MM2 Approximate Evaluation of Functions

- kl.8.15-9.00 review of MM1 and some examples
- kl.9.00 – 10.30 exercise (see notes)
- kl.10.40-11.30 MM2 lecture (I)

1

What does NM concern?(MM1)

- Number representations and errors
- Accuracy and precision
- Computation speed
- Required computation power
- Efficiency
- Robustness
- Robust, efficient computing algorithms with prescribed (acceptable) accuracy

2

Floating-point representation (MM1)

- ***significant digits × base^{exponent}***

- Base (radix):
  decimal (10), binary (2), hexadecimal (16)
- Mantissa (significant digits, significand)
- Exponent
- Normalization

- floating-point numbers achieve their greater range at the expense of precision.

3

Measures of errors (MM1)

- Regarding to a number
  - Absolute error
  - Relative error
- Regarding to a function
  - L_infty norm
  - L_1 norm
  - L_2 norm

4

Matlab Basics (MM1)

- Format short, long, short e,
- String, print (fprintf)
- Arithemetic operations
- Mathematical functions
- Vectors (column, row)
- Colon, semi-colon
- Linspace, logspace
- Array arithmetic
- String functions

See appendix A 1-11

5

Conversion methods:

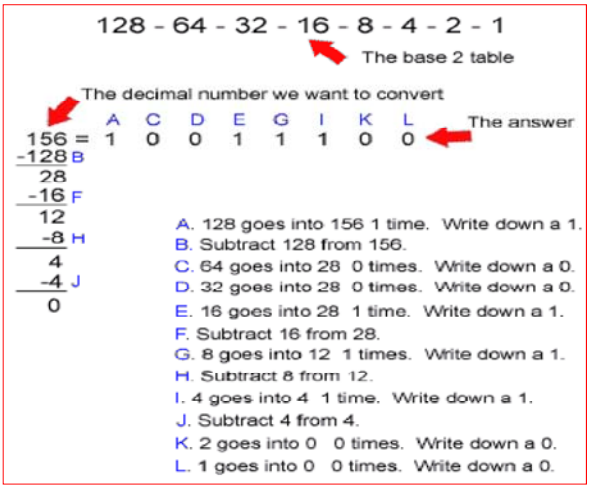
- from Decimal to Binary

http://www.wikihow.com/Convert-from-Decimal-to-Binary

6

Method-1: Comparison with descending powers of two and subtraction

- List the powers of two in a "base 2 table" from right to left.



7

Method-2: Division by two with remainder

- Write the integer answer (quotient) under the long division symbol, and write the remainder (0 or 1) to the right of the dividend.
- Continue downwards, dividing each new quotient by two and writing the remainders to the right of each dividend. Stop when the quotient is 0
- Starting with the bottom remainder, read the sequence of remainders upwards to the top. You should have 10011100. This is the binary equivalent of the decimal number
- This method can be modified to convert from decimal to *any* base

2)156	0
2)78	0
2)39	1
2)19	1
2)9	1
2)4	0
2)2	0
2)1	1
0	

8

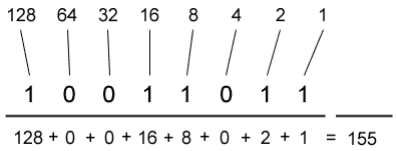
Conversion methods:
Convert from Binary to Decimal

http://www.wikihow.com/Convert-from-Binary-to-Decimal

9

Method-1: Positional notation method

- List the powers of two from right to left. Start at 2^0. Increment the exponent by one for each power. Stop when the amount of elements in the list is equal to the amount of digits in the binary number.



10

Method-2: Doubling method

- Start with the left-most digit of the given binary number. For each digit as you move to the right, double your previous total and add the current digit.

1. **1**011001 → *0**2+**1** = 1
2. **10**11001 → 1*2+**0** = 2
3. **101**1001 → 2*2+**1** = 5
4. **1011**001 → 5*2+**1** = 11
5. **10110**01 → 11*2+**0** – 22
6. **101100**1 → 22*2+**0** – 44
7. **1011001** → 44*2+**1** – 89_{10}

11

Conversion methods:
Convert Decimal Fractions to Binary

- Step 1: Begin with the decimal fraction and multiply by 2. The whole number part of the result is the first binary digit to the right of the point.
- Step 2: Next we disregard the whole number part of the previous result and multiply by 2 once again. The whole number part of this new result is the *second* binary digit to the right of the point. We will continue this process until we get a zero as our decimal part or until we recognize an infinite repeating pattern.

