2} \right)u_3 + \left(\frac{1}{0.6^2} + \frac{1}{(6.8)(0.6)} \right)u_4 = 0$$

$$2.7778u_2 - 5.8223u_3 + 3.0229u_4 = 0$$

Step 5 At node $i = 4$, $r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4"$

$$\frac{1}{0.6^2}u_3 + \left(-\frac{2}{0.6^2} - \frac{1}{(7.4)(0.6)} - \frac{1}{(7.4)^2} \right)u_4 + \left(\frac{1}{0.6^2} + \frac{1}{(7.4)(0.6)} \right)u_5 = 0$$

$$2.7778u_3 - 5.7990u_4 + 3.0030u_5 = 0$$

Step 6 At node $i = 5$, $r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8$

$$u_5 = u|_{r=b} = 0.0030769"$$

Solving system of equations

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.7778 & -5.8851 & 3.0754 & 0 & 0 & 0 \\ 0 & 2.7778 & -5.8504 & 3.0466 & 0 & 0 \\ 0 & 0 & 2.7778 & -5.8223 & 3.0229 & 0 \\ 0 & 0 & 0 & 2.7778 & -5.7990 & 3.0030 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

$$u_0 = 0.0038731 \quad u_3 = 0.0032743$$

$$u_1 = 0.0036165 \quad u_4 = 0.0031618$$

$$u_2 = 0.0034222 \quad u_5 = 0.0030769$$

Solution Cont

$$\frac{du}{dr} \Big|_{r=a} \approx \frac{u_1 - u_0}{\Delta r} = \frac{0.0036165 - 0.0038731}{0.6} = -0.00042767$$

$$\sigma_{\max} = \frac{30 \times 10^6}{1 - 0.3^2} \left(\frac{0.0038731}{5} + 0.3(-0.00042767) \right) = 21307 \text{ psi}$$

$$FS = \frac{36 \times 10^3}{21307} = 1.6896$$

$$E_t = 20538 - 21307 = -768.59$$

$$|\epsilon_t| = \left| \frac{20538 - 21307}{20538} \right| \times 100 = 3.744 \%$$

Solution Cont

Using the approximation of

$$\frac{d^2y}{dx^2} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2} \quad \text{and} \quad \frac{dy}{dx} \approx \frac{y_{i+1} - y_{i-1}}{2(\Delta x)}$$

Gives you

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i} \frac{u_{i+1} - u_{i-1}}{2(\Delta r)} - \frac{u_i}{r_i^2} = 0$$

$$\left(-\frac{1}{2r_i(\Delta r)} + \frac{1}{(\Delta r)^2} \right) u_{i-1} + \left(-\frac{2}{(\Delta r)^2} - \frac{1}{r_i^2} \right) u_i + \left(\frac{1}{(\Delta r)^2} + \frac{1}{2r_i \Delta r} \right) u_{i+1} = 0$$

Solution Cont

Step 1 At node $i = 0, r_0 = a = 5$

$$u_0 = 0.0038731$$

Step 2 At node $i = 1, r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6"$

$$\left(-\frac{1}{2(5.6)(0.6)} + \frac{1}{(0.6)^2} \right) u_0 + \left(-\frac{2}{(0.6)^2} - \frac{1}{(5.6)^2} \right) u_1 + \left(\frac{1}{0.6^2} + \frac{1}{2(5.6)(0.6)} \right) u_2 = 0$$

$$2.6297u_0 - 5.5874u_1 + 2.9266u_2 = 0$$

Step 3 At node $i = 2, r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2$

$$\left(-\frac{1}{2(6.2)(0.6)} + \frac{1}{0.6^2} \right) u_1 + \left(-\frac{2}{0.6^2} - \frac{1}{6.2^2} \right) u_2 + \left(\frac{1}{0.6^2} + \frac{1}{2(6.2)(0.6)} \right) u_3 = 0$$

$$2.6434u_1 - 5.5816u_2 + 2.9122u_3 = 0$$

Solution Cont

Step 4 At node $i = 3$, $r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8$

$$\left(-\frac{1}{2(6.8)(0.6)} + \frac{1}{0.6^2} \right) u_2 + \left(-\frac{2}{0.6^2} - \frac{1}{6.8^2} \right) u_3 + \left(\frac{1}{0.6^2} + \frac{1}{2(6.8)(0.6)} \right) u_4 = 0$$
$$2.6552u_2 - 5.5772u_3 + 2.9003u_4 = 0$$

Step 5 At node $i = 4$, $r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4$

$$\left(-\frac{1}{2(7.4)(0.6)} + \frac{1}{0.6^2} \right) u_3 + \left(-\frac{2}{0.6^2} - \frac{1}{(7.4)^2} \right) u_4 + \left(\frac{1}{0.6^2} + \frac{1}{2(7.4)(0.6)} \right) u_5 = 0$$
$$2.6651u_3 - 5.5738u_4 + 2.8903u_5 = 0$$

Step 6 At node $i = 5$, $r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8"$

$$u_5 = u|_{r=b} = 0.0030769"$$

Solving system of equations

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.6297 & -5.5874 & 2.9266 & 0 & 0 & 0 \\ 0 & 2.6434 & -5.5816 & 2.9122 & 0 & 0 \\ 0 & 0 & 2.6552 & -5.5772 & 2.9003 & 0 \\ 0 & 0 & 0 & 2.6651 & -5.5738 & 2.8903 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

$$u_0 = 0.0038731 \quad u_3 = 0.0032689$$

$$u_1 = 0.0036115 \quad u_4 = 0.0031586$$

$$u_2 = 0.0034159 \quad u_5 = 0.0030769$$

Solution Cont

$$\frac{du}{dr} \Big|_{r=a} \approx \frac{-3u_0 + 4u_1 - u_2}{2(\Delta r)} = \frac{-3 \times 0.0038731 + 4 \times 0.0036115 - 0.0034159}{2(0.6)} = -0.0004925$$

$$\sigma_{\max} = \frac{30 \times 10^6}{1 - 0.3^2} \left(\frac{0.0038731}{5} + 0.3(-0.0004925) \right) = 20666 \text{ psi}$$

$$FS = \frac{36 \times 10^3}{20666} = 1.7420$$

$$E_t = 20538 - 20666 = -128$$

$$|\epsilon_t| = \left| \frac{20538 - 20666}{20538} \right| \times 100 = 0.62323 \%$$

Comparison of radial displacements

Table 1 Comparisons of radial displacements from two methods

r	u_{exact}	$u_{\text{1st order}}$	$ \epsilon_t $	$u_{\text{2nd order}}$	$ \epsilon_t $
5	0.0038731	0.0038731	0.0000	0.0038731	0.0000
5.6	0.0036110	0.0036165	$1.5160 \cdot 10^{-1}$	0.0036115	$1.4540 \cdot 10^{-2}$
6.2	0.0034152	0.0034222	$2.0260 \cdot 10^{-1}$	0.0034159	$1.8765 \cdot 10^{-2}$
6.8	0.0032683	0.0032743	$1.8157 \cdot 10^{-1}$	0.0032689	$1.6334 \cdot 10^{-2}$
7.4	0.0031583	0.0031618	$1.0903 \cdot 10^{-1}$	0.0031586	$9.5665 \cdot 10^{-3}$
8	0.0030769	0.0030769	0.0000	0.0030769	0.0000

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

[http://numericalmethods.eng.usf.edu/topics/finite difference method.html](http://numericalmethods.eng.usf.edu/topics/finite_difference_method.html)

THE END

<http://numericalmethods.eng.usf.edu>

Numerical Methods

Golden Section Search Method - Theory

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Golden Section Search Method

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Equal Interval Search Method

- Choose an interval $[a, b]$ over which the optima occurs
- Compute $f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right)$ and $f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right)$
- If $f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right) > f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right)$
then the interval in
which the maximum
occurs is $\left[\frac{a+b}{2} - \frac{\varepsilon}{2}, b\right]$
otherwise it occurs in
 $\left[a, \frac{a+b}{2} + \frac{\varepsilon}{2}\right]$

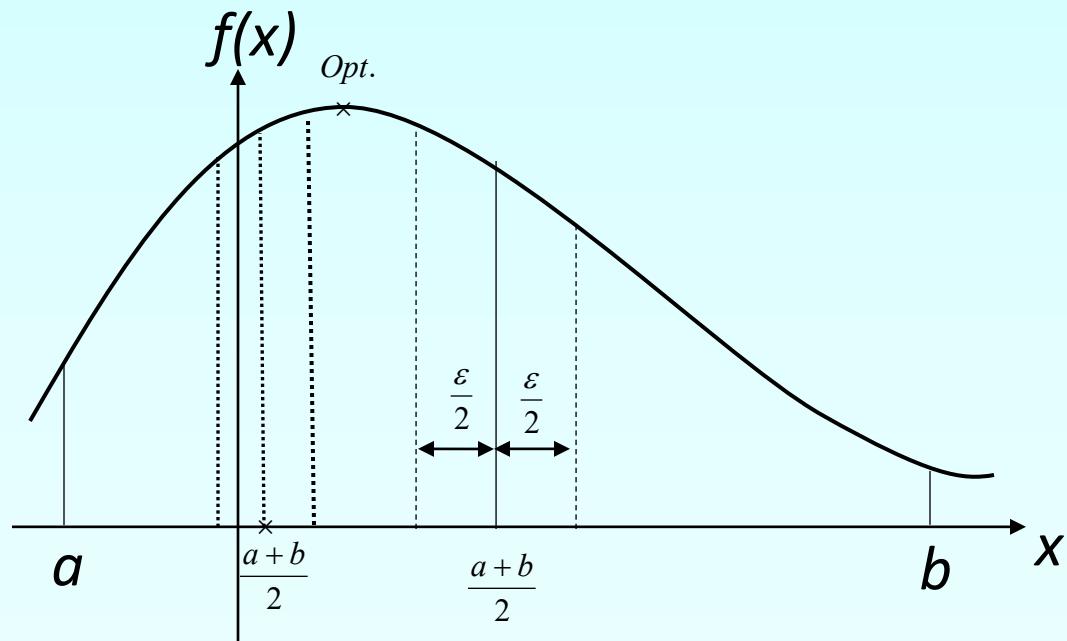


Figure 1 Equal interval search method.

Golden Section Search Method

- The Equal Interval method is inefficient when ε is small. **Also, we need to compute 2 interior points !**
- The Golden Section Search method divides the search more efficiently closing in on the optima in fewer iterations.

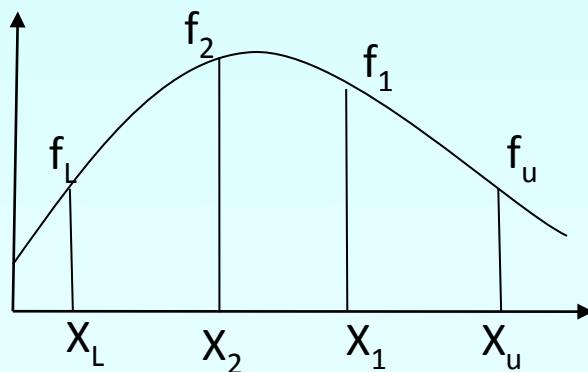
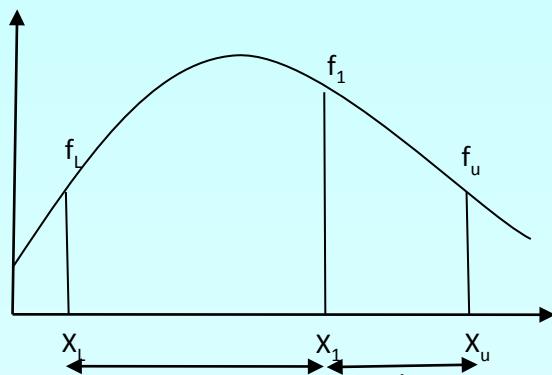


Figure 2. Golden Section Search method

Golden Section Search Method- Selecting the Intermediate Points



Determining the first intermediate point

$$X_1 = X_l + a = X_u - b$$

$$\frac{a}{(a+b)} = \frac{b}{a} = 0.618(\text{why?}); \text{ hence}$$

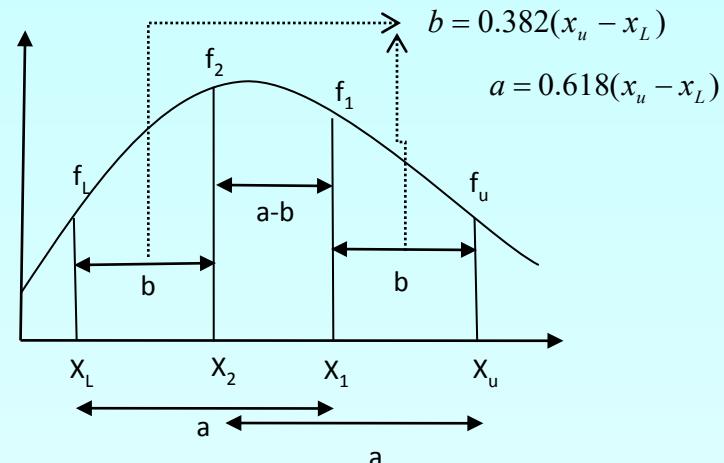
$$a = 0.618 * (X_u - X_l), \text{ and } b = 0.382 * (X_u - X_l)$$

$$\frac{a}{b} = \frac{a+b}{a} = 1 + \frac{b}{a}$$

$$\text{Let } R = \frac{b}{a}, \text{ hence}$$

$$\frac{1}{R} = 1 + R \Rightarrow R^2 + R - 1 = 0 \Rightarrow R = \frac{(\sqrt{5}-1)}{2} \Rightarrow R = 0.61803$$

$$\text{Golden Ratio} \Rightarrow \frac{b}{a} = 0.618\dots$$



Determining the second intermediate point

$$X_2 = X_u - a = X_l + b$$

Golden Section Search Method

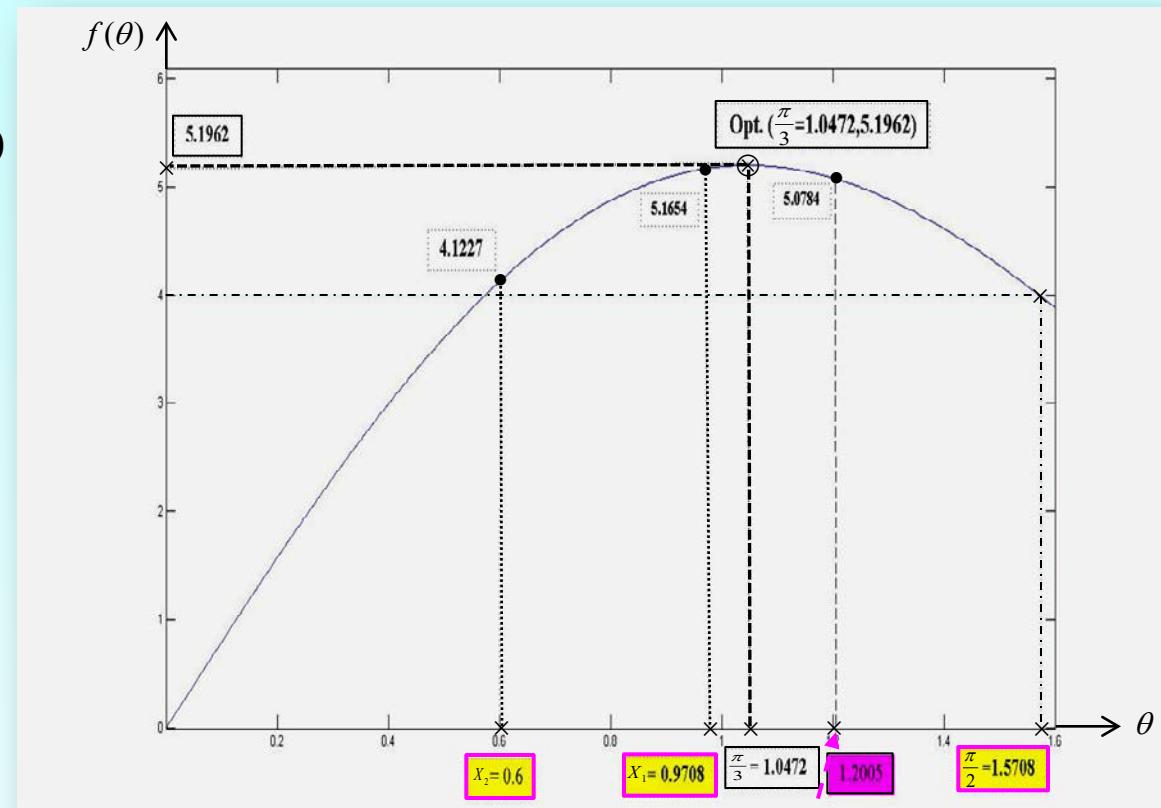
$$f(\theta) = 4 \sin \theta (1 + \cos \theta)$$

$$f(\theta) = 4 \sin \theta + 2 \sin(2\theta)$$

$$f'(\theta) = 4 \cos \theta + 4 \cos(2\theta) = 0$$

$$\Rightarrow 4 \cos \theta + 4[2 \cos^2 \theta - 1] = 0$$

Hence, $\theta_{opt.} = \frac{\pi}{3}$ after solving quadratic equation, with initial guess = (0, 1.5708 rad)



1st = Initial Iteration

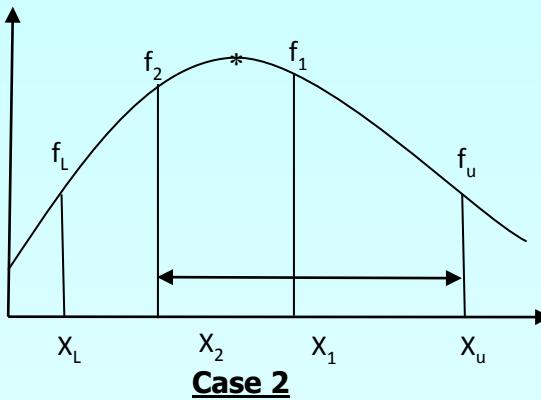
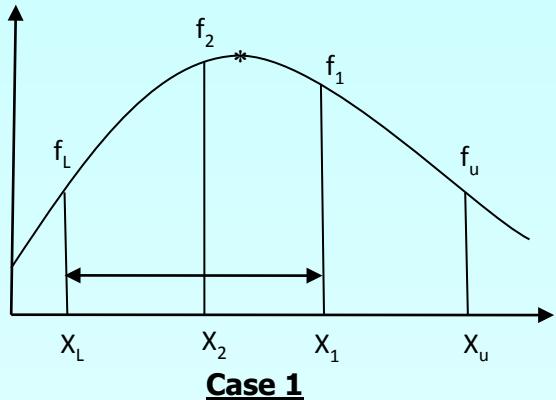


Second Iteration



Only 1 new inserted location need to be completed!

Golden Section Search- Determining the new search region



- **Case1:**

If $f(x_2) > f(x_1)$ then the new interval is $[x_L, x_2, x_1]$

- **Case2:**

If $f(x_2) < f(x_1)$ then the new interval is $[x_2, x_1, x_u]$

Golden Section Search- Determining the new search region

- At each new interval ,one needs to determine only 1(not 2) new inserted location (**either compute the new x_1 ,or new x_2**)
- Max. $f(\theta) = 4 \sin \theta(1 + \cos \theta) \Leftrightarrow$ Min. $\bar{f}(\theta) = -4 \sin \theta(1 + \cos \theta)$
- It is desirable to have automated procedure to compute x_L and x_u initially.

Golden Section Search- (1-D) Line Search Method

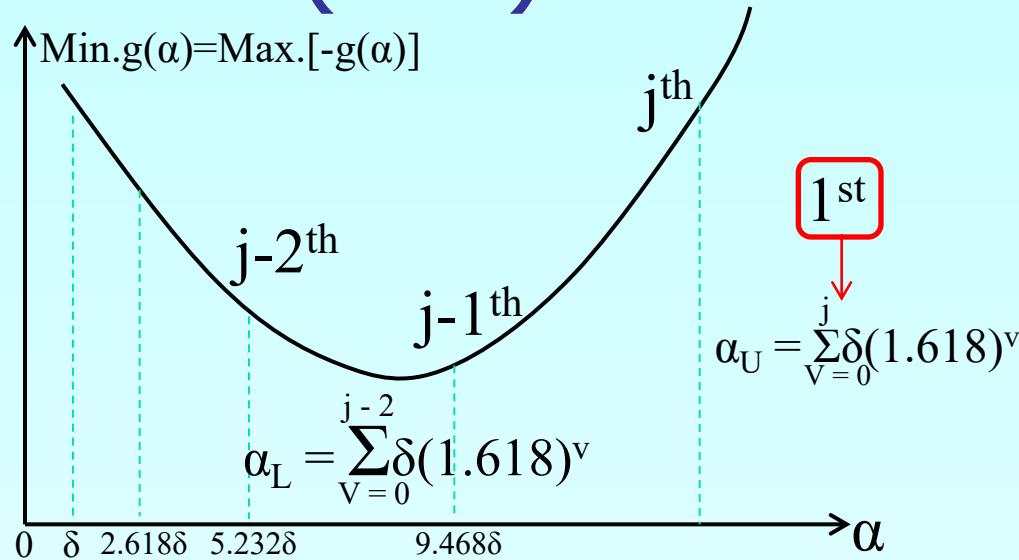


Figure 2.4 Bracketing the minimum point.

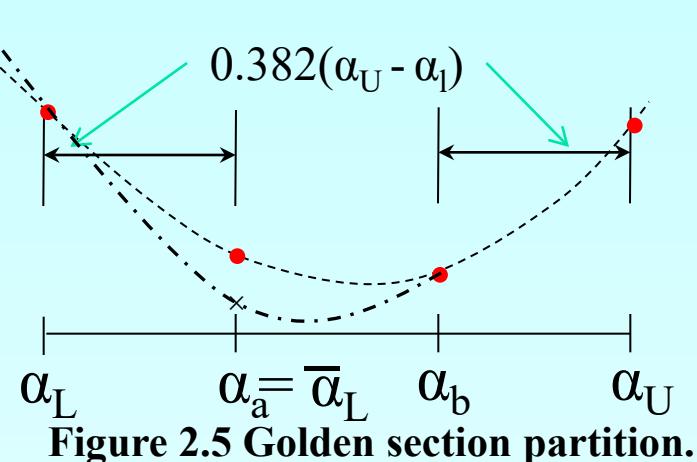
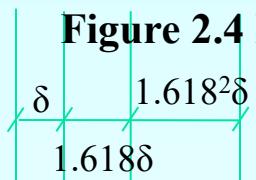


Figure 2.5 Golden section partition.

$$\alpha_a = \alpha_L + 0.382(\alpha_U - \alpha_L) = \sum_{v=0}^{j-2} \delta(1.618)^v + 0.382\delta(1.618)^{j-1}(1+1.618)$$

← 3rd

$$\alpha_a = \sum_{v=0}^{j-2} \delta(1.618)^v + 1\delta(1.618)^{j-1} = \sum_{v=0}^{j-1} \delta(1.618)^v = \text{already known !}$$

← 4th

Golden Section Search- (1-D) Line Search Method

- If $g(\alpha_a) = g(\alpha_b)$, Then the minimum will be between α_a & α_b .
- If $g(\alpha_a) > g(\alpha_b)$ as shown in Figure 2.5, Then the minimum will be between α_a & $\alpha_U \Rightarrow \bar{\alpha}_L = \alpha_a$ and $\bar{\alpha}_U = \alpha_U$.

Notice that: $\bar{\alpha}_U - \bar{\alpha}_L = \alpha_U - \alpha_a = \delta(1.618)^j$

And

$$\begin{aligned}\alpha_b - \bar{\alpha}_L &= \alpha_b - \alpha_a = (1 - 2 \times 0.382)(\alpha_U - \alpha_L) = (0.236)(\delta[1.618]^{j-1} + \delta[1.618]^j) \\ &= (0.236)(\delta[1.618]^{j-1} \times [1 + 1.618]) = 0.618(\delta[1.618]^{j-1}) \times \frac{1.618}{1.618}\end{aligned}$$

$$\alpha_b - \bar{\alpha}_L = (0.382) \times (\delta[1.618]^j) = 0.382(\bar{\alpha}_U - \bar{\alpha}_L)$$

Thus α_b (wrt $\bar{\alpha}_U$ & $\bar{\alpha}_L$) plays same role as α_a (wrt α_U & α_L) !!

Golden Section Search- (1-D) Line Search Method

Step 1 : For a chosen small step size δ in a , say $\delta = +10^{-2} \rightarrow 10^{-1}$, let j be the smallest integer such that $g(\sum_{V=0}^j \delta(1.618)^V) > g(\sum_{V=0}^{j-1} \delta(1.618)^V)$

The upper and lower bound on a^i are $\alpha_U = \sum_{V=0}^J \delta(1.618)^V$ and $\alpha_L = \sum_{V=0}^{j-2} \delta(1.618)^V$.

Step 2: Compute $g(\alpha_b)$, where $\alpha_a = \alpha_L + 0.382(\alpha_U - \alpha_L)$, and $\alpha_b = \alpha_L + 0.618(\alpha_U - \alpha_L)$.

Note that $\alpha_a = \sum_{V=0}^{j-1} \delta(1.618)^V$, so $g(\alpha_a)$ is already known.

Step 3: Compare $g(\alpha_a)$ and $g(\alpha_b)$ and go to Step 4, 5, or 6.

Step 4: If $g(\alpha_a) < g(\alpha_b)$, then $\alpha_L \leq a^i \leq \alpha_b$. By the choice of α_a and α_b , the new points $\bar{\alpha}_L = \alpha_L$ and $\bar{\alpha}_u = \alpha_b$ have $\bar{\alpha}_b = \alpha_a$.

Compute $g(\bar{\alpha}_a)$, where $\bar{\alpha}_a = \bar{\alpha}_L + 0.382(\bar{\alpha}_u - \bar{\alpha}_L)$ and go to Step 7.

Golden Section Search- (1-D) Line Search Method

Step 5: If $g(\alpha_a) > g(\alpha_b)$, then $\alpha_a \leq \alpha^i \leq \alpha_u$. Similar to the procedure in Step 4, put $\bar{\alpha}_L = \alpha_a$ and $\bar{\alpha}_u = \alpha_u$.

Compute $g(\bar{\alpha}_b)$, where $\bar{\alpha}_b = \bar{\alpha}_L + 0.618(\bar{\alpha}_u - \bar{\alpha}_L)$ and go to Step 7.

Step 6: If $g(\alpha_a) = g(\alpha_b)$ put $\alpha_L = \alpha_a$ and $\alpha_u = \alpha_b$ and return to Step 2.

Step 7: If $\bar{\alpha}_u - \bar{\alpha}_L$ is suitably small, put $\alpha^i = \frac{1}{2}(\bar{\alpha}_u + \bar{\alpha}_L)$ and stop.

Otherwise, delete the bar symbols on $\bar{\alpha}_L, \bar{\alpha}_a, \bar{\alpha}_b$, and $\bar{\alpha}_u$ and return to Step 3.



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

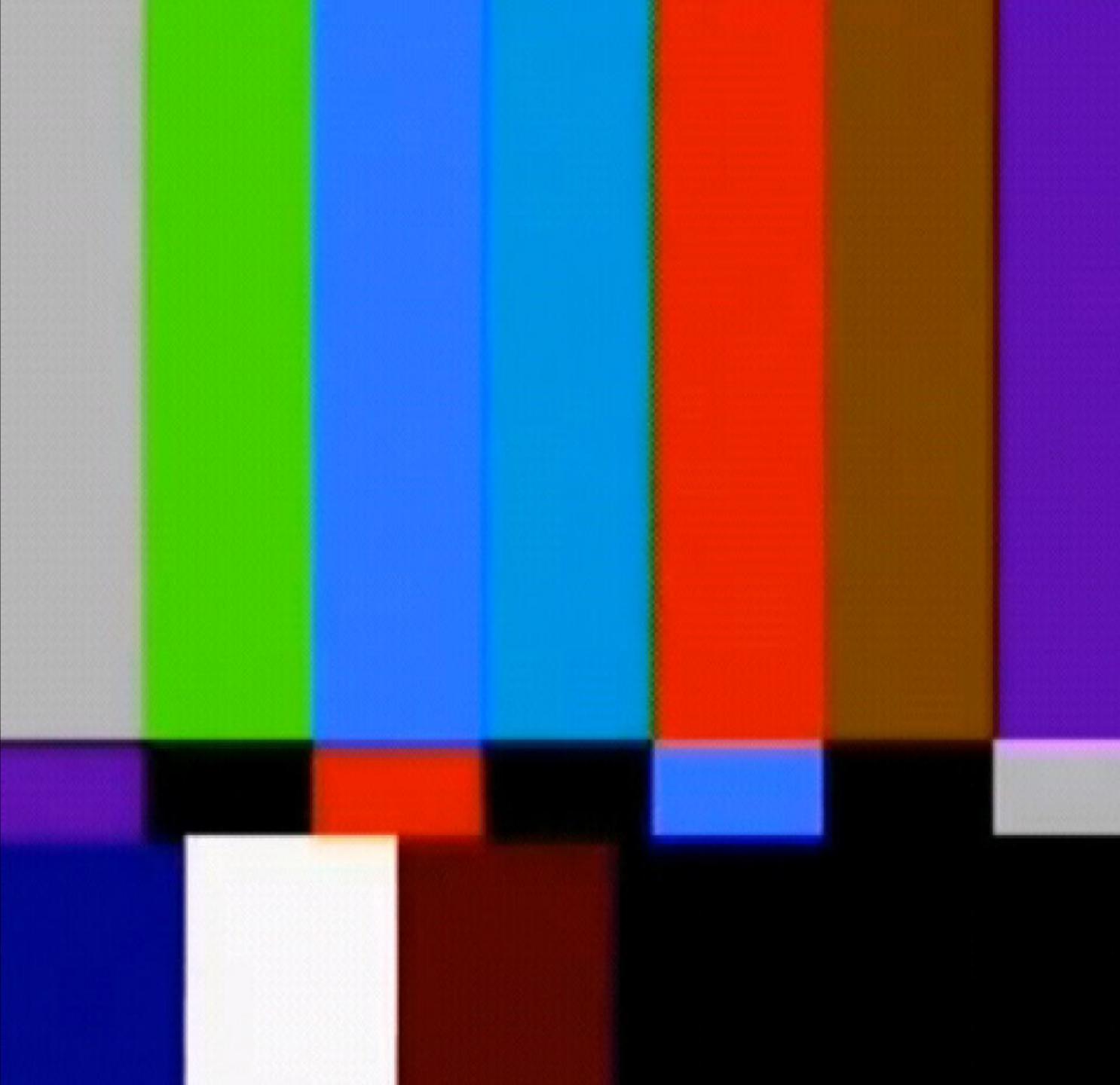
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Golden Section Search Method - Example

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Golden Section Search Method

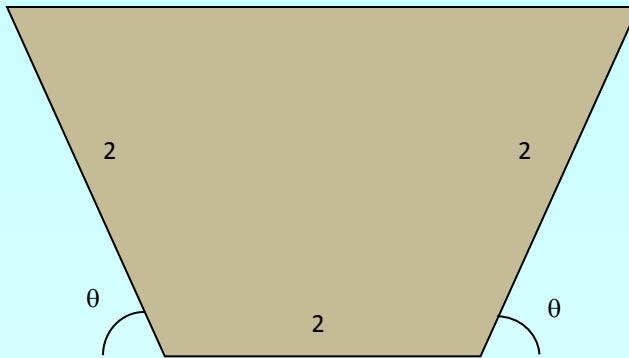
You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Example



The cross-sectional area A of a gutter with equal base and edge length of 2 is given by (**trapezoidal** area):

$$\text{Max. } f(\theta) = A = 4 \sin \theta (1 + \cos \theta) = 4 \sin \theta + 2 \sin(2\theta)$$

Find the angle θ which maximizes the cross-sectional area of the gutter. Using an initial interval of $[0, \frac{\pi}{2}]$ find the solution after 2 iterations.

Convergence achieved if "interval length" is within $\varepsilon = 0.05$

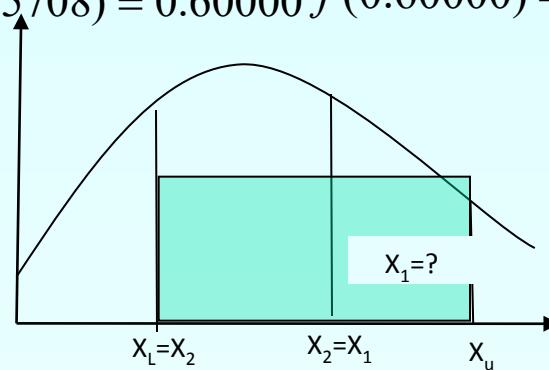
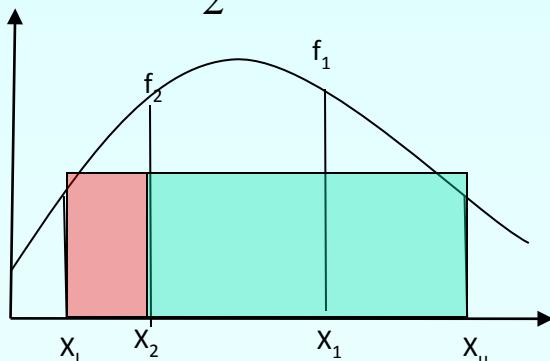
Solution

The function to be maximized is $f(\theta) = 4 \sin \theta(1 + \cos \theta)$

Iteration 1: Given the values for the boundaries of $x_L = 0$ and $x_u = \pi/2$ we can calculate the initial intermediate points as follows:

$$x_1 = x_L + \frac{\sqrt{5}-1}{2}(x_u - x_L) = 0 + \frac{\sqrt{5}-1}{2}(1.5708) = 0.97080 \quad f(0.97080) = 5.1654$$

$$x_2 = x_u - \frac{\sqrt{5}-1}{2}(x_u - x_L) = 1.5708 - \frac{\sqrt{5}-1}{2}(1.5708) = 0.60000 \quad f(0.60000) = 4.1227$$



Solution Cont

$$x_1 = x_L + \frac{\sqrt{5} - 1}{2}(x_u - x_L) = 0.60000 + \frac{\sqrt{5} - 1}{2}(1.5708 - 0.60000) = 1.2000$$

To check the stopping criteria the difference between x_u and x_L is calculated to be

$$x_u - x_L = 1.5708 - 0.60000 = 0.97080$$

Solution Cont

Iteration 2

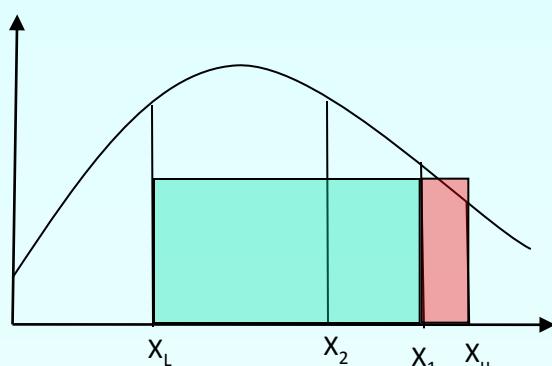
$$x_L = 0.60000$$

$$x_u = 1.5708$$

$$x_1 = 1.2000 \quad f(1.2000) = 5.0791$$

$$x_2 = 0.97080 \quad f(0.97080) = 5.1654$$

$$f(x_1) < f(x_2)$$



$$x_L = 0.60000$$

$$x_u = 1.2000$$

$$x_1 = 0.97080$$

$$x_2 = x_u - \frac{\sqrt{5}-1}{2}(x_u - x_L) = 1.2000 - \frac{\sqrt{5}-1}{2}(1.2000 - 0.6000) = 0.82918$$

$$\frac{x_u + x_L}{2} = 1.2000 + 0.6000 = 0.9000$$

Theoretical Solution and Convergence

Iteration	x_l	x_u	x_1	x_2	$f(x_1)$	$f(x_2)$	ε
1	0.0000	1.5714	0.9712	0.6002	5.1657	4.1238	1.5714
2	0.6002	1.5714	1.2005	0.9712	5.0784	5.1657	0.9712
3	0.6002	1.2005	0.9712	0.8295	5.1657	4.9426	0.6002
4	0.8295	1.2005	1.0588	0.9712	5.1955	5.1657	0.3710
5	0.9712	1.2005	1.1129	1.0588	5.1740	5.1955	0.2293
6	0.9712	1.1129	1.0588	1.0253	5.1955	5.1937	0.1417
7	1.0253	1.1129	1.0794	1.0588	5.1908	5.1955	0.0876
8	1.0253	1.0794	1.0588	1.0460	5.1955	5.1961	0.0541
9	1.0253	1.0588	1.0460	1.0381	5.1961	5.1957	0.0334

$$\frac{x_u + x_L}{2} = \frac{1.0253 + 1.0588}{2} = 1.0420 \quad f(1.0420) = 5.1960$$

The theoretically optimal solution to the problem happens at exactly 60 degrees which is 1.0472 radians and gives a maximum cross-sectional area of 5.1962.



THE END

<http://nm.mathforcollege.com>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

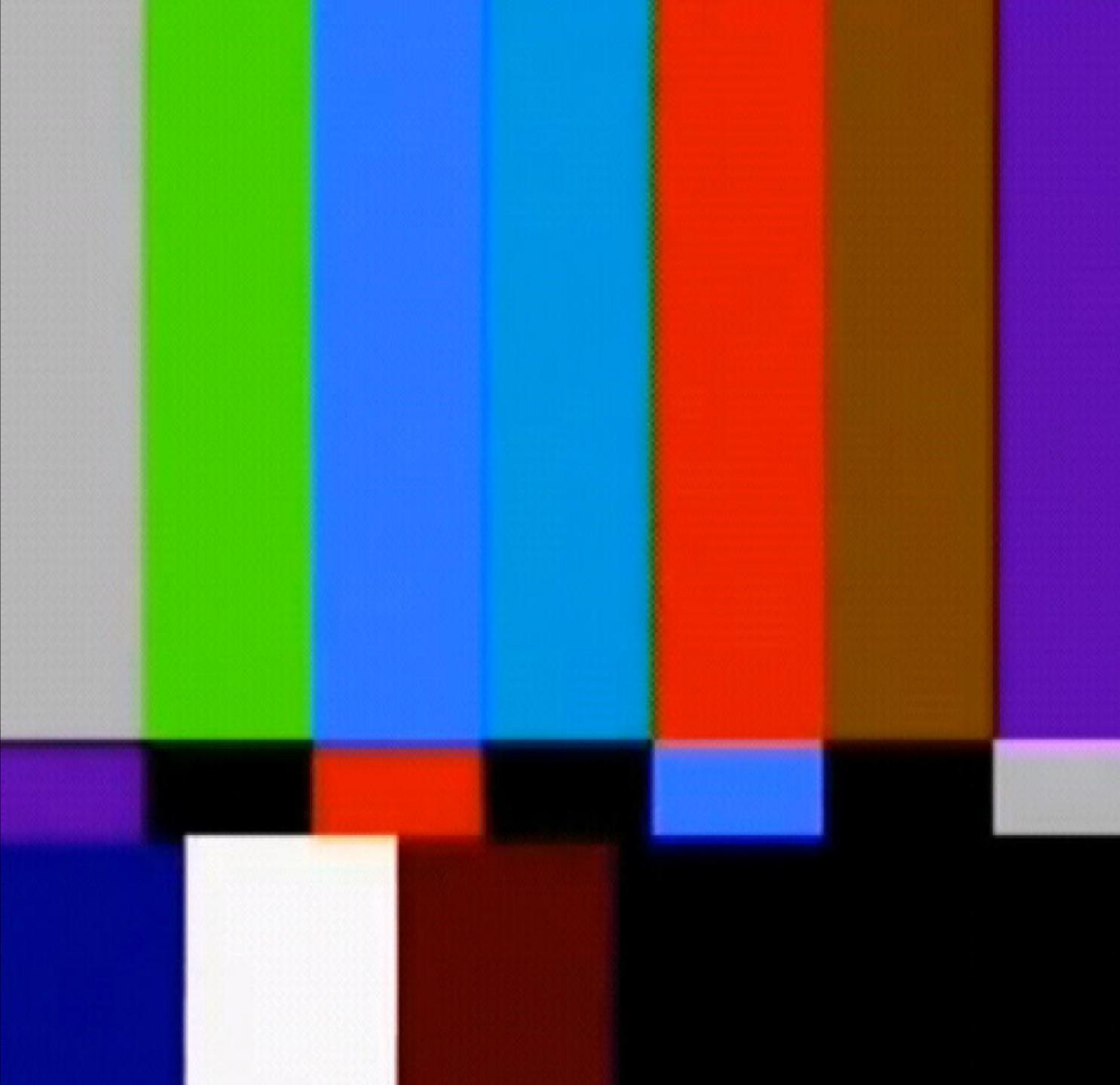
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.





Numerical Methods

Newton's Method for One -
Dimensional Optimization -
Theory

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Newton's Method for One-Dimensional Optimization

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Newton's Method-Overview

- Open search method
- A good initial estimate of the solution is required
- The objective function must be twice differentiable
- Unlike Golden Section Search method
 - Lower and upper search boundaries are not required (open vs. bracketing)
 - May not converge to the optimal solution

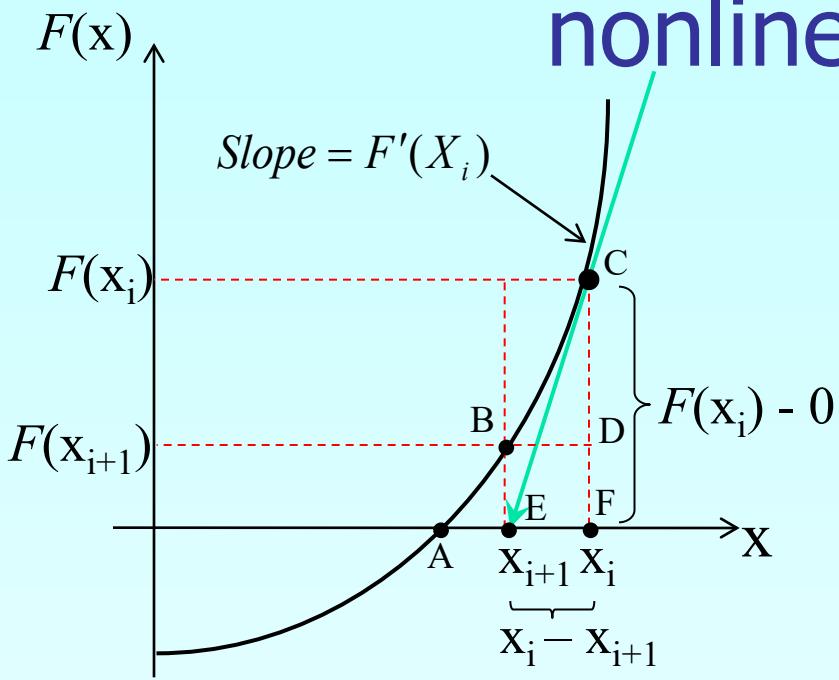
Newton's Method-How it works

- The derivative of the function $f_{Opt.}(x)$, Nonlinear root finding equation $f'(x) = 0 = F(x)$ at the function's maximum and minimum.
- The minima and the maxima can be found by applying the Newton-Raphson method to the derivative, essentially obtaining

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

- Next slide will explain how to get/derive the above formula

Newton's Method-To find root of a nonlinear equation



Slope @ pt. C $\approx \frac{F(X_i) - F(X_{i+1})}{X_i - X_{i+1}}$

We "wish" that in the next iteration x_{i+1} will be the root, or $F(X_{i+1}) = 0$.

Thus:

$$\text{Slope @ pt. C} = \frac{F(X_i) - 0}{X_i - X_{i+1}}$$

$$\text{Or } F'(X_i) = \frac{F(X_i)}{X_i - X_{i+1}}$$

Hence:

$$X_{i+1} = X_i - \frac{F(X_i)}{F'(X_i)}$$

N-R Equation

Newton's Method-To find root of a nonlinear equation

- If $F(x) \equiv f'(x)$, then $X_{i+1} = X_i - \frac{f'(X_i)}{f''(X_i)}$.
- For Multi-variable case ,then N-R method becomes

$$\vec{X}_{i+1} = \vec{X}_i - [f''(\vec{X}_i)]^{-1} \times \nabla \vec{f}(\vec{X}_i)$$

Newton's Method-Algorithm

Initialization: Determine a reasonably good estimate for the maxima or the minima of the function $f(x)$.

