CSCC51H – Numerical Approximation, Integration and Ordinary Differential Equations

Wayne Enright

enright@cs.utoronto.ca

Department of Computer Science

University of Toronto

General Information

- What is Numerical Analysis ?
- The need to approximate
- Mathematical Review :
  - Floating point arithmetic
  - Notation
  - Theorems from Calculus

What is Numerical Analysis?

- Consider the investigation of a well defined mathematical model arising in any application area. Assume the model is ‘well defined’ in the sense that there exists a ‘solution’ and it is unique. Examples include modelling the spread of an infectious disease, modelling cancer treatments, or modelling the pricing of ‘options’.
- We are interested in the ‘Conditioning’ (or sensitivity) of the underlying mathematical problem to ‘small’ changes in the problem definition.
- For virtually all mathematical models of practical interest one cannot determine a useful ‘closed form’ expression for the exact solution and one must approximate the exact solution.

Scientific Computing

1. Formulate a mathematical model of the problem.
2. Approximate the solution of the model.
3. Visualize the <u>approximate</u> solution.
4. Verify that the approximate solution is consistent with the model.
5. Verify that the model is well-posed.

In this course we will focus on developing, analysing and evaluating software/methods for addressing 2.

Focus of Numerical Analysis

The emphasis is on the development and analysis of algorithms to approximate the exact solution to mathematical models.

- Algorithms must be constructive and finite .
- We must analyse the errors in the approximation.
- We must also quantify the stability and efficiency of the algorithms.

Numerical Analysis (cont)

We will be concerned with the intelligent use of existing algorithms embedded in widely used numerical software. We will not spend much time on developing algorithms or on writing code.

- How to interpret the numerical (approximate) results.
- What method (algorithm) should be used.
- What methods are available in the usual ‘Problem Solving Environments’ that scientists, engineers and students work in. For example in MATLAB, MAPLE or Mathematica.
- In order to appreciate the limitations of the methods we must analyse and understand the underlying algorithms on which the methods are based.

A Review of Relevant Mathematics

Floating Point Arithmetic

- Recall that a floating point number system, Z, can be characterized by four parameters, (β, s, m, M), and each element of Z is defined by:

  $$z = .d_1 d_2 \cdots d_s \times \beta^e,$$

  where $d_1 \neq 0$, $0 \leq d_i \leq (\beta - 1)$, and $m \leq e \leq M$.

- The floating point representation mapping, $fl(x)$, is a mapping from the Reals to Z that satisfies:

  $$fl(x) = x(1+\epsilon), \ \ \text{with} \ \ |\epsilon| \leq \mu.$$

  where μ is the ‘unit roundoff’ and is defined to be $1/2\ \beta^{1-s}$.

FP Arithmetic (cont)

- For any standard elementary arithmetic operation (+, -, × and /), we have the corresponding F.P. approximation (denoted by $\oplus, \ominus, \otimes$ and $\oslash$) which satisfies, for any $a, b \in Z$,

  $$a \odot b = fl(a \cdot b) = (a \cdot b)(1 + \epsilon),$$

  where $|\epsilon| \leq \mu$ and $\cdot$ is any elementary operation.

- For any real-valued function, $F(a_1, a_2, \cdots a_n)$, the most we can expect is that the floating point implementation $\bar{F}$, will return (when invoked) the value $\bar{y}$ satisfying:

  $$\begin{aligned} \bar{y} &= \bar{F}(fl(a_1), fl(a_2), \cdots fl(a_n)), \\ &= \bar{F}(a_1(1+\epsilon_1), a_2(1+\epsilon_2), \cdots a_n(1+\epsilon_n)), \\ &= fl(F(a_1(1+\epsilon_1), a_2(1+\epsilon_2), \cdots a_n(1+\epsilon_n)). \end{aligned}$$

FP Function Evaluation

In this case,

$$\begin{aligned} \bar{y} - y &= [fl(F(a_1(1+\epsilon_1), a_2(1+\epsilon_2), \cdots a_n(1+\epsilon_n)) \\ &\quad -F(a_1(1+\epsilon_1), a_2(1+\epsilon_2), \cdots a_n(1+\epsilon_n))] \\ &\quad +[F(a_1(1+\epsilon_1), a_2(1+\epsilon_2), \cdots a_n(1+\epsilon_n)) \\ &\quad -F(a_1, a_2 \cdots a_n)]. \\ &\equiv A + B, \end{aligned}$$

where $\frac{|A|}{|y|} < \mu$ and $|B|$ can be bounded using the MVT for multivariate functions.