- Infinite Binary Fractions

Example: 0.625_{10} = 0.101 (base 2)
$1/10_{10}$ = ??? (base 2)

12

Approximate Error Analysis

- Determine the number of terms of the exponential series required to estimate e to three decimal places (example 1.4, p.9)

$$e = \exp(1) = \sum_{n=0}^{\infty} \frac{1}{n!}, \quad \hat{e} = \exp(1) = \sum_{n=0}^{N} \frac{1}{n!},$$

$$e - \hat{e} = \sum_{n=N}^{\infty} \frac{1}{n!} = \frac{1}{N!} \sum_{k=0}^{\infty} \frac{1}{N^k} = \frac{1}{N!} \frac{N}{N-1} = \frac{1}{(N-1)(N-1)!}$$

$$N = 6, \quad \frac{1}{5(5!)} = 1.6667 \times 10^{-3}, \quad N = 7, , \quad \frac{1}{6(6!)} = 2.3148 \times 10^{-4}$$

$$\hat{e} = \exp(1) = \sum_{n=0}^{6} \frac{1}{n!} = 2.7181 \approx 2.718$$

$$2.7181 = \sum_{n=0}^{6} \frac{1}{n!} \le e \le \sum_{n=0}^{6} \frac{1}{n!} + \sum_{n=7}^{\infty} \frac{1}{n!} = 2.7181 + 2.3148 \times 10^{-4} = 2.7183$$

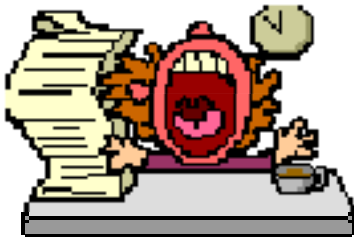
13

Matlab Codes for error analysis (example)

```
clear clc    % clear the workspace and move the cursor at the beginning

y=1          % n=0  term
i=0          % counter
sum=1    % sum all terms so far obtained

while y>10e-4      % while loop controlled by first truncated term

    i=i+1
    y=y*1/i;
   sum=sum+ y
end

% matlab built-in function for exponational
ym=exp(1)

% comparison between our calculation and Matlab function

error=sum-ym
```

14

Exercises (MM1)



15

Question One:

(Exercise 1.2.1, page 7) Express the base of natural logarithms e as a normalized floating-point number, using both chopping and symmetric rounding, for each of the following systems:

- (a) base 10 with 4 significant digits;
- (b) base 10 with 7 significant digits;
- (c) base 2 with 10 significant bits.

Question Two:

(Exercise 1.2.2, page 7) Write down the normalized binary floating-point representations of 1/3, 1/5 and 1/6. Use enough bits in the mantissa to see the recurring patterns.

Question Three:

(Exercise 1.3.3, page 11) How many terms of the series expression

$$cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}$$

are needed to estimate $cosh(1/2)$ with a truncation error less than 10^{-8}? check your answer by comparing with Matlab built-in cosh function.

Question Four:

(Exercise 1.5.1, page 19) Let $x = 1.3576$, $y = 1.3754$. For a hypothetical four decimal digit machine, write down the representations $\hat{x}$ and $\hat{y}$ of x, y. Find the relative errors in the stored results of $x + y$, $x - y$,xy, and x/y using

- (a) chopping, and
- (b) symmetric rounding.

MM2 Approximate Evaluation of Functions

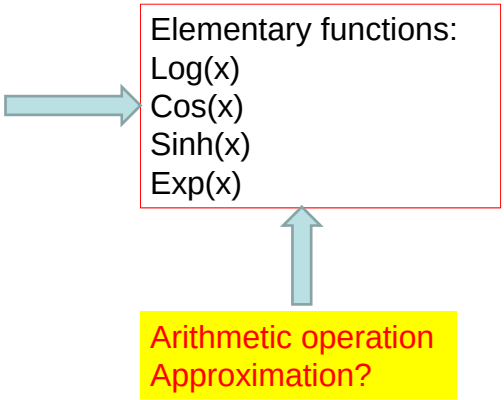
Reading material:

Subsection 3,1, 3,2, and 3.4

17

Question

- How does machine compute values of some functions? Such as
- >> log(1.5)
- ans = 0.4055
- >> cos(100)
- ans = 0.8623
- >> sinh(29)
- ans = 1.9657e+012
- >> exp(-5)
- ans = 0.0067



18

Series Expansions

geometric series :

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots, \qquad |x| < 1$$

exp *onential series* :

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \qquad all\ x$$

trigonomet ric functions :

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots, \qquad all\ x$$

$$\sin x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots, \qquad all\ x$$

19

Concerns of Using Series Approximate

- Radius of convergence of the series expansion, e.g., geometric series
- Truncation error / precision
- Efficient approximate algorithms
  - Approximate of pi using arctan expansion at arctan(1)=pi/4 for IEEE double precision requires 10^16 number of terms, it will take nearly 4 months to obtain the value (see Example 3.2, page 55)
  - Same task using arctan expansion at arctan(1/sqrt(3))=pi/6 only requires 30 terms, it takes one-millionth of a second (see Example 3.5, page 59)

- **CORDIC algorithms**

20

General Series Expansion

Recall from calculus the Taylor's series for a function, $f(x)$, expanded about some number, c, is written as

$$f(x) \sim a_0 + a_1 (x-c) + a_2 (x-c)^2 + \dots .$$

Here the symbol $\sim$ is used to denote a "formal series," meaning that convergence is not guaranteed in general. The constants a_i are related to the function f and its derivatives evaluated at c. When $c = 0$, this is a MacLaurin series.