Step 1. Determine $f'(x)$ and $f''(x)$.

Step 2. Substitute x_i (initial estimate x_0 for the first iteration) and $f'(x)$ into $f''(x)$

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

to determine x_{i+1} and the function value in iteration i .

Step 3. If the value of the first derivative of the function is zero then you have reached the optimum (maxima or minima). Otherwise, repeat Step 2 with the new value of x_i



THE END

<http://nm.mathforcollege.com>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

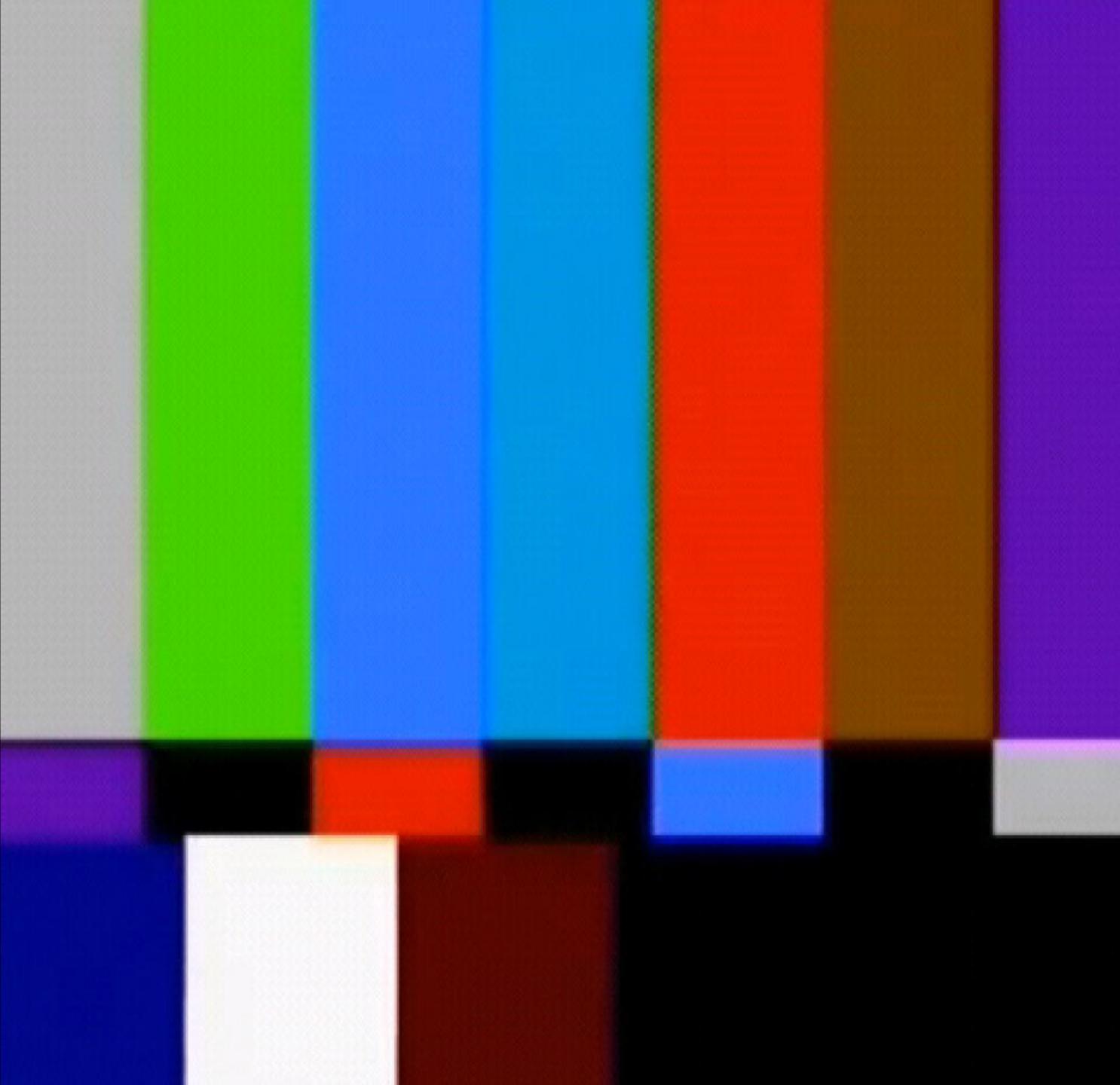
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Newton's Method for One - Dimensional Optimization - Example

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Newton's Method for One-Dimensional Optimization

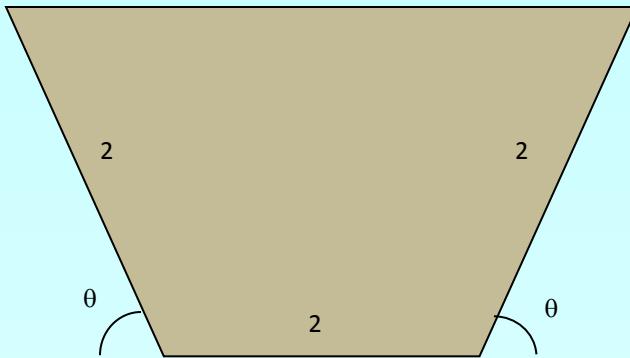
You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Example



The cross-sectional area A of a gutter with equal base and edge length of 2 is given by

$$A = 4 \sin \theta(1 + \cos \theta)$$

Find the angle θ which maximizes the cross-sectional area of the gutter.

Solution

The function to be maximized is $f(\theta) = 4 \sin \theta(1 + \cos \theta)$

$$f'(\theta) = 4(\cos \theta + \cos^2 \theta - \sin^2 \theta)$$

$$f''(\theta) = -4 \sin \theta(1 + 4 \cos \theta)$$

Iteration 1: Use $\theta_0 = \frac{\pi}{4} = 0.7854 \text{ rad}$ as the initial estimate of the solution

$$\theta_1 = \frac{\pi}{4} - \frac{4(\cos \frac{\pi}{4} + \cos^2 \frac{\pi}{4} - \sin^2 \frac{\pi}{4})}{-4 \sin \frac{\pi}{4}(1 + 4 \cos \frac{\pi}{4})} = 1.0466$$

$$f(1.0466) = 5.196151$$

Solution Cont.

Iteration 2:

$$\theta_2 = 1.0466 - \frac{4(\cos 1.0466 + \cos^2 1.0466 - \sin^2 1.0466)}{-4 \sin 1.0466(1 + 4 \cos 1.0466)} = 1.0472$$

Summary of iterations

Iteration	θ	$f'(\theta)$	$f''(\theta)$	θ estimate	$f(\theta)$
1	0.7854	2.8284	-10.8284	1.0466	5.1962
2	1.0466	0.0062	-10.3959	1.0472	5.1962
3	1.0472	1.06E-06	-10.3923	1.0472	5.1962
4	1.0472	3.06E-14	-10.3923	1.0472	5.1962
5	1.0472	1.3322E-15	-10.3923	1.0472	5.1962

Remember that the actual solution to the problem is at 60 degrees or 1.0472 radians.



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

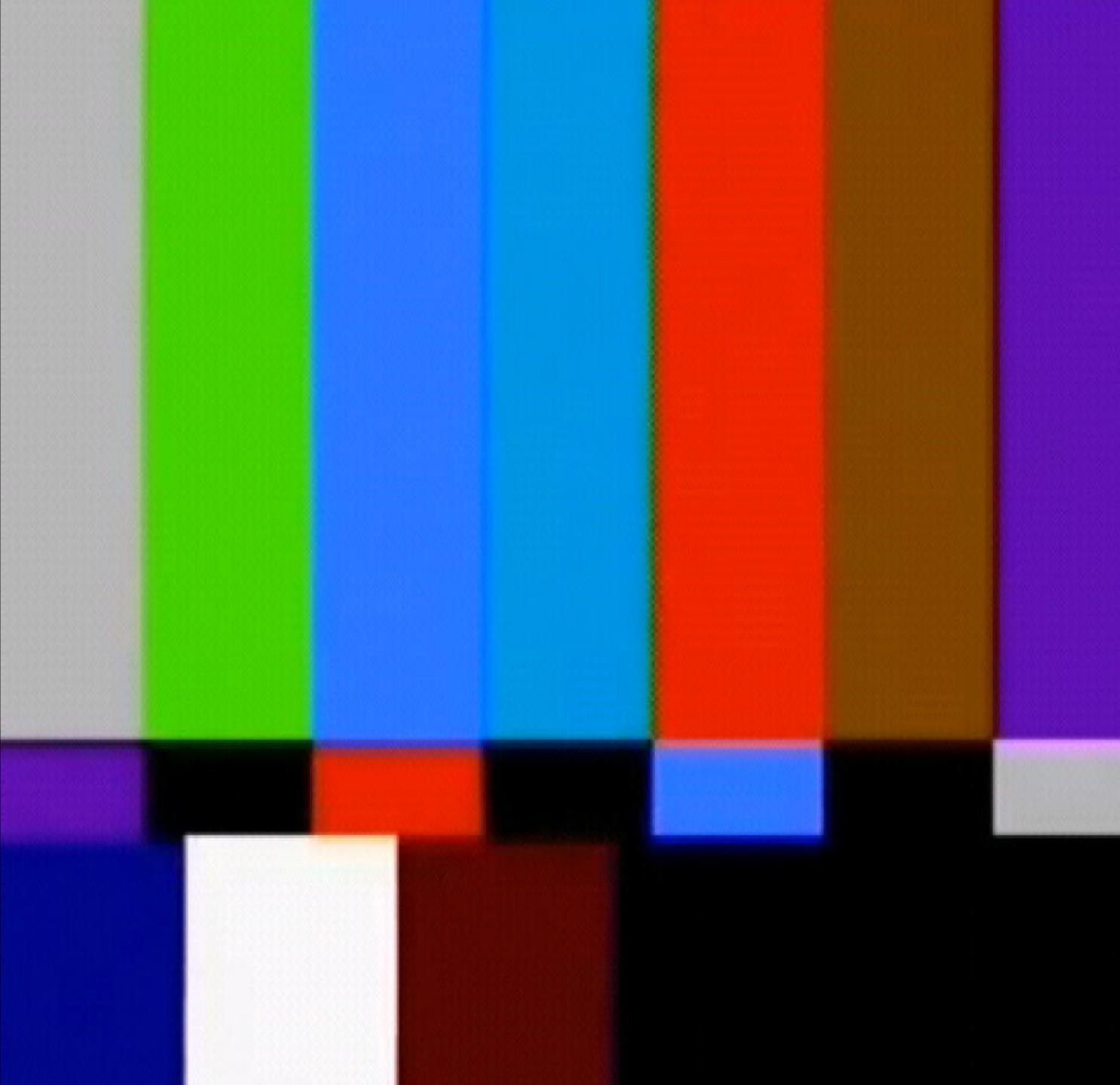
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Multi Dimensional Direct Search Methods - Theory

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Multi Dimensional Direct Search Methods

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Multi Dimensional Direct Search Methods Method-Overview

- Obvious approach is to enumerate all possible solutions and find the min or the max.
- Very generally applicable but computationally complex
- Direct search methods are open
- A good initial estimate of the solution is required
- The objective function need not be differentiable

Coordinate Cycling Method

- Starts from an initial point and looks for an optimal solution along each coordinate direction iteratively.
- For a function with two independent variables x and y , starting at an initial point (x_0, y_0) , the first iteration will first move along direction $(1, 0)$ until an optimal solution is found for the function .
- The next search involves searching along the direction $(0,1)$ to determine the optimal value for the function.
- Once searches in all directions are completed, the process is repeated in the next iteration and iterations continue until convergence occurs.
- The search along each coordinate direction can be conducted using anyone of the one-dimensional search techniques previously covered.

Multi Dimensional Direct Search

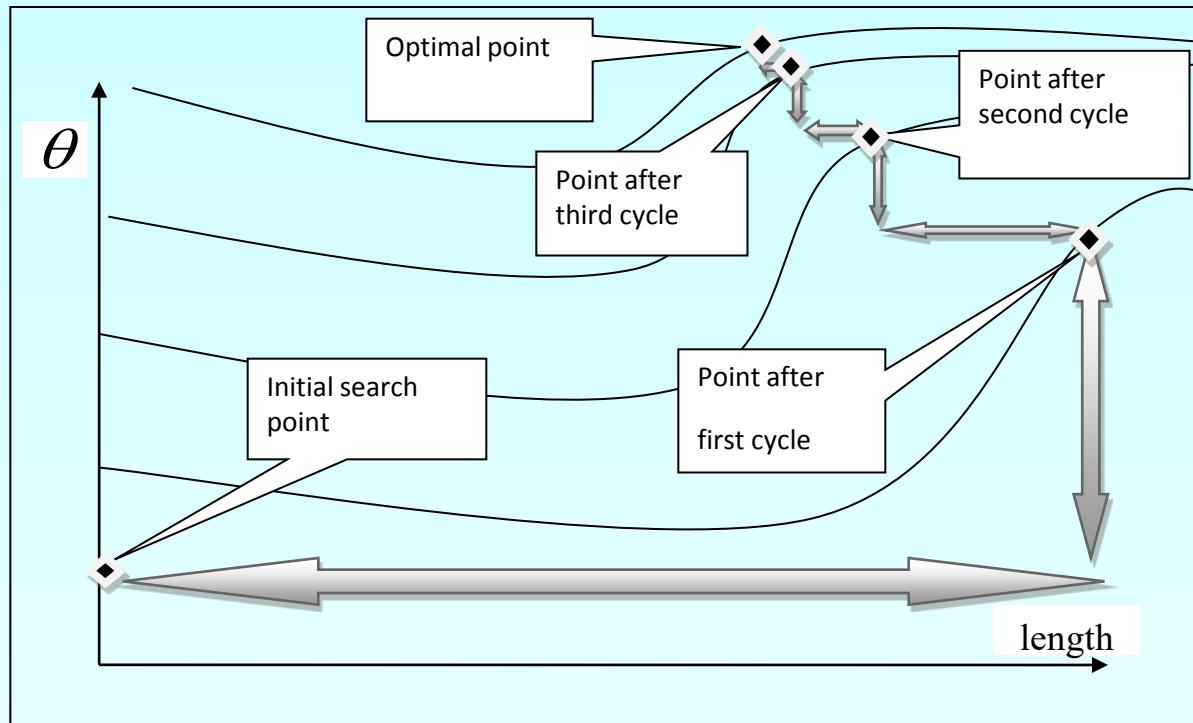


Figure 1: Visual Representation of a Multidimensional Search



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

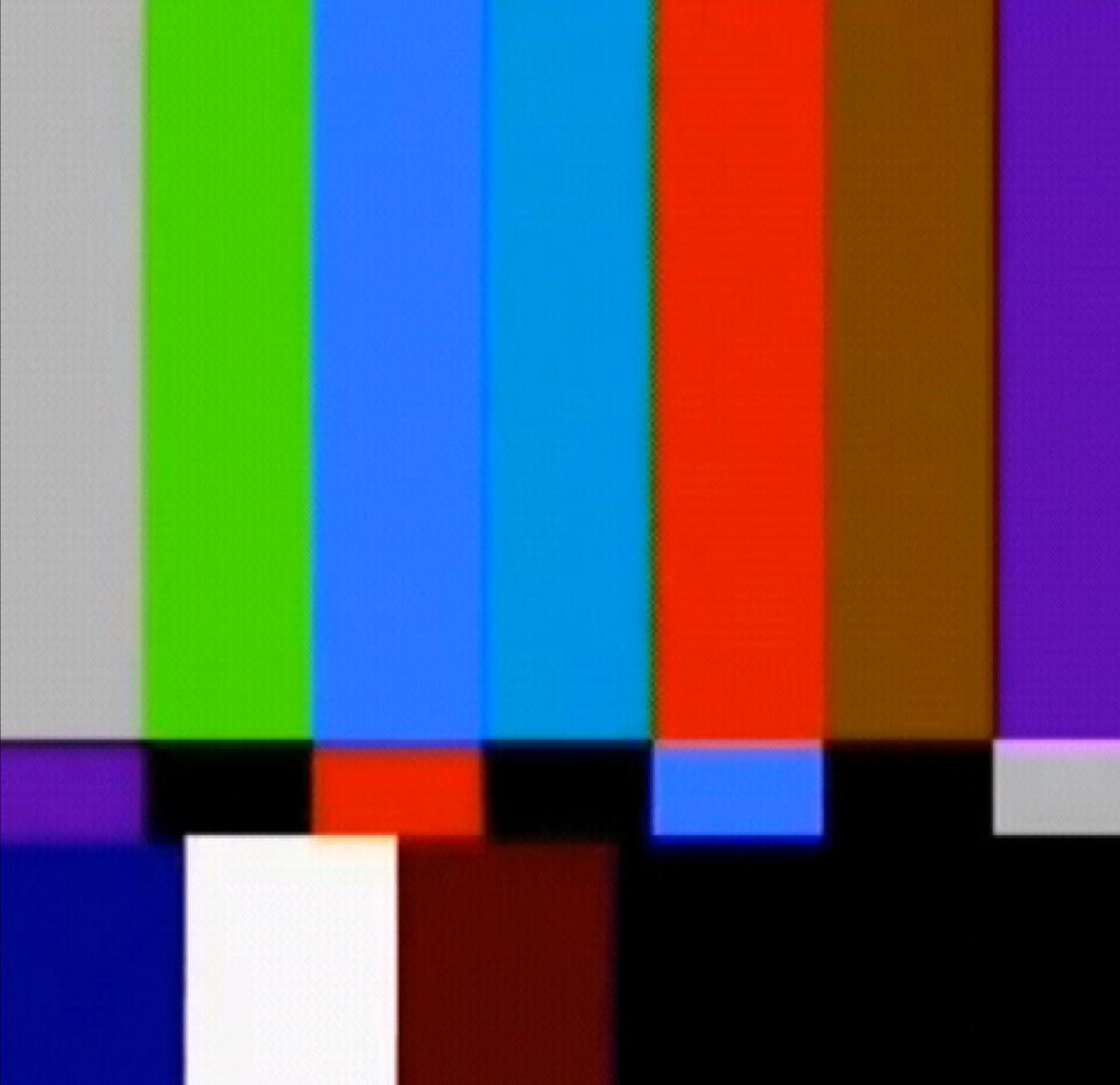
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Multi Dimensional Direct Search Methods - Example

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Multi Dimensional Direct Search Methods

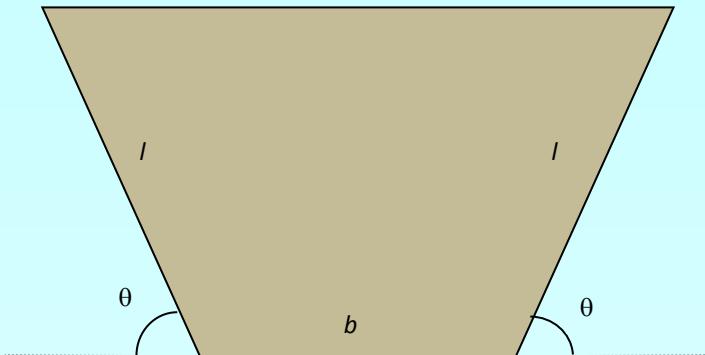
You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Example



The cross-sectional area A of a gutter with base length b and edge length of l is given by

$$A = \frac{1}{2}(b + b + 2l \cos \theta)l \sin \theta$$

Assuming that the width of material to be bent into the gutter shape is 6 ($6 = b + 2l$), find the angle θ and edge length l which maximizes the cross-sectional area of the gutter.

Notes :

- To get the maximum cross-sectional area $0 \leq \theta \leq \frac{\pi}{2} \approx 1.5708$
- To have a physical meaning, $l_{\max} \leq 3$ (otherwise "b" will have a negative value!)

Solution

- Recognizing that the base length b can be expressed as $b = 6 - 2l$, we can re-write the area function as

$$f(l, \theta) = (6 - 2l + l \cos \theta)l \sin \theta$$

- Use $(0, \frac{\pi}{6} = 0.5236)$ as the initial estimate of the solution and use Golden Search method to determine optimal solution in each dimension.
- To use the golden search method we will use $l = 0 \rightarrow 3$ as the lower and upper bounds for the search region

Solution Cont.

Iteration 1 along (1,0)

$$f(l, \theta = 0.5236\text{rad}) = [6 - 2l + l \cos(0.5236)]l \sin(0.5236)$$

Iteration	x_1	x_u	x_1	x_2	$f(x_1)$	$f(x_2)$	ε
1	0.0000	3.0000	1.8541	1.1459	3.6143	2.6941	3.0000
2	1.1459	3.0000	2.2918	1.8541	3.8985	3.6143	1.8541
3	1.8541	3.0000	2.5623	2.2918	3.9655	3.8985	1.1459
4	2.2918	3.0000	2.7295	2.5623	3.9654	3.9655	0.7082
5	2.2918	2.7295	2.5623	2.4590	3.9655	3.9497	0.4377
6	2.4590	2.7295	2.6262	2.5623	3.9692	3.9655	0.2705
7	2.5623	2.7295	2.6656	2.6262	3.9692	3.9692	0.1672
8	2.5623	2.6656	2.6262	2.6018	3.9692	3.9683	0.1033
9	2.6018	2.6656	2.6412	2.6262	3.9694	3.9692	0.0639
10	2.6262	2.6656	2.6506	2.6412	3.9694	3.9694	0.0395

The maximum area of 3.6964 is obtained at point (2.6459,0.5236) by using either Golden Section Method (see above table) or analytical method (set $\frac{df}{dl} = 0 \Rightarrow l = 2.6459$).

Solution Cont.

Iteration 1 along (0,1)

$$f(l = 2.6459, \theta) = [6 - 2 \times 2.6459 + 2.6459 \cos \theta] \times 2.6459 \sin \theta$$

Iteration	x_1	x_u	x_1	x_2	$f(x_1)$	$f(x_2)$	ε
1	0.0000	1.5714 = $\frac{\pi}{2}$	0.9712	0.6002	4.8084	4.3215	1.5714
2	0.6002	1.5714	1.2005	0.9712	4.1088	4.8084	0.9712
3	0.6002	1.2005	0.9712	0.8295	4.8084	4.8689	0.6002
4	0.6002	0.9712	0.8295	0.7419	4.8689	4.7533	0.3710
5	0.7419	0.9712	0.8836	0.8295	4.8816	4.8689	0.2293
6	0.8295	0.9712	0.9171	0.8836	4.8672	4.8816	0.1417
7	0.8295	0.9171	0.8836	0.8630	4.8816	4.8820	0.0876
8	0.8295	0.8836	0.8630	0.8502	4.8820	4.8790	0.0541
9	0.8502	0.8836	0.8708	0.8630	4.8826	4.8820	0.0334

The maximum area of 4.8823 is obtained at point (2.6459, 0.87), by using either Golden Section Method (see above table) or analytical method (set $\frac{df}{d\theta} = 0 \Rightarrow \theta = 0.87$).

Solution Cont.

- Since this is a two-dimensional search problem, the two searches along the two dimensions completes the first iteration.
- In the next iteration we return to the first dimension for which we conducted a search and start the second iteration with a search along this dimension.
- After the fifth cycle, the optimal solution of (2.0016, 1.0420) with an area of 5.1960 is obtained.
- The optimal solution to the problem is exactly 60 degrees which is 1.0472 radians and an edge and base length of 2 inches. The area of the gutter at this point is 5.1962.



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

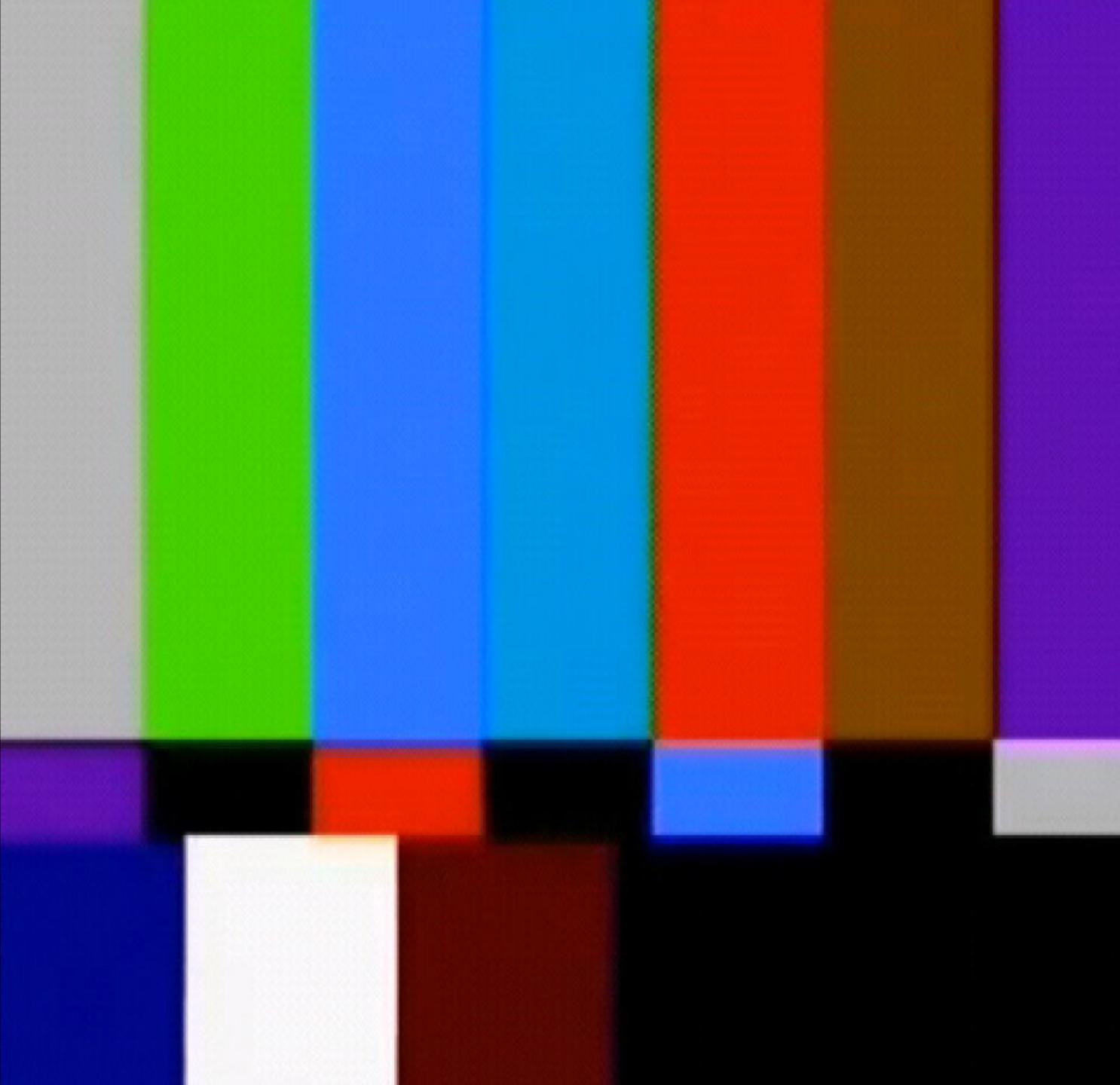
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.





Numerical Methods

Multidimensional Gradient Methods in Optimization- Theory

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Multidimensional Gradient Methods in Optimization

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Multidimensional Gradient Methods - Overview

- Use information from the derivatives of the optimization function to guide the search
- Finds solutions quicker compared with direct search methods
- A good initial estimate of the solution is required
- The objective function needs to be differentiable

Gradients

- The *gradient* is a vector operator denoted by ∇ (referred to as “del”)
- When applied to a function , it represents the functions directional derivatives
- The gradient is the special case where the direction of the gradient is the direction of most or the *steepest ascent/descent*
- The gradient is calculated by

$$\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j}$$

Gradients-Example

Calculate the gradient to determine the direction of the steepest slope at point (2, 1) for the function

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

Solution: To calculate the gradient we would need to calculate

$$\frac{\partial f}{\partial x} = 2xy^2 = 2(2)(1)^2 = 4 \quad \frac{\partial f}{\partial y} = 2x^2y = 2(2)^2(1) = 8$$

which are used to determine the gradient at point (2,1) as

$$\nabla f = 4\mathbf{i} + 8\mathbf{j}$$

Hessians

- The *Hessian* matrix or just the *Hessian* is the Jacobian matrix of second-order partial derivatives of a function.
- The determinant of the Hessian matrix is also referred to as the Hessian.
- For a two dimensional function the Hessian matrix is simply

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

Hessians cont.

The determinant of the Hessian matrix denoted by $|H|$ can have three cases:

1. If $|H| > 0$ and $\partial^2 f / \partial^2 x^2 > 0$ then $f(x, y)$ has a local minimum.
2. If $|H| > 0$ and $\partial^2 f / \partial^2 x^2 < 0$ then $f(x, y)$ has a local maximum.
3. If $|H| < 0$ then $f(x, y)$ has a saddle point.

Hessians-Example

Calculate the hessian matrix at point (2, 1) for the function $f(x, y) = x^2y^2$

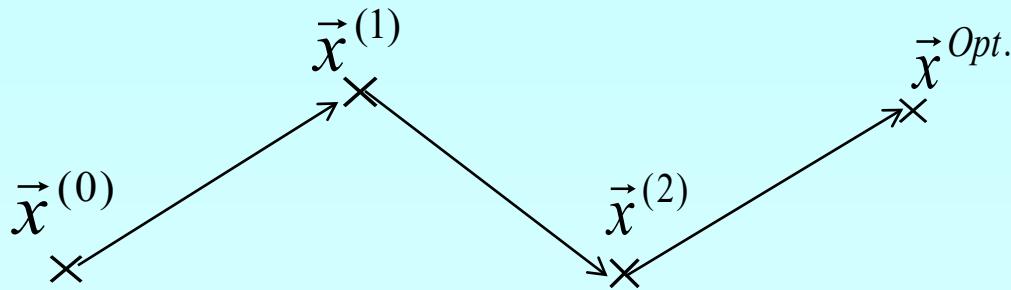
Solution: To calculate the Hessian matrix; the partial derivatives must be evaluated as

$$\frac{\partial^2 f}{\partial x^2} = 2y^2 = 2(1)^2 = 2 \quad \frac{\partial^2 f}{\partial y^2} = 2x^2 = 2(2)^2 = 8 \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = 4xy = 4(2)(1) = 8$$

resulting in the Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 8 \\ 8 & 8 \end{bmatrix}$$

Steepest Ascent/Descent Method



- **Step** Starts from an initial guessed point $\vec{x}^{(i=0)}$ and looks for a local optimal solution along a gradient.
- **Step2** The gradient at the initial solution is calculated(or finding the direction to travel),compute $\nabla \vec{f}_{\min} = \frac{\partial f_{\min}}{\partial x_k} = \left[\frac{\partial f_{\min}}{\partial x_1}, \frac{\partial f_{\min}}{\partial x_2}, \dots, \frac{\partial f_{\min}}{\partial x_k}, \dots \right]$ •

Steepest Ascent/Descent Method

- **Step3** Find the step size “h” along the Calculated (gradient) direction (using Golden Section Method or Analytical Method).
- **Step4:** A new solution is found at the local optimum along the gradient ,compute
$$\vec{x}^{i+1} = \vec{x}^{(i)} + h \vec{\nabla f}_{\min} \Big|_{\vec{x}_{(i)}}$$
- **Step5:** If “converge”,such as $\vec{\nabla f}_{x^{i+1}} \leq (\varepsilon_{tol} = 10^{-5})$ then stop. Else, return to step 2 (using the newly computed point $\vec{x}^{(i+1)}$).



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate
<http://nm.mathforcollege.com>

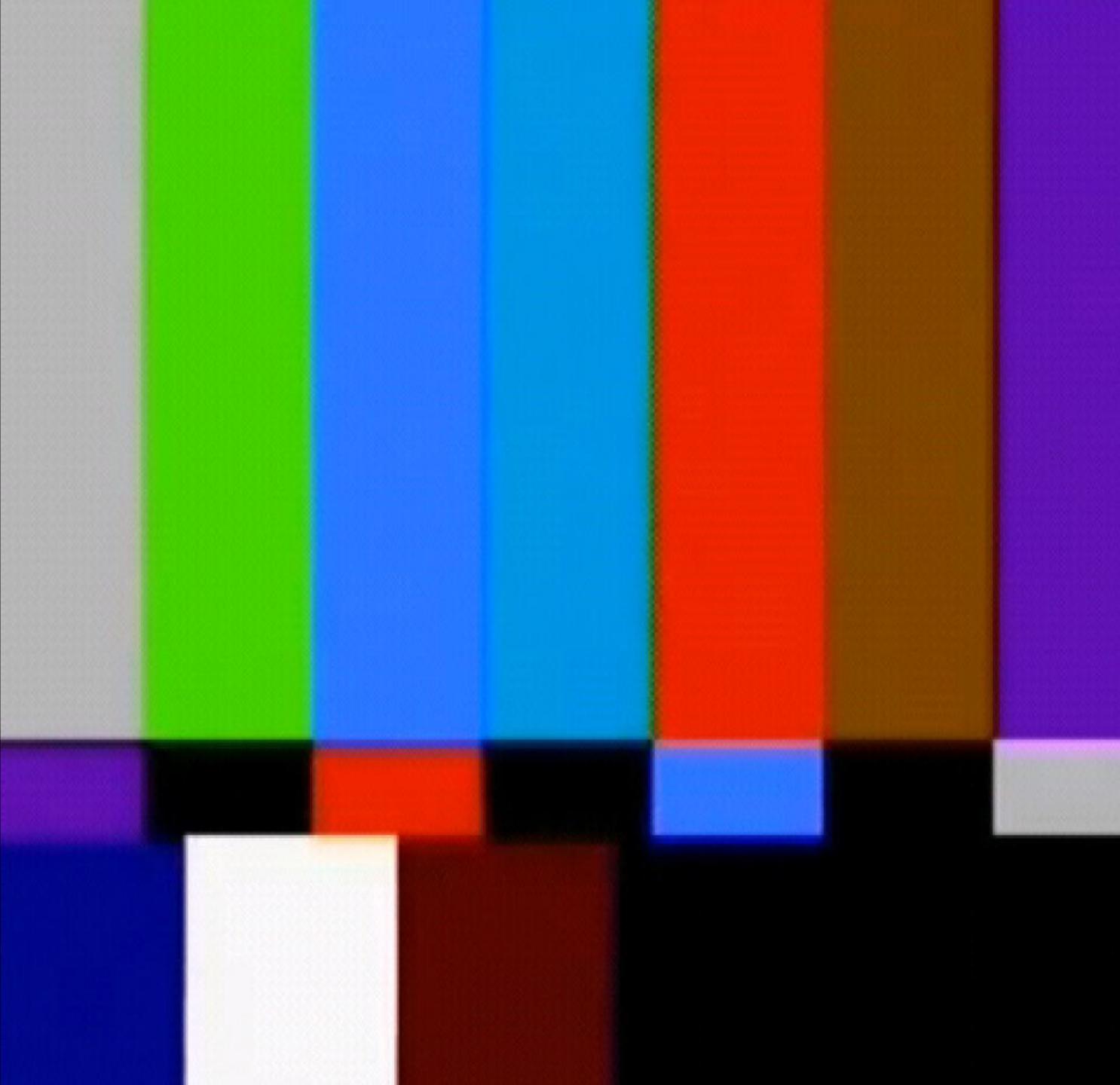
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.





Numerical Methods

Multidimensional Gradient Methods in Optimization- Example

<http://nm.mathforcollege.com>

For more details on this topic

- Go to <http://nm.mathforcollege.com>
- Click on Keyword
- Click on Multidimensional Gradient Methods in Optimization

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Example

Determine the minimum of the function

$$f(x, y) = x^2 + y^2 + 2x + 4$$

Use the point $\vec{x}^{(0)} = \begin{Bmatrix} x^{(0)} \\ y^{(0)} \end{Bmatrix} = (2, 1)$ as the initial estimate of the optimal solution.

Solution

Iteration 1: To calculate the gradient; the partial derivatives must be evaluated as

Recalled that $f(x, y) = x^2 + y^2 + 2x + 4$

$$\frac{\partial f}{\partial x} = 2x + 2 = 2(2) + 2 = 6$$

$$\frac{\partial f}{\partial y} = 2y = 2(1) = 2$$

$$\nabla f = 6\mathbf{i} + 2\mathbf{j}$$

$$\vec{x}^{(i+1)} = \vec{x}^{(i)} + h \vec{\nabla f}$$

$$\vec{x}^{(i+1)} = \begin{Bmatrix} 2 \\ 1 \end{Bmatrix} + h \begin{Bmatrix} 6 \\ 2 \end{Bmatrix} = \begin{Bmatrix} 2 + 6h \\ 1 + 2h \end{Bmatrix}$$

Solution

Now the function $f(x, y)$ can be expressed along the direction of gradient as

$$f(\vec{x}^{i+1}) = (2 + 6h)^2 + (1 + 2h)^2 + 2(2 + 6h) + 4 \equiv g(h)$$

$$g(h) = 40h^2 + 40h + 13$$

To get g_{\min} , we set $\frac{dg}{dh} = 0 = 80h + 40 \Rightarrow h^* = -0.5$

Solution Cont.

Iteration 1 continued:

This is a simple function and it is easy to determine $h^* = -0.50$ by taking the first derivative and solving for its roots.

This means that traveling a step size of $h = -0.5$ along the gradient reaches a minimum value for the function in this direction. These values are substituted back to calculate a new value for x and y as follows:

$$x = 2 + 6(-0.5) = -1$$

$$y = 1 + 2(-0.5) = 0$$

Note that $f(2,1) = 13$ $f(-1,0) = 3.0$

Solution Cont.

Iteration 2: The new initial point is $(-1, 0)$. We calculate the gradient at this point as

$$\frac{\partial f}{\partial x} = 2x + 2 = 2(-1) + 2 = 0$$

$$\frac{\partial f}{\partial y} = 2y = 2(0) = 0$$

$$\nabla f = (0) \hat{i} + (0) \hat{j}$$

Solution Cont.

This indicates that the current location is a local optimum along this gradient and no improvement can be gained by moving in any direction. The minimum of the function is at point (-1,0), and $f_{\min} = (-1)^2 + (0)^2 + 2(-1) + 4 = 3$.



THE END

<http://nm.mathforcollege.com>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://nm.mathforcollege.com>

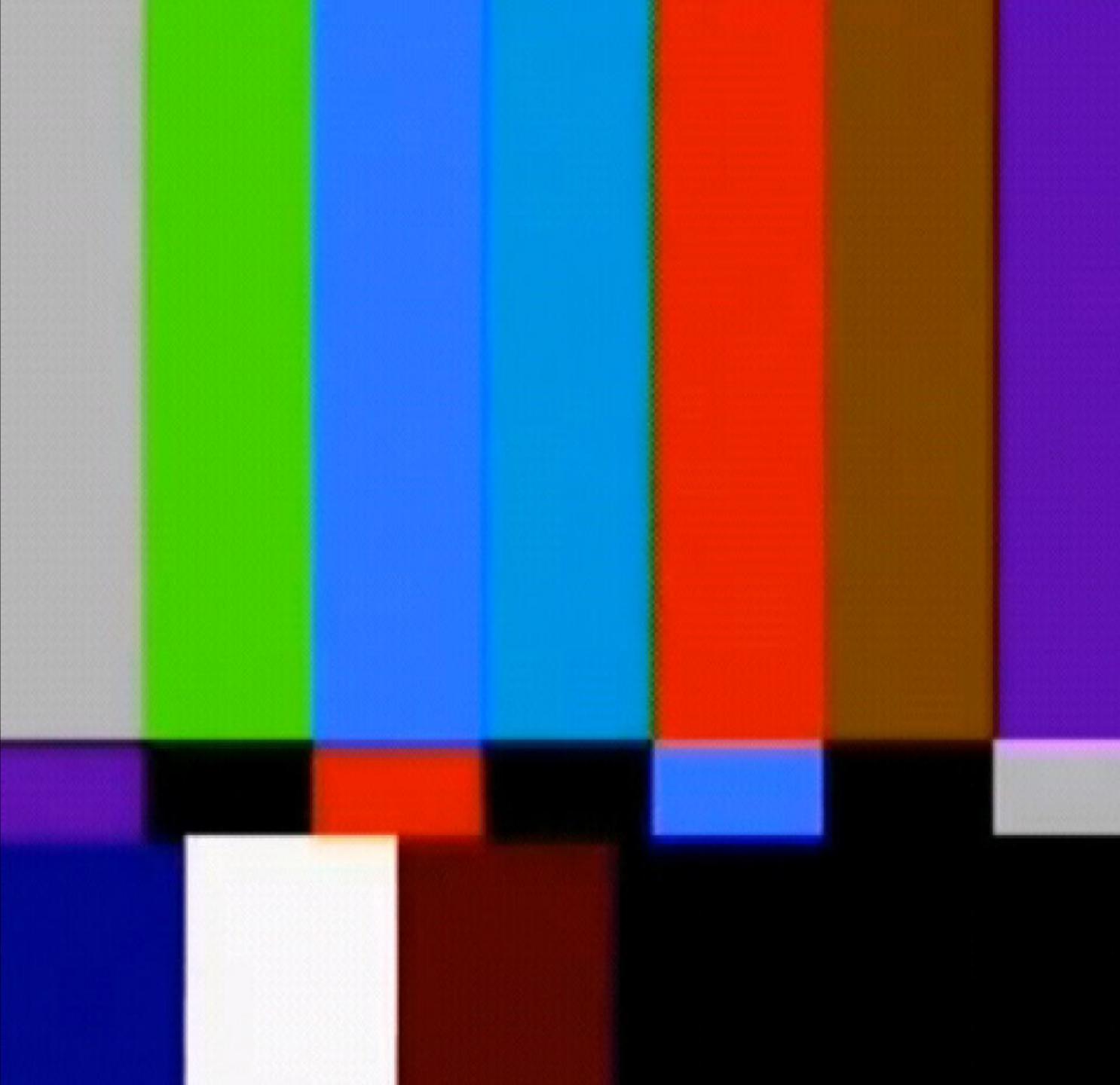
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://nm.mathforcollege.com>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Introduction to Partial Differential Equations

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM Undergraduates

What is a Partial Differential Equation ?

- Ordinary Differential Equations have only one independent variable

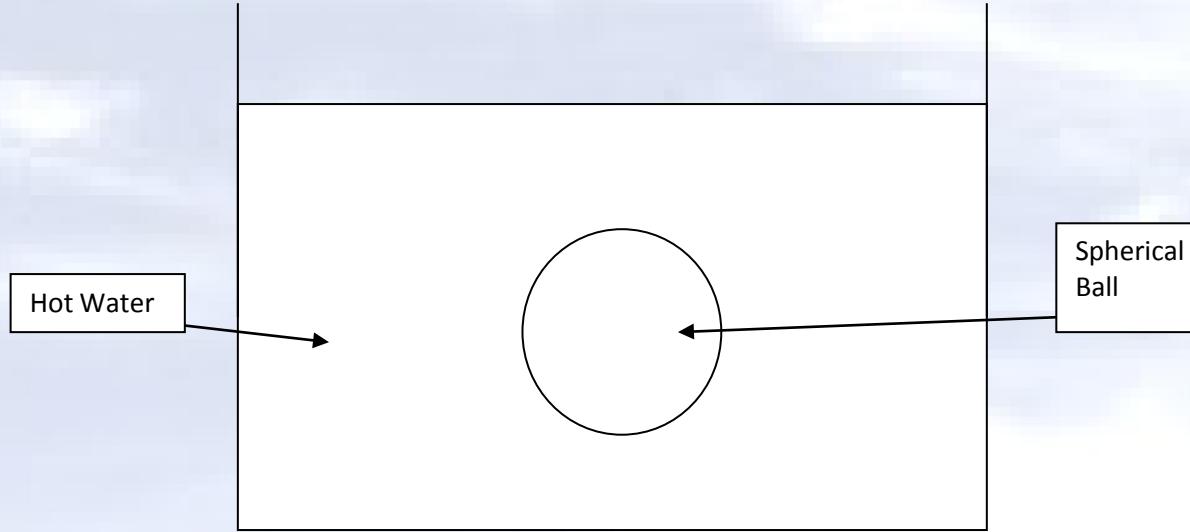
$$3 \frac{dy}{dx} + 5y^2 = 3e^{-x}, y(0) = 5$$

- Partial Differential Equations have more than one independent variable

$$3 \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = x^2 + y^2$$

- subject to certain conditions: where u is the dependent variable, and x and y are the independent variables.