FP Error Bound

If $y = F(a_1, a_2, \cdots a_n)$ is the desired result (defined by exact arithmetic over the Reals), the computed value, $\bar{y}$, will at best satisfy:

$$\begin{aligned} \frac{|\bar{y} - y|}{|y|} &\le \mu + \frac{\|(\frac{\partial F}{\partial \underline{x}})^T (a_1\epsilon_1, a_2\epsilon_2 \cdots a_n\epsilon_n)^T\|}{\|F\|}, \\ &\le \mu + \frac{\|\frac{\partial F}{\partial \underline{x}}\| \, \|a\|\mu}{\|F\|}, \end{aligned}$$

where

$$(\frac{\partial F}{\partial \underline{x}})^T = [\frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \cdots \frac{\partial F}{\partial x_n}],$$

evaluated at $\underline{x} = a = (a_1, a_2, \cdots a_n)$. That is, the relative errors can be large (independent of the approximation used) whenever

$$\frac{\|\frac{\partial F}{\partial \underline{x}}\|\|a\|}{\|F\|} \quad \text{is large.}$$

Notation

- $[a, b]$ is the closed interval, ($x \in R$, such that $a \leq x \leq b$).
- (a, b) is the open interval, ($x \in R$, such that $a < x < b$).
- $f^n(x) = \frac{d^n}{dx^n} f(x)$.
- $f \in C^n[a, b] \;\Rightarrow\; f$ is n times differentiable on $[a, b]$ and $f^n(x)$ is continuous on (a, b).
- $g_x(x, y) \equiv \frac{\partial}{\partial x} g(x, y)$, $g_y(x, y) \equiv \frac{\partial}{\partial y} g(x, y)$, $g_{xy}(x, y) \equiv \frac{\partial^2}{\partial x \partial y} g(x, y)$ etc.
- $g(h) = O(h^n)$ as $h \rightarrow 0 \Leftrightarrow \exists h_0 > 0 \; and \; K > 0 \ni |g(h)| < Kh^n \; \forall \; 0 < h < h_0$.

Theorems From Calculus

- **Intermediate Value Theorem**

  Let $f(x)$ be continuous on $[a, b]$. If $f(x_1) < \alpha < f(x_2)$ for some α and $x_1,\ x_2 \in [a, b]$, then $\alpha = f(\eta)$ for some $\eta \in [a, b]$.

- **Max-Min Theorem**

  Let $f(x)$ be continuous on $[a, b]$. Then $f(x)$ assumes its maximum and minimum values on $[a, b]$. (That is, $\exists \underline{x}\ and\ \bar{x}\ \in [a, b] \ni \forall x \in [a, b]$, we have $f(\underline{x}) \leq f(x) \leq f(\bar{x})$.)

- **Mean Value Theorem for Integrals**

  Let $g(x)$ be a non-negative (or non-positive) integrable function on $[a, b]$. If $f(x)$ is continuous on $[a, b]$ then

$$\int_a^b f(x)g(x)dx = f(\eta) \int_a^b g(x)dx,$$

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

Theorems (cont)

- **Mean Value Theorem for Sums**

  Let $f(x) \in C^1[a,b]$, let $x_1, x_2, \cdots, x_n$ be points in $[a,b]$ and let $w_1, w_2, \cdots, w_n$ be real numbers of one sign, then

  $$\sum_{i=1}^{n} w_i f(x_i) = f(\eta) \sum_{i=1}^{n} w_i,$$

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

- **Rolle's Theorem**

  Let $f(x) \in C^1[a,b]$. If $f(a) = f(b) = 0$ then $f'(\eta) = 0$ for some $\eta \in (a,b)$.

Theorems (cont)

- **Mean Value Theorem for Derivatives**

  If $f(x) \in C^1[a,b]$ then

  $$\frac{f(b)-f(a)}{b-a} = f'(\eta),$$

  for some $\eta \in (a,b)$.

- **Fundamental Theorem of Calculus**

  If $f(x) \in C^1[a,b]$ then $\forall x \in [a,b]$ and any $c \in [a,b]$ we have

  $$f(x) = f(c) + \int_c^x f'(s)ds.$$

Theorems (cont)

- **Taylor's Theorem (with remainder)**

  If $f(x) \in C^{n+1}[a,b]$ and $c \in [a,b]$, then for $x \in [a,b]$,

$$\begin{aligned} f(x) &= f(c) + f'(c)(x-c) + \cdots + f^n(c)\frac{(x-c)^n}{n!} \\ &\quad + R_{n+1}(x), \end{aligned}$$

  where $R_{n+1}(x) = \frac{1}{n!}\int_c^x (x-u)^n f^{n+1}(u)du$.

Note that Taylor's Theorem is particularly relevant to this course. We can observe that, since $(x-u)^n$ is of constant sign for $u \in [c,x]$,

$$R_{n+1}(x) = \frac{1}{n!}\int_c^x (x-u)^n f^{n+1}(u)du = f^{n+1}(\eta)\frac{(x-c)^{n+1}}{(n+1)!},$$

for some $\eta \in [c,x]$.

Taylors Theorem (cont)

We can also observe the first few terms of the Taylor Series provides an accurate approximation to $f(c+h)$ for small h since we have for $h = x - c$,

$$\begin{aligned} f(c+h) &= f(c) + hf'(c) + \cdots \frac{h^n}{n!} f^n(c) \\ &+ \frac{h^{n+1}}{(n+1)!} f^{n+1}(\eta). \end{aligned}$$

where the error term, $E(h)$ is $O(h^{n+1})$.