For example we have the following Taylor's series (with $c = 0$):

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \tag{1.1}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \tag{1.2}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots \tag{1.3}$$

Taylor's Theorem

Theorem 1.1 (Taylor's Theorem). If $f(x)$ has derivatives of order $0, 1, 2, \dots, n+1$ on the closed interval $[a, b]$, then for any x and c in this interval

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

where ξ is some number between x and c, and $f^k(x)$ is the k^{th} derivative of f at x.

We will use this theorem again and again in this class. The main usage is to approximate a function by the first few terms of its Taylor's series expansion; the theorem then tells us the approximation is "as good" as the final term, also known as the *error term*. That is, we can make the following manipulation:

$$\left| f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(c)\,(x-c)^k}{k!} \right| = \frac{\left| f^{(n+1)}(\xi) \right| |x-c|^{n+1}}{(n+1)!} \leq M |x-c|^{n+1}.$$

22

Example-1 of Taylor Series

Example Problem 1.2. Find an approximation for $f(x) = \sin x$, expanded about $c = 0$, using $n = 3$.

Solution: Solving for $f^{(k)}$ is fairly easy for this function. We find that

$$\begin{aligned} f(x) = \sin x &= \sin(0) + \frac{\cos(0)\,x}{1!} + \frac{-\sin(0)\,x^2}{2!} + \frac{-\cos(0)\,x^3}{3!} + \frac{\sin(\xi)\,x^4}{4!} \\ &= x - \frac{x^3}{6} + \frac{\sin(\xi)\,x^4}{24}, \end{aligned}$$

so

$$\left|\sin x - \left(x - \frac{x^3}{6}\right)\right| = \left|\frac{\sin(\xi)\,x^4}{24}\right| \leq \frac{x^4}{24},$$

because $|\sin(\xi)| \leq 1$. ⊣

23

Example-2 of Taylor Series

Example Problem 1.4. Apply Taylor's Theorem to expand $f(x) = x^3 - 21x^2 + 17$ around $c = 1$.

Solution: Simple calculus gives us

$$\begin{aligned} f^{(0)}(x) &= x^3 - 21x^2 + 17, \\ f^{(1)}(x) &= 3x^2 - 42x, \\ f^{(2)}(x) &= 6x - 42, \\ f^{(3)}(x) &= 6, \\ f^{(k)}(x) &= 0. \end{aligned}$$

with the last holding for $k > 3$. Evaluating these at $c = 1$ gives

$$f(x) = -3 + -39(x-1) + \frac{-36\,(x-1)^2}{2} + \frac{6\,(x-1)^3}{6}.$$

Note there is no error term, since the higher order derivatives are identically zero. By carrying out simple algebra, you will find that the above expansion is, in fact, the function $f(x)$. ⊣

24

Alternative Form

Theorem 1.5 (Taylor's Theorem, Alternative Form). If $f(x)$ has derivatives of order $0, 1, \ldots, n+1$ on the closed interval $[a, b]$, then for any x in this interval and any h such that $x + h$ is in this interval,

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

where ξ is some number between x and $x + h$.

We generally apply this form of the theorem with $h \to 0$. This leads to a discussion on the matter of *Orders of Convergence*. The following definition will suffice for this class

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

25

Big-Oh of h^k

Definition 1.6. We say that a function $f(h)$ is in the class $\mathcal{O}\left(h^k\right)$ (pronounced "big-Oh of h^k") if there is some constant C such that

$$|f(h)| \leq C\,|h|^k$$

for all h "sufficiently small," *i.e.*, smaller than some h^* in absolute value.

For a function $f \in \mathcal{O}\left(h^k\right)$ we sometimes write $f = \mathcal{O}\left(h^k\right)$. We sometimes also write $\mathcal{O}\left(h^k\right)$, meaning some function which is a member of this class.

Roughly speaking, through use of the "Big-O" function we can write an expression without "sweating the small stuff." This can give us an intuitive understanding of how an approximation works, without losing too many of the details.

26

One Example

Example 1.7. Consider the Taylor expansion of $\ln x$:

$$\ln(x+h) = \ln x + \frac{(1/x)\ h}{1} + \frac{\left(-1/x^2\right)\ h^2}{2} + \frac{\left(2/\xi^3\right)\ h^3}{6}$$

Letting $x = 1$, we have

$$\ln(1+h) = h - \frac{h^2}{2} + \frac{1}{3\xi^3}h^3.$$

Using the fact that ξ is between 1 and $1+h$, as long as h is relatively small (say smaller than $\frac{1}{2}$), the term $\frac{1}{3\xi^3}$ can be bounded by a constant, and thus

$$\ln(1+h) = h - \frac{h^2}{2} + \mathcal{O}\left(h^3\right).$$

Thus we say that $h - \frac{h^2}{2}$ is a $\mathcal{O}\left(h^3\right)$ approximation to $\ln(1+h)$. For example

$$\ln(1+0.01) \approx 0.009950331 \approx 0.00995 = 0.01 - \frac{0.01^2}{2}.$$

27