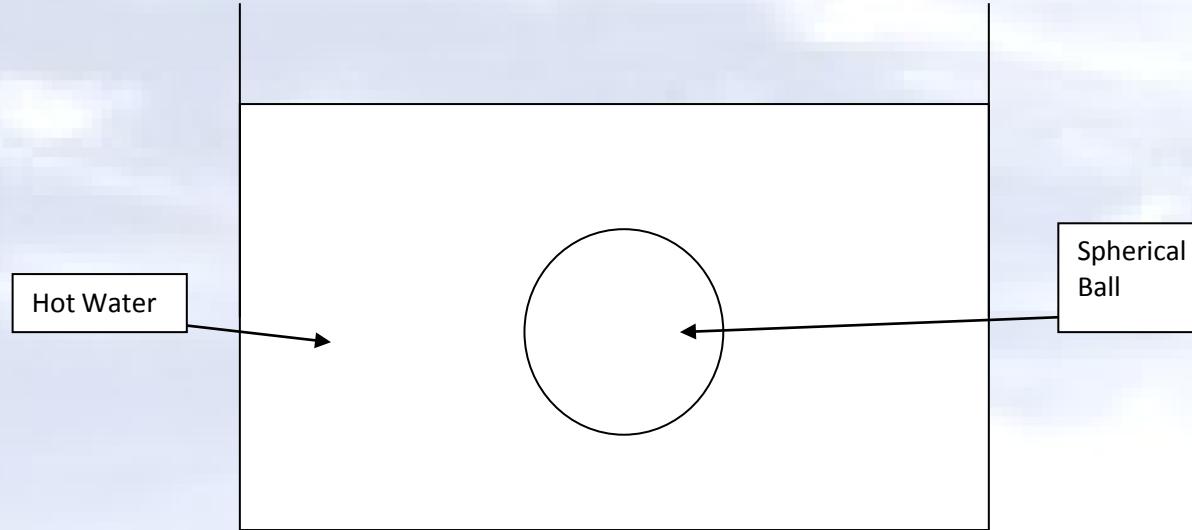
Example of an Ordinary Differential Equation



$$hA(\theta - \theta_a) = mC \frac{d\theta}{dt}$$

- **Assumption: Ball is a lumped system.**
- **Number of Independent variables:**
One (t)

Example of an Partial Differential Equation



$$\frac{k}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial T}{\partial r} \right) + \frac{k}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial T}{\partial \theta} \right) + \frac{k}{r^2 \sin^2 \theta} \frac{\partial^2 T}{\partial \phi^2} = \rho C \frac{\partial T}{\partial t}, \quad t \geq 0, \quad T(r, \theta, \phi, 0) = T_a$$

- **Assumption: Ball is not a lumped system.**
- **Number of Independent variables:**
Four (r, θ, φ, t)

Classification of 2nd Order Linear PDE's

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

where A , B , and C are functions of x and y , and D is a function of x, y, u and $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}$.

Classification of 2nd Order Linear PDE's

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

Can Be:

- Elliptic
- Parabolic
- Hyperbolic

Classification of 2nd Order Linear PDE's: Elliptic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

If $B^2 - 4AC < 0$, then equation
is elliptic.

Classification of 2nd Order Linear PDE's: Elliptic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

Example: $\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$

where, $A = 1, B = 0, C = 1$ giving

$$B^2 - 4AC = 0 - 4(1)(1) = -4 < 0$$

therefore the equation is elliptic.

Classification of 2nd Order Linear PDE's: Parabolic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

If $B^2 - 4AC = 0$, then the
equation is parabolic.

Classification of 2nd Order Linear PDE's: Parabolic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

Example: $\frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2}$

where, $A = k, B = 0, C = 0, D = -1$ giving

$$B^2 - 4AC = 0 - 4(0)(k) = 0$$

therefore the equation is parabolic.

Classification of 2nd Order Linear PDE's: Hyperbolic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

If $B^2 - 4AC > 0$, then the
equation is hyperbolic.

Classification of 2nd Order Linear PDE's: Hyperbolic

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

Example: $\frac{\partial^2 y}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 y}{\partial t^2}$

where, $A = 1, B = 0, C = -\frac{1}{c^2}$ giving

$$B^2 - 4AC = 0 - 4(1)\left(\frac{-1}{c^2}\right) = \frac{4}{c^2} > 0$$

therefore the equation is hyperbolic.

THE END

Parabolic Partial Differential Equations

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM Undergraduates

Defining Parabolic PDE's

- The general form for a second order linear PDE with two independent variables and one dependent variable is

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

- Recall the criteria for an equation of this type to be considered parabolic
 $B^2 - 4AC = 0$
- For example, examine the heat-conduction equation given by

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \quad , \text{where} \quad A = \alpha, B = 0, C = 0, D = -1$$

Then

$$\begin{aligned} B^2 - 4AC &= 0 - 4(\alpha)(0) \\ &= 0 \end{aligned}$$

thus allowing us to classify this equation as parabolic.

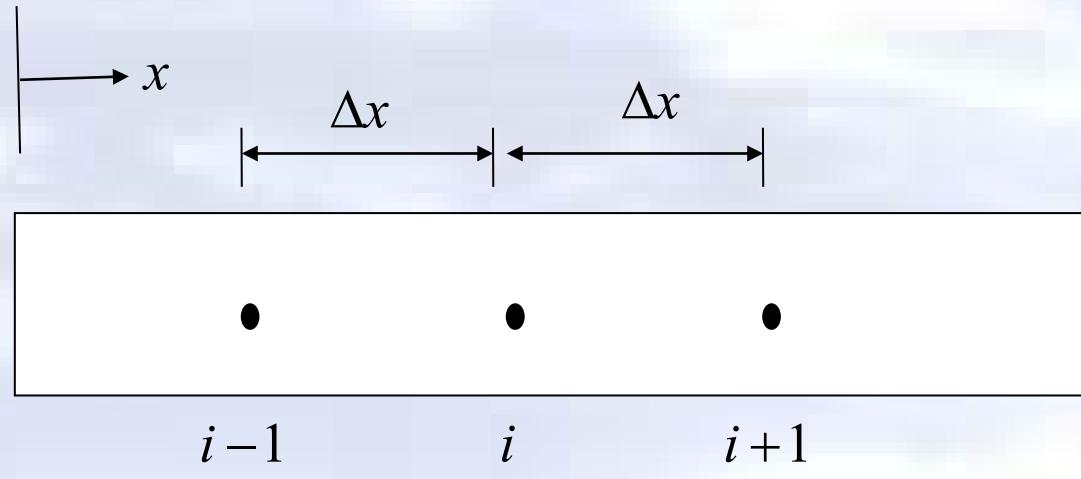
Physical Example of an Elliptic PDE



The internal temperature of a metal rod exposed to two different temperatures at each end can be found using the heat conduction equation.

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t}$$

Discretizing the Parabolic PDE



Schematic diagram showing interior nodes

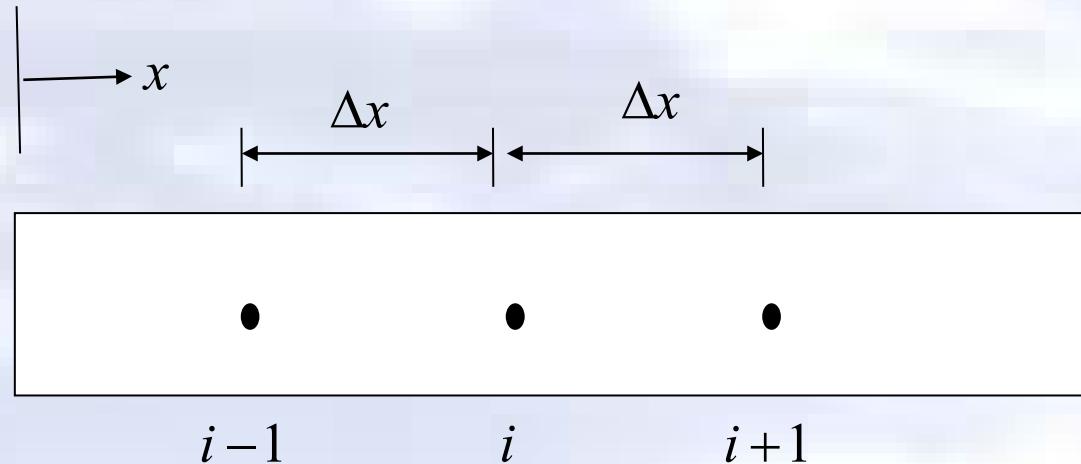
For a rod of length L divided into $n+1$ nodes $\Delta x = \frac{L}{n}$

The time is similarly broken into time steps of Δt

Hence T_i^j corresponds to the temperature at node i , that is,

$$x = (i)(\Delta x) \text{ and time } t = (j)(\Delta t)$$

The Explicit Method

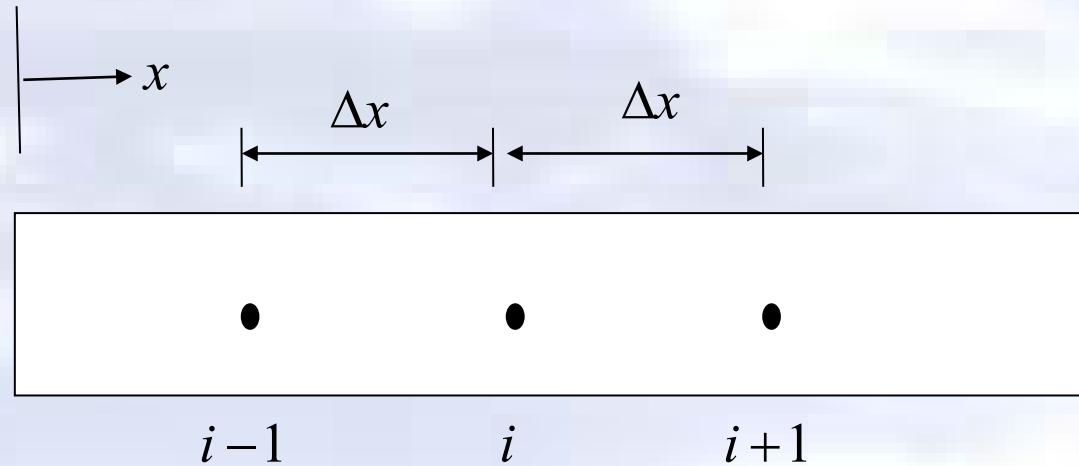


If we define $\Delta x = \frac{L}{n}$ we can then write the finite central divided difference approximation of the left hand side at a general interior node (i) as

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j} \cong \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2}$$

where (j) is the node number along the time.

The Explicit Method



The time derivative on the right hand side is approximated by the forward divided difference method as,

$$\left. \frac{\partial T}{\partial t} \right|_{i,j} \cong \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

The Explicit Method

Substituting these approximations into the governing equation yields

$$\alpha \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

Solving for the temp at the time node $j+1$ gives

$$T_i^{j+1} = T_i^j + \alpha \frac{\Delta t}{(\Delta x)^2} (T_{i+1}^j - 2T_i^j + T_{i-1}^j)$$

choosing,

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

we can write the equation as,

$$T_i^{j+1} = T_i^j + \lambda (T_{i+1}^j - 2T_i^j + T_{i-1}^j).$$

The Explicit Method

$$T_i^{j+1} = T_i^j + \lambda(T_{i+1}^j - 2T_i^j + T_{i-1}^j)$$

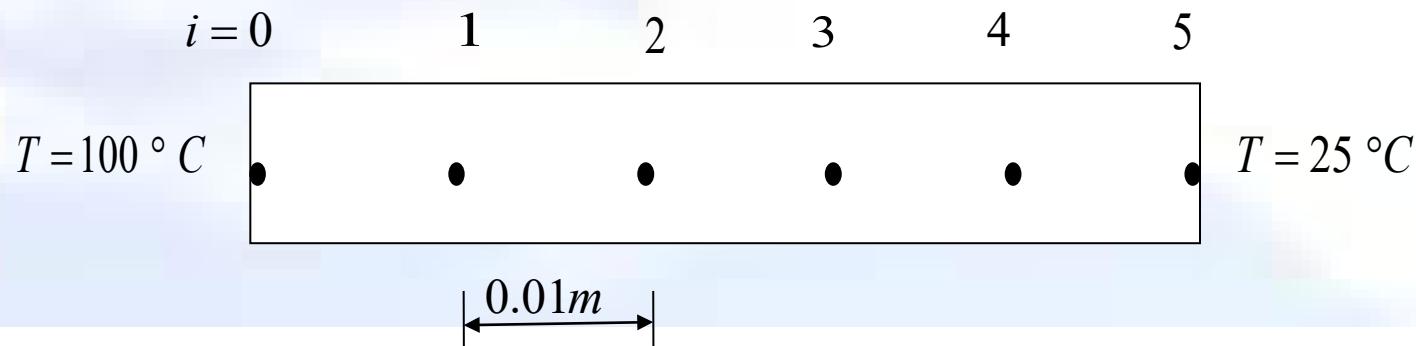
- This equation can be solved explicitly because it can be written for each internal location node of the rod for time node $j+1$ in terms of the temperature at time node j .
- In other words, if we know the temperature at node $j = 0$, and the boundary temperatures, we can find the temperature at the next time step.
- We continue the process by first finding the temperature at all nodes $j = 1$, and using these to find the temperature at the next time node, $j = 2$. This process continues until we reach the time at which we are interested in finding the temperature.

Example I: Explicit Method

Consider a steel rod that is subjected to a temperature of $100^{\circ}C$ on the left end and $25^{\circ}C$ on the right end. If the rod is of length $0.05m$, use the explicit method to find the temperature distribution in the rod from $t = 0$ and $t = 9$ seconds. Use $\Delta x = 0.01m$, $\Delta t = 3s$.

Given: $k = 54 \frac{W}{m \cdot K}$, $\rho = 7800 \frac{kg}{m^3}$, $C = 490 \frac{J}{kg \cdot K}$

The initial temperature of the rod is $20^{\circ}C$.



Example I: Explicit Method

Recall,

$$\alpha = \frac{k}{\rho C}$$

therefore,

$$\begin{aligned}\alpha &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ } m^2 / \text{s.}\end{aligned}$$

Then,

$$\begin{aligned}\lambda &= \alpha \frac{\Delta t}{(\Delta x)^2} \\ &= 1.4129 \times 10^{-5} \frac{3}{(0.01)^2} \\ &= 0.4239.\end{aligned}$$

Number of time steps,

$$\begin{aligned}&= \frac{t_{final} - t_{initial}}{\Delta t} \\ &= \frac{9 - 0}{3} \\ &= 3.\end{aligned}$$

Boundary Conditions

$$\left. \begin{aligned}T_0^j &= 100^\circ C \\ T_5^j &= 25^\circ C\end{aligned} \right\} \text{ for all } j = 0, 1, 2, 3$$

All internal nodes are at $20^\circ C$ for $t = 0$ sec. This can be represented as,

$$T_i^0 = 20^\circ C, \text{ for all } i = 1, 2, 3, 4$$

Example I: Explicit Method

Nodal temperatures when $t = 0 \text{ sec}$, $j = 0$:

$$\begin{aligned} T_0^0 &= 100^\circ C \\ T_1^0 &= 20^\circ C \\ T_2^0 &= 20^\circ C \\ T_3^0 &= 20^\circ C \\ T_4^0 &= 20^\circ C \\ T_5^0 &= 25^\circ C \end{aligned} \quad \left. \begin{array}{l} \\ \\ \\ \\ \\ \end{array} \right\} \text{Interior nodes}$$

We can now calculate the temperature at each node explicitly using the equation formulated earlier,

$$T_i^{j+1} = T_i^j + \lambda(T_{i+1}^j - 2T_i^j + T_{i-1}^j)$$

Example I: Explicit Method

Nodal temperatures when $t = 3 \text{ sec}$ (Example Calculations)

$i = 0$ $T_0^1 = 100^\circ C$ – Boundary Condition
 setting $j = 0$

$i = 1$
$$\begin{aligned} T_1^1 &= T_1^0 + \lambda(T_2^0 - 2T_1^0 + T_0^0) \\ &= 20 + 0.4239(20 - 2(20) + 100) \\ &= 20 + 0.4239(80) \\ &= 20 + 33.912 \\ &= 53.912^\circ C \end{aligned}$$

$i = 2$
$$\begin{aligned} T_2^1 &= T_2^0 + \lambda(T_3^0 - 2T_2^0 + T_1^0) \\ &= 20 + 0.4239(20 - 2(20) + 20) \\ &= 20 + 0.4239(0) \\ &= 20 + 0 \\ &= 20^\circ C \end{aligned}$$

Nodal temperatures when $t = 3 \text{ sec}$, $j = 1$:

$$T_0^1 = 100^\circ C \text{ – Boundary Condition}$$

$$\left. \begin{array}{l} T_1^1 = 53.912^\circ C \\ T_2^1 = 20^\circ C \\ T_3^1 = 20^\circ C \\ T_4^1 = 22.120^\circ C \end{array} \right\} \text{Interior nodes}$$

$$T_5^1 = 25^\circ C \text{ – Boundary Condition}$$

Example I: Explicit Method

Nodal temperatures when $t = 6 \text{ sec}$ (Example Calculations)

$i = 0$ $T_0^2 = 100^\circ\text{C}$ – Boundary Condition
 setting $j = 1$,

$i = 1$
$$\begin{aligned} T_1^2 &= T_1^1 + \lambda(T_2^1 - 2T_1^1 + T_0^1) \\ &= 53.912 + 0.4239(20 - 2(53.912) + 100) \\ &= 53.912 + 0.4239(12.176) \\ &= 53.912 + 5.1614 \\ &= 59.073^\circ\text{C} \end{aligned}$$

$i = 2$
$$\begin{aligned} T_2^2 &= T_2^1 + \lambda(T_3^1 - 2T_2^1 + T_1^1) \\ &= 20 + 0.4239(20 - 2(20) + 53.912) \\ &= 20 + 0.4239(33.912) \\ &= 20 + 14.375 \\ &= 34.375^\circ\text{C} \end{aligned}$$

Nodal temperatures when $t = 6 \text{ sec}$, $j = 2$:

$$T_0^2 = 100^\circ\text{C} \text{ – Boundary Condition}$$

$$\left. \begin{array}{l} T_1^2 = 59.073^\circ\text{C} \\ T_2^2 = 34.375^\circ\text{C} \\ T_3^2 = 20.889^\circ\text{C} \\ T_4^2 = 22.442^\circ\text{C} \end{array} \right\} \text{Interior nodes}$$

$$T_5^2 = 25^\circ\text{C} \text{ – Boundary Condition}$$

Example I: Explicit Method

Nodal temperatures when $t = 9 \text{ sec}$ (Example Calculations)

$i = 0$ $T_0^3 = 100^\circ\text{C}$ – Boundary Condition
 setting $j = 2$,

$$\begin{aligned} i &= 1 \\ T_1^3 &= T_1^2 + \lambda(T_2^2 - 2T_1^2 + T_0^2) \\ &= 59.073 + 0.4239(34.375 - 2(59.073) + 100) \\ &= 59.073 + 0.4239(16.229) \\ &= 59.073 + 6.8795 \\ &= 65.953^\circ\text{C} \end{aligned}$$

$$\begin{aligned} i &= 2 \\ T_2^3 &= T_2^2 + \lambda(T_3^2 - 2T_2^2 + T_1^2) \\ &= 34.375 + 0.4239(20.899 - 2(34.375) + 59.073) \\ &= 34.375 + 0.4239(11.222) \\ &= 34.375 + 4.7570 \\ &= 39.132^\circ\text{C} \end{aligned}$$

Nodal temperatures when $t = 9 \text{ sec}, j = 3$:

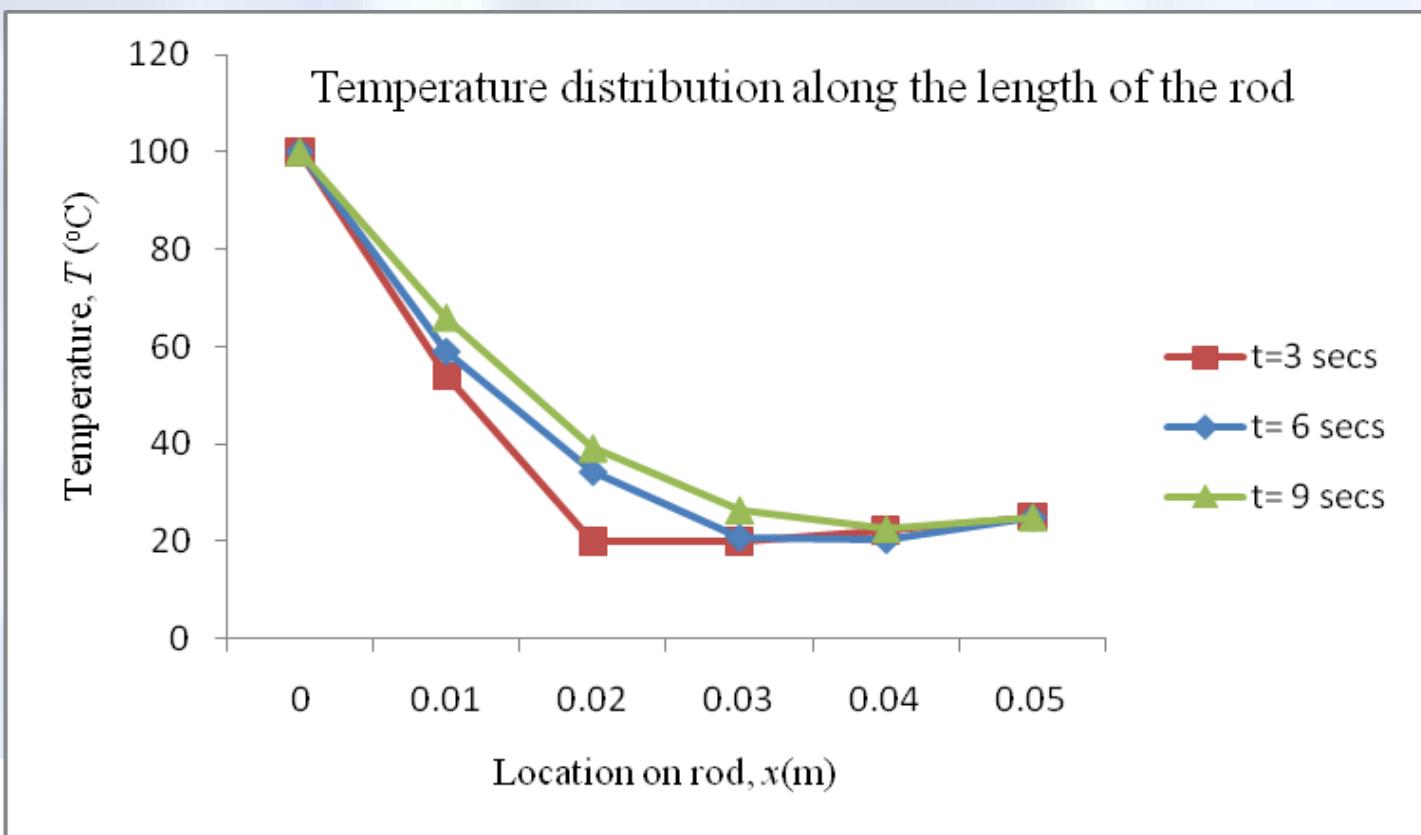
$$T_0^3 = 100^\circ\text{C} \text{ – Boundary Condition}$$

$$\left. \begin{array}{l} T_1^3 = 65.953^\circ\text{C} \\ T_2^3 = 39.132^\circ\text{C} \\ T_3^3 = 27.266^\circ\text{C} \\ T_4^3 = 22.872^\circ\text{C} \end{array} \right\} \text{Interior nodes}$$

$$T_5^3 = 25^\circ\text{C} \text{ – Boundary Condition}$$

Example I: Explicit Method

To better visualize the temperature variation at different locations at different times, the temperature distribution along the length of the rod at different times is plotted below.



The Implicit Method

WHY:

- Using the explicit method, we were able to find the temperature at each node, one equation at a time.
- However, the temperature at a specific node was only dependent on the temperature of the neighboring nodes from the previous time step. This is contrary to what we expect from the physical problem.
- The implicit method allows us to solve this and other problems by developing a system of simultaneous linear equations for the temperature at all interior nodes at a particular time.

The Implicit Method

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t}$$

The second derivative on the left hand side of the equation is approximated by the CDD scheme at time level $j+1$ at node (i) as

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j+1} \approx \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2}$$

The Implicit Method

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t}$$

The first derivative on the right hand side of the equation is approximated by the BDD scheme at time level $j+1$ at node (i) as

$$\left. \frac{\partial T}{\partial t} \right|_{i,j+1} \approx \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

The Implicit Method

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t}$$

Substituting these approximations into the heat conduction equation yields

$$\alpha \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

The Implicit Method

From the previous slide,

$$\alpha \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

Rearranging yields

$$-\lambda T_{i-1}^{j+1} + (1 + 2\lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = T_i^j$$

given that,

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

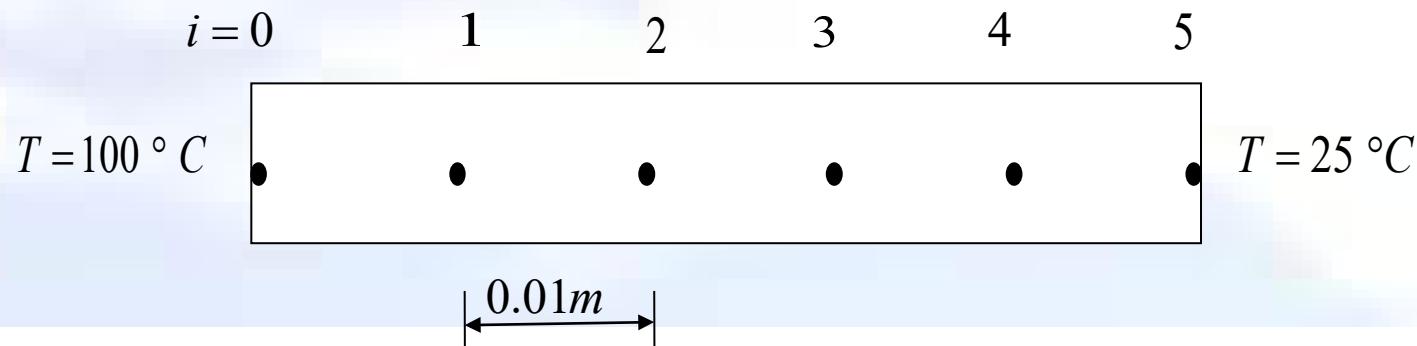
The rearranged equation can be written for every node during each time step. These equations can then be solved as a simultaneous system of linear equations to find the nodal temperatures at a particular time.

Example 2: Implicit Method

Consider a steel rod that is subjected to a temperature of $100^{\circ}C$ on the left end and $25^{\circ}C$ on the right end. If the rod is of length $0.05m$, use the implicit method to find the temperature distribution in the rod from $t = 0$ and $t = 9$ seconds. Use $\Delta x = 0.01m$, $\Delta t = 3s$.

Given: $k = 54 \frac{W}{m \cdot K}$, $\rho = 7800 \frac{kg}{m^3}$, $C = 490 \frac{J}{kg \cdot K}$

The initial temperature of the rod is $20^{\circ}C$.



Example 2: Implicit Method

Recall,

$$\alpha = \frac{k}{\rho C}$$

therefore,

$$\begin{aligned}\alpha &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ } m^2 / s.\end{aligned}$$

Then,

$$\begin{aligned}\lambda &= \alpha \frac{\Delta t}{(\Delta x)^2} \\ &= 1.4129 \times 10^{-5} \frac{3}{(0.01)^2} \\ &= 0.4239.\end{aligned}$$

Number of time steps,

$$\begin{aligned}&= \frac{t_{final} - t_{initial}}{\Delta t} \\ &= \frac{9 - 0}{3} \\ &= 3.\end{aligned}$$

Boundary Conditions

$$\left. \begin{array}{l} T_0^j = 100^\circ C \\ T_5^j = 25^\circ C \end{array} \right\} \text{ for all } j = 0, 1, 2, 3$$

All internal nodes are at $20^\circ C$ for $t = 0$ sec. This can be represented as,

$$T_i^0 = 20^\circ C, \text{ for all } i = 1, 2, 3, 4$$

Example 2: Implicit Method

Nodal temperatures when $t = 0 \text{ sec}$, $j = 0$:

$$\left. \begin{array}{l} T_0^0 = 100^\circ C \\ T_1^0 = 20^\circ C \\ T_2^0 = 20^\circ C \\ T_3^0 = 20^\circ C \\ T_4^0 = 20^\circ C \\ T_5^0 = 25^\circ C \end{array} \right\} \text{Interior nodes}$$

We can now form our system of equations for the first time step by writing the approximated heat conduction equation for each node.

$$-\lambda T_{i-1}^{j+1} + (1 + 2\lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = T_i^j$$

Example 2: Implicit Method

Nodal temperatures when $t = 3 \text{ sec}$, (Example Calculations)

$i = 0 \quad T_0^1 = 100^\circ\text{C}$ – Boundary Condition

For the interior nodes setting $j = 0$ and $i = 1, 2, 3, 4$ gives the following,

$$\begin{aligned} i = 1 \quad -\lambda T_0^1 + (1 + 2\lambda)T_1^1 - \lambda T_2^1 &= T_1^0 \\ (-0.4239 \times 100) + (1 + 2 \times 0.4239)T_1^1 - (0.4239T_2^1) &= 20 \\ -42.39 + 1.8478T_1^1 - 0.4239T_2^1 &= 20 \\ 1.8478T_1^1 - 0.4239T_2^1 &= 62.390 \end{aligned}$$

$$\begin{aligned} i = 2 \quad -\lambda T_1^1 + (1 + 2\lambda)T_2^1 - \lambda T_3^1 &= T_2^0 \\ -0.4239T_1^1 + 1.8478T_2^1 - 0.4239T_3^1 &= 20 \end{aligned}$$

For the first time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 62.390 \\ 20 \\ 20 \\ 30.598 \end{bmatrix}$$

Example 2: Implicit Method

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 62.390 \\ 20 \\ 20 \\ 30.598 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

$$\begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 39.451 \\ 24.792 \\ 21.438 \\ 21.477 \end{bmatrix}$$

Hence, the nodal temps at $t = 3$ sec are

$$\begin{bmatrix} T_0^1 \\ T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \\ T_5^1 \end{bmatrix} = \begin{bmatrix} 100 \\ 39.451 \\ 24.792 \\ 21.438 \\ 21.477 \\ 25 \end{bmatrix}$$

Example 2: Implicit Method

Nodal temperatures when $t = 6 \text{ sec}$, (Example Calculations)

$i = 0 \quad T_0^2 = 100^\circ\text{C}$ – Boundary Condition

For the interior nodes setting $j = 1$ and $i = 1, 2, 3, 4$ gives the following,

$$\begin{aligned} \underline{i = 1} \quad & -\lambda T_0^2 + (1 + 2\lambda)T_1^2 - \lambda T_2^2 = T_1^1 \\ & (-0.4239 \times 100) + (1 + 2 \times 0.4239)T_1^2 - 0.4239T_2^2 = 39.451 \\ & -42.39 + 1.8478T_1^2 - 0.4239T_2^2 = 39.451 \\ & 1.8478T_1^2 - 0.4239T_2^2 = 81.841 \end{aligned}$$

$$\begin{aligned} \underline{i = 2} \quad & -\lambda T_1^2 + (1 + 2\lambda)T_2^2 - \lambda T_3^2 = T_2^1 \\ & -0.4239T_1^2 + 1.8478T_2^2 - 0.4239T_3^2 = 24.792 \end{aligned}$$

For the second time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 81.841 \\ 24.792 \\ 21.438 \\ 32.075 \end{bmatrix}$$

Example 2: Implicit Method

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 81.841 \\ 24.792 \\ 21.438 \\ 32.075 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

$$\begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 51.326 \\ 30.669 \\ 23.876 \\ 22.836 \end{bmatrix}$$

Hence, the nodal temps at $t = 6$ sec are

$$\begin{bmatrix} T_0^2 \\ T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \\ T_5^2 \end{bmatrix} = \begin{bmatrix} 100 \\ 51.326 \\ 30.669 \\ 23.876 \\ 22.836 \\ 25 \end{bmatrix}$$

Example 2: Implicit Method

Nodal temperatures when $t = 9 \text{ sec}$, (Example Calculations)

$i = 0 \quad T_0^3 = 100^\circ\text{C}$ – Boundary Condition

For the interior nodes setting $j = 2$ and $i = 1, 2, 3, 4$ gives the following,

$$\begin{aligned} \underline{i = 1} \quad & -\lambda T_0^3 + (1 + 2\lambda)T_1^3 - \lambda T_2^3 = T_1^2 \\ & (-0.4239 \times 100) + (1 + 2 \times 0.4239)T_1^3 - (0.4239 T_2^3) = 51.326 \\ & -42.39 + 1.8478 T_1^3 - 0.4239 T_2^3 = 51.326 \\ & 1.8478 T_1^3 - 0.4239 T_2^3 = 93.716 \end{aligned}$$

$$\begin{aligned} \underline{i = 2} \quad & -\lambda T_1^3 + (1 + 2\lambda)T_2^3 - \lambda T_3^3 = T_2^2 \\ & -0.4239 T_1^3 + 1.8478 T_2^3 - 0.4239 T_3^3 = 30.669 \end{aligned}$$

For the third time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 93.716 \\ 30.669 \\ 23.876 \\ 33.434 \end{bmatrix}$$

Example 2: Implicit Method

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 93.716 \\ 30.669 \\ 23.876 \\ 33.434 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

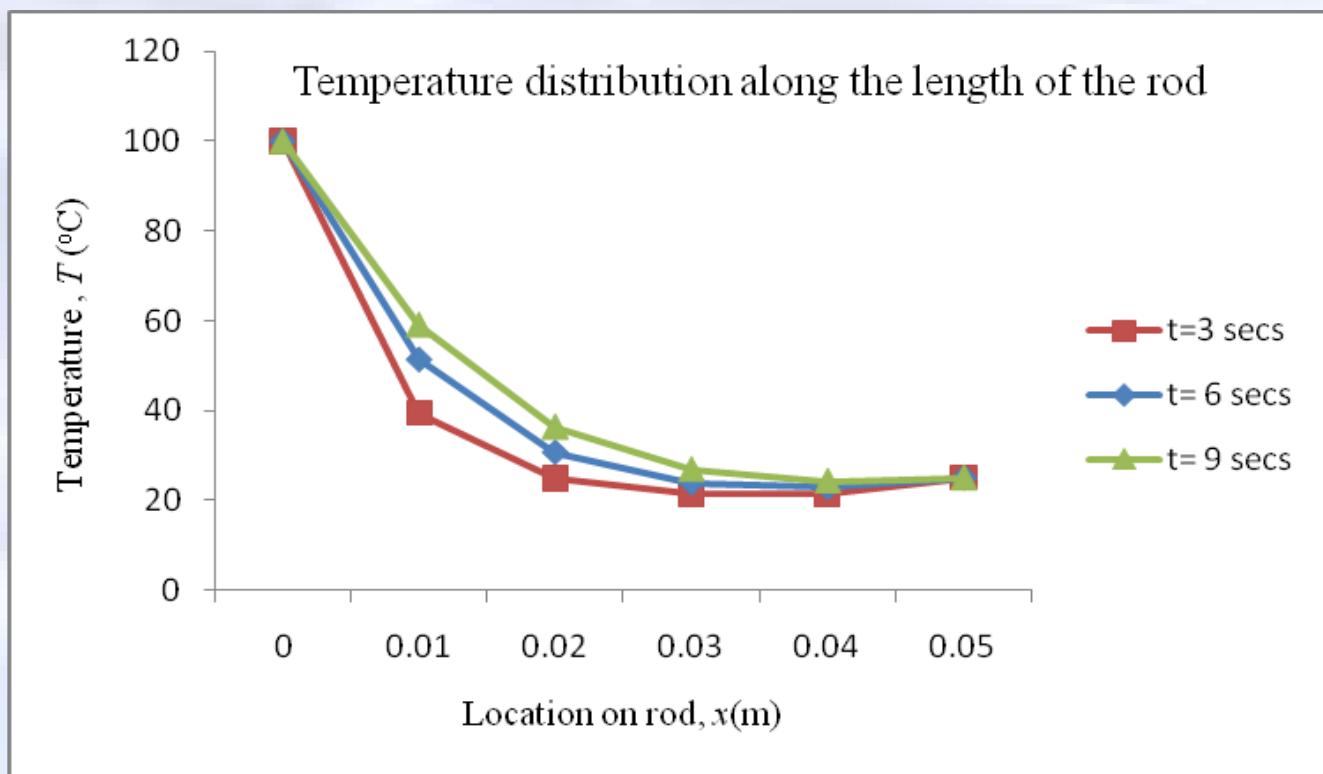
$$\begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 59.043 \\ 36.292 \\ 26.809 \\ 24.243 \end{bmatrix}$$

Hence, the nodal temps at $t = 9$ sec are

$$\begin{bmatrix} T_0^3 \\ T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \\ T_5^3 \end{bmatrix} = \begin{bmatrix} 100 \\ 59.043 \\ 36.292 \\ 26.809 \\ 24.243 \\ 25 \end{bmatrix}$$

Example 2: Implicit Method

To better visualize the temperature variation at different locations at different times, the temperature distribution along the length of the rod at different times is plotted below.



The Crank-Nicolson Method

WHY:

Using the implicit method our approximation of $\frac{\partial^2 T}{\partial x^2}$ was of $O(\Delta x)^2$ accuracy, while our approximation of $\frac{\partial T}{\partial t}$ was of $O(\Delta t)$ accuracy.

The Crank-Nicolson Method

One can achieve similar orders of accuracy by approximating the second derivative, on the left hand side of the heat equation, at the midpoint of the time step. Doing so yields

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j} \approx \frac{\alpha}{2} \left[\frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} + \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} \right]$$

The Crank-Nicolson Method

The first derivative, on the right hand side of the heat equation, is approximated using the forward divided difference method at time level $j+1$,

$$\frac{\partial T}{\partial t} \Big|_{i,j} \approx \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

The Crank-Nicolson Method

- Substituting these approximations into the governing equation for heat conductance yields

$$\frac{\alpha}{2} \left[\frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} + \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} \right] = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

giving

$$-\lambda T_{i-1}^{j+1} + 2(1+\lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = \lambda T_{i-1}^j + 2(1-\lambda)T_i^j + \lambda T_{i+1}^j$$

where

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

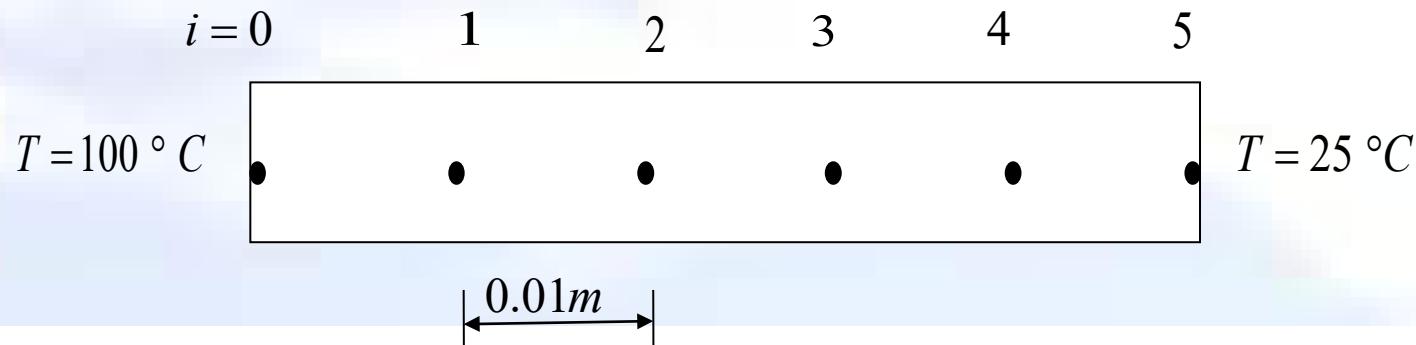
- Having rewritten the equation in this form allows us to describe the physical problem. We then solve a system of simultaneous linear equations to find the temperature at every node at any point in time.

Example 3: Crank-Nicolson

Consider a steel rod that is subjected to a temperature of $100^{\circ}C$ on the left end and $25^{\circ}C$ on the right end. If the rod is of length $0.05m$, use the Crank-Nicolson method to find the temperature distribution in the rod from $t = 0$ to $t = 9$ seconds. Use $\Delta x = 0.01m$, $\Delta t = 3s$.

Given: $k = 54 \frac{W}{m-K}$, $\rho = 7800 \frac{kg}{m^3}$, $C = 490 \frac{J}{kg-K}$

The initial temperature of the rod is $20^{\circ}C$.



Example 3: Crank-Nicolson

Recall,

$$\alpha = \frac{k}{\rho C}$$

therefore,

$$\begin{aligned}\alpha &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ } m^2 / s.\end{aligned}$$

Then,

$$\begin{aligned}\lambda &= \alpha \frac{\Delta t}{(\Delta x)^2} \\ &= 1.4129 \times 10^{-5} \frac{3}{(0.01)^2} \\ &= 0.4239.\end{aligned}$$

Number of time steps,

$$\begin{aligned}&= \frac{t_{final} - t_{initial}}{\Delta t} \\ &= \frac{9 - 0}{3} \\ &= 3.\end{aligned}$$

Boundary Conditions

$$\left. \begin{array}{l} T_0^j = 100^\circ C \\ T_5^j = 25^\circ C \end{array} \right\} \text{ for all } j = 0, 1, 2, 3$$

All internal nodes are at $20^\circ C$ for $t = 0$ sec. This can be represented as,

$$T_i^0 = 20^\circ C, \text{ for all } i = 1, 2, 3, 4$$

Example 3: Crank-Nicolson

Nodal temperatures when $t = 0 \text{ sec}$, $j = 0$:

$$\left. \begin{array}{l} T_0^0 = 100^\circ C \\ T_1^0 = 20^\circ C \\ T_2^0 = 20^\circ C \\ T_3^0 = 20^\circ C \\ T_4^0 = 20^\circ C \\ T_5^0 = 25^\circ C \end{array} \right\} \text{Interior nodes}$$

We can now form our system of equations for the first time step by writing the approximated heat conduction equation for each node.

$$-\lambda T_{i-1}^{j+1} + 2(1 + \lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = \lambda T_{i-1}^j + 2(1 - \lambda)T_i^j + \lambda T_{i+1}^j$$

Example 3: Crank-Nicolson

Nodal temperatures when $t = 3 \text{ sec}$, (Example Calculations)

$$\underline{i=0} \quad T_0^1 = 100^\circ\text{C} - \text{Boundary Condition}$$

For the interior nodes setting $j = 0$ and $i = 1, 2, 3, 4$ gives the following

$$\underline{i=1}$$

$$-\lambda T_0^1 + 2(1+\lambda)T_1^1 - \lambda T_2^1 = \lambda T_0^0 + 2(1-\lambda)T_1^0 + \lambda T_2^0$$

$$(-0.4239 \times 100) + 2(1 + 0.4239)T_1^1 - 0.4239T_2^1 = (0.4239)100 + 2(1 - 0.4239)20 + (0.4239)20$$

$$-42.39 + 2.8478T_1^1 - 0.4239T_2^1 = 42.39 + 23.044 + 8.478$$

$$2.8478T_1^1 - 0.4239T_2^1 = 116.30$$

For the first time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 116.30 \\ 40.000 \\ 40.000 \\ 52.718 \end{bmatrix}$$

Example 3: Crank-Nicolson

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 116.30 \\ 40.000 \\ 40.000 \\ 52.718 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

$$\begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 44.372 \\ 23.746 \\ 20.797 \\ 21.607 \end{bmatrix}$$

Hence, the nodal temps at $t = 3$ sec are

$$\begin{bmatrix} T_0^1 \\ T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \\ T_5^1 \end{bmatrix} = \begin{bmatrix} 100 \\ 44.372 \\ 23.746 \\ 20.797 \\ 21.607 \\ 25 \end{bmatrix}$$

Example 3: Crank-Nicolson

Nodal temperatures when $t = 6 \text{ sec}$, (Example Calculations)

$i = 0$ $T_0^2 = 100^\circ\text{C}$ – Boundary Condition

For the interior nodes setting $j = 1$ and $i = 1, 2, 3, 4$ gives the following,
 $i = 1$

$$-\lambda T_0^2 + 2(1+\lambda)T_1^2 - \lambda T_2^2 = \lambda T_0^1 + 2(1-\lambda)T_1^1 + \lambda T_2^1$$

$$(-0.4239 \times 100) + 2(1 + 0.4239)T_1^2 - 0.4239T_2^2 =$$

$$(0.4239)100 + 2(1 - 0.4239)44.372 + (0.4239)23.746$$

$$-42.39 + 2.8478T_1^2 - 0.4239T_2^2 = 42.39 + 51.125 + 10.066$$

$$2.8478T_1^2 - 0.4239T_2^2 = 145.971$$

For the second time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 145.971 \\ 54.985 \\ 43.187 \\ 54.908 \end{bmatrix}$$

Example 3: Crank-Nicolson

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 145.971 \\ 54.985 \\ 43.187 \\ 54.908 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

$$\begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 55.883 \\ 31.075 \\ 23.174 \\ 22.730 \end{bmatrix}$$

Hence, the nodal
temps at $t = 6$ sec are

$$\begin{bmatrix} T_0^2 \\ T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \\ T_5^2 \end{bmatrix} = \begin{bmatrix} 100 \\ 55.883 \\ 31.075 \\ 23.174 \\ 22.730 \\ 25 \end{bmatrix}$$

Example 3: Crank-Nicolson

Nodal temperatures when $t = 9 \text{ sec}$, (Example Calculations)

$$i = 0 \quad T_0^3 = 100^\circ\text{C} - \text{Boundary Condition}$$

For the interior nodes setting $j = 2$ and $i = 1, 2, 3, 4$ gives the following,
 $i = 1$

$$-\lambda T_0^3 + 2(1 + \lambda)T_1^3 - \lambda T_2^3 = \lambda T_0^2 + 2(1 - \lambda)T_1^2 + \lambda T_2^2$$

$$(-0.4239 \times 100) + 2(1 + 0.4239)T_2^3 - 0.4239 T_2^3 =$$

$$(0.4239)100 + 2(1 - 0.4239)55.883 + (0.4239)31.075$$

$$-42.39 + 2.8478T_1^3 - 0.4239 T_2^3 = 42.39 + 64.388 + 13.173$$

$$2.8478T_1^3 - 0.4239 T_2^3 = 162.34$$

For the third time step we can write four such equations with four unknowns, expressing them in matrix form yields

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 162.34 \\ 69.318 \\ 49.509 \\ 57.210 \end{bmatrix}$$

Example 3: Crank-Nicolson

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 162.34 \\ 69.318 \\ 49.509 \\ 57.210 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

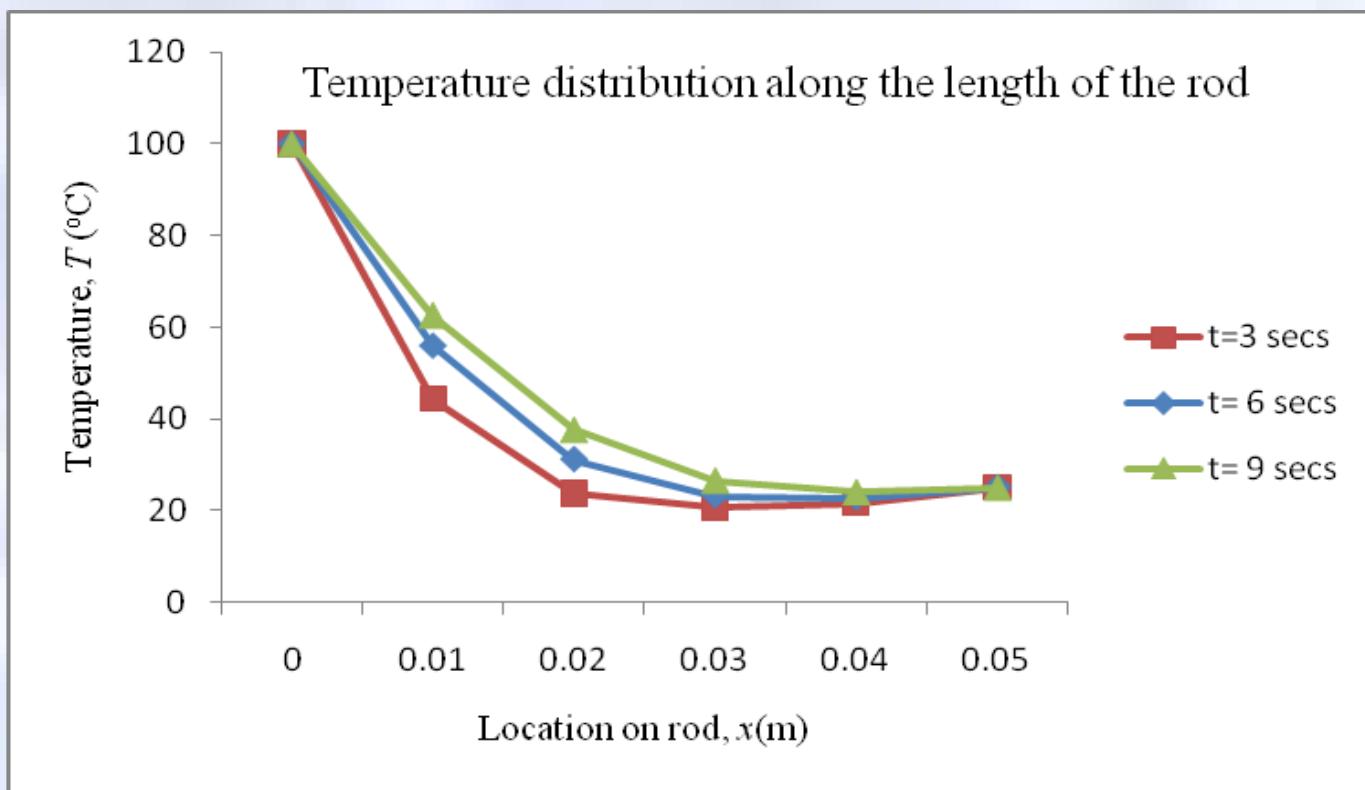
$$\begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 62.604 \\ 37.613 \\ 26.562 \\ 24.042 \end{bmatrix}$$

Hence, the nodal temps at $t = 9$ sec are

$$\begin{bmatrix} T_0^3 \\ T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \\ T_5^3 \end{bmatrix} = \begin{bmatrix} 100 \\ 62.604 \\ 37.613 \\ 26.562 \\ 24.042 \\ 25 \end{bmatrix}$$

Example 3: Crank-Nicolson

To better visualize the temperature variation at different locations at different times, the temperature distribution along the length of the rod at different times is plotted below.



Internal Temperatures at 9 sec.

The table below allows you to compare the results from all three methods discussed in juxtaposition with the analytical solution.

Node	Explicit	Implicit	Crank-Nicolson	Analytical
T_1^3	65.953	59.043	62.604	62.510
T_2^3	39.132	36.292	37.613	37.084
T_3^3	27.266	26.809	26.562	25.844
T_4^3	22.872	24.243	24.042	23.610

THE END

Elliptic Partial Differential Equations

<http://numericalmethods.eng.usf.edu>

Transforming Numerical Methods Education for STEM Undergraduates

Defining Elliptic PDE's

- The general form for a second order linear PDE with two independent variables (x, y) and one dependent variable (u) is

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

- Recall the criteria for an equation of this type to be considered elliptic
 $B^2 - 4AC < 0$
- For example, examine the Laplace equation given by

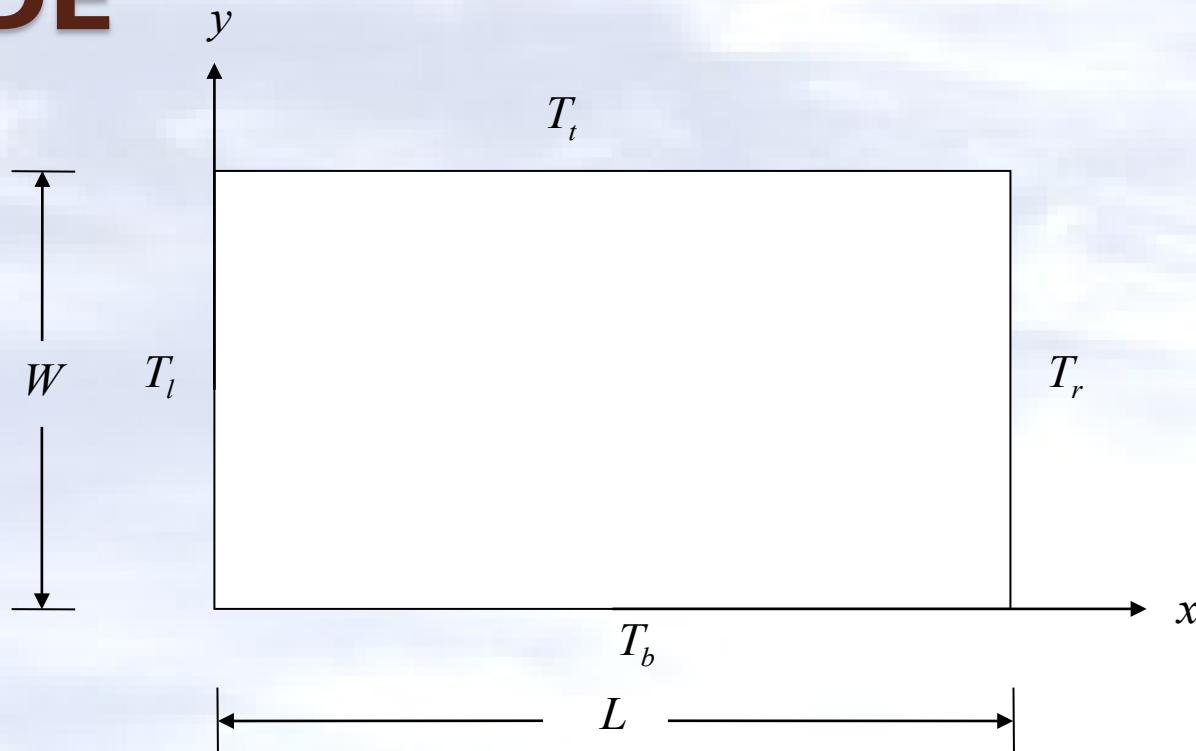
$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0, \text{ where } A = 1, B = 0, C = 1 \text{ and } D = 0$$

then

$$\begin{aligned} B^2 - 4AC &= 0 - 4(1)(1) \\ &= -4 < 0 \end{aligned}$$

thus allowing us to classify this equation as elliptic.

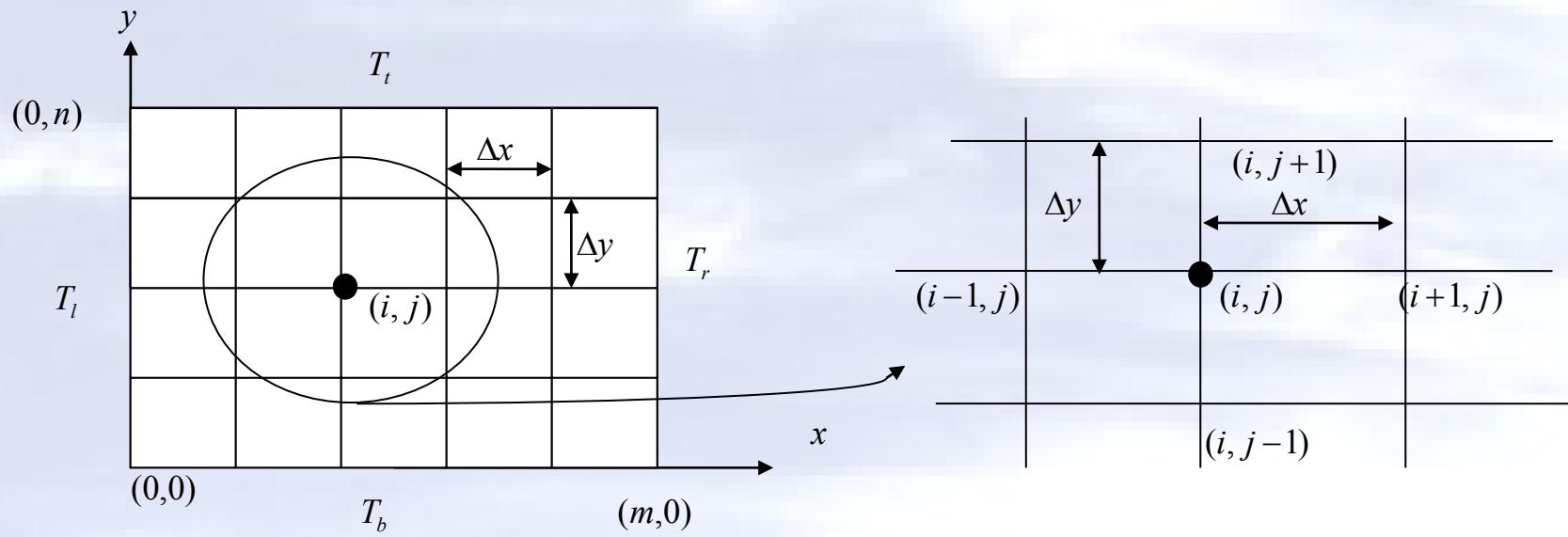
Physical Example of an Elliptic PDE



Schematic diagram of a plate with specified temperature boundary conditions

The Laplace equation governs the temperature: $\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$

Discretizing the Elliptic PDE



If we define $\Delta x = \frac{L}{m}$ and $\Delta y = \frac{W}{n}$, we can then write the finite difference

approximation of the partial derivatives at a general interior node (i, j) as

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j} \cong \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{(\Delta x)^2} \quad \text{and} \quad \left. \frac{\partial^2 T}{\partial y^2} \right|_{i,j} \cong \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{(\Delta y)^2}$$

Discretizing the Elliptic PDE

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

Substituting these approximations into the Laplace equation yields:

$$\frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{(\Delta x)^2} + \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{(\Delta y)^2} = 0$$

if,

$$\Delta x = \Delta y$$

the Laplace equation can be rewritten as

$$T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0$$

Discretizing the Elliptic PDE

$$T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0$$

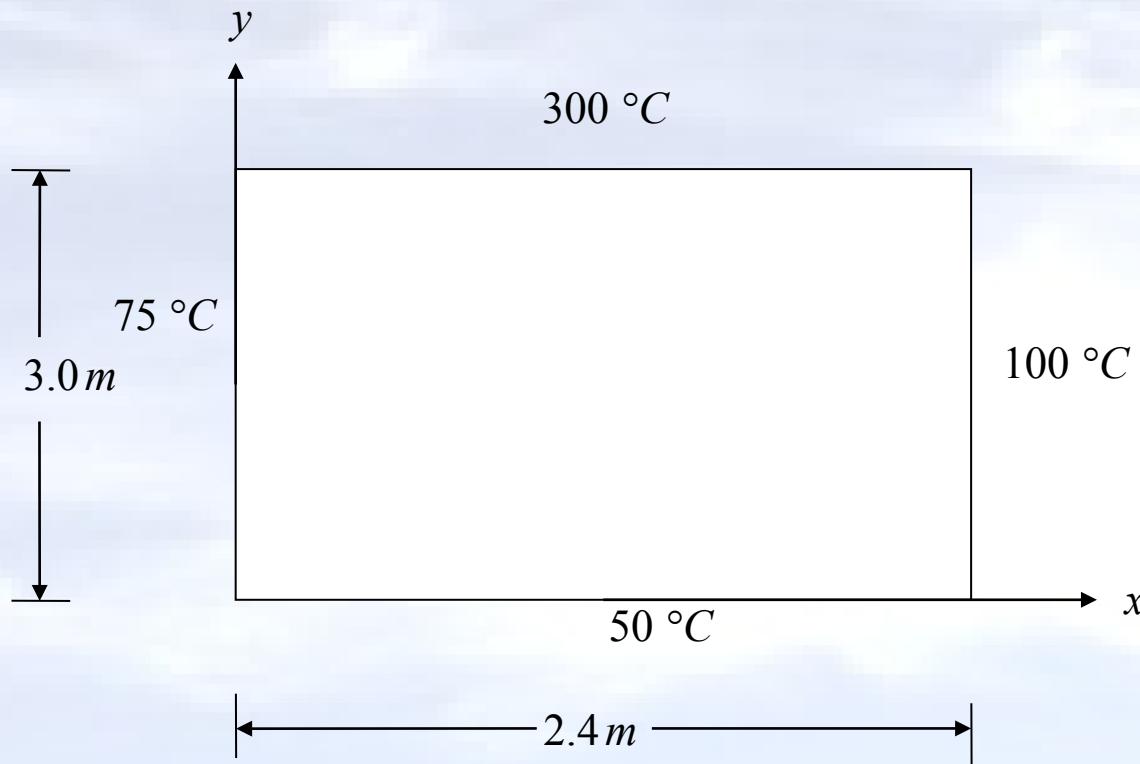
Once the governing equation has been discretized there are several numerical methods that can be used to solve the problem.

We will examine the:

- Direct Method
- Gauss-Seidel Method
- Lieberman Method

Example I: Direct Method

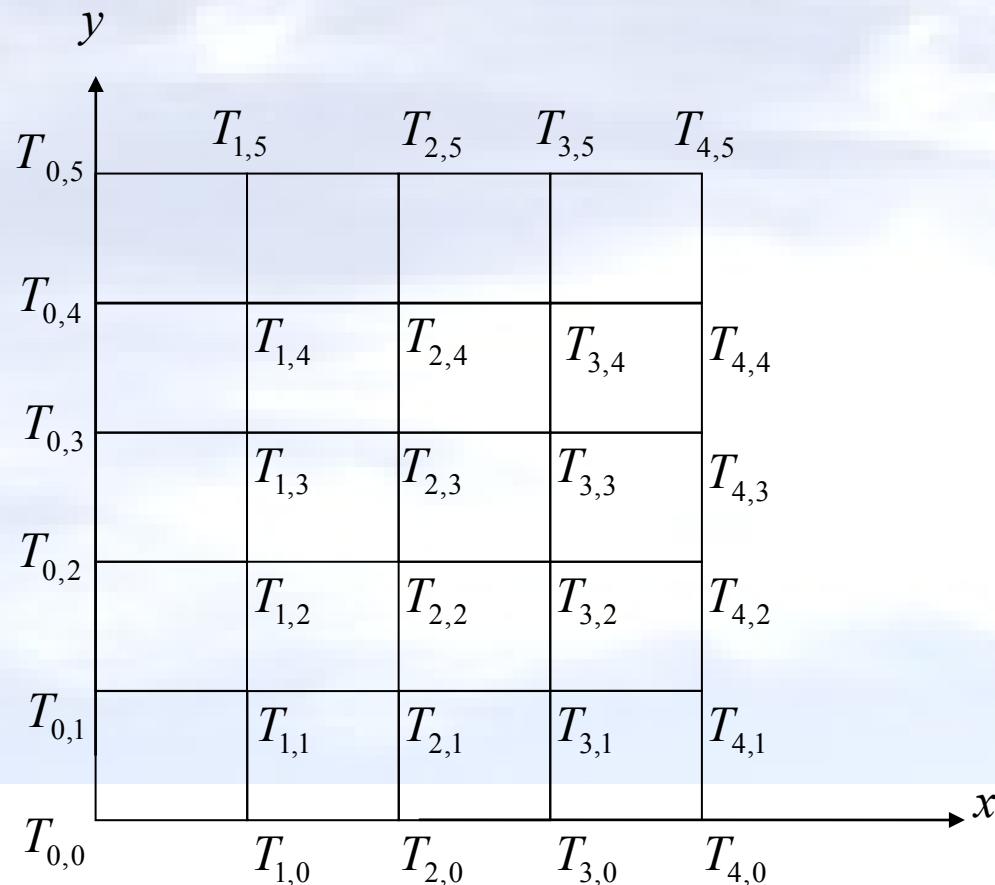
Consider a plate $2.4\text{ m} \times 3.0\text{ m}$ that is subjected to the boundary conditions shown below. Find the temperature at the interior nodes using a square grid with a length of 0.6 m by using the direct method.



Example I: Direct Method

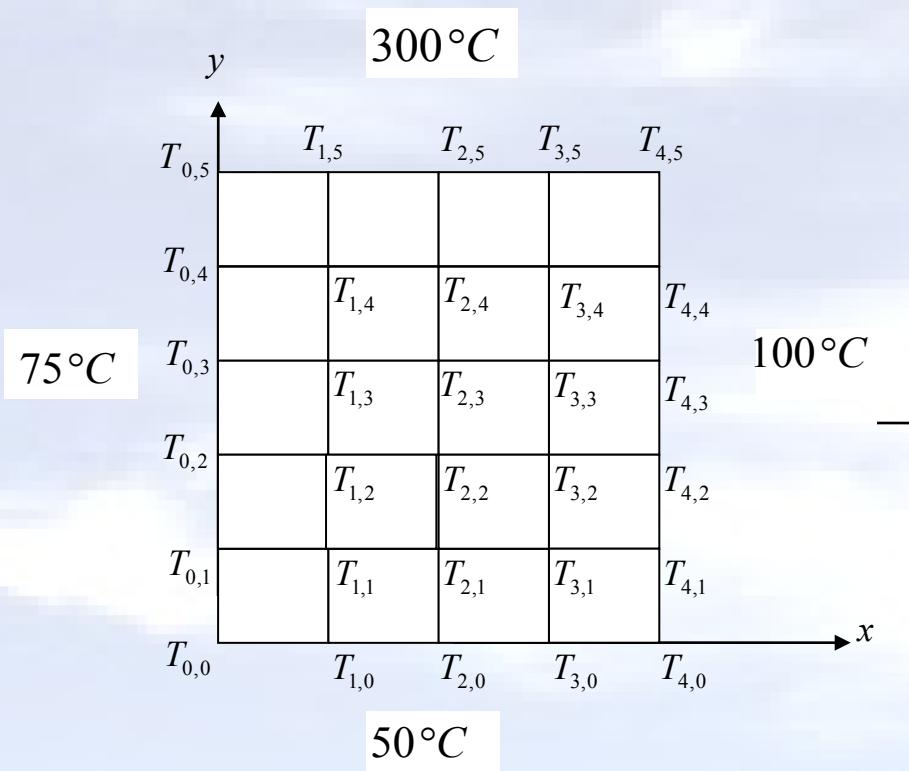
We can discretize the plate by taking,

$$\Delta x = \Delta y = 0.6m$$



Example I: Direct Method

The nodal temperatures at the boundary nodes are given by:



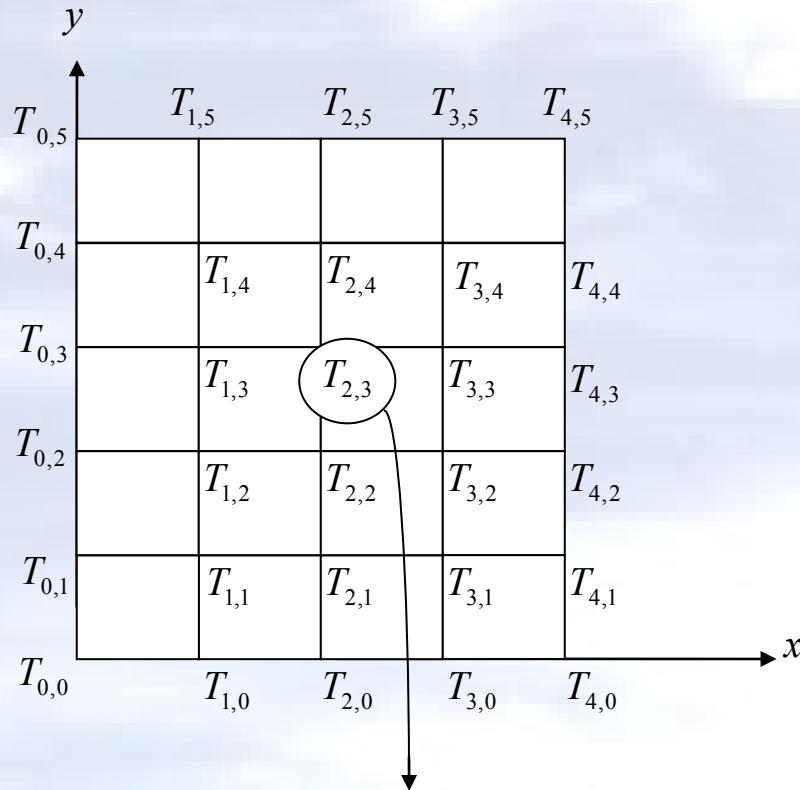
$$T_{0,j} = 75, j = 1, 2, 3, 4$$

$$T_{4,j} = 100, j = 1, 2, 3, 4$$

$$T_{i,0} = 50, i = 1, 2, 3$$

$$T_{i,5} = 300, i = 1, 2, 3$$

Example I: Direct Method



Here we develop the equation for the temperature at the node (2,3)

$$\underline{i=2 \text{ and } j=3} \quad T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0$$

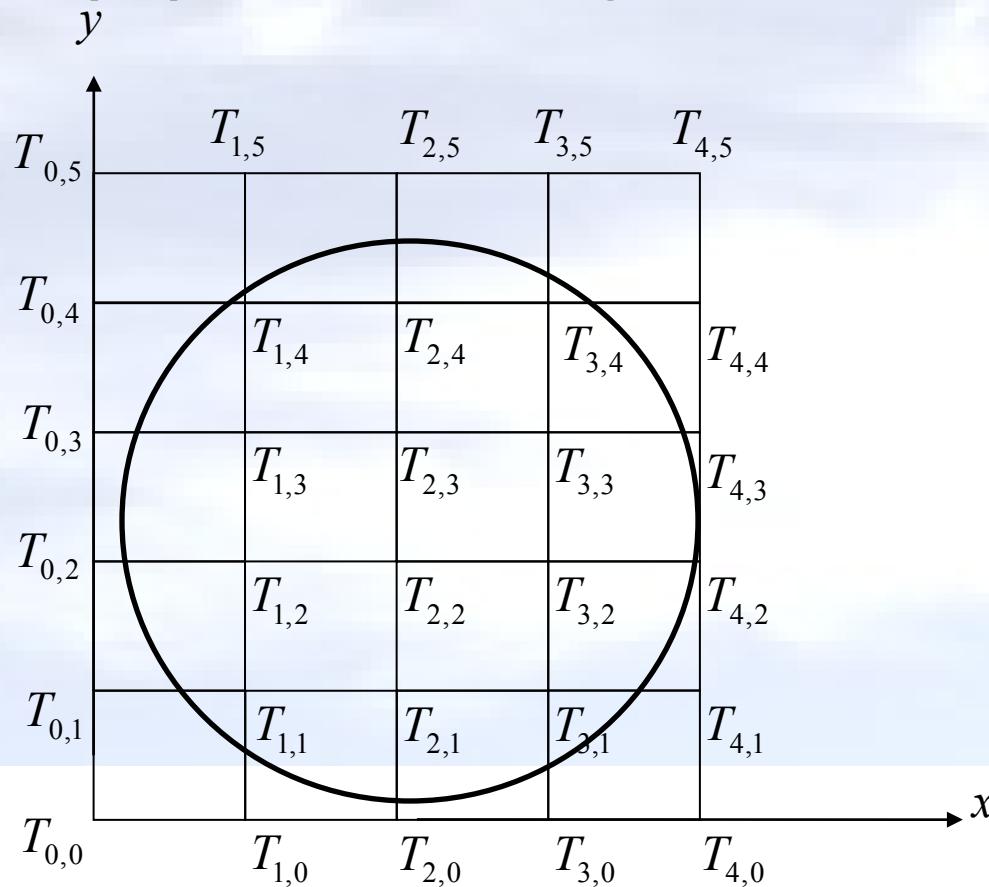
$$T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2} - 4T_{2,3} = 0$$

$$T_{1,3} + T_{2,2} - 4T_{2,3} + T_{2,4} + T_{3,3} = 0$$

Example I: Direct Method

We can develop similar equations for every interior node leaving us with an equal number of equations and unknowns.

Question: How many equations would this generate?



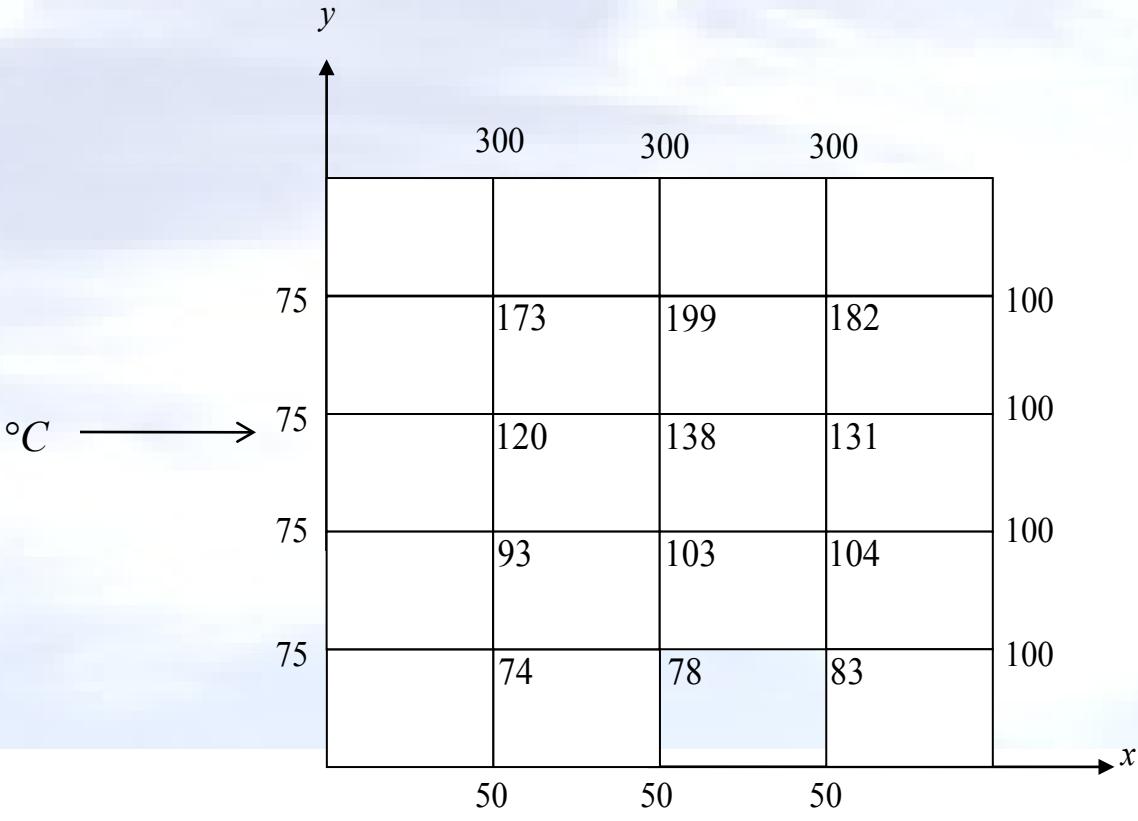
Example I: Direct Method

We can develop similar equations for every interior node leaving us with an equal number of equations and unknowns.

Question: How many equations would this generate? **Answer:** 12

Solving yields:

$$\begin{bmatrix} T_{1,1} \\ T_{1,2} \\ T_{1,3} \\ T_{1,4} \\ T_{2,1} \\ T_{2,2} \\ T_{2,3} \\ T_{2,4} \\ T_{3,1} \\ T_{3,2} \\ T_{3,3} \\ T_{3,4} \end{bmatrix} = \begin{bmatrix} 73.8924 \\ 93.0252 \\ 119.907 \\ 173.355 \\ 77.5443 \\ 103.302 \\ 138.248 \\ 198.512 \\ 82.9833 \\ 104.389 \\ 131.271 \\ 182.446 \end{bmatrix} {}^{\circ}\text{C}$$



The Gauss-Seidel Method

- Recall the discretized equation

$$T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0$$

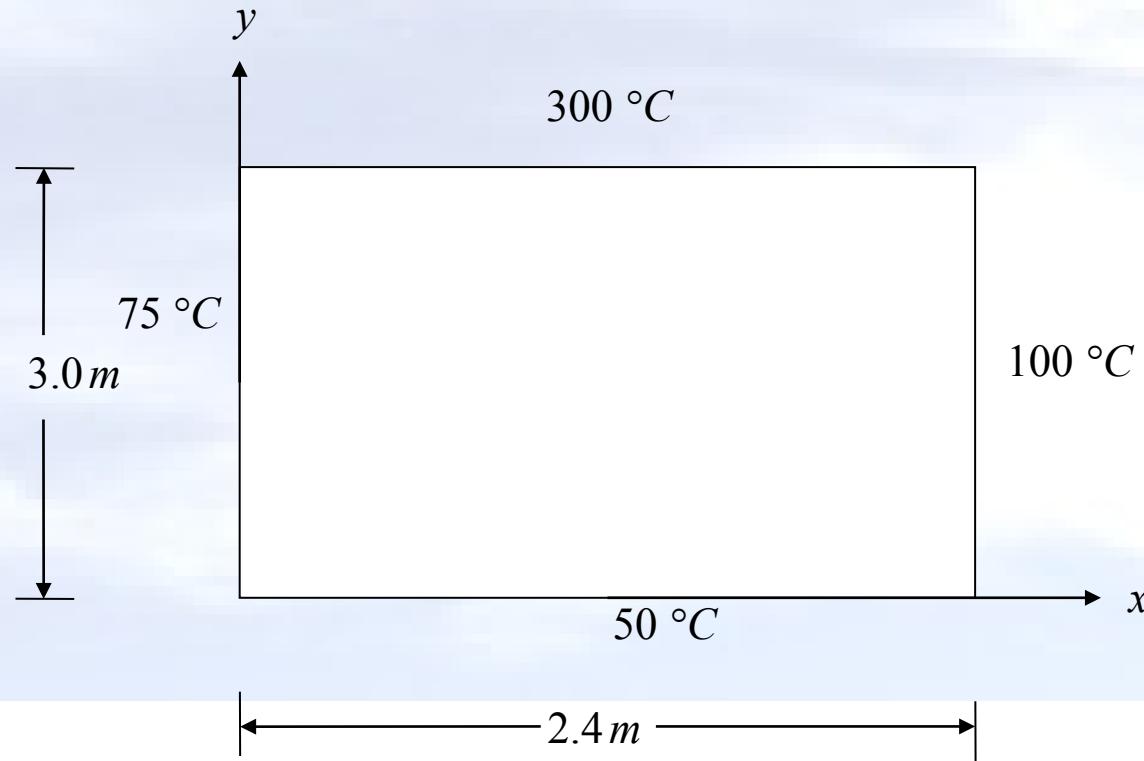
- This can be rewritten as

$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}}{4}$$

- For the Gauss-Seidel Method, this equation is solved iteratively for all interior nodes until a pre-specified tolerance is met.

Example 2: Gauss-Seidel Method

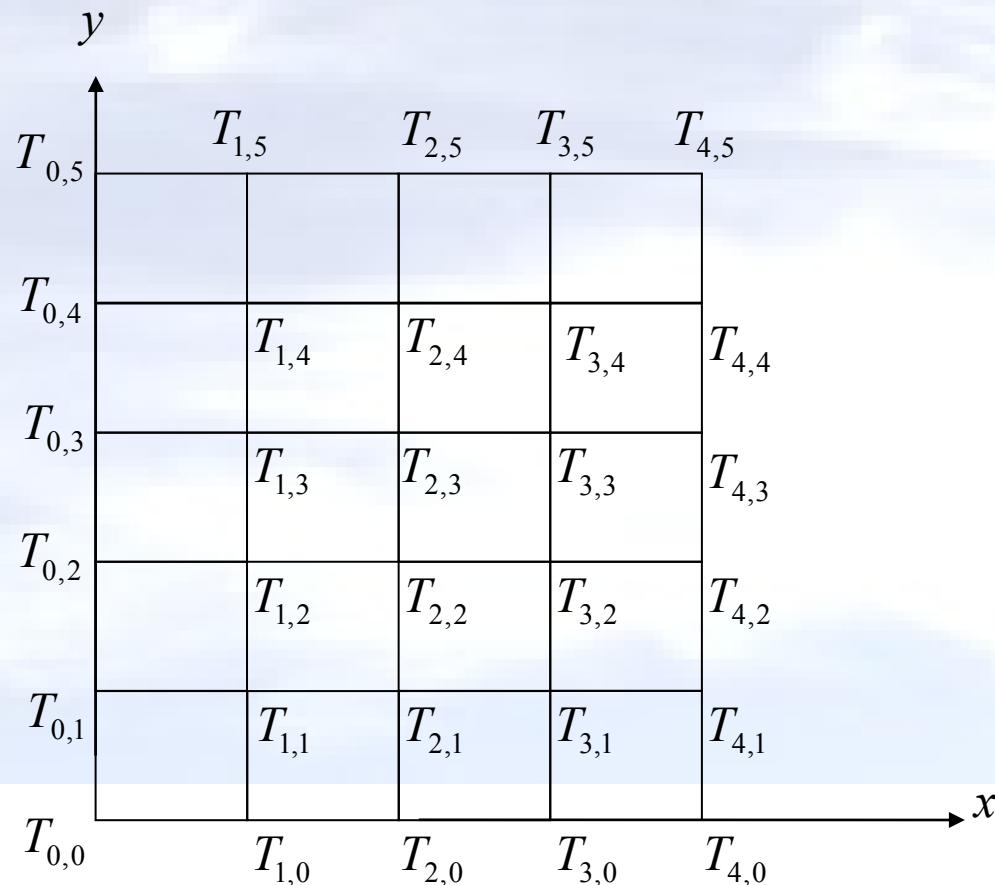
Consider a plate $2.4\text{ m} \times 3.0\text{ m}$ that is subjected to the boundary conditions shown below. Find the temperature at the interior nodes using a square grid with a length of 0.6 m using the Gauss-Siedel method. Assume the initial temperature at all interior nodes to be 0°C .



Example 2: Gauss-Seidel Method

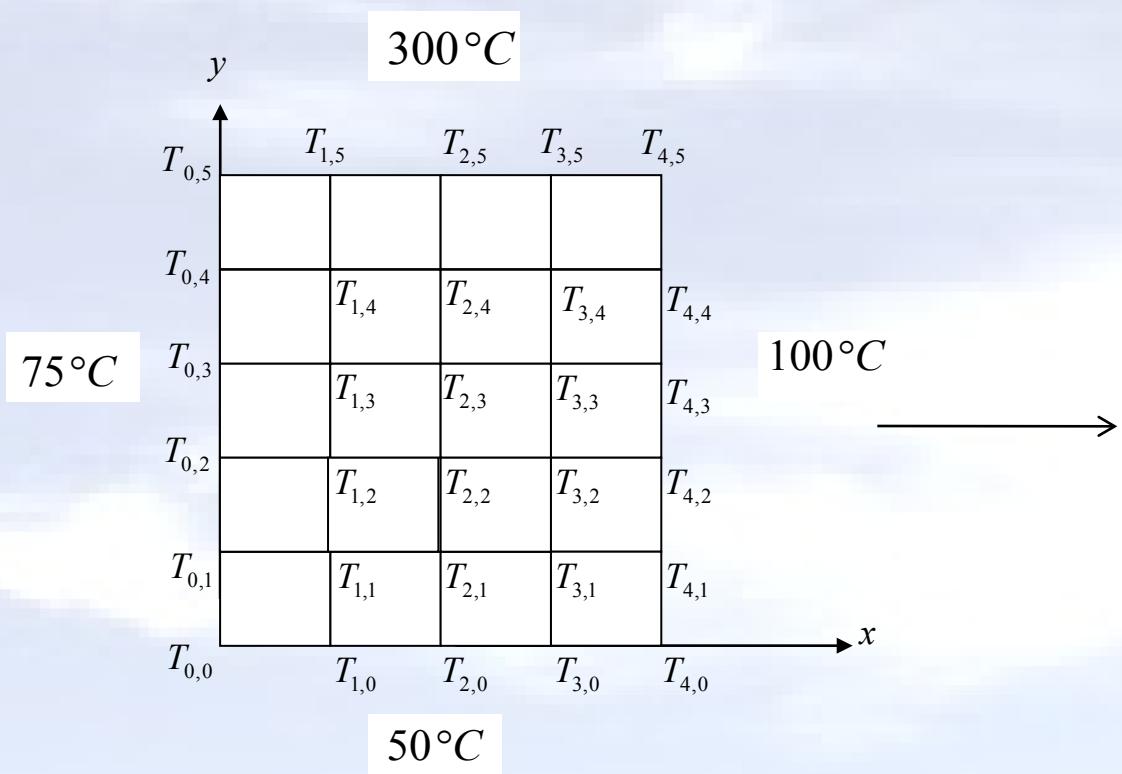
We can discretize the plate by taking

$$\Delta x = \Delta y = 0.6m$$



Example 2: Gauss-Seidel Method

The nodal temperatures at the boundary nodes are given by:



$$T_{0,j} = 75, j = 1, 2, 3, 4$$

$$T_{4,j} = 100, j = 1, 2, 3, 4$$

$$T_{i,0} = 50, i = 1, 2, 3$$

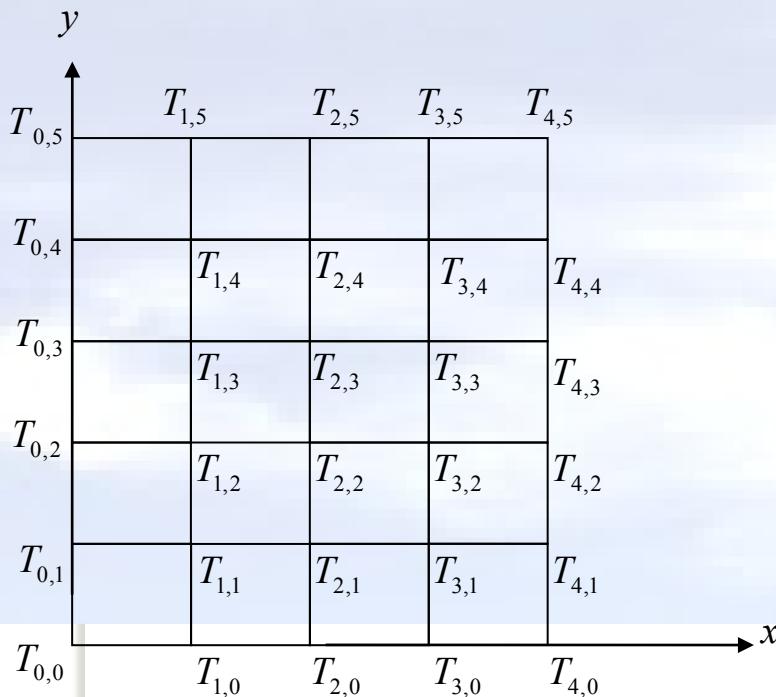
$$T_{i,5} = 300, i = 1, 2, 3$$

Example 2: Gauss-Seidel Method

- Now we can begin to solve for the temperature at each interior node using

$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}}{4}$$

- Assume all internal nodes to have an initial temperature of zero.



Iteration #1

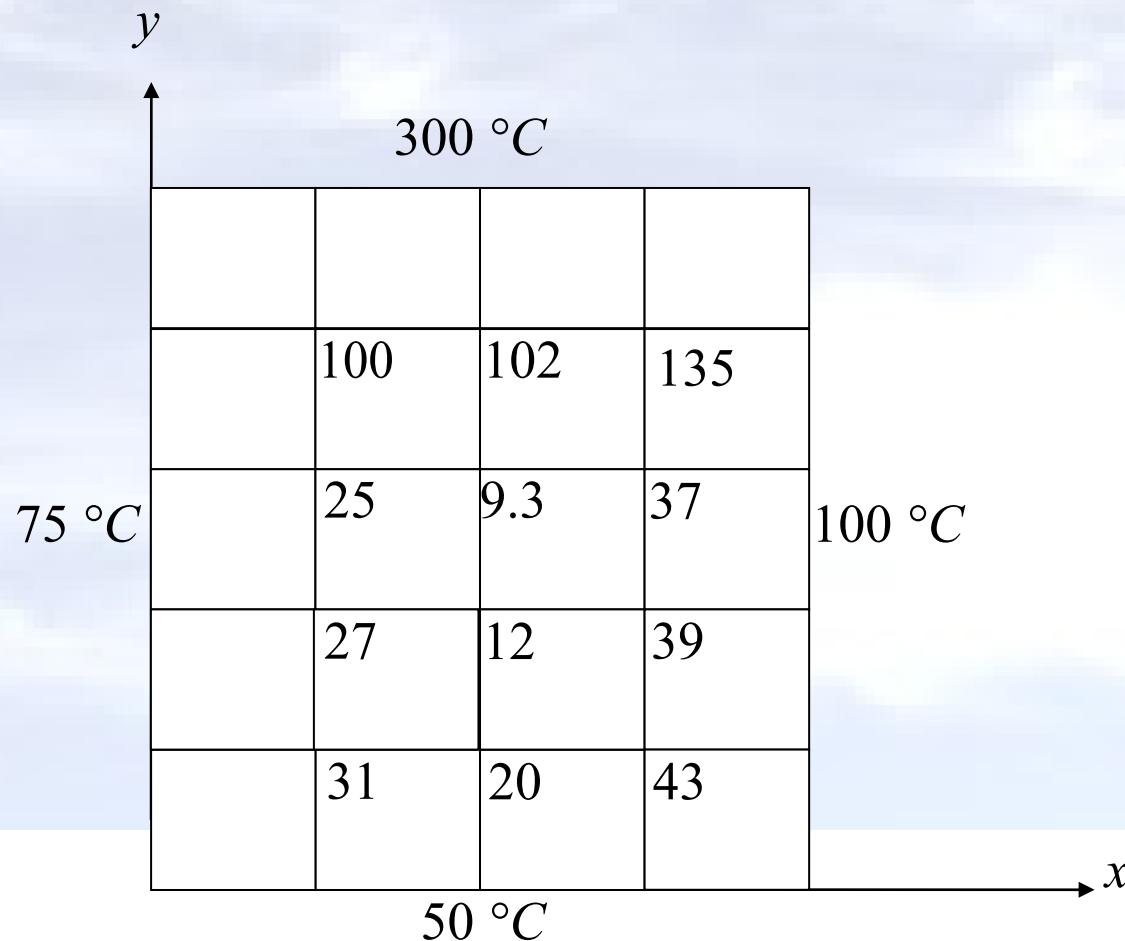
i=1 and j=1 $T_{1,1} = \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4}$
 $= \frac{0 + 75 + 0 + 50}{4}$
 $= 31.2500^{\circ}\text{C}$

i=1 and j=2

$$T_{1,2} = \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4}$$
$$= \frac{0 + 75 + 0 + 31.2500}{4}$$
$$= 26.5625^{\circ}\text{C}$$

Example 2: Gauss-Seidel Method

After the first iteration, the temperatures are as follows. These will now be used as the nodal temperatures for the second iteration.



Example 2: Gauss-Seidel Method

Iteration #2

i=1 and j=1

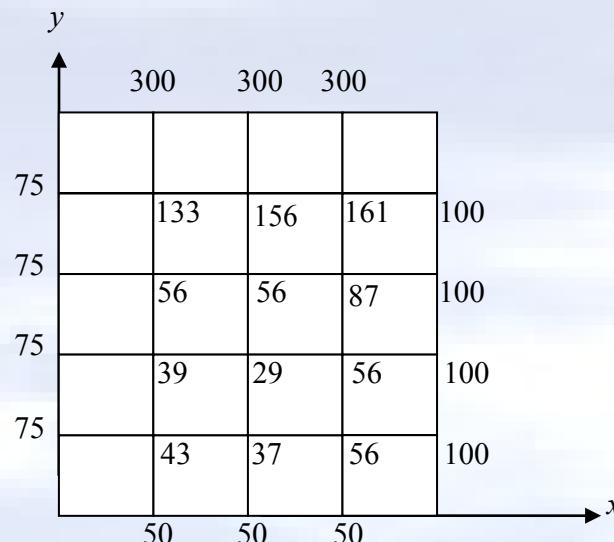
$$\begin{aligned}T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\&= \frac{20.3125 + 75 + 26.5625 + 50}{4} \\&= \underline{\underline{42.9688^\circ C}}\end{aligned}$$

$$\begin{aligned}|\epsilon_a|_{1,1} &= \left| \frac{T_{1,1}^{present} - T_{1,1}^{previous}}{T_{1,1}^{present}} \right| \times 100 \\&= \left| \frac{42.9688 - 31.2500}{42.9688} \right| \times 100 \\&= \underline{\underline{27.27\%}}\end{aligned}$$

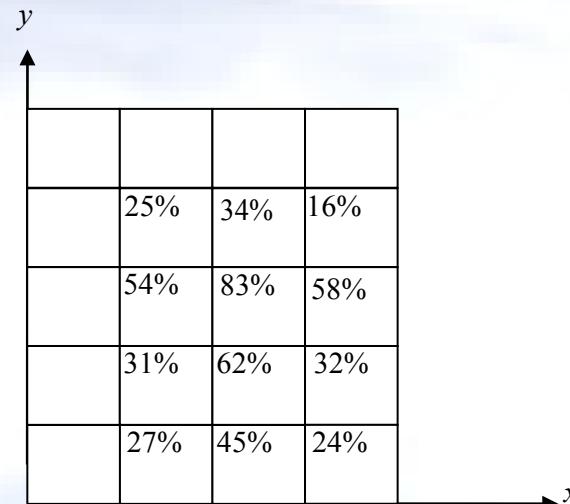
Example 2: Gauss-Seidel Method

The figures below show the temperature distribution and absolute relative error distribution in the plate after two iterations:

Temperature Distribution



Absolute Relative Approximate Error Distribution



Example 2: Gauss-Seidel Method

Node	Temperature Distribution in the Plate (°C)			
	Number of Iterations			Exact
	1	2	10	
$T_{1,1}$	31.2500	42.9688	73.0239	
$T_{1,2}$	26.5625	38.7695	91.9585	
$T_{1,3}$	25.3906	55.7861	119.0976	
$T_{1,4}$	100.0977	133.2825	172.9755	
$T_{2,1}$	20.3125	36.8164	76.6127	
$T_{2,2}$	11.7188	30.8594	102.1577	
$T_{2,3}$	9.2773	56.4880	137.3802	
$T_{2,4}$	102.3438	156.1493	198.1055	
$T_{3,1}$	42.5781	56.3477	82.4837	
$T_{3,2}$	38.5742	56.0425	103.7757	
$T_{3,3}$	36.9629	86.8393	130.8056	
$T_{3,4}$	134.8267	160.7471	182.2278	

The Lieberman Method

- Recall the equation used in the Gauss-Siedel Method,

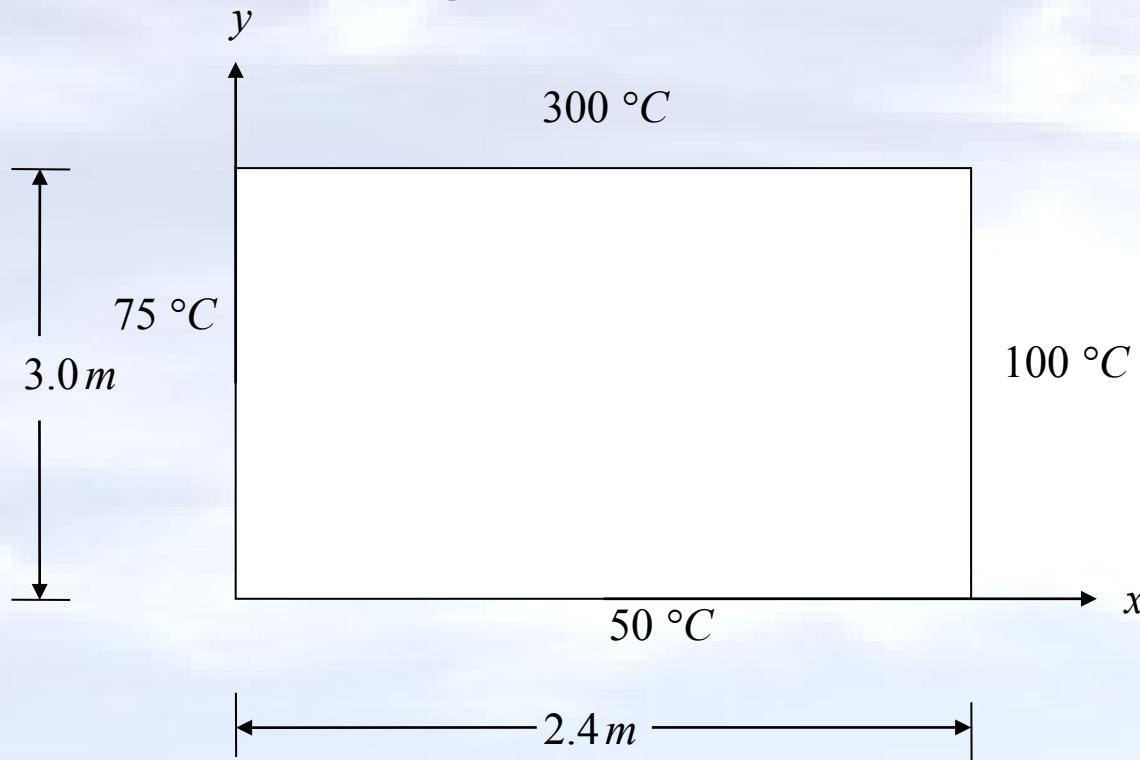
$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}}{4}$$

- Because the Guass-Siedel Method is guaranteed to converge, we can accelerate the process by using over- relaxation. In this case,

$$T_{i,j}^{relaxed} = \lambda T_{i,j}^{new} + (1 - \lambda)T_{i,j}^{old}$$

Example 3: Lieberman Method

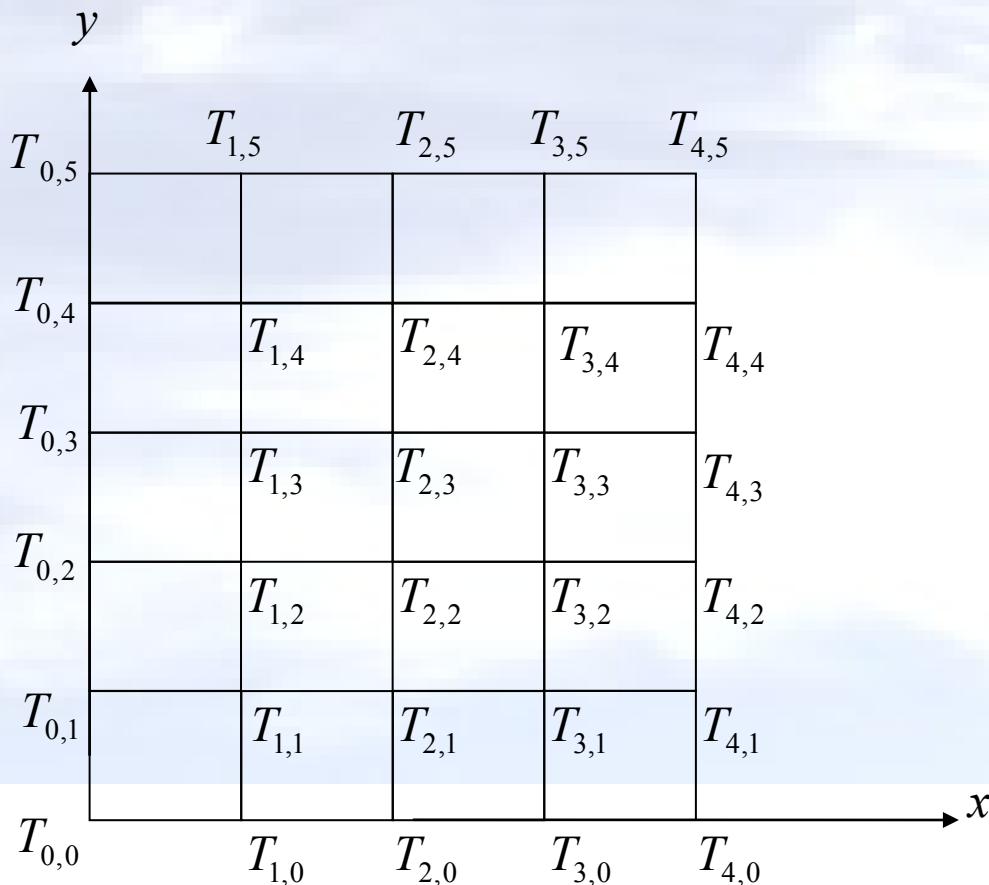
Consider a plate $2.4\text{ m} \times 3.0\text{ m}$ that is subjected to the boundary conditions shown below. Find the temperature at the interior nodes using a square grid with a length of 0.6 m . Use a weighting factor of 1.4 in the Lieberman method. Assume the initial temperature at all interior nodes to be 0°C .



Example 3: Lieberman Method

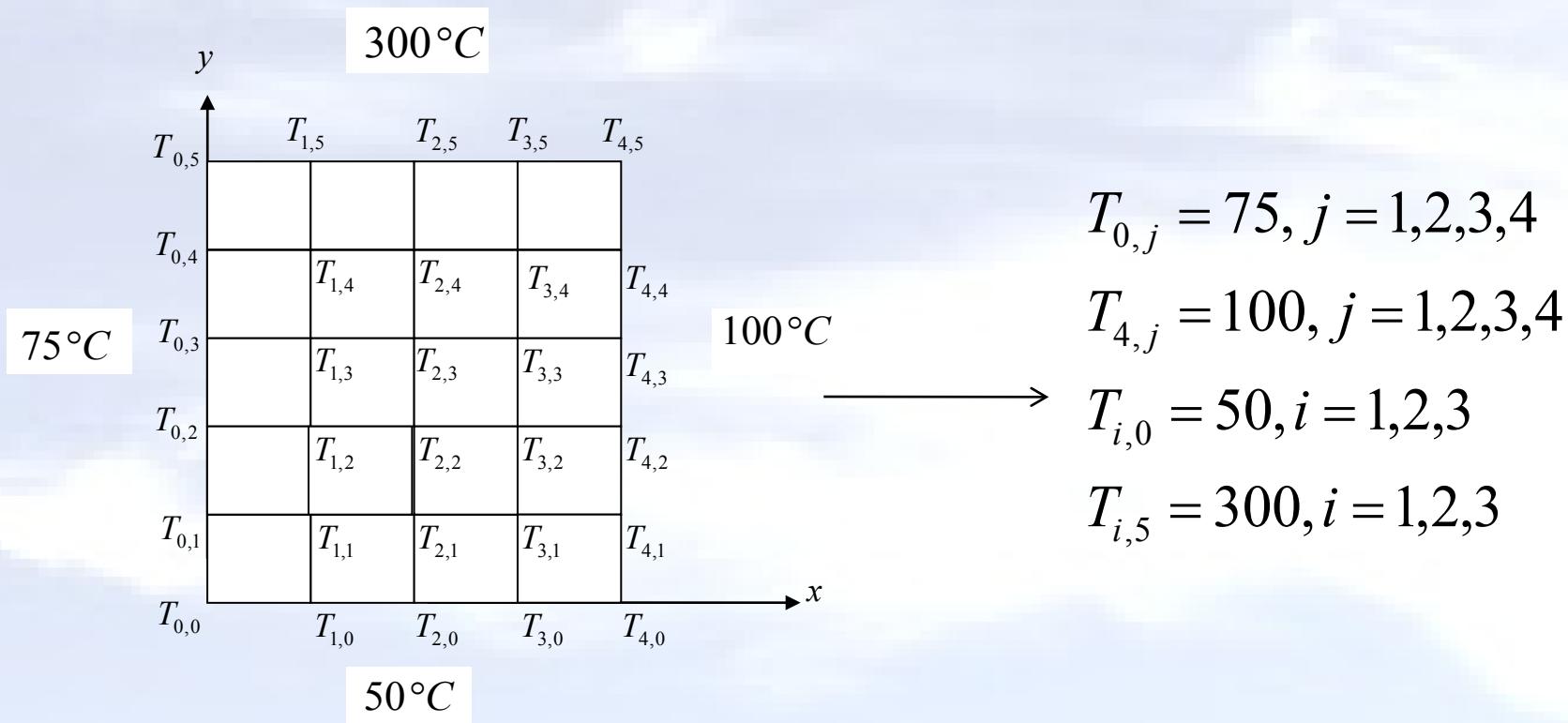
We can discretize the plate by taking

$$\Delta x = \Delta y = 0.6m$$



Example 3: Lieberman Method

We can also develop equations for the boundary conditions to define the temperature of the exterior nodes.



Example 3: Lieberman Method

- Now we can begin to solve for the temperature at each interior node using the rewritten Laplace equation from the Gauss-Siedel method.
- Once we have the temperature value for each node we will apply the over relaxation equation of the Lieberman method
- Assume all internal nodes to have an initial temperature of zero.

Iteration #1

i=1 and j=1 $T_{1,1} = \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4}$

$$= \frac{0 + 75 + 0 + 50}{4}$$
$$= 31.2500^{\circ}\text{C}$$

$$\begin{aligned}T_{1,1}^{relaxed} &= \lambda T_{1,1}^{new} + (1 - \lambda) T_{1,1}^{old} \\&= 1.4(31.2500) + (1 - 1.4)0 \\&= 43.7500^{\circ}\text{C}\end{aligned}$$

Iteration #2

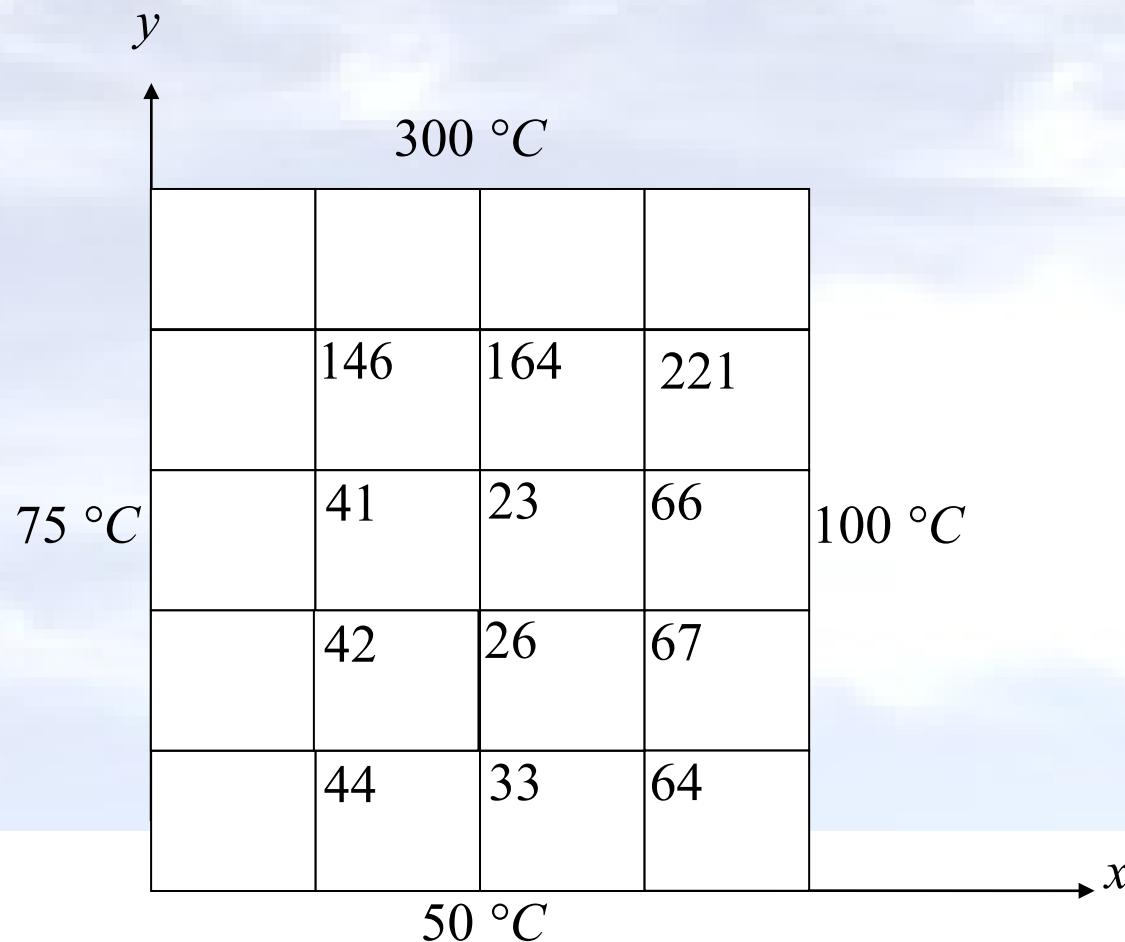
i=1 and j=2 $T_{1,2} = \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4}$

$$= \frac{0 + 75 + 0 + 43.75}{4}$$
$$= 29.6875^{\circ}\text{C}$$

$$\begin{aligned}T_{1,1}^{relaxed} &= \lambda T_{1,1}^{new} + (1 - \lambda) T_{1,1}^{old} \\&= 1.4(29.6875) + (1 - 1.4)0 \\&= 41.5625^{\circ}\text{C}\end{aligned}$$

Example 3: Lieberman Method

After the first iteration the temperatures are as follows. These will be used as the initial nodal temperatures during the second iteration.



Example 3: Lieberman Method

Iteration #2

i=1 and j=1

$$\begin{aligned} T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\ &= \frac{32.8125 + 75 + 41.5625 + 50}{4} \\ &= \underline{49.8438^\circ C} \end{aligned}$$

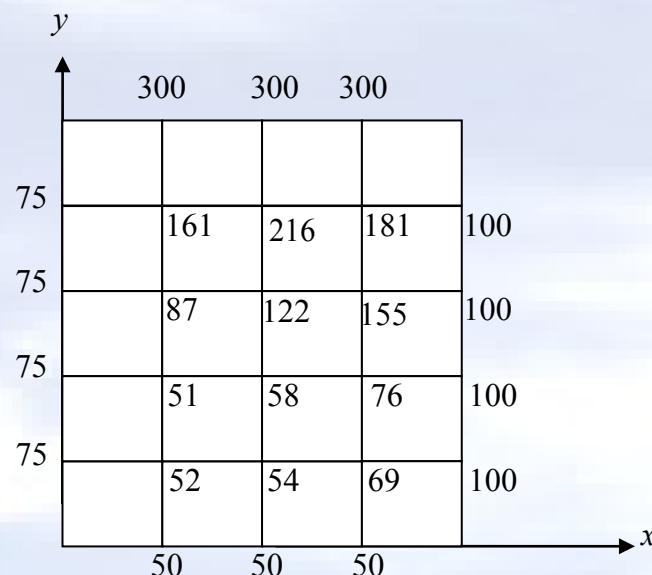
$$\begin{aligned} T_{1,1}^{relaxed} &= \lambda T_{1,1}^{new} + (1 - \lambda) T_{1,1}^{old} \\ &= 1.4(49.8438) + (1 - 1.4)43.75 \\ &= 52.2813^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{1,1} &= \left| \frac{T_{1,1}^{present} - T_{1,1}^{previous}}{T_{1,1}^{present}} \right| \times 100 \\ &= \left| \frac{52.2813 - 43.7500}{52.2813} \right| \times 100 \\ &= \underline{16.32\%} \end{aligned}$$

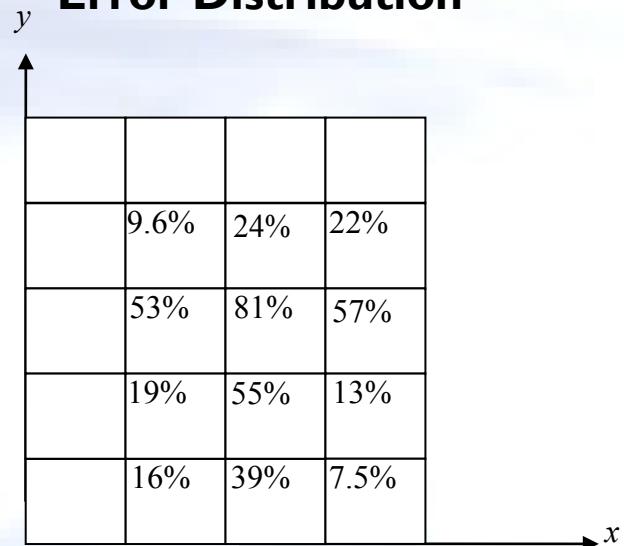
Example 3: Lieberman Method

The figures below show the temperature distribution and absolute relative error distribution in the plate after two iterations:

Temperature Distribution



Absolute Relative Approximate Error Distribution

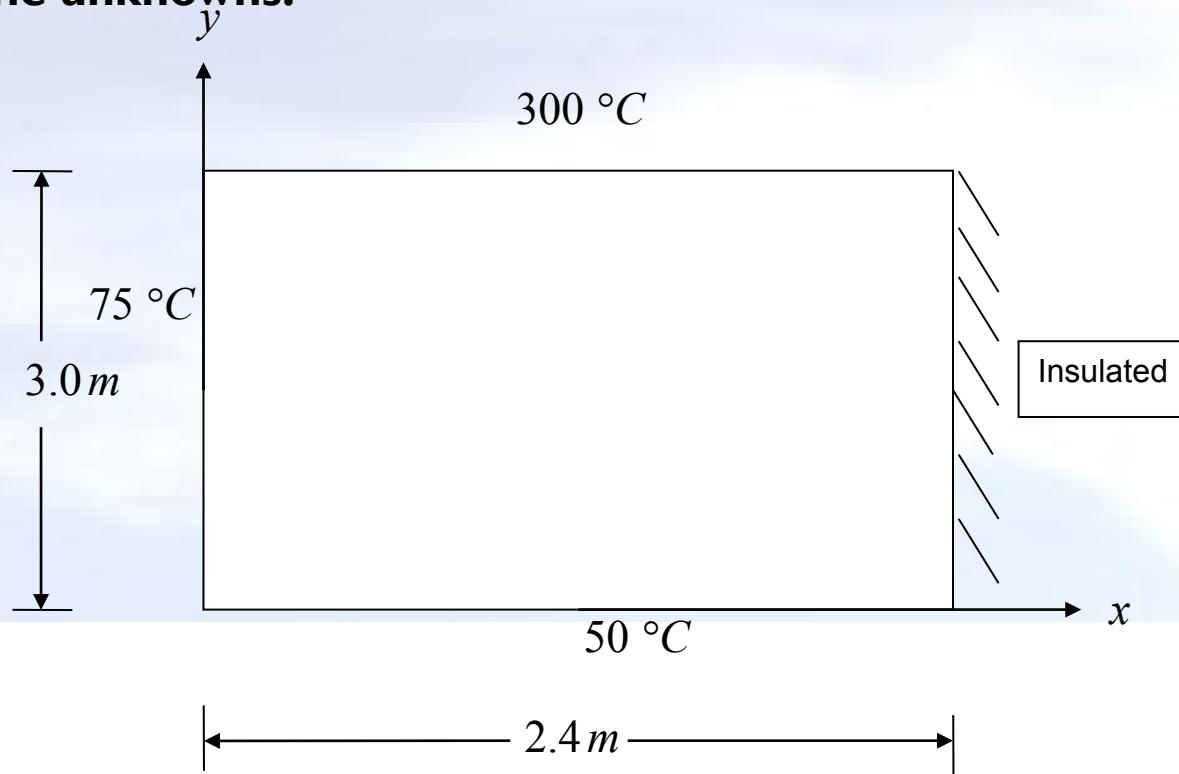


Example 3: Lieberman Method

Node	Temperature Distribution in the Plate (°C)			
	Number of Iterations			
	1	2	9	Exact
$T_{1,1}$	43.7500	52.2813	73.7832	
$T_{1,2}$	41.5625	51.3133	92.9758	
$T_{1,3}$	40.7969	87.0125	119.9378	
$T_{1,4}$	145.5289	160.9353	173.3937	
$T_{2,1}$	32.8125	54.1789	77.5449	
$T_{2,2}$	26.0313	57.9731	103.3285	
$T_{2,3}$	23.3898	122.0937	138.3236	
$T_{2,4}$	164.1216	215.6582	198.5498	
$T_{3,1}$	63.9844	69.1458	82.9805	
$T_{3,2}$	66.5055	76.1516	104.3815	
$T_{3,3}$	66.4634	155.0472	131.2525	
$T_{3,4}$	220.7047	181.4650	182.4230	

Alternative Boundary Conditions

- In Examples 1-3, the boundary conditions on the plate had a specified temperature on each edge. What if the conditions are different? For example, what if one of the edges of the plate is insulated.
- In this case, the boundary condition would be the derivative of the temperature. Because if the right edge of the plate is insulated, then the temperatures on the right edge nodes also become unknowns.



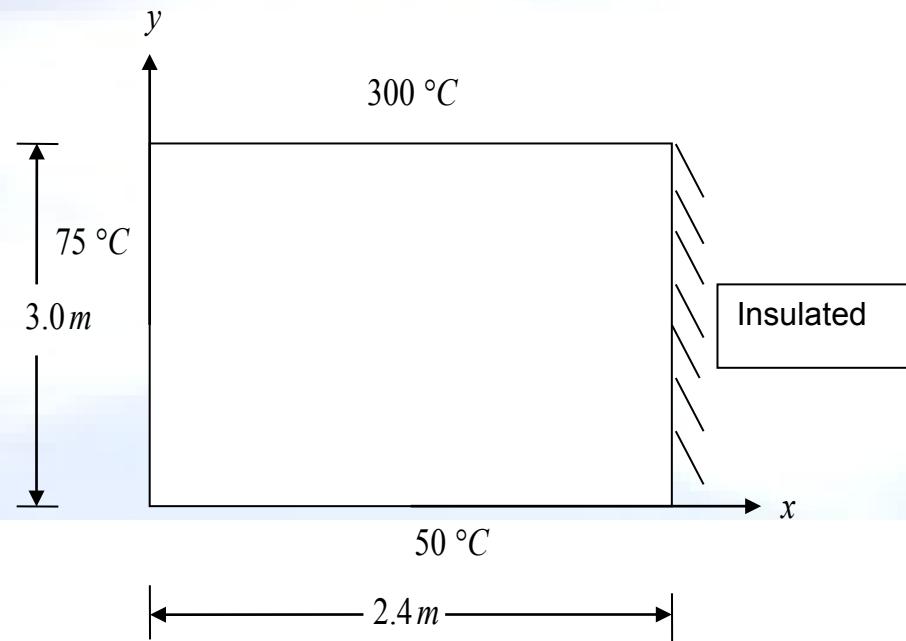
Alternative Boundary Conditions

- The finite difference equation in this case for the right edge for the nodes (m, j) for $j = 2, 3, \dots, n - 1$

$$T_{m+1,j} + T_{m-1,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0$$

- However the node $(m + 1, j)$ is not inside the plate. The derivative boundary condition needs to be used to account for these additional unknown nodal temperatures on the right edge. This is done by approximating the derivative at the edge node as (m, j)

$$\left. \frac{\partial T}{\partial x} \right|_{m,j} \approx \frac{T_{m+1,j} - T_{m-1,j}}{2(\Delta x)}$$



Alternative Boundary Conditions

- Rearranging this approximation gives us,

$$T_{m+1,j} = T_{m-1,j} + 2(\Delta x) \frac{\partial T}{\partial x} \Big|_{m,j}$$

- We can then substitute this into the original equation gives us,

$$2T_{m-1,j} + 2(\Delta x) \frac{\partial T}{\partial x} \Big|_{m,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0$$

- Recall that if the edge is insulated then,

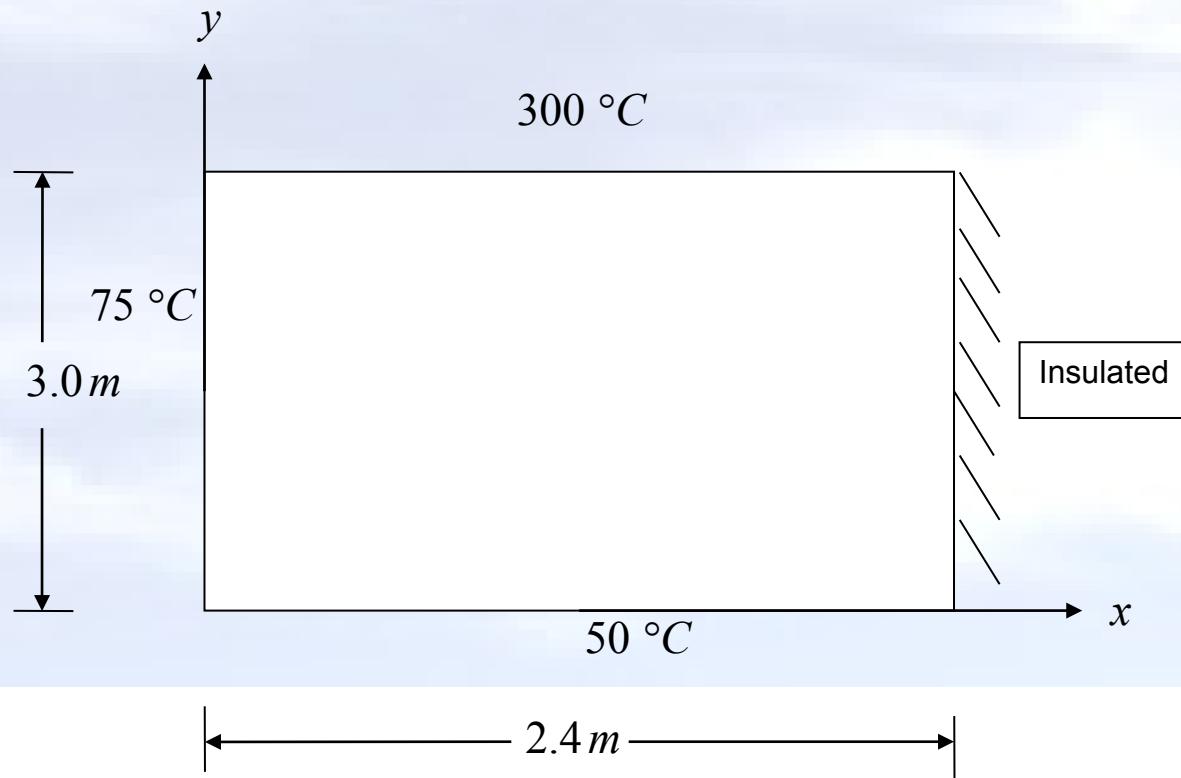
$$\frac{\partial T}{\partial x} \Big|_{m,j} = 0$$

- Substituting this again yields,

$$2T_{m-1,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0$$

Example 3: Alternative Boundary Conditions

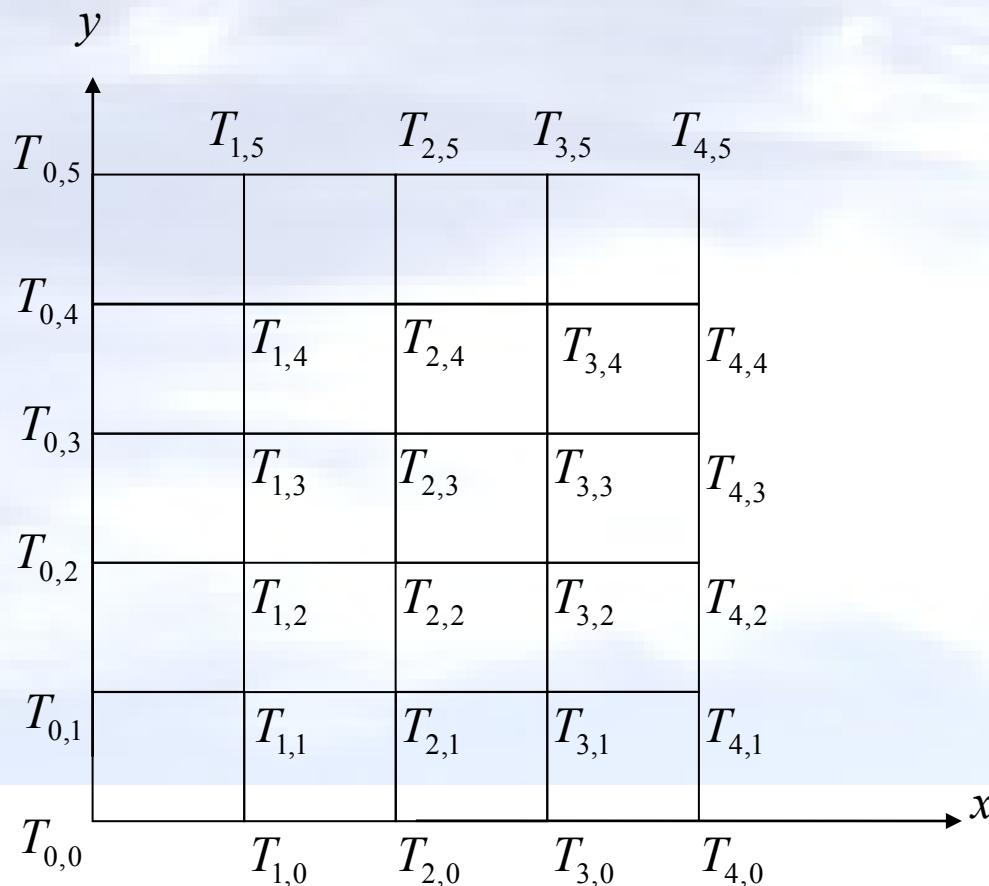
A plate $2.4\text{ m} \times 3.0\text{ m}$ is subjected to the temperatures and insulated boundary conditions as shown in Fig. 12. Use a square grid length of 0.6 m . Assume the initial temperatures at all of the interior nodes to be 0°C . Find the temperatures at the interior nodes using the direct method.



Example 3: Alternative Boundary Conditions

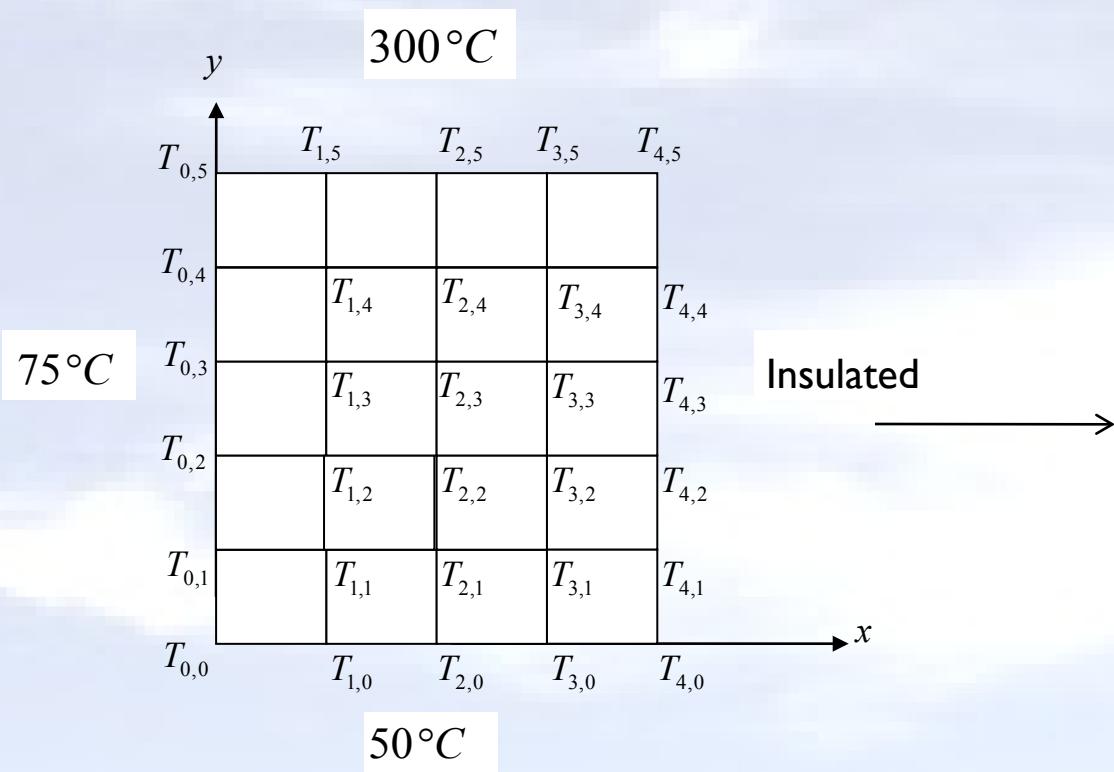
We can discretize the plate taking,

$$\Delta x = \Delta y = 0.6m$$



Example 3: Alternative Boundary Conditions

We can also develop equations for the boundary conditions to define the temperature of the exterior nodes.



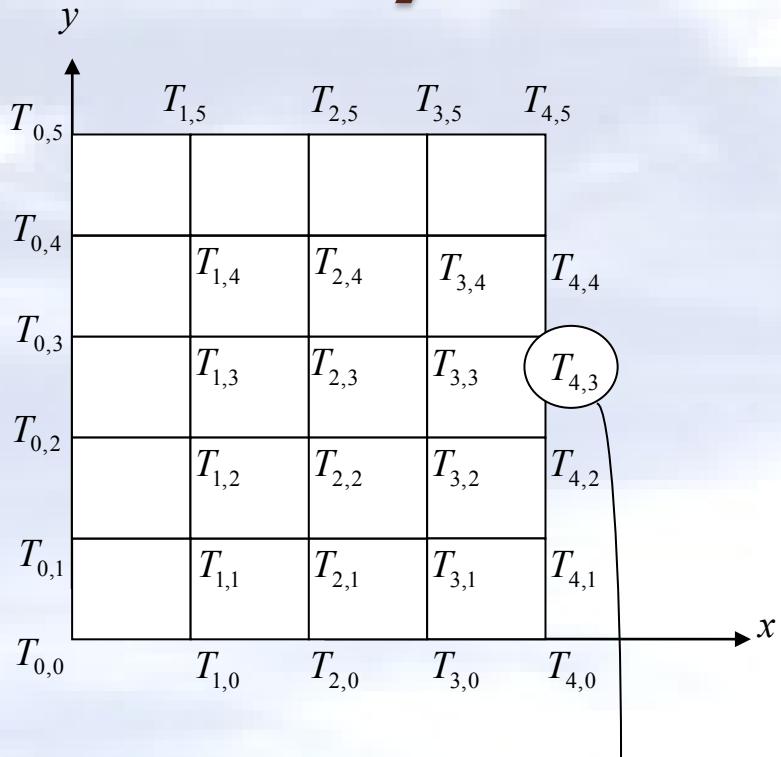
$$T_{0,j} = 75; j = 1, 2, 3, 4$$

$$T_{i,0} = 50; i = 1, 2, 3, 4$$

$$T_{i,5} = 300; i = 1, 2, 3, 4$$

$$\left. \frac{\partial T}{\partial x} \right|_{4,j} = 0; j = 1, 2, 3, 4$$

Example 3: Alternative Boundary Conditions



Here we develop the equation for the temperature at the node (4,3), to show the effects of the alternative boundary condition.

$$\underline{i=4 \text{ and } j=3} \quad 2T_{3,3} + T_{4,2} + T_{4,4} - 4T_{4,3} = 0$$

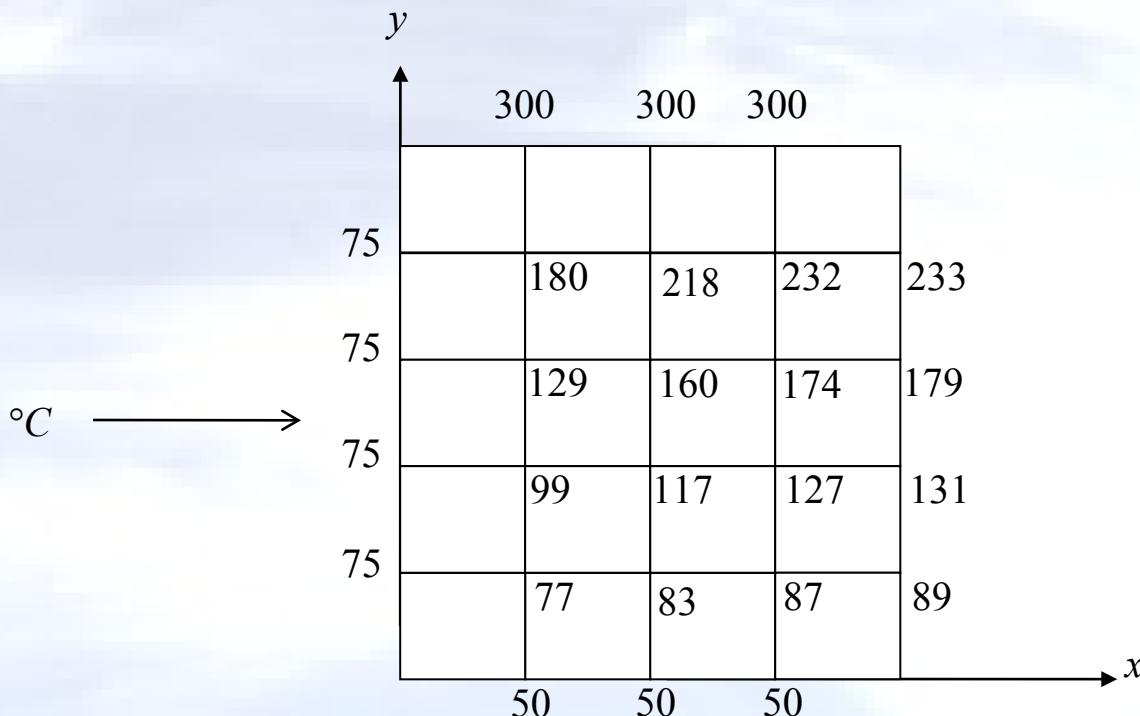
$$2T_{3,3} + T_{4,2} - 4T_{4,3} + T_{4,4} = 0$$

Example 3: Alternative Boundary Conditions

The addition of the equations for the boundary conditions gives us a system of 16 equations with 16 unknowns.

Solving yields:

$$\begin{bmatrix} T_{1,1} \\ T_{1,2} \\ T_{1,3} \\ T_{1,4} \\ T_{2,1} \\ T_{2,2} \\ T_{2,3} \\ T_{2,4} \\ T_{3,1} \\ T_{3,2} \\ T_{3,3} \\ T_{3,4} \\ T_{4,1} \\ T_{4,2} \\ T_{4,3} \\ T_{4,4} \end{bmatrix} = \begin{bmatrix} 76.8254 \\ 99.4444 \\ 128.617 \\ 180.410 \\ 82.8571 \\ 117.335 \\ 159.614 \\ 218.021 \\ 87.2678 \\ 127.426 \\ 174.483 \\ 232.060 \\ 88.7882 \\ 130.617 \\ 178.830 \\ 232.738 \end{bmatrix} {}^{\circ}\text{C}$$



The background of the image is a soft-focus photograph of a sky filled with white and light blue clouds. A bright, yellowish-white sun is positioned in the upper left quadrant, casting a warm glow and creating lens flare effects.

THE END

Introduction to Fourier Series

Part: Introduction to Fourier Series

<http://numericalmethods.eng.usf.edu>

For more details on this topic

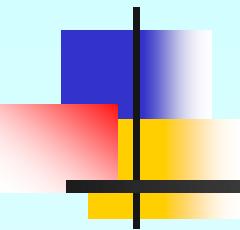
- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Introduction to Fourier Series

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 1

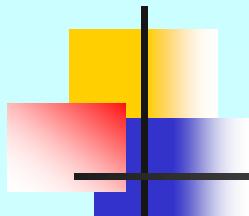
Chapter 11.01: Introduction to Fourier Series

Major: All Engineering Majors

Authors: Duc Nguyen

<http://numericalmethods.eng.usf.edu>

Numerical Methods for STEM undergraduates

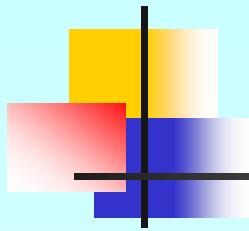


Background

The following relationships can be readily established

$$\int_0^T \sin(kw_0 t) dt = \int_0^T \cos(kw_0 t) dt = 0 \quad (1)$$

$$\int_0^T \sin^2(kw_0 t) dt = \int_0^T \cos^2(kw_0 t) dt = \frac{T}{2} \quad (2)$$

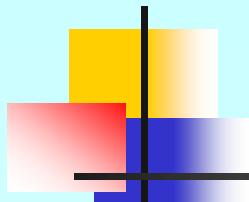


Background cont.

$$\int_0^T \cos(kw_0 t) \sin(gw_0 t) dt = 0 \quad (3)$$

$$\int_0^T \sin(kw_0 t) \sin(gw_0 t) dt = 0 \quad (4)$$

$$\int_0^T \cos(kw_0 t) \cos(gw_0 t) dt = 0 \quad (5)$$



Background cont.

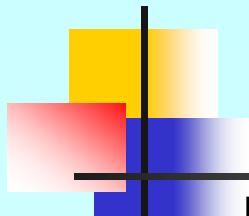
$$\omega_0 = 2\pi f \quad (6)$$

$$f = \frac{1}{T} \quad (7)$$

Where f and T represents the frequency in (cycles/time) and period (in seconds) respectively.

A periodic function with a period T should satisfy the following equation:

$$f(t + T) = f(t) \quad (8)$$



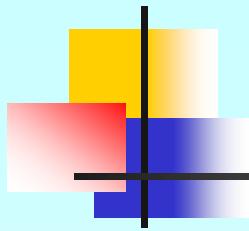
Background cont.

Example 1

Let

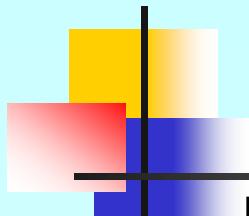
$$A = \int_0^T \sin(kw_0 t) dt \quad (9)$$

$$= -\left(\frac{1}{kw_0}\right) [\cos(kw_0 t)]_0^T$$



Background cont.

$$\begin{aligned} A &= \left(\frac{-1}{kw_0} \right) [\cos(kw_0 T) - \cos(0)] \quad (10) \\ &= \left(\frac{-1}{kw_0} \right) [\cos(k2\pi) - 1] \\ &= 0 \end{aligned}$$



Background cont.

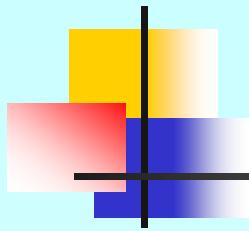
Example 2

$$\text{Let } B = \int_0^T \sin^2(kw_0 t) dt \quad (11)$$

Recall

$$\sin^2(\alpha) = \frac{1 - \cos(2\alpha)}{2} \quad (12)$$

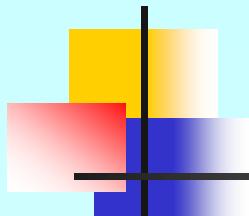
$$B = \int_0^T \left[\frac{1}{2} - \frac{1}{2} \cos(2kw_0 t) \right] dt \quad (13)$$



Background cont.

$$= \left[\left(\frac{1}{2} \right) t - \left(\frac{1}{2} \right) \left(\frac{1}{2k\omega_0} \right) \sin(2k\omega_0 t) \right]_0^T$$

$$B = \left[\frac{T}{2} - \frac{1}{4k\omega_0} \sin(2k\omega_0 T) \right] - [0] \quad (14)$$



Background cont.

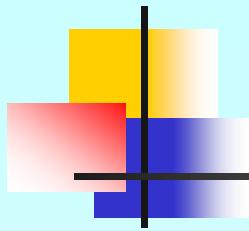
$$= \frac{T}{2} - \left(\frac{1}{4kw_0} \right) \sin(2k * 2\pi)$$

$$= \frac{T}{2}$$

Example 3

Let

$$C = \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt \quad (15)$$

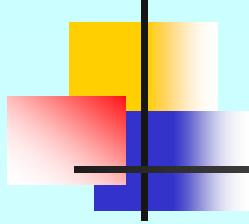


Background cont.

Recall that

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \sin(\beta)\cos(\alpha) \quad (16)$$

$$C = \int_0^T [\sin[(g + k)w_0 t] - \sin(kw_0 t)\cos(gw_0 t)] dt \quad (17)$$



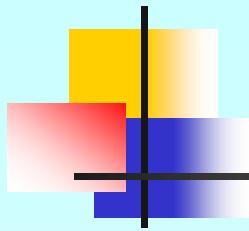
Background cont.

$$= \int_0^T \sin[(g + k)w_0 t] dt - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \quad (18)$$

$$C = 0 - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \quad (19)$$

Adding Equations (15), (19),

$$\begin{aligned} 2C &= \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \\ &= \int_0^T \sin[(gw_0 t) - (kw_0 t)] dt = \int_0^T \sin[(g - k)w_0 t] dt \end{aligned} \quad (20)$$



Background cont.

$$2C = 0,$$

since the right side of the above equation is zero
Thus,

$$C = \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt = 0 \quad (21)$$



THE END

<http://numericalmethods.eng.usf.edu>

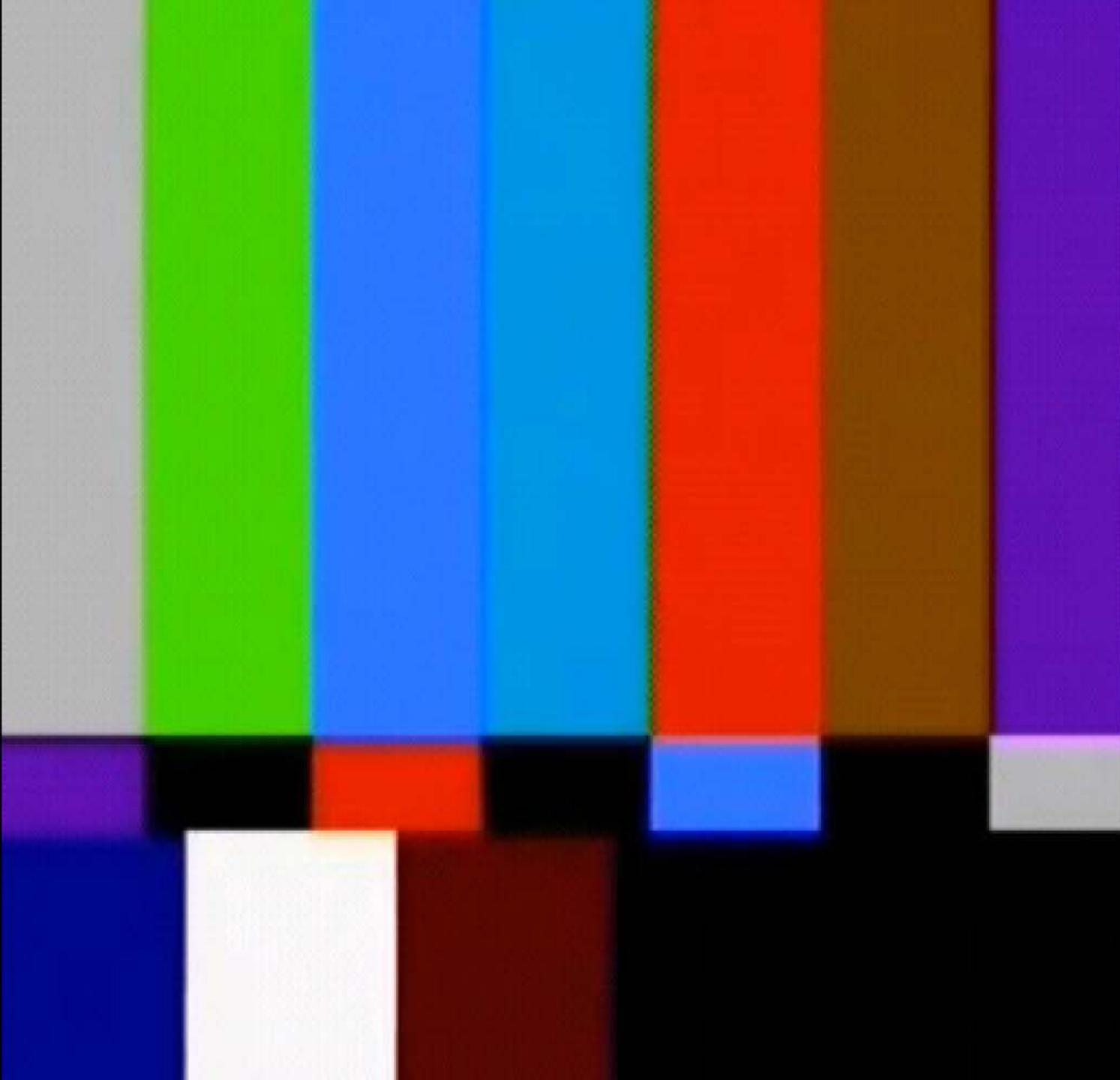
Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate
<http://numericalmethods.eng.usf.edu>
Committed to bringing numerical methods to the
undergraduate

For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Fourier Transform Pair

Part: Frequency and Time
Domain

<http://numericalmethods.eng.usf.edu>

For more details on this topic

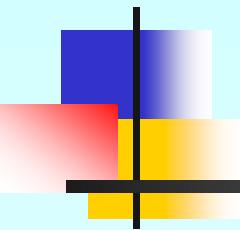
- Go to <http://numericalmethods.eng.usf.edu>
- Click on keyword
- Click on Fourier Transform Pair

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 5

Chapter 11.03: Fourier Transform Pair: Frequency and Time Domain

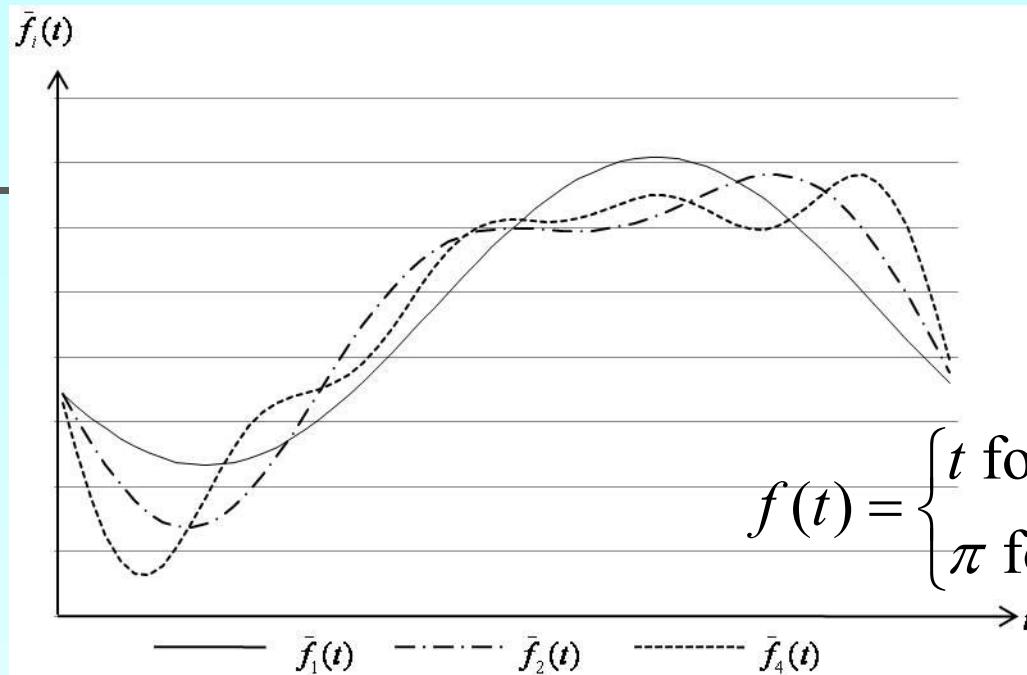
Major: All Engineering Majors

Authors: Duc Nguyen

<http://numericalmethods.eng.usf.edu>

Numerical Methods for STEM undergraduates

Example 1



$$\bar{f}_1(t) \approx a_0 + a_1 \cos(t) + b_1 \sin(t)$$

$$\bar{f}_2(t) \approx a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t)$$

$$\begin{aligned} \bar{f}_4(t) \approx & a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t) \\ & + a_3 \cos(3t) + b_3 \sin(3t) + a_4 \cos(4t) + b_4 \sin(4t) \end{aligned}$$

Frequency and Time Domain

The amplitude (vertical axis) of a given periodic function can be plotted versus time (horizontal axis), but it can also be plotted in the frequency domain as shown in Figure 2.

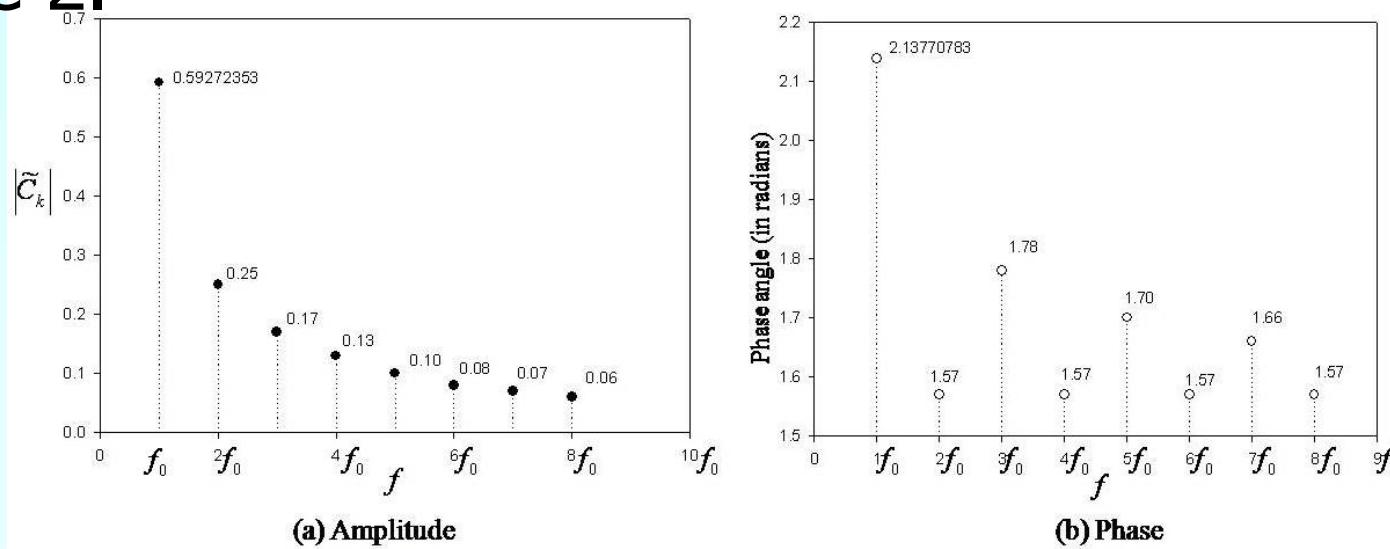
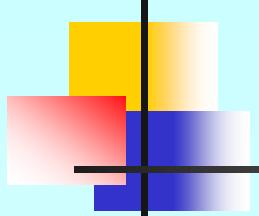


Figure 2 Periodic function (see Example 1 in Chapter 11.02 Continuous Fourier Series) in frequency domain.



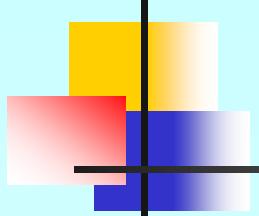
Frequency and Time Domain cont.

Figures 2(a) and 2(b) can be described with the following equations from chapter 11.02,

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t} \quad (39, \text{ repeated})$$

where

$$\tilde{C}_k = \left(\frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\} \quad (41, \text{ repeated})$$

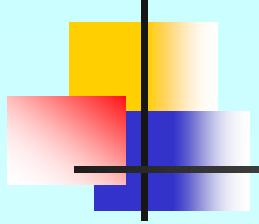


Frequency and Time Domain cont.

For the periodic function shown in Example 1 of Chapter 11.02 (Figure 1), one has:

$$w_0 = 2\pi f = \frac{2\pi}{T} = \frac{2\pi}{2\pi} = 1$$

$$\tilde{C}_k = \left(\frac{1}{T} \right) \left\{ \int_0^{\pi} t \times e^{-ikt} dt + \int_{\pi}^{2\pi} \pi \times e^{-ikt} dt \right\}$$



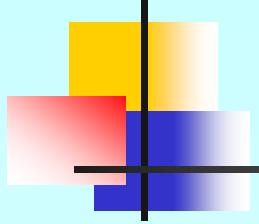
Frequency and Time Domain cont.

Define:

$$A \equiv \int_0^{\pi} t \times e^{-ikt} dt = \left[t \times \left(\frac{-1}{ik} \right) e^{-ikt} \right]_0^{\pi} + \int_0^{\pi} \left(\frac{1}{ik} \right) e^{-ikt} dt$$

or

$$\begin{aligned} A &= \left[\left(\frac{-\pi}{ik} \right) e^{-ik\pi} \right] + \left(\frac{1}{k^2} \right) [e^{-ik\pi} - 1] \\ &= \left[\left(\left(\frac{\pi i}{k} \right) e^{-ik\pi} + \left(\frac{1}{k^2} \right) e^{-ik\pi} - \frac{1}{k^2} \right) \right] \end{aligned}$$

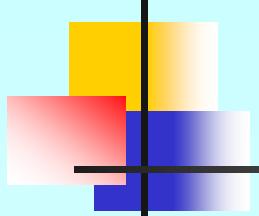


Frequency and Time Domain cont.

Also,

$$B \equiv \pi \int_{-\pi}^{2\pi} e^{-ikt} dt = \left[\left(e^{-ikt} \right) \left(\frac{-\pi}{ik} \right) \right]_{-\pi}^{2\pi}$$

$$B = \left(\frac{-\pi}{ik} \right) [e^{-ik2\pi} - e^{-ik\pi}] = \left(\frac{\pi i}{k} \right) [e^{-ik2\pi} - e^{-ik\pi}]$$



Frequency and Time Domain cont.

Thus:

$$\tilde{C}_k = \left(\frac{1}{2\pi} \right) \{A + B\}$$

$$\tilde{C}_k = \left(\frac{1}{2\pi} \right) \left\{ e^{-ik\pi} \left(\frac{\pi i}{k} + \frac{1}{k^2} - \frac{\pi i}{k} \right) - \frac{1}{k^2} + \left(\frac{\pi i}{k} \right) e^{-ik2\pi} \right\}$$

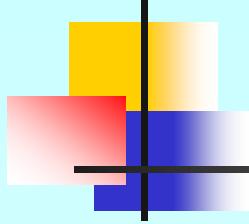
Using the following Euler identities

$$e^{-ik\pi} = \cos(-k\pi) + i \sin(-k\pi)$$

$$= \cos(k\pi) - i \sin(k\pi)$$

$$= \cos(k\pi)$$

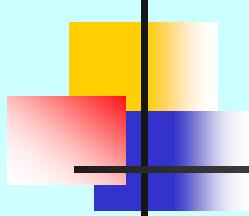
$$e^{-ik2\pi} = \cos(k2\pi) - i \sin(k2\pi) = \cos(k2\pi)$$



Frequency and Time Domain cont.

Noting that $\cos(k2\pi) = 1$ for any integer k

$$\tilde{C}_k = \left(\frac{1}{2\pi} \right) \left\{ \cos(k\pi) \times \left(\frac{1}{k^2} \right) - \frac{1}{k^2} + \left(\frac{\pi i}{k} \right) \cos(k2\cancel{\pi}_1) \right\}$$



Frequency and Time Domain cont.

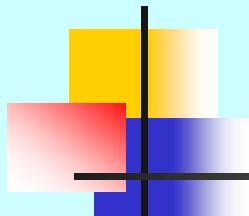
Also,

$$\cos(k\pi) = \begin{cases} -1 & \text{for } k = \text{odd number } (= 1, 3, 5, 7, \dots) \\ +1 & \text{for } k = \text{even number } (= 2, 4, 6, 8, \dots) \end{cases}$$

Thus,

$$\tilde{C}_k = \left(\frac{1}{2\pi} \right) \left\{ \frac{(-1)^k}{k^2} - \frac{1}{k^2} + \frac{\pi i}{k} \right\}$$

$$\tilde{C}_k = \left(\frac{1}{2\pi k^2} \right) [(-1)^k - 1] + \left(\frac{1}{2k} \right) i$$



Frequency and Time Domain cont.

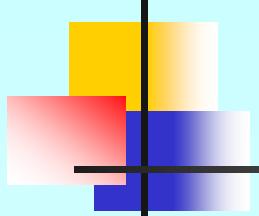
From Equation (36, Ch. 11.02), one has

$$\tilde{C}_k = \frac{a_k - ib_k}{2} \quad (36, \text{ repeated})$$

Hence; upon comparing the previous 2 equations, one concludes:

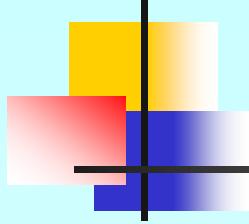
$$a_k \equiv \left(\frac{1}{\pi k^2} \right) [(-1)^k - 1]$$

$$b_k = \left(\frac{-1}{k} \right)$$



Frequency and Time Domain cont.

For $k = 1, 2, 3, 4 \dots 8$; the values for a_k and b_k (based on the previous 2 formulas) are exactly identical as the ones presented earlier in Example 1 of Chapter 11.02.

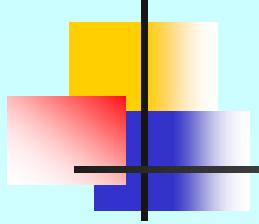


Frequency and Time Domain cont.

Thus:

$$\tilde{C}_1 = \frac{a_1 - ib_1}{2} = \frac{\frac{-2}{\pi} - i(-1)}{2} = \frac{-1}{\pi} + \frac{1}{2}i$$

$$\tilde{C}_2 = \frac{a_2 - ib_2}{2} = \frac{0 - i\left(-\frac{1}{2}\right)}{2} = 0 + \frac{1}{4}i$$

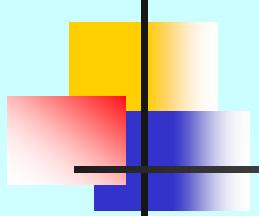


Frequency and Time Domain cont.

$$\tilde{C}_3 = \frac{a_3 - ib_3}{2} = \frac{\left(\frac{-2}{9\pi}\right) - i\left(\frac{-1}{3}\right)}{2} = \left(\frac{-1}{9\pi}\right) + \frac{1}{6}i$$

$$\tilde{C}_4 = \frac{a_4 - ib_4}{2} = \frac{0 - i\left(\frac{-1}{4}\right)}{2} = 0 + \frac{1}{8}i$$

$$\tilde{C}_5 = \frac{a_5 - ib_5}{2} = \frac{\left(\frac{-2}{25\pi}\right) - i\left(\frac{-1}{5}\right)}{2} = \left(\frac{-1}{25\pi}\right) + \frac{1}{10}i$$

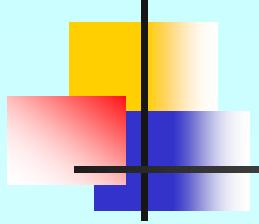


Frequency and Time Domain cont.

$$\tilde{C}_6 = \frac{a_6 - ib_6}{2} = \frac{0 - i\left(\frac{-1}{6}\right)}{2} = 0 + \frac{1}{12}i$$

$$\tilde{C}_7 = \frac{a_7 - ib_7}{2} = \frac{\left(\frac{-2}{49\pi}\right) - i\left(\frac{-1}{7}\right)}{2} = \left(\frac{-1}{49\pi}\right) + \frac{1}{14}i$$

$$\tilde{C}_8 = \frac{a_8 - ib_8}{2} = \frac{0 - i\left(\frac{-1}{8}\right)}{2} = 0 + \frac{1}{16}i$$



Frequency and Time Domain cont.

In general, one has

$$\tilde{C}_k = \begin{cases} \frac{-1}{k^2\pi} + \left(\frac{1}{2k}\right)i & \text{for } k = 1, 3, 5, 7, \dots = \text{odd number} \\ \left(\frac{1}{2k}\right)i & \text{for } k = 2, 4, 6, 8, \dots = \text{even number} \end{cases}$$



THE END

<http://numericalmethods.eng.usf.edu>

Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate
<http://numericalmethods.eng.usf.edu>
Committed to bringing numerical methods to the
undergraduate

For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Fourier Transform Pair

Part: Complex Number in Polar Coordinates

<http://numericalmethods.eng.usf.edu>

For more details on this topic

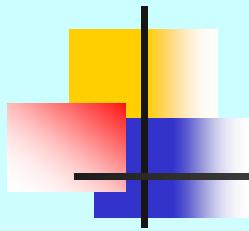
- Go to <http://numericalmethods.eng.usf.edu>
- Click on keyword
- Click on Fourier Transform Pair

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 6

Chapter 11.03: Complex number in polar coordinates (Contd.)

In Cartesian (Rectangular) Coordinates, a complex number \tilde{C}_k can be expressed as:

$$\tilde{C}_k = R_k + (I_k)i$$

In Polar Coordinates, a complex number \tilde{C}_k can be expressed as:

$$\tilde{C}_k = Ae^{i\theta} = A\{\cos(\theta) + i \sin(\theta)\} = \{A \cos(\theta)\} + \{A \sin(\theta)\}i$$

Complex number in polar coordinates cont.

Thus, one obtains the following relations between the Cartesian and polar coordinate systems:

$$R_k = A \cos(\theta) \quad I_k = A \sin(\theta)$$

This is represented graphically in Figure 3.

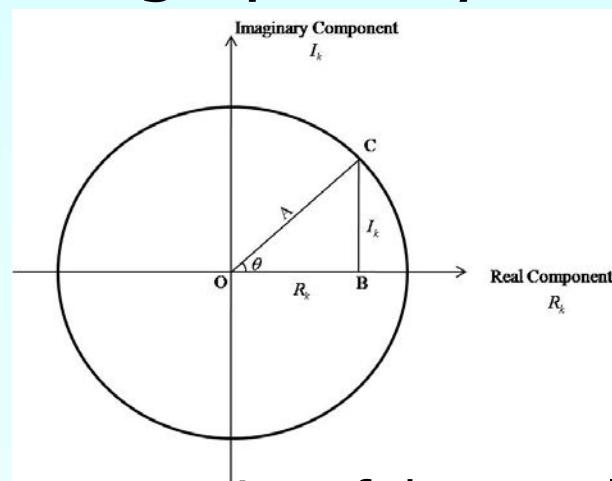
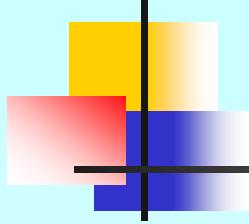


Figure 3. Graphical representation of the complex number system in polar coordinates.



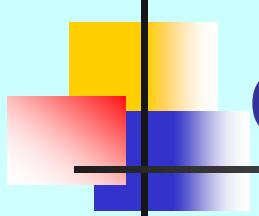
Complex number in polar coordinates cont.

Hence

$$R_k^2 + I_k^2 = A^2 \cos^2(\theta) + A^2 \sin^2(\theta) = A^2 [\cos^2(\theta) + \sin^2(\theta)]$$

$$\cos(\theta) = \frac{R_k}{A} \Rightarrow \theta = \cos^{-1}\left(\frac{R_k}{A}\right) \quad \text{and}$$

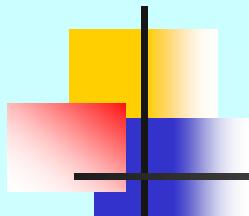
$$\sin(\theta) = \frac{I_k}{A} \Rightarrow \theta = \sin^{-1}\left(\frac{I_k}{A}\right)$$



Complex number in polar coordinates cont.

Based on the above 3 formulas, the complex numbers \tilde{C}_k can be expressed as:

$$\tilde{C}_1 = \frac{-1}{\pi} + \left(\frac{1}{2} \right) i = (0.59272353)e^{i(2.13770783)}$$



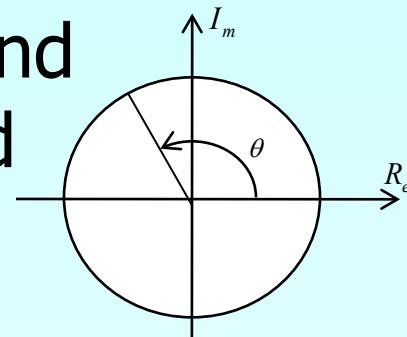
Complex number in polar coordinates cont.

Notes:

(a) The amplitude and angle \tilde{C}_1 are 0.59 and 2.14 respectively (also see Figures 2a, and 2b in chapter 11.03).

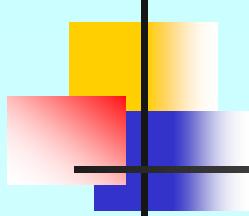
(b) The angle θ (in radian) obtained from

$$\cos(\theta) = \frac{R_k}{A} \text{ will be } 2.138 \text{ radians } (=122.48^\circ).$$



However based on $\sin(\theta) = \frac{I_k}{A}$

Then $\theta = 1.004$ radians ($=57.52^\circ$).

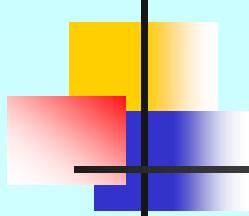


Complex number in polar coordinates cont.

Since the Real and Imaginary components of θ are negative and positive, respectively, the proper selection for θ should be 2.1377 radians.

$$\tilde{C}_2 = 0 + \frac{1}{4}i = (0.25)e^{i\left(\frac{\pi}{2}\right)} = (0.25)e^{i(1.57079633)}$$

$$\tilde{C}_3 = \left(\frac{-1}{9\pi}\right) + \frac{1}{6}i = (0.17037798)e^{i(1.77990097)}$$

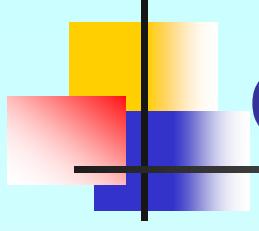


Complex number in polar coordinates cont.

$$\tilde{C}_4 = 0 + \frac{1}{8}i = (0.125)e^{i\left(\frac{\pi}{2}\right)} = (0.125)e^{i(1.57079633)}$$

$$\tilde{C}_5 = \left(\frac{-1}{25\pi} \right) + \frac{1}{10}i = (0.100807311)e^{i(1.69743886)}$$

$$\tilde{C}_6 = 0 + \frac{1}{12}i = (0.08333333)e^{i\left(\frac{\pi}{2}\right)} = (0.08333333)e^{i(1.57079633)}$$



Complex number in polar coordinates cont.

$$\tilde{C}_7 = \left(\frac{-1}{49\pi} \right) + \frac{1}{14}i = (0.07172336)e^{i(1.66149251)}$$

$$\tilde{C}_8 = 0 + \frac{1}{16}i = (0.0625)e^{i\left(\frac{\pi}{2}\right)}$$



THE END

<http://numericalmethods.eng.usf.edu>

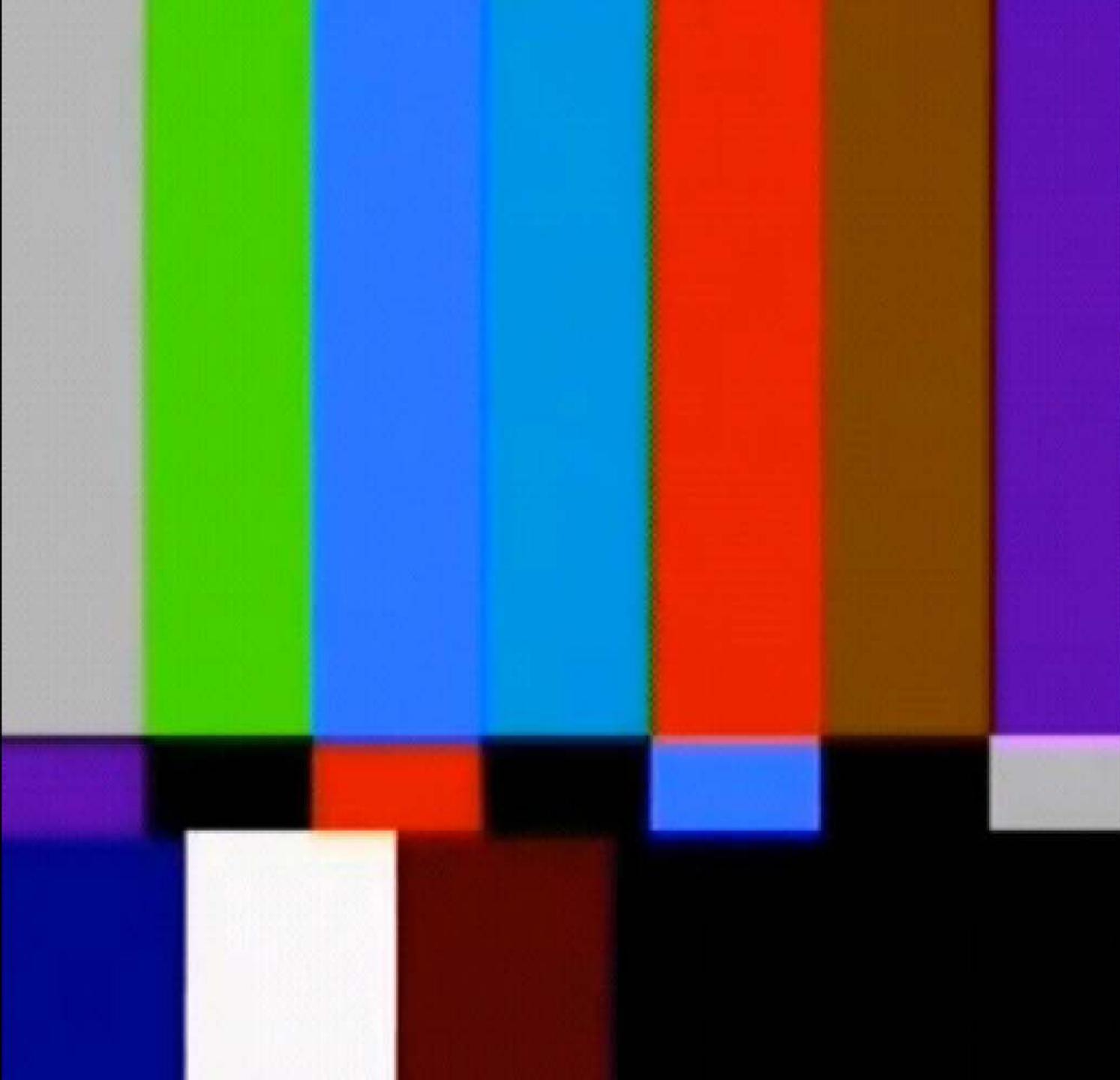
Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate
<http://numericalmethods.eng.usf.edu>
Committed to bringing numerical methods to the
undergraduate

For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Fourier Transform Pair

Part: Non-Periodic Functions

<http://numericalmethods.eng.usf.edu>

For more details on this topic

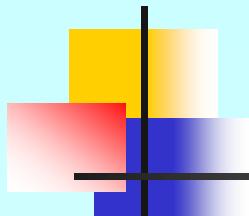
- Go to <http://numericalmethods.eng.usf.edu>
- Click on keyword
- Click on Fourier Transform Pair

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 7

Chapter 11. 03: Non-Periodic Functions (Contd.)

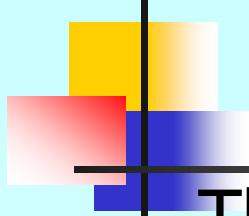
Recall

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t} \quad (39, \text{ repeated})$$

$$\tilde{C}_k = \left(\frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\} \quad (41, \text{ repeated})$$

Define

$$\hat{F}(ikw_0) = \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-ikw_0 t} dt \quad (1)$$



Non-Periodic Functions

Then, Equation (41) can be written as

$$\tilde{C}_k = \left(\frac{1}{T} \right) \times \hat{F}(ik\omega_0)$$

And Equation (39) becomes

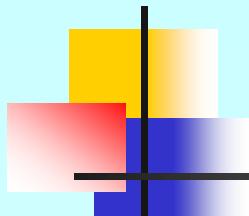
$$f(t) = \sum_{k=-\infty}^{\infty} \left(\frac{1}{T} \right) \times \hat{F}(ik\omega_0) e^{ik\omega_0 t}$$

From above equation

$$f_{np}(t) = \lim_{\substack{T \rightarrow \infty \\ \text{or } \Delta f \rightarrow 0}} f(t) = \lim_{\Delta f \rightarrow 0} \sum_{k=-\infty}^{\infty} (\Delta f) \times \hat{F}(ik\omega_0) e^{ik\omega_0 t}$$

or

$$f_{np}(t) = \lim_{\Delta f \rightarrow 0} \sum_{k=-\infty}^{\infty} (\Delta f) \times \hat{F}(ik2\pi\Delta f) e^{ik2\pi\Delta ft}$$



Non-Periodic Functions cont.

From Figure 4,

$$k\Delta f = f$$

$$f_{np}(t) = \int df \times \hat{F}(i2\pi f) e^{i2\pi ft}$$

$$f_{np}(t) = \int \hat{F}(i2\pi f) e^{i2\pi ft} df$$

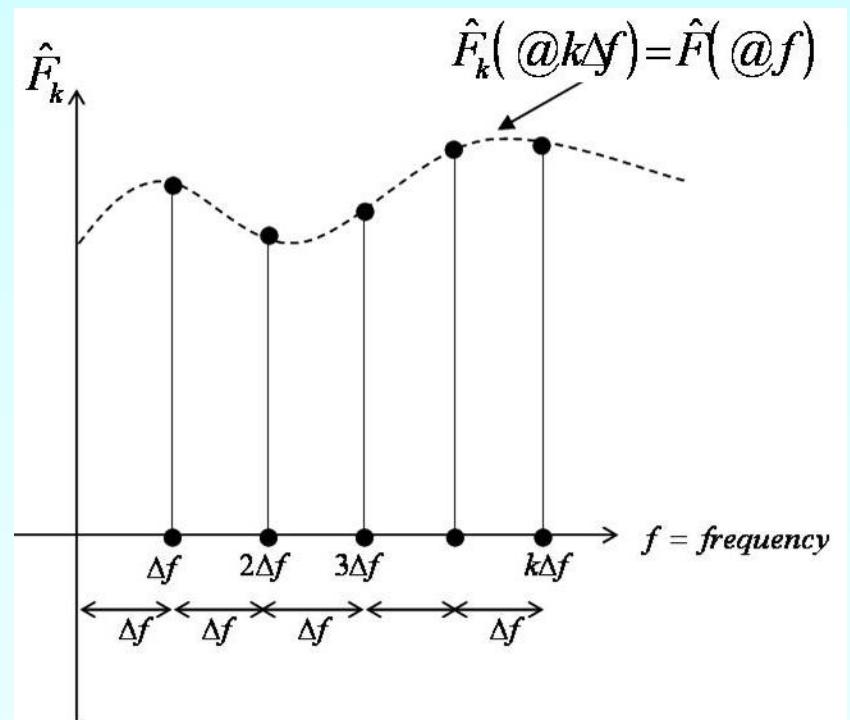
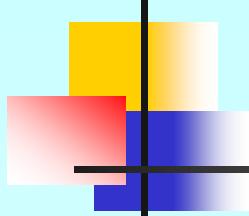


Figure 4. Frequency are discretized.



Non-Periodic Functions cont.

Multiplying and dividing the right-hand-side of the equation by 2π , one obtains

$$f_{np}(t) = \left(\frac{1}{2\pi} \right) \int_{-\infty}^{\infty} \hat{F}(iw_0) e^{iw_0 t} d(w_0); \text{ inverse Fourier transform}$$

Also, using the definition stated in Equation (1), one gets

$$\hat{F}(iw_0) = \int_{-\infty}^{\infty} f_{np}(t) e^{-iw_0 t} d(t); \text{ Fourier transform}$$



THE END

<http://numericalmethods.eng.usf.edu>

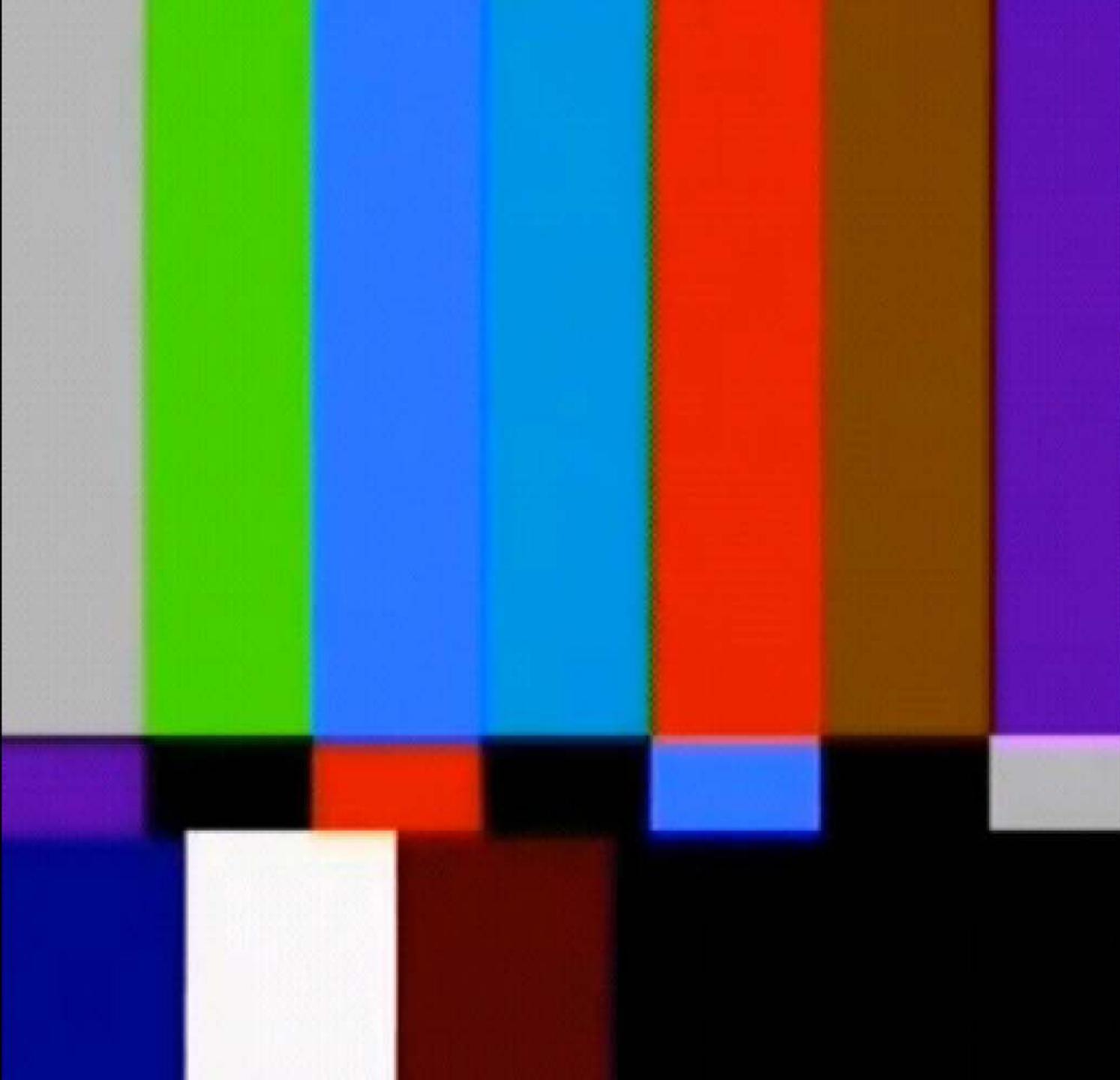
Acknowledgement

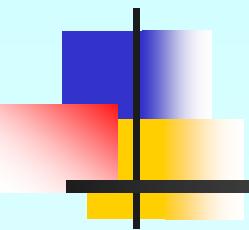
This instructional power point brought to you by
Numerical Methods for STEM undergraduate
<http://numericalmethods.eng.usf.edu>
Committed to bringing numerical methods to the
undergraduate

For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.





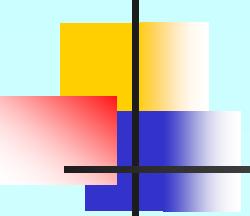
Discrete Fourier Transform (DFT)

Major: All Engineering Majors

Authors: Duc Nguyen

<http://numericalmethods.eng.usf.edu>

Numerical Methods for STEM undergraduates



Discrete Fourier Transform

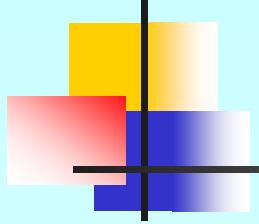
Recalled the exponential form of Fourier series (see Eqs. 26, 28 in Ch. 11.01), one gets:

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t} \quad (26, \text{ repeated})$$

$$\tilde{C}_k = \left(\frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\} \quad (28, \text{ repeated})$$

If time " t " is discretized at $t_1 = \Delta t, t_2 = 2\Delta t, t_3 = 3\Delta t, \dots, t_n = n\Delta t$, then Eq. (26) becomes:

$$f(t_n) = \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 t_n} \quad (1)$$



Discrete Fourier Transform cont.

To simplify the notation, define:

$$t_n = n \quad (2)$$

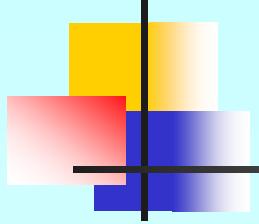
Then, Eq. (2) can be written as:

$$f(n) = \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 n} \quad (3)$$

Multiplying both sides of Eq. (3) by $e^{-ilw_0 n}$, and performing the summation on “ n ”, one obtains (note: $/ =$ integer number)

$$\sum_{n=0}^{N-1} f(n) \times e^{-ilw_0 n} = \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 n} \times e^{-ilw_0 n} \quad (4)$$

$$= \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{C}_k e^{i(k-l)\frac{2\pi}{N} n} \quad (5)$$



Discrete Fourier Transform cont.

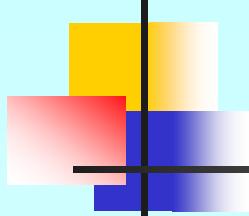
Switching the order of summations on the right-hand-side of Eq.(5), one obtains:

$$\sum_{n=0}^{N-1} f(n) \times e^{-il\left(\frac{2\pi}{N}\right)n} = \sum_{k=0}^{N-1} \tilde{C}_k \sum_{n=0}^{N-1} e^{i(k-l)\left(\frac{2\pi}{N}\right)n} \quad (6)$$

Define:

$$A = \sum_{n=0}^{N-1} e^{i(k-l)\left(\frac{2\pi}{N}\right)n} \quad (7)$$

There are 2 possibilities for $(k - l)$ to be considered in Eq. (7)



Discrete Fourier Transform—Case 1

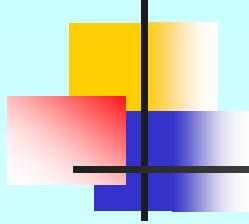
Case(1): $(k-l)$ is a multiple integer of N , such as: $(k-l) = mN$; or $k = \pm mN$ where $m = 0, \pm 1, \pm 2, \dots$

Thus, Eq. (7) becomes:

$$A = \sum_{n=0}^{N-1} e^{im2\pi n} = \sum_{n=0}^{N-1} \cos(mn2\pi) + i \sin(mn2\pi) \quad (8)$$

Hence:

$$A = N \quad (9)$$



Discrete Fourier Transform—Case 2

Case(2): $(k - l)$ is NOT a multiple integer of N . In this case, from Eq. (7) one has:

$$A = \sum_{n=0}^{N-1} \left\{ e^{i(k-l)\left(\frac{2\pi}{N}\right)} \right\}^n \quad (10)$$

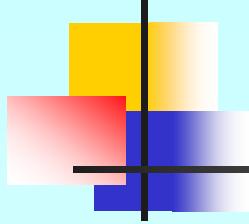
Define:

$$a = e^{i(k-l)\frac{2\pi}{N}} = \cos\left\{(k-l)\frac{2\pi}{N}\right\} + i \sin\left\{(k-l)\frac{2\pi}{N}\right\} \quad (11)$$

$a \neq 1$; because $(k - l)$ is “NOT” a multiple integer of N

Then, Eq. (10) can be expressed as:

$$A = \sum_{n=0}^{N-1} \{a\}^n \quad (12)$$



Discrete Fourier Transform—Case 2

From mathematical handbooks, the right side of Eq. (12) represents the “geometric series”, and can be expressed as:

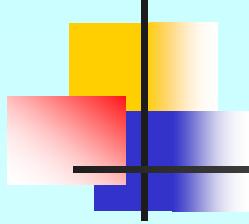
$$A = \sum_{n=0}^{N-1} \{a\}^n = N; \text{ if } a = 1 \quad (13)$$

$$= \frac{1-a^N}{1-a}; \text{ if } a \neq 1 \quad (14)$$

Because of Eq. (11), hence Eq. (14) should be used to compute A . Thus:

$$A = \frac{1-a^N}{1-a} = \frac{1-e^{i(k-l)2\pi}}{1-a} \quad (\text{See Eq. (10)}) \quad (15)$$

$$e^{i(k-l)2\pi} \equiv \cos\{(k-l)2\pi\} + i \sin\{(k-l)2\pi\} = 1 \quad (16)$$



Discrete Fourier Transform—Case 2

Substituting Eq. (16) into Eq. (15), one gets

$$A = 0 \quad (17)$$

Thus, combining the results of case 1 and case 2, we get

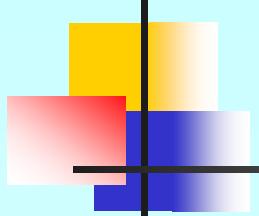
$$A = N + 0 = N \quad (18)$$

Substituting Eq.(18) into Eq.(7), and then referring to Eq.(6), one gets:

$$\sum_{n=0}^{N-1} f(n)e^{-ilw_0n} = \sum_{k=0}^{N-1} \tilde{C}_k \times N \quad (18A)$$

Recall $k = l + mN$ (where l, m are integer numbers), and since k must be in the range $0 \rightarrow N - 1$, $m=0$. Thus:

$$k = l + mN \text{ becomes } k = l$$



Discrete Fourier Transform—Case 2

Eq. (18A) can, therefore, be simplified to

$$\sum_{n=0}^{N-1} f(n)e^{-ilw_0 n} = \tilde{C}_l \times N \quad (18B)$$

Thus:

$$\tilde{C}_k = \left(\frac{1}{N} \right) \sum_{n=0}^{N-1} f(n)e^{-ikw_0 n} = \left(\frac{1}{N} \right) \sum_{n=0}^{N-1} f(n) \{ \cos(lw_0 n) - i \sin(lw_0 n) \} \quad (19)$$

where $n \equiv t_n$ and

$$f(n) = \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 n} = \sum_{k=0}^{N-1} \tilde{C}_k \{ \cos(kw_0 n) + i \sin(kw_0 n) \} \quad (1, \text{ repeated})$$

Aliasing Phenomenon, Nyquist samples, Nyquist rate

When a function $f(t)$, which may represent the signals from some real-life phenomenon (shown in Figure 1), is sampled, it basically converts that function into a sequence $\tilde{f}(k)$ at discrete locations of t .

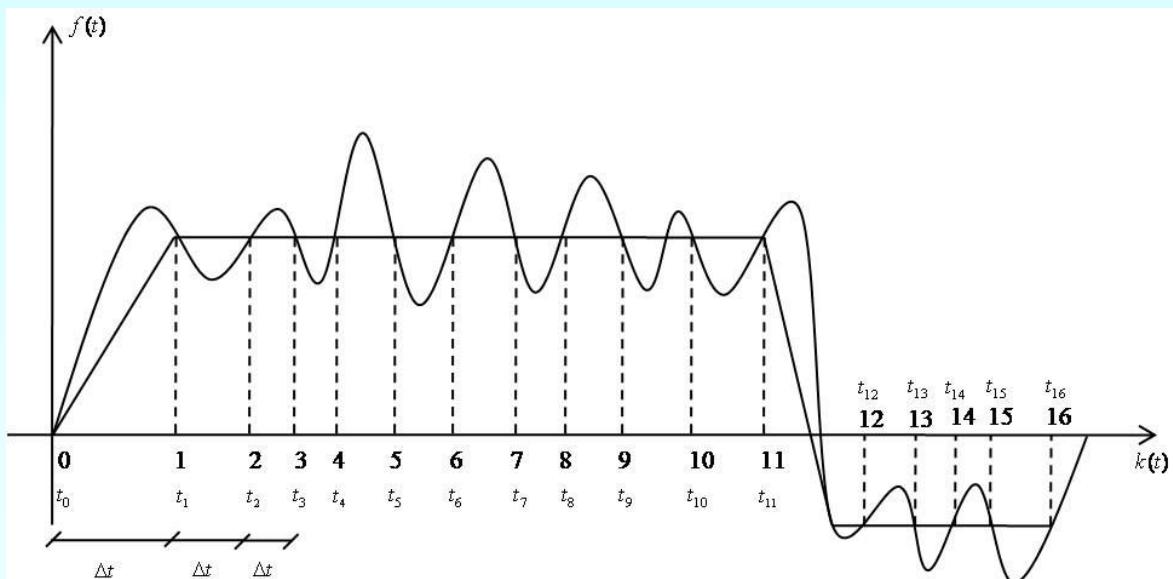
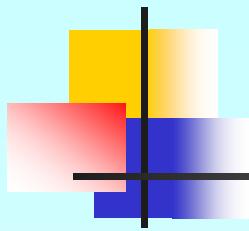


Figure 1 Function to be sampled and "Aliased" sample problem.

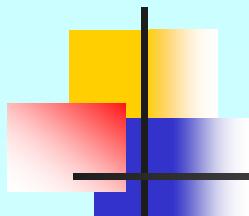


Aliasing Phenomenon, Nyquist samples, Nyquist rate cont.

Thus, $\tilde{f}(k)$ represents the value of $f(t)$ at $t = t_0 + k\Delta t$, where t_0 is the location of the first sample (at $k = 0$).

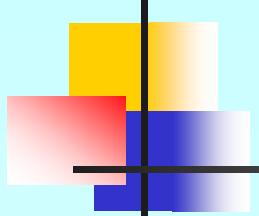
In Figure 1, the samples have been taken with a fairly large Δt . Thus, these sequence of discrete data will not be able to recover the original signal function $f(t)$.

For example, if all discrete values of $f(t)$ were connected by piecewise linear fashion, then a nearly horizontal straight line will occur between t_1 through t_8 and t_9 through t_{12} respectively (See Figure 1).



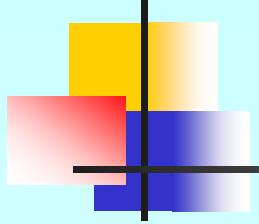
Aliasing Phenomenon, Nyquist samples, Nyquist rate cont.

These piecewise linear interpolation (or other interpolation schemes) will NOT produce a curve which closely resembles the original function $f(t)$. This is the case where the data has been “ALIASED”.



“Windowing” phenomenon

Another potential difficulty in sampling the function is called “windowing” problem. As indicated in Figure 2, while Δt is small enough so that a piecewise linear interpolation for connecting these discrete values will adequately resemble the original function $f(t)$, however, only a portion of the function has been sampled (from t_0 through t_{17}) rather than the entire one. In other words, one has placed a “window” over the function.



“Windowing” phenomenon cont.

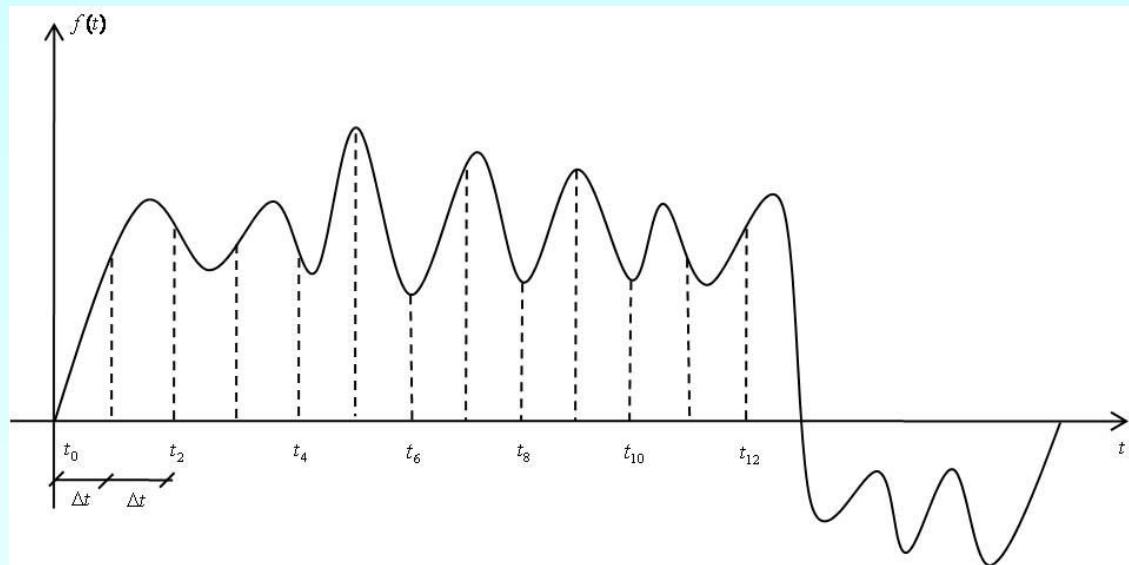
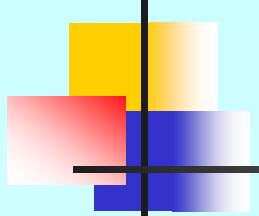


Figure 2. Function to be sampled and “windowing” sample problem.



“Windowing” phenomenon cont.

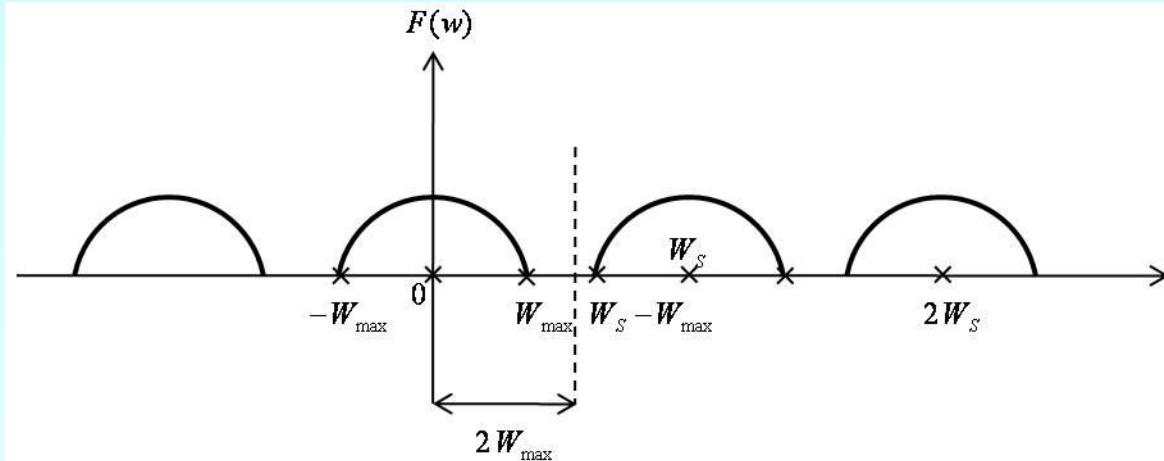
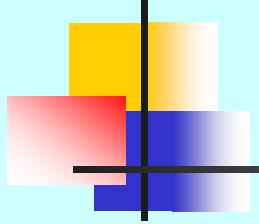


Figure 3. Frequency of sampling rate (w_s) versus maximum frequency content (w_{\max}).

In order to satisfy $F(w)=0$ for $|w| \geq w_{\max}$ the frequency (w) should be between points A and B of Figure 3.



“Windowing” phenomenon cont.

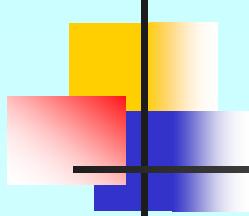
Hence:

$$w_{\max} \leq w \leq w_s - w_{\max}$$

which implies:

$$w_s \geq 2w_{\max}$$

Physically, the above equation states that one must have at least 2 samples per cycle of the highest frequency component present (Nyquist samples, Nyquist rate).



“Windowing” phenomenon cont.

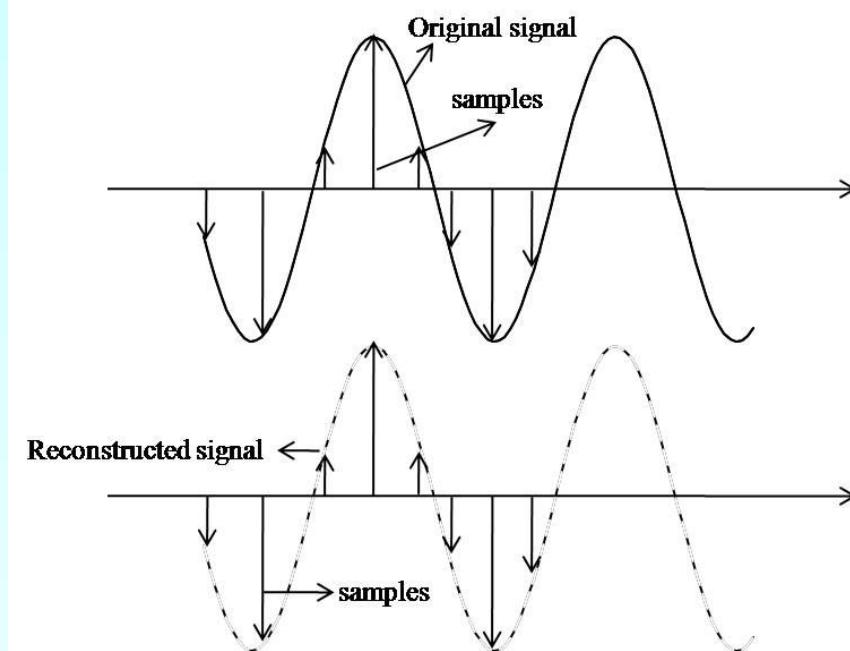
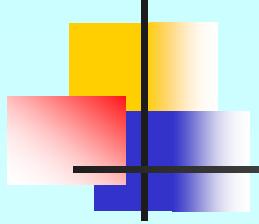


Figure 4. Correctly reconstructed signal.



“Windowing” phenomenon cont.

In Figure 4, a sinusoidal signal is sampled at the rate of 6 samples per 1 cycle (or $w_s = 6w_0$). Since this sampling rate does satisfy the sampling theorem requirement of ($w_s \geq 2w_{\max}$), the reconstructed signal does correctly represent the original signal.

“Windowing” phenomenon cont.

In Figure 5 a sinusoidal signal is sampled at the rate of 6 samples per 4 cycles $\left(\text{or } w_s = \frac{6}{4}w_0\right)$

Since this sampling rate does NOT satisfy the requirement $(w_s \geq 2w_{\max})$, the reconstructed signal was wrongly represent the original signal!

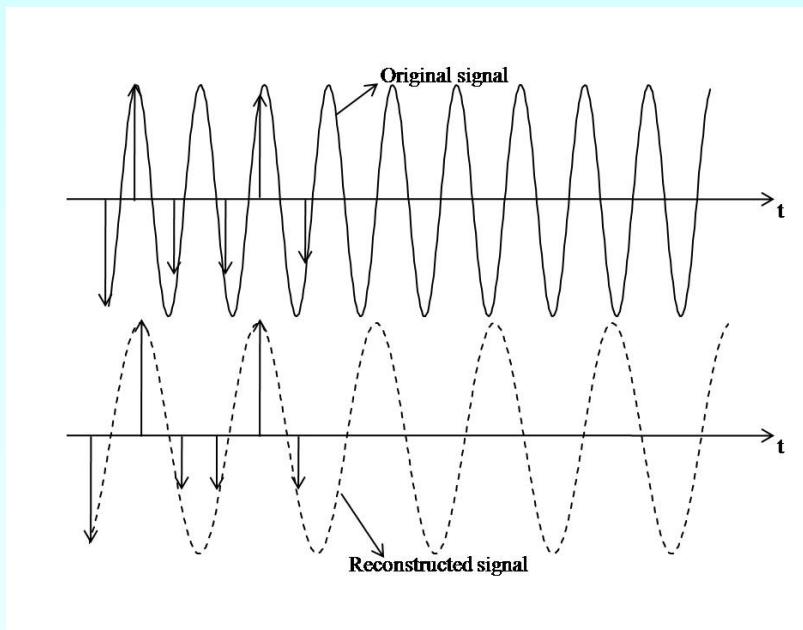
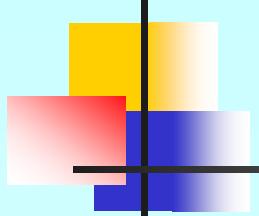


Figure 5. Wrongly reconstructed signal.



Discrete Fourier Transform cont.

Equations (19) and (1) can be rewritten as

$$\tilde{C}_n = \sum_{k=0}^{N-1} f(k) e^{-ik\left(w_0 = \frac{2\pi}{N}\right)n} \quad (20)$$

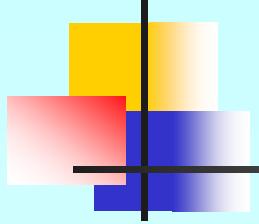
$$f(k) = \left(\frac{1}{N}\right) \sum_{n=0}^{N-1} \tilde{C}_n e^{ik\left(w_0 = \frac{2\pi}{N}\right)n} \quad (21)$$

To avoid computation with “complex numbers”, Equation (20) can be expressed as

$$\tilde{C}_n^R + i\tilde{C}_n^I = \sum_{k=0}^{N-1} \left\{ f^R(k) + i f^I(k) \right\} \times \{\cos(\theta) - i \sin(\theta)\} \quad (20A)$$

where

$$\theta = k \left(w_0 = \frac{2\pi}{N} \right) n$$



Discrete Fourier Transform cont.

$$\tilde{C}_n^R + i\tilde{C}_n^I = \sum_{k=0}^{N-1} \left\{ f^R(k) \cos(\theta) + f^I(k) \sin(\theta) \right\} + i \left\{ f^I(k) \cos(\theta) - f^R(k) \sin(\theta) \right\} \quad (20B)$$

The above “complex number” equation is equivalent to the following 2 “real number” equations:

$$\tilde{C}_n^R = \sum_{k=0}^{N-1} \left\{ f^R(k) \cos(\theta) + f^I(k) \sin(\theta) \right\} \quad (20C)$$

$$\tilde{C}_n^I = \sum_{k=0}^{N-1} \left\{ f^I(k) \cos(\theta) - f^R(k) \sin(\theta) \right\} \quad (20D)$$

Fast Fourier Transform

Part: Informal Development of Fast Fourier Transform

<http://numericalmethods.eng.usf.edu>

For more details on this topic

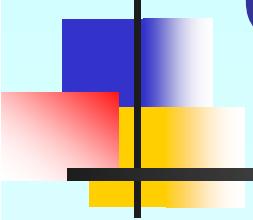
- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Fast Fourier Transform

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 11

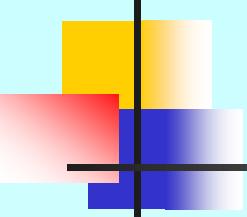
Chapter 11.05: Informal Development of Fast Fourier Transform

Major: All Engineering Majors

Authors: Duc Nguyen

<http://numericalmethods.eng.usf.edu>

Numerical Methods for STEM undergraduates



Informal Development of Fast Fourier Transform

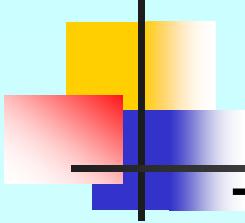
Recall the DFT pairs of Equations (20) and (21) of Chapter 11.04 and swapping the indexes n, k one obtains

$$\tilde{C}_n = \sum_{k=0}^{N-1} f(k) e^{-in\left(\omega_0 = \frac{2\pi}{N}\right)k} \quad (1)$$

$$f(k) = \left(\frac{1}{N}\right) \sum_{n=0}^{N-1} \tilde{C}_n e^{in\left(\omega_0 = \frac{2\pi}{N}\right)k} \quad (2)$$

$$\text{where } n, k = 0, 1, 2, 3, \dots, N-1 \quad (3)$$

$$\text{Let } E = e^{-i\frac{2\pi}{N}} \quad (\text{hence } E^N = e^{-i2\pi} = 1) \quad (4)$$



Informal Development cont.

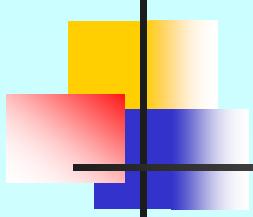
Then Eq. (1) and Eq. (2) become

$$\tilde{C}_n = \tilde{C}(n) = \sum_{k=0}^{N-1} f(k) E^{nk} \quad (5)$$

$$f(k) = \left(\frac{1}{N} \right) \sum_{n=0}^{N-1} \tilde{C}_n E^{-nk}$$

Assuming $N = 4 = 2^{(r=2)}$, then

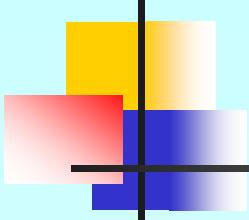
$$\left(\frac{1}{N} \right) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^{-1} & E^{-2} & E^{-3} \\ 1 & E^{-2} & E^{-4} & E^{-6} \\ 1 & E^{-3} & E^{-6} & E^{-9} \end{bmatrix} \begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{Bmatrix} = \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (5A)$$



Informal Development cont.

To obtain the above unknown vector $\{\tilde{C}\}$ for a given vector $\{f\}$, the coefficient matrix can be easily converted as

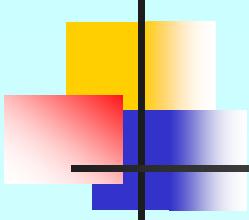
$$\left[\left(\frac{1}{N} \right) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^{-1} & E^{-2} & E^{-3} \\ 1 & E^{-2} & E^{-4} & E^{-6} \\ 1 & E^{-3} & E^{-6} & E^{-9} \end{bmatrix} \right]^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^1 & E^2 & E^3 \\ 1 & E^2 & E^4 & E^6 \\ 1 & E^3 & E^6 & E^9 \end{bmatrix}$$



Informal Development cont.

Hence, the unknown vector $\{\tilde{C}\}$ can be computed as with matrix vector operations, as following

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} E^{(0)(0)} & E^{(0)(1)} & E^{(0)(2)} & E^{(0)(3)} \\ E^{(1)(0)} & E^{(1)(1)} & E^{(1)(2)} & E^{(1)(3)} \\ E^{(2)(0)} & E^{(2)(1)} & E^{(2)(2)} & E^{(2)(3)} \\ E^{(3)(0)} & E^{(3)(1)} & E^{(3)(2)} & E^{(3)(3)} \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (6)$$

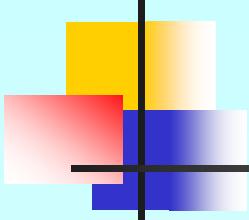


Informal Development cont.

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} E^0 & E^0 & E^0 & E^0 \\ E^0 & E^1 & E^2 & E^3 \\ E^0 & E^2 & E^4 & E^6 \\ E^0 & E^3 & E^6 & E^9 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (7)$$

For $N = 4$, $k = 3$ and $n = 2$, then:

$$E^{nk} = E^6 = [E^{(N=4)}]E^2 = \left(e^{\frac{-i2\pi}{N}} \right)^N E^2 = [e^{-i2\pi}]E^2 = E^2$$

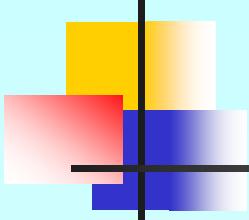


Informal Development cont.

Thus, in general (for $nk \geq N$)

$$E^{nk} = E^U \text{ where } U = \text{mod}(nk, N) \quad (8)$$

$$U = \text{remainder} \left(\frac{nk}{N} \right)$$



Informal Development cont.

Remarks:

- a) Matrix times vector, shown in Eq. (7), will require 16 (or N^2) complex multiplications and 12 (or $N(N - 1)$) complex additions.

- b) Use of Eq. (8) will help to reduce the number of operation counts, as explained in the next section.



THE END

<http://numericalmethods.eng.usf.edu>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://numericalmethods.eng.usf.edu>

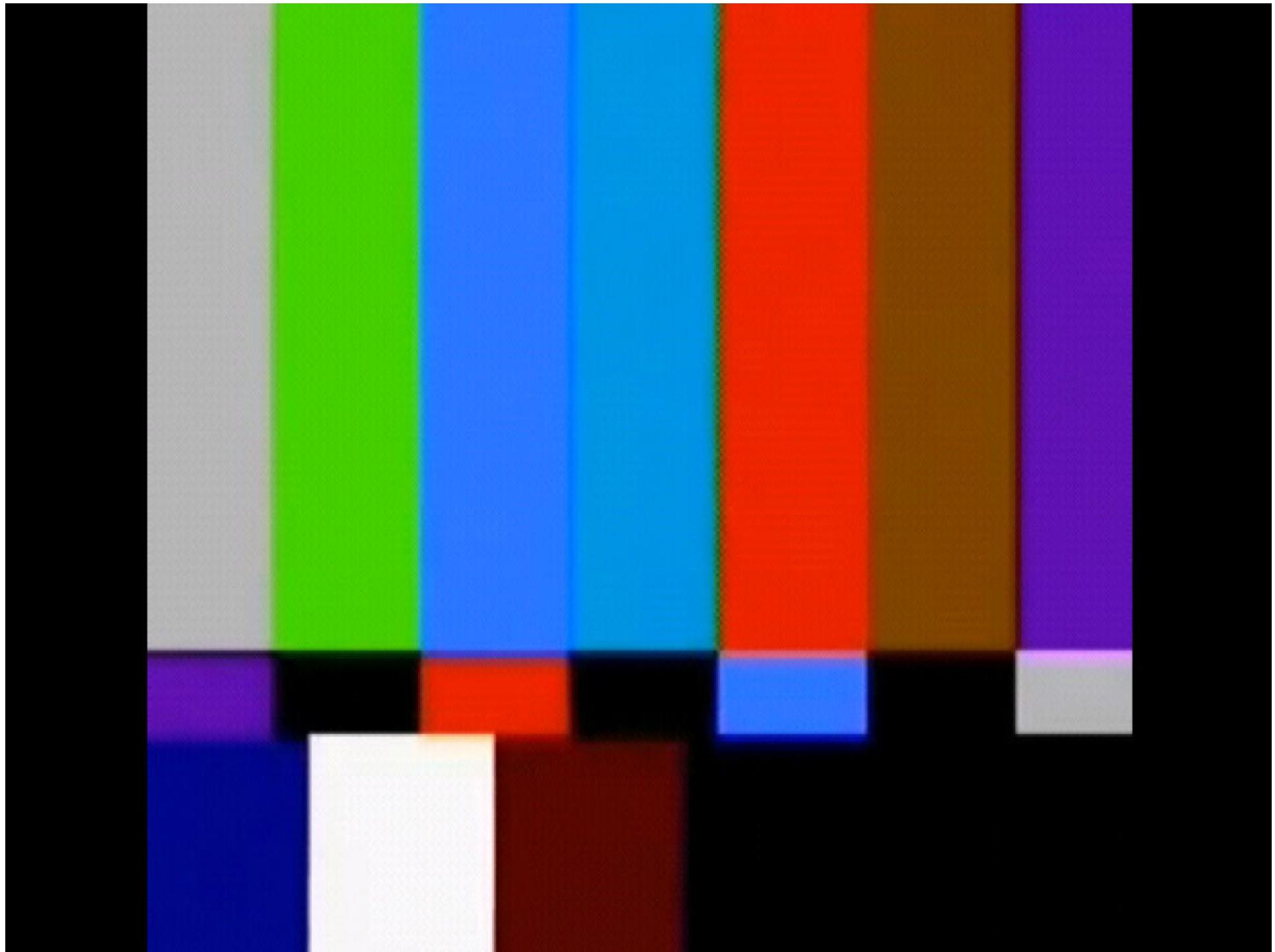
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Fast Fourier Transform

Part: Factorized Matrix and Further Operation Count

<http://numericalmethods.eng.usf.edu>

For more details on this topic

- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Fast Fourier Transform

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Lecture # 12

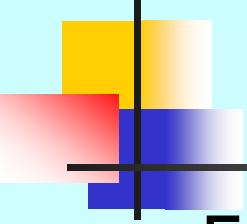
Chapter 11.05: Factorized Matrix and Further Operation Count (Contd.)

Equation (7) can be factorized as

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(2) \\ \tilde{C}(1) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} 1 & E^0 & 0 & 0 \\ 1 & E^2 & 0 & 0 \\ 0 & 0 & 1 & E^1 \\ 0 & 0 & 1 & E^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & E^0 & 0 \\ 0 & 1 & 0 & E^0 \\ 1 & 0 & E^2 & 0 \\ 0 & 1 & 0 & E^2 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (9)$$

Let's define the following "inner – product"

$$\begin{Bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{Bmatrix} = \begin{bmatrix} 1 & 0 & E^0 & 0 \\ 0 & 1 & 0 & E^0 \\ 1 & 0 & E^2 & 0 \\ 0 & 1 & 0 & E^2 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (10)$$



Factorized Matrix cont.

From Eq. (9) and (10) we obtain

$$f_1(0) = f(0) + E^0 f(2) \quad (11A)$$

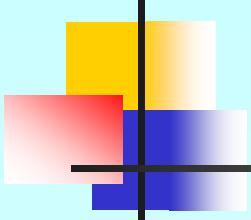
$$f_1(1) = f(1) + E^0 f(3) \quad (11B)$$

$$\begin{aligned} f_1(2) &= f(0) + E^2 f(2) \\ &= f(0) - E^0 f(2) \end{aligned} \quad (11C)$$

with

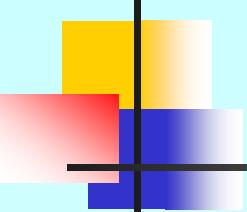
$$E^2 = e^{-i\frac{2\pi}{4}*2} = e^{-i\pi} = -1 = -E^0$$

$$\begin{aligned} f_1(3) &= f(1) + E^2 f(3) \\ &= f(1) - E^0 f(3) \end{aligned} \quad (11D)$$



Factorized Matrix cont.

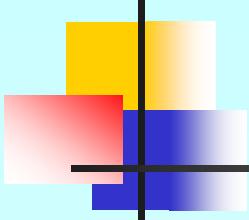
Equations(11A through 11D) for the “inner” matrix times vector requires 2 complex multiplications and 4 complex additions.



Factorized Matrix cont.

Finally, performing the “outer” product (matrix times vector) on the RHS of Equation(9), one obtains

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(2) \\ \tilde{C}(1) \\ \tilde{C}(3) \end{Bmatrix} = \begin{Bmatrix} f_2(0) \\ f_2(1) \\ f_2(2) \\ f_2(3) \end{Bmatrix} = \begin{bmatrix} 1 & E^0 & 0 & 0 \\ 1 & E^2 & 0 & 0 \\ 0 & 0 & 1 & E^1 \\ 0 & 0 & 1 & E^3 \end{bmatrix} \begin{Bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{Bmatrix} \quad (12)$$



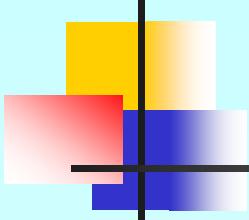
Factorized Matrix cont.

$$f_2(0) = f_1(0) + E^0 f_1(1) \quad (13A)$$

$$f_2(1) = f_1(0) + E^2 f_1(1) = f_1(0) - E^0 f_1(1) \quad (13B)$$

$$f_2(2) = f_1(2) + E^1 f_1(3) \quad (13C)$$

$$\begin{aligned} f_2(3) &= f_1(2) + E^3 f_1(3) = f_1(2) + E^2 E^1 f_1(3) \\ &= f_1(2) - E^1 f_1(3) \end{aligned} \quad (13D)$$

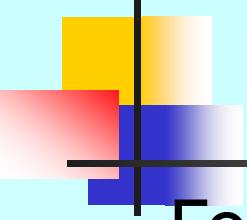


Factorized Matrix cont.

Again, Eqs (13A-13D) requires 2 complex multiplications And 4 complex additions. Thus, the complete RHS of Eq. (9) Can be computed by only 4 complex multiplications (or $N \frac{r}{2} = 4 \frac{2}{2}$) and 8 complex additions (or $Nr = 4 * 2$),

where $N = 2^r$.

Since computational time is mainly controlled by the number of multiplications, implementing Eq. (9) will significantly reduce the number of multiplication operations, as compared to a direct matrix times vector operations. (as shown in Eq. (7)).



Factorized Matrix cont.

For a large number of data points,

$$Ratio = \frac{N^2}{\left(\frac{Nr}{2}\right)} = \left(\frac{2N}{r}\right) \quad (14)$$

For $N = 2048 = 2^{(r=11)}$, Equation (14) gives:

$$Ratio = \frac{2(2048)}{11} = 372.36$$

This implies that the number of complex multiplications involved in Eq. (9) is about 372 times less than the one involved in Eq. (7).

Graphical Flow of Eq. 9

Consider the case $N = 2^r = 2^2 = 4$

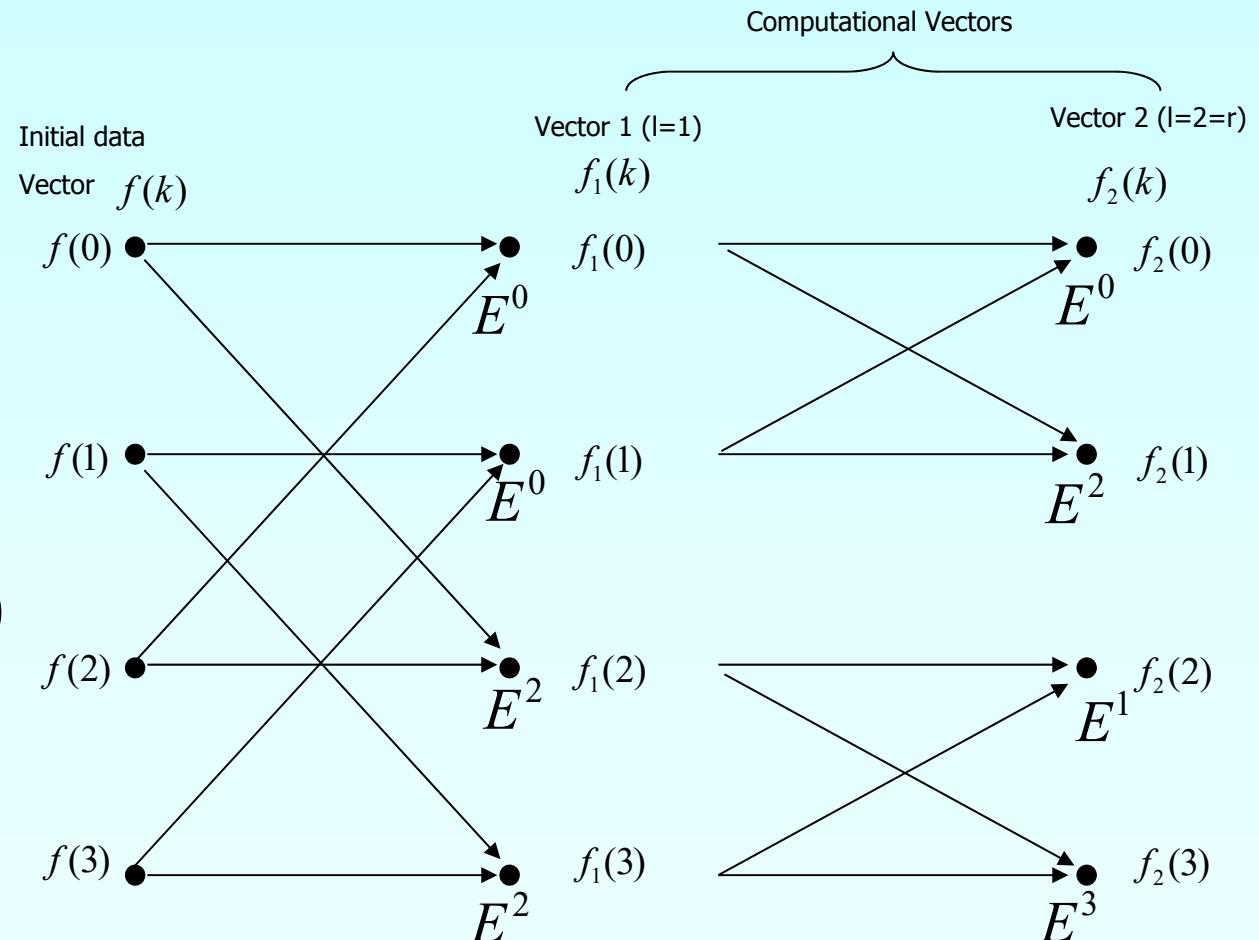


Figure 1. Graphical form of FFT (Eq. 9) for the case

$$N = 2^r = 2^2 = 4$$

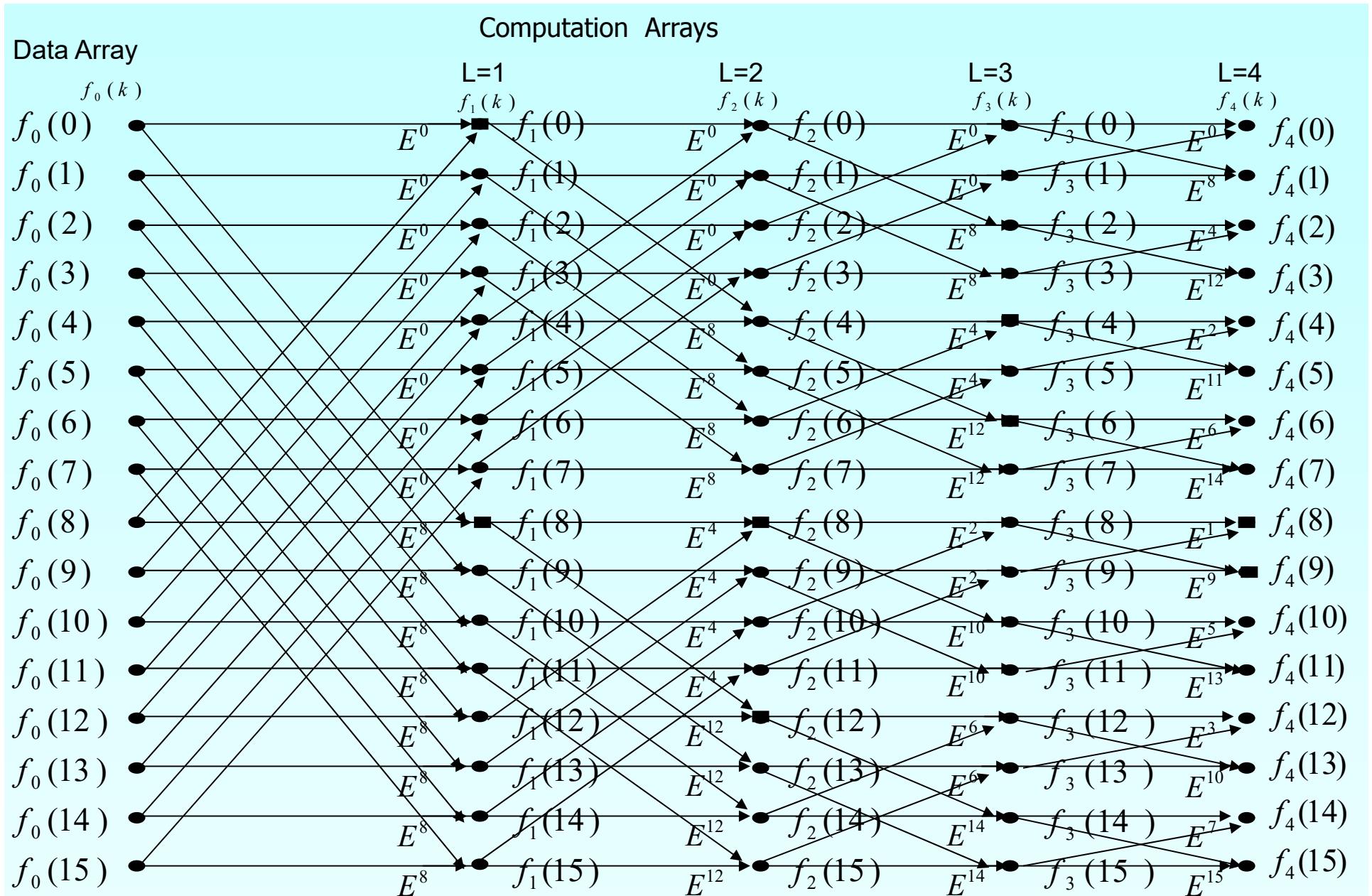


Figure 2. Graphical Form of FFT (Eq. 9) for the case

$$N = 2^r = 2^4 = 16$$



THE END

<http://numericalmethods.eng.usf.edu>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://numericalmethods.eng.usf.edu>

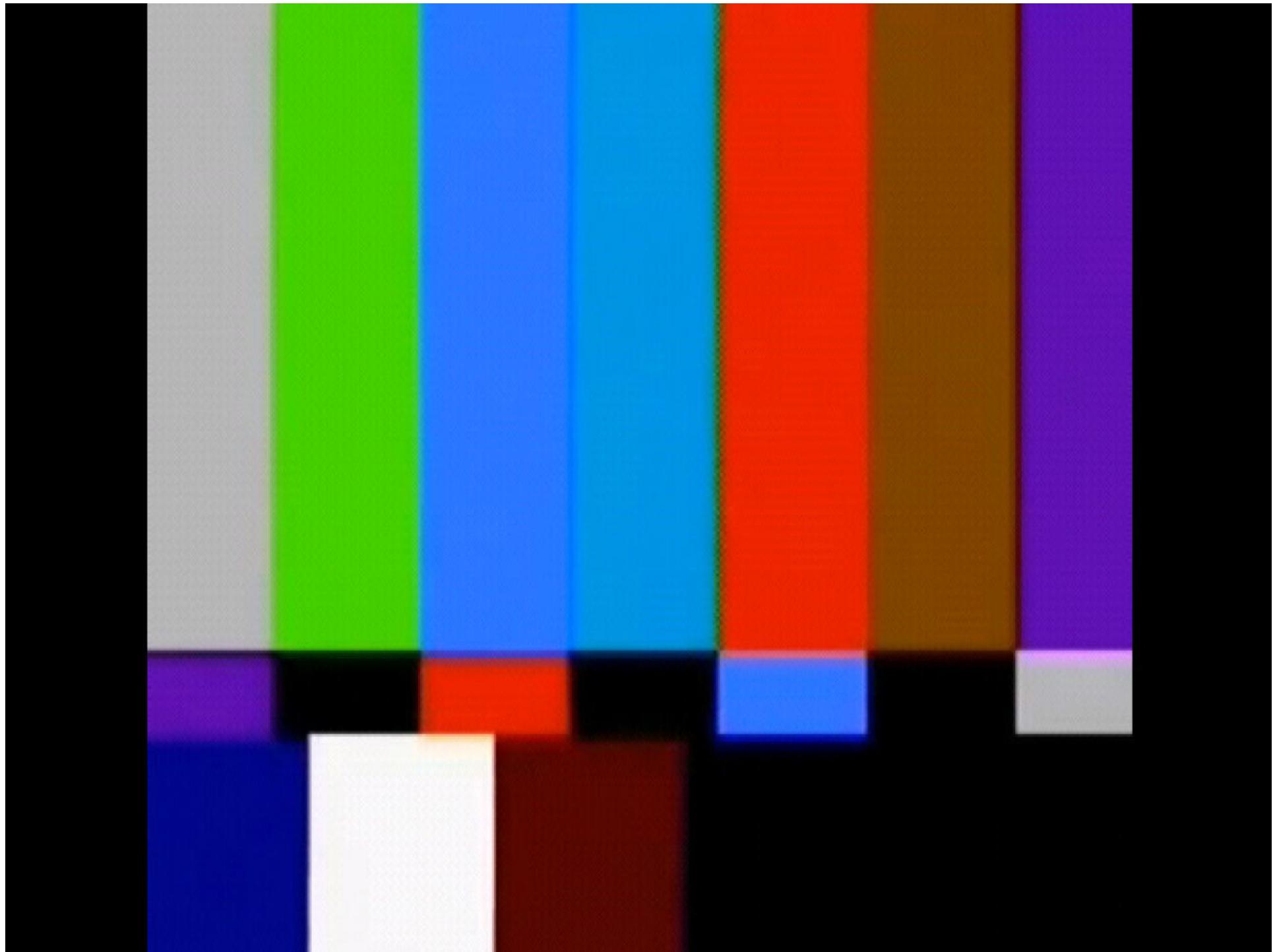
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Fast Fourier Transform

Part: Companion Node Observation

<http://numericalmethods.eng.usf.edu>

For more details on this topic

- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Fast Fourier Transform

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Chapter 11.05 : Companion Node Observation (Contd.)

Careful observation of Figure 2 has revealed that each computed l^{th} vector (where $l = 1, 2, \dots, r$ and $N = 2^r = 2^4 = 16$) we can always find two (companion) nodes which came from the same pair of nodes in the previous vector.

For example, $f_1(0)$ and $f_1(8)$ are computed in terms of $f(0)$ and $f(8)$.

Similarly, the companion nodes $f_2(8)$ and $f_2(12)$ are computed from the same pair of nodes as $f_1(8)$ and $f_1(12)$.

Data Array

$$f_0(k)$$

$$f_0(0)$$

$$f_0(1)$$

$$f_0(2)$$

$$f_0(3)$$

$$f_0(4)$$

$$f_0(5)$$

$$f_0(6)$$

$$f_0(7)$$

$$f_0(8)$$

$$f_0(9)$$

$$f_0(10)$$

$$f_0(11)$$

$$f_0(12)$$

$$f_0(13)$$

$$f_0(14)$$

$$f_0(15)$$

Computation Arrays

L=1

$$f_1(k)$$

$$f_1(0)$$

$$f_1(1)$$

$$f_1(2)$$

$$f_1(3)$$

$$f_1(4)$$

$$f_1(5)$$

$$f_1(6)$$

$$f_1(7)$$

$$f_1(8)$$

$$f_1(9)$$

$$f_1(10)$$

$$f_1(11)$$

$$f_1(12)$$

$$f_1(13)$$

$$f_1(14)$$

$$f_1(15)$$

L=2

$$f_2(k)$$

$$f_2(0)$$

$$f_2(1)$$

$$f_2(2)$$

$$f_2(3)$$

$$f_2(4)$$

$$f_2(5)$$

$$f_2(6)$$

$$f_2(7)$$

$$f_2(8)$$

$$f_2(9)$$

$$f_2(10)$$

$$f_2(11)$$

$$f_2(12)$$

$$f_2(13)$$

$$f_2(14)$$

$$f_2(15)$$

L=3

$$f_3(k)$$

$$f_3(0)$$

$$f_3(1)$$

$$f_3(2)$$

$$f_3(3)$$

$$f_3(4)$$

$$f_3(5)$$

$$f_3(6)$$

$$f_3(7)$$

$$f_3(8)$$

$$f_3(9)$$

$$f_3(10)$$

$$f_3(11)$$

$$f_3(12)$$

$$f_3(13)$$

$$f_3(14)$$

$$f_3(15)$$

L=4

$$f_4(k)$$

$$f_4(0)$$

$$f_4(1)$$

$$f_4(2)$$

$$f_4(3)$$

$$f_4(4)$$

$$f_4(5)$$

$$f_4(6)$$

$$f_4(7)$$

$$f_4(8)$$

$$f_4(9)$$

$$f_4(10)$$

$$f_4(11)$$

$$f_4(12)$$

$$f_4(13)$$

$$f_4(14)$$

$$f_4(15)$$

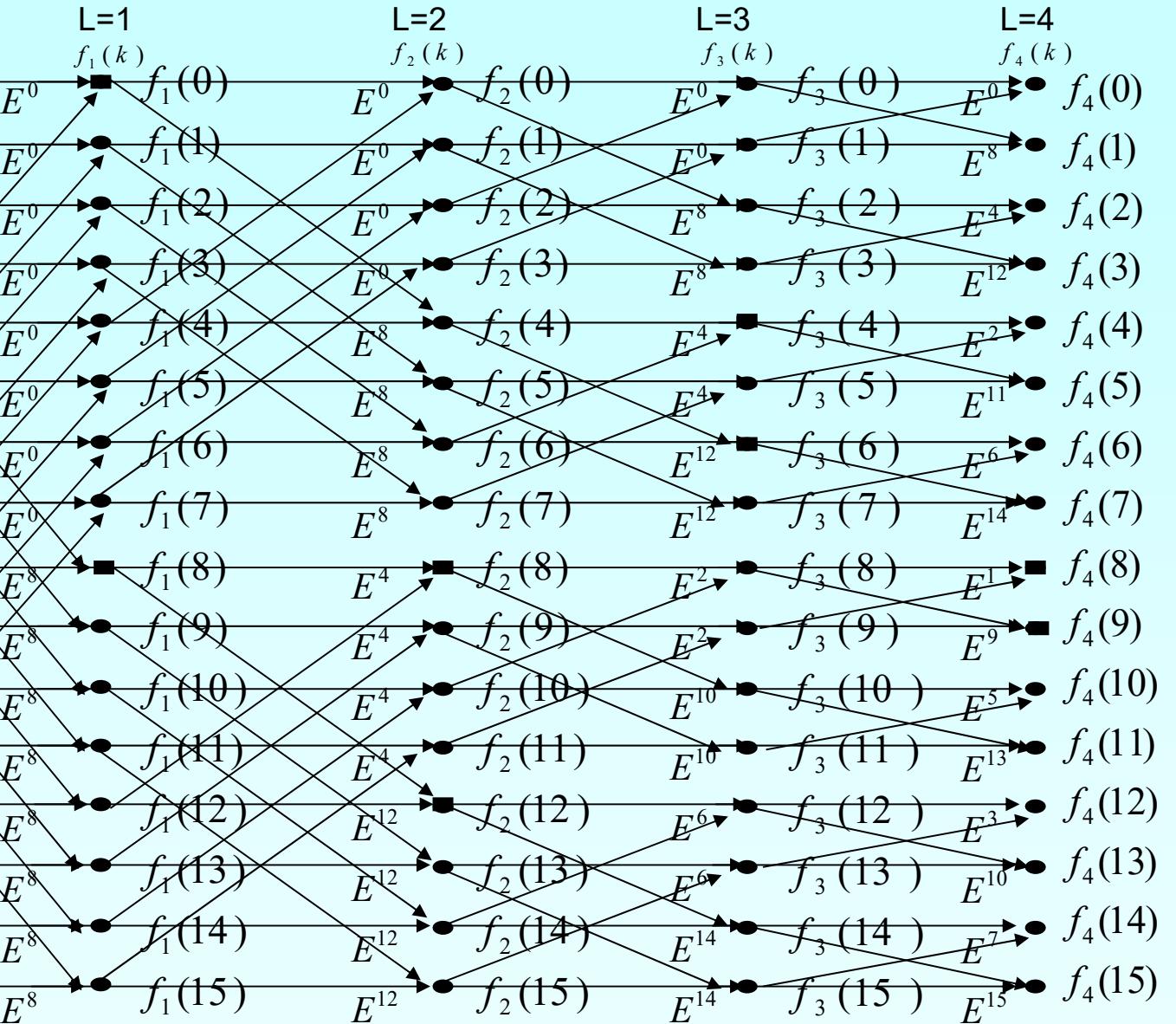
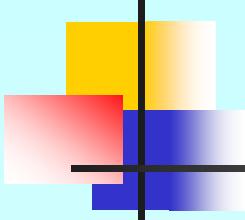


Figure 2. Graphical Form of FFT (Eq. 9) for the case

$$N = 2^r = 2^4 = 16$$

<http://numericalmethods.eng.usf.edu>

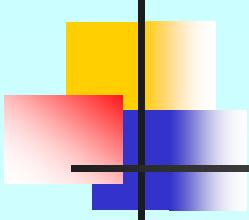


Companion Node Observation cont.

Furthermore, the computation of companion nodes are independent of other nodes (within the l^{th} –vector). Therefore, the computed $f_1(0)$ and $f_1(8)$ will override the original space of $f(0)$ and $f(8)$.

Similarly, the computed $f_2(8)$ and $f_2(12)$ will override the space occupied by $f_1(8)$ and $f_1(12)$ which in turn, will occupy the original space of $f(8)$ and $f(12)$.

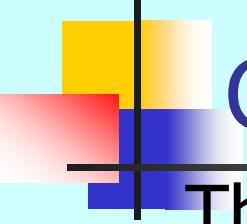
Hence, only one complex vector (or 2 real vectors) of length N are needed for the entire FFT process.



Companion Node Spacing

Observing Figure 2, the following statements can be made:

- a) in the first vector ($l = 1$), the companion nodes $f_1(8)$ and $f_1(12)$ are separated by $k = 8$ (or $\frac{N}{2^l} = \frac{16}{2^1}$)
- b) in the second vector ($l = 2$), the companion nodes $f_2(8)$ and $f_2(12)$ are separated by $k = 4$.
(or $\frac{N}{2^l} = \frac{16}{2^2} = \frac{16}{4}$), etc.

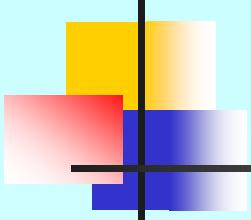


Companion Node Computation

The operation counts in any companion nodes (of the $l^{th} = 2^{nd}$ vector), such as $f_2(8)$ and $f_2(12)$ can be explained as (see Figure 2).

$$\begin{aligned} f_2(8) &= f_1(8) + f_1(12) \times E^4 \\ f_2(12) &= f_1(8) + f_1(12) \times E^{12} \\ &= f_1(8) + f_1(12) \times E^8 E^4 \\ &= f_1(8) + f_1(12) \left[e^{-i \frac{2\pi}{(N=16)}} \right]^8 E^4 \\ &= f_1(8) + f_1(12) [e^{-i\pi}]^4 E^4 \end{aligned} \tag{15}$$

$$f_2(12) = f_1(8) - f_1(12) \times E^4 \tag{16}$$



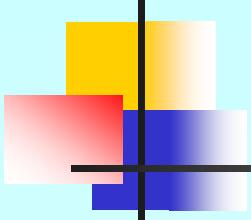
Companion Node Computation cont.

Thus, the companion nodes $f_2(8)$ and $f_2(12)$ computation will require 1 complex multiplication and 2 complex additions (see Eq. (15) and (16)). The weighting factors for the companion nodes ($f_2(8)$ and $f_2(12)$) are E^4 (or E^U) and E^{12} (or $E^{U+N/2}$), respectively.

$$f_l(k) = f_{l-1}(k) + E^U f_{l-1}\left(k + \frac{N}{2^l}\right) \quad (17)$$

48

$$f_l\left(k + \frac{N}{2^l}\right) = f_{l-1}(k) - E^U f_{l-1}\left(k + \frac{N}{2^l}\right) \quad (18)$$



Skipping Computation of Certain Nodes

Because the pair of companion nodes k and $k + \frac{N}{2^L}$ are separated by the “distance” $\frac{N}{2^L}$, at the L^{th} level, after every $\frac{N}{2^L}$ node computation, then the next $\frac{N}{2^L}$ nodes will be skipped. (see Figure 2)



THE END

<http://numericalmethods.eng.usf.edu>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://numericalmethods.eng.usf.edu>

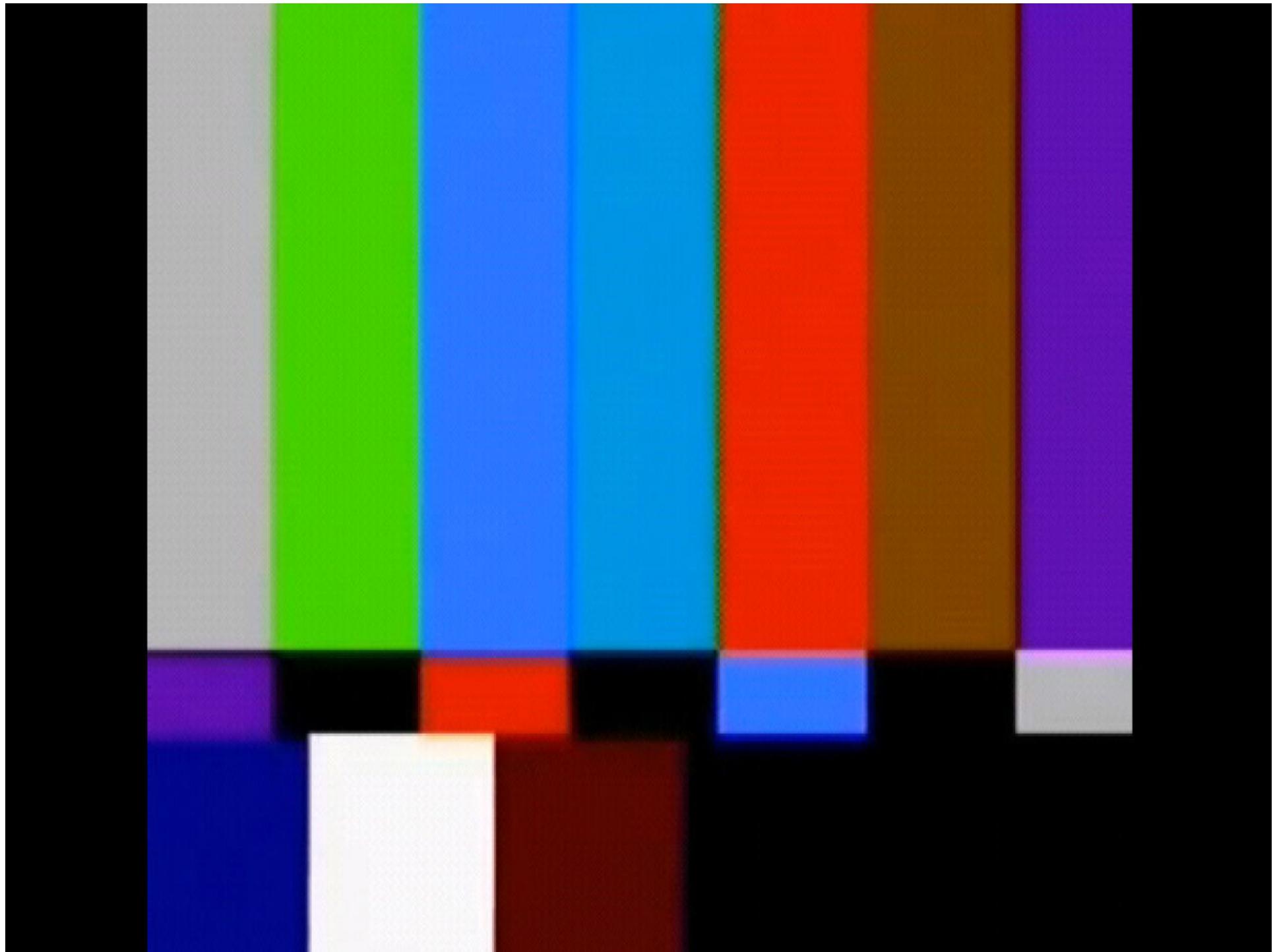
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.





Numerical Methods

Fast Fourier Transform

Part: Determination of E^U

<http://numericalmethods.eng.usf.edu>

For more details on this topic

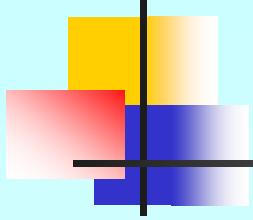
- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Fast Fourier Transform

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Lecture # 14

Chapter 11.05: Determination of E^U

The values of " U "

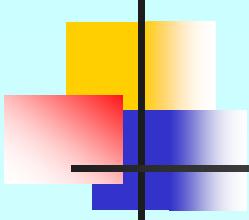
$$f_l(k) = f_{l-1}(k) + E^U f_{l-1}\left(k + \frac{N}{2^l}\right);$$

$$f_l\left(k + \frac{N}{2^l}\right) = f_{l-1}(k) - E^U f_{l-1}\left(k + \frac{N}{2^l}\right)$$

can be determined by the following steps:

Step 1: Express the index $k (= 0, 1, 2, \dots, N - 1)$ in binary form, using r bits. For $k = 8, L = 2, r = 4$, and $N = 2^r = 2^4 = 16$, one obtains:

$$k = 8 = 1,0,0,0 = (1)2^{r-1=3} + (0)2^2 + (0)2^1 + (0)2^0$$



Determination of E^U cont.

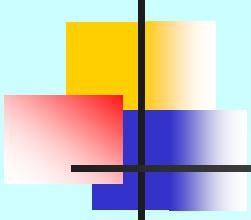
Step 2: Sliding this binary number $r - L = 4 - 2 = 2$ positions to the right, and fill in zeros, the results are

$$1,0,0,0 \rightarrow X,X,1,0 \rightarrow 0,0,1,0$$

It is important to realize that the results of Step 2 (0,0,1,0) are equivalent to expressing an integer

$$M = \frac{k}{2^{r-L}} = \frac{8}{2^{4-2}} = 2 \text{ in binary format. In other words}$$

$$M = 2 = (0,0,1,0)$$



Determination of E^U cont.

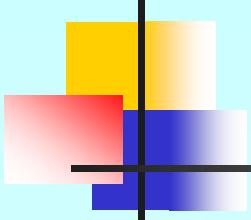
Step 3: Reverse the order of the bits,
then $(0,0,1,0)$ becomes $(0,1,0,0) = U$.

Thus,

$$U = (0)2^3 + (1)2^2 + (0)2^1 + (0)2^0 = 4$$

It is “NOT” really necessary to perform Step 3, since the results of Step 2 can be used to compute “ U ” as following

$$U = (0)2^0 + (0)2^1 + (1)2^2 + (0)2^3 = 4$$



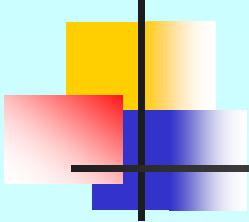
Computer Implementation to find E^U

Based on the previous discussions (with the 3-step procedures), to find the value of “ U ”, one only needs a procedure to express an integer

$$M = \frac{k}{2^{r-L}} \text{ in binary format, with } r \text{ bits.}$$

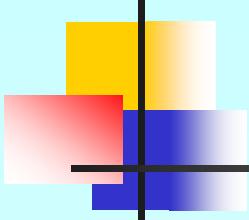
Assuming M (a base 10 number) can be expressed as (assuming $r = 4$ bits)

$$M = a_4 a_3 a_2 a_1 = J_1 \quad (19)$$



Computer Implementation cont.

Divide M by 2, $J_2 = J_1/2$ then multiply the truncated result by 2 ($JJ_2 = J_2 \times 2$), and compute the difference between the original number and the new number.



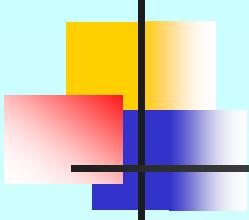
Computer Implementation cont.

Compute the difference between the original number and the new number ($= M = J_1 \& \& JJ_2$) :

$$IDIFF = J_1 - JJ_2 \left\{ = M - \left(\frac{M}{2} \right)_{Truncated} \times 2 \right\} \quad (20)$$

If $IDIFF = 0$, then the bit $a_1 = 0$

If $IDIFF \neq 0$, then the bit $a_1 = 1$



Computer Implementation cont.

Once the bit a_1 has been determined, the value of J_1 is set to J_2 (or value of J_1 is reduced by a factor of 2; since the previous .

$$J_1 = M = a_4 a_3 a_2 a_1$$

$$J_1 = M = a_4 (2^3) + a_3 (2^2) + a_2 (2^1) + a_1 (2^0)$$

A similar process can be used to determine the value of process can be used to determine the next bit a_2 etc.

Example 1

For $k = 8, N = 16 = 2^r, r = 4$ bits and $L = 2$
Find the value of U .

$$M = \frac{k}{2^{r-L}} = \frac{8}{2^{4-2}} = 2 = J_1$$

Determine the bit a_1 (Index $I = 1$)

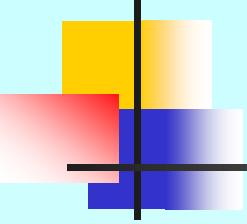
Initialize $U = 0$

$$J_2 = \frac{J_1}{2} = \frac{2}{2} = 1$$

$$IDIFF = J_1 - (J_2 \times 2) = 2 - (1)(2) = 0$$

Thus $a_1 = 0$

$$U = U \times 2 + IDIFF = 0 \times 2 + 0 = 0$$



Example 1 cont.

Determine the bit a_2 (Index $I = 2$)

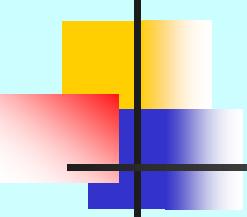
$$J_1 = J_2 = 1$$

$$J_2 = \frac{J_1}{2} = \frac{1}{2} = 0$$

$$IDIFF = J_1 - (JJ_2 = J_2 \times 2) = 1 - (0 \times 2) = 1$$

Thus $a_2 = 1$

$$U = U \times 2 + IDIFF = 0 \times 2 + 1 = 1$$



Example 1 cont.

Determine the bit a_3 (Index $I = 3$)

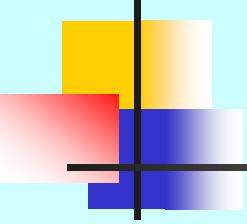
$$J_1 = J_2 = 0$$

$$J_2 = \frac{J_1}{2} = \frac{0}{2} = 0$$

$$IDIFF = J_1 - (JJ_2 = J_2 \times 2) = 0 - (0)(2) = 0$$

Thus $a_3 = 0$

$$U = U \times 2 + IDIFF = 1 \times 2 + 0 = 2$$



Example 1 cont.

Determine the bit a_4 (Index $I = 4$)

$$J_1 = J_2 = 0$$

$$J_2 = \frac{J_1}{2} = \frac{1}{2} = 0$$

$$IDIFF = J_1 - (JJ_2 = J_2 \times 2) = 0 - (0)(2) = 0$$

Thus $a_4 = 0$

$$U = U \times 2 + IDIFF = 2 \times 2 + 0 = 4$$



THE END

<http://numericalmethods.eng.usf.edu>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://numericalmethods.eng.usf.edu>

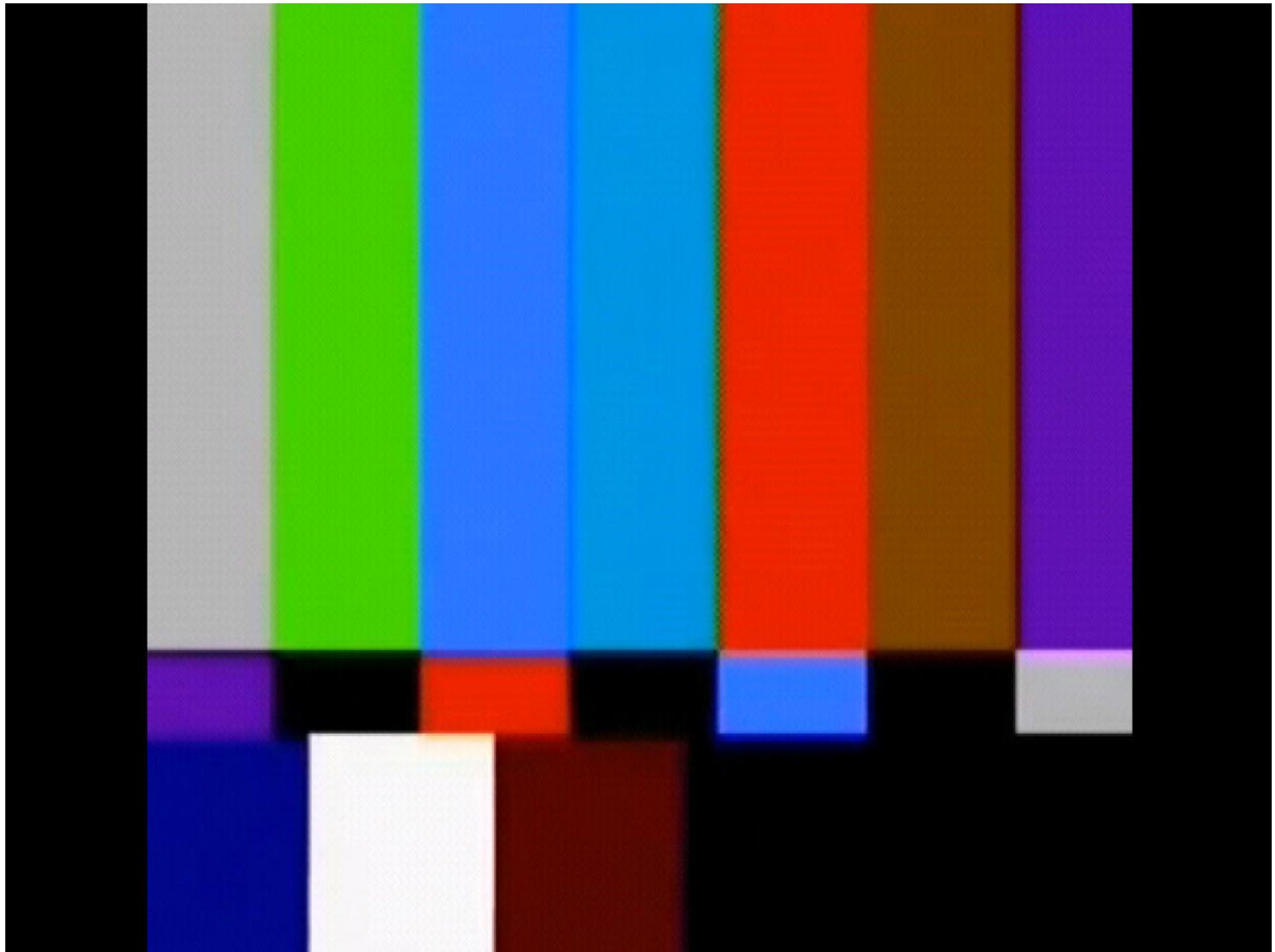
Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Numerical Methods

Fast Fourier Transform

Part: Unscrambling the FFT

<http://numericalmethods.eng.usf.edu>

For more details on this topic

- Go to <http://numericalmethods.eng.usf.edu>
- Click on Keyword
- Click on Fast Fourier Transform

You are free

- to **Share** – to copy, distribute, display and perform the work
- to **Remix** – to make derivative works

Under the following conditions

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** — You may not use this work for commercial purposes.
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Lecture # 15

Chapter 11.05: Unscrambling the FFT (Contd.)

For the case

$$N = 16 = 2^{r=4}$$

, (see Figure 2), the final “bit-reversing” operation for FFT is shown in Figure 3.

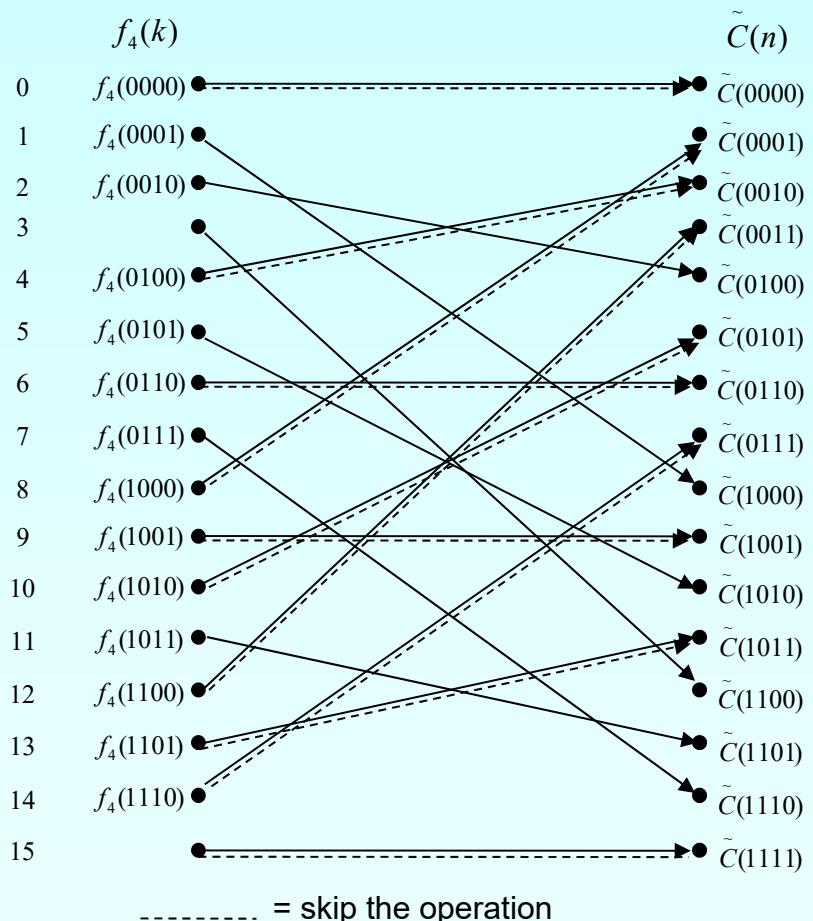
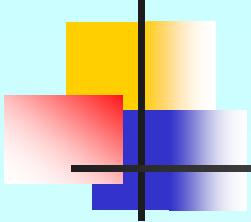


Figure 3. Final “bit-reversing” for FFT (with $N = 2^r = 2^4 = 16$)
<http://numericalmethods.eng.usf.edu>



For do-loop index $k = 0 = (0, 0, 0, 0) \Rightarrow i = (0, 0, 0, 0) = \text{bit-reversion} = 0$

If (i.GT.k) Then

$T = f_4(k)$

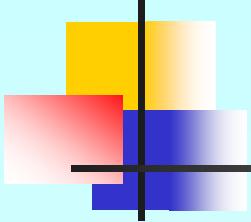
$f_4(k) = f_4(i)$

$f_4(i) = T$

Endif

Hence, $f_4(0) = f_4(0)$ no swapping.

55



For $k = 1 = (0,0,0,1) \Rightarrow i = (1,0,0,0)$
= bit-reversion = 8

If (i.GT.k) Then

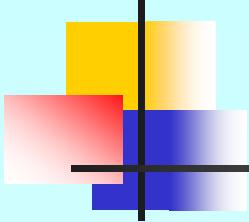
$T = f_4(k=1)$

$f_4(k=1) = f_4(i=8)$

$f_4(i=8) = T$

Endif

Hence, $f_4(1) = f_4(8)$ are swapped.



.For $k=2=(0,0,1,0) \Rightarrow i = (0,1,0,0) = 4$

Hence, $f_4(2) = f_4(4)$; are swapped.

.For $k=3=(0,0,1,1) \Rightarrow i = (1,1,0,0) = 12$

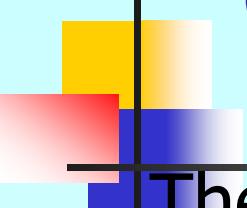
Hence, $f_4(3) = f_4(12)$; are swapped.

. For $k=4=(0,1,0,0) \Rightarrow i=(0,0,1,0)=2$

In this case, since “i” is not greater than “k”.

Hence, no swapping, since $f_4 (k = 2)$ and $f_4 (i = 4)$; had already been swapped earlier!

.etc.

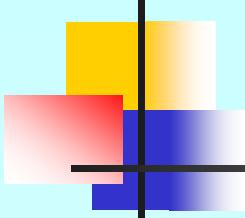


Computer Implementation of FFT case for $N=2^r$

The pair of companion nodes computation are given by Eqs.(17) and (18). To avoid “complex number” operations, Eq.(17) can be computed based on “real number” operations, as following

$$\begin{aligned} \left\{ f_L^R(k) + if_L^I(k) \right\} &= \left\{ f_{L-1}^R(k) + if_{L-1}^I(k) \right\} \\ &+ \left\{ E^{U,R} + iE^{U,I} \right\} \times \left\{ f_{L-1}^R\left(k + \frac{N}{2^L}\right) + if_{L-1}^I\left(k + \frac{N}{2^L}\right) \right\} \end{aligned} \quad (21)$$

In Eq. (21), the superscripts R and I denote real and imaginary components, respectively.

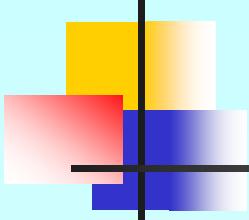


Computer Implementation cont.

Multiplying the last 2 complex numbers, one obtains

$$\begin{aligned} \left\{ f_L^R(k) + i f_L^I(k) \right\} &= \left\{ f_{L-1}^R(k) + i f_{L-1}^I(k) \right\} \\ &+ \left\{ E^{U,R} \times f_{L-1}^R(k + \frac{N}{2^L}) - E^{U,I} \times f_{L-1}^I(k + \frac{N}{2^L}) \right\} \\ &+ i \left\{ E^{U,R} \times f_{L-1}^I(k + \frac{N}{2^L}) + E^{U,I} \times f_{L-1}^R(k + \frac{N}{2^L}) \right\} \quad (22) \end{aligned}$$

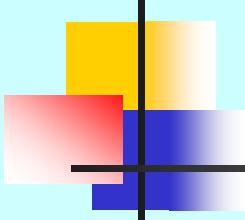
Equating the real (and then, imaginary) components on the Left-Hand-Side (LHS), and the Right-Hand-Side (RHS) of Eq. (22), one obtains



Computer implementation cont.

$$\{f_L^R(k)\} = \{f_{L-1}^R(k)\} + \left\{ E^{U,R} \times f_{L-1}^R(k + \frac{N}{2^L}) - E^{U,I} \times f_{L-1}^I(k + \frac{N}{2^L}) \right\} \quad (23A)$$

$$\{f_L^I(k)\} = \{f_{L-1}^I(k)\} + \left\{ E^{U,R} \times f_{L-1}^I(k + \frac{N}{2^L}) + E^{U,I} \times f_{L-1}^R(k + \frac{N}{2^L}) \right\} \quad (23B)$$



Computer implementation cont.

Recall Eq. (4)

$$E = e^{-i\frac{2\pi}{N}}$$

Hence

$$E^U = \left(e^{-i\frac{2\pi}{N}} \right)^U = e^{-i\frac{2\pi U}{N}} = e^{-i\theta} = \cos(\theta) - i \sin(\theta) \quad (24)$$

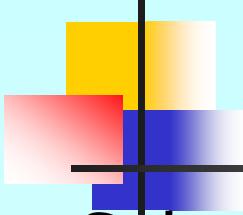
where

$$\theta = \frac{2\pi U}{N} = \frac{6.28U}{N} \quad (25)$$

Thus:

$$E^{U,R} = \cos(\theta) \quad (26A)$$

$$E^{U,I} = -\sin(\theta) \quad (26B)$$



Computer Implementation cont.

Substituting Eqs. (26A) and (26B) into Eqs. (23A) and (23B), one gets

$$\{f_L^R(k)\} = \{f_{L-1}^R(k)\} + \left\{ \cos(\theta) \times f_{L-1}^R\left(k + \frac{N}{2^L}\right) + \sin(\theta) \times f_{L-1}^I\left(k + \frac{N}{2^L}\right) \right\} \quad (27A)$$

$$\{f_L^I(k)\} = \{f_{L-1}^I(k)\} + \left\{ \cos(\theta) \times f_{L-1}^I\left(k + \frac{N}{2^L}\right) - \sin(\theta) \times f_{L-1}^R\left(k + \frac{N}{2^L}\right) \right\} \quad (27B)$$

Similarly, the single (complex number) Eq. (18) can be expressed as 2 equivalent (real number) Eqs. Like Eqs. (27A) and (27B).



THE END

<http://numericalmethods.eng.usf.edu>



Acknowledgement

This instructional power point brought to you by
Numerical Methods for STEM undergraduate

<http://numericalmethods.eng.usf.edu>

Committed to bringing numerical methods to the
undergraduate



For instructional videos on other topics, go to

<http://numericalmethods.eng.usf.edu/videos/>

This material is based upon work supported by the National Science Foundation under Grant # 0717624. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

