Numerical Analysis
Math 128A, Summer 2019
Lecture 1 6/24/2019

Remark: This set of notes is made public in hopes of being helpful in consolidating understanding. It is unofficial and guaranteed to contain errors. If you spot any such inaccuracies or have any comments, please let me know at dsuryakusuma@berkeley.edu.

Lecturer: Professor John A. Strain
`https://math.berkeley.edu/~strain/128a.m19/index.html`

Class: M/Tu/W/Th 11-1pm (289 Cory)
OH: M,Tu 1-2pm (891 Evans)

Textbook: Burden & Faires (BFB) (req'd); Gander, Gander, & Kwok (GGK)

Grading:
- Homework: 30% ; Due weekly on Wednesdays
- MT: 30% ; Wed July 17 ; 11am - 1pm (289 Cory)
- Final: 40% ; Thurs Aug 15 ; 11am - 1pm (289 Cory)

$\geq$ Matlab required (or something better than Matlab)
- vector, matrix, for-loop, functions, plotting

1 Review of Calculus

Theorem 1.1. Intermediate Value Theorem -

For continuous function $f : [a, b] \to \mathbb{R}$,
for any y, there exists $a \leq b, \quad f(c) = y$.

$$\min\{f(a), f(b)\} \leq y \leq \max\{f(a), f(b)\}$$

Example:
For the first few lectures, we'll be looking at a two-step process:
(1) Solve $f(x) = 0$ by bracketing $a \leq x \leq b$ where:

$$f(a) \leq 0 \leq f(b) \text{ or } f(b) \leq 0 \leq f(a),$$

i.e. sign $f(a) \leq$ sign $f(b)$. For example, consider:

$$\ln x = \begin{cases} 1 & x = e \\ -1 & x = 1/e \end{cases}$$

Then there exists $\frac{1}{e} = \frac{1}{2.72} < x < e = 2.72$. So $x \ln x = 0$.
(2) Shrink bracket somehow, with the goal of getting a better solution of $x \ln x = 0$.

Theorem 1.2. Fundamental Theorem of Calculus (FTC)
For f, f' continuous, we have:

$$f(t) = f(a) + \int_a^t f'(s)\ ds$$

Note: We use s here to not get confused with x or t.

Now we derive the Mean Value Theorem (and "everything else we're going to need") using the FTC (hence 'fundamental theorem of calculus').

Theorem 1.3. Mean Value Theorem:
Averaging

$$\text{IVT: } \underbrace{\min_{a\le x\le b} f(x)}_{\text{constant}} \le f(s) \le \underbrace{\max_{a\le x\le b} f(x)}_{\text{constant}}$$

$$0 < a(x) \le b(x)$$
$$0 < c(x) \le d(x)$$

$$a \le b \implies -a \ge -b \implies -b \le -a$$

Example:

$$\int_a^b \left[\min_{a\le x\le b} f(x)\right]\ dx \le \int_a^b f(s)\ ds \le (b-a)\max_{a\le x\le b} f(x)$$

$$\min_x f(x) \le \frac{1}{b-a}\int_a^b f(s)\ ds = \overline{f} = \overline{y} \le \max_x f(x)$$

So there is some number c with $a \le c \le b$ such that

$$f(c) = \overline{f} = \frac{1}{b-a}\int_a^b f(s)\ ds.$$

MVT for Integrals:

$$\int_a^b f(s)\ ds = (b-a)f(c)$$

for some $c \in [a, b]$, c is unknown, but usually enough to know c exists.

Example: $a \le b \to ga \le gb$ (if $g > 0$)

$$g(s)[\min f(x)] \le f(s) \le g(s)[\max f(x)]$$

$$\int_a^b g(s)[\min f(x)] \le \int_a^b f(s) \le \int_a^b g(s)[\max f(x)]$$

$$\min f(x) \le \frac{\int_a^b g(s)f(s)\ ds}{\int_a^b g(s)\ ds} \le \max f(x)$$

Then the 'expectation' of f is $\overline{f}$ or $\langle f \rangle$ or $E(f)$.

MVT for Integrals:
The IVT tells us the same thing: $\exists c \in [a, b]$ such that

$$\int_a^b g(s)f(s)\ ds = f(c) \int_a^b g(s)\ ds$$

$$\begin{aligned}\frac{f(x) - f(a)}{x - a} &= \frac{1}{x - a}\int_a^x f'(s)\ ds \\ &= f'(c) \quad \text{for some unknown } c \text{ between } a \text{ and } x.\end{aligned}$$

Another way of saying this is:

$$f(x) - f(a) = f'(c)(x - a)$$

If you know the derivative of a function, then you can control how far apart the values are.

Example: $f(x) = \sin(x) \qquad f'(x) = \cos(x)$
$\sin(x) - \sin(a) = \cos(c)(x - a)$
So if x is close to a, then cos is always less than or equal to 1.

$$|\sin(x) - \sin(a)| = |\cos(c)(x - a)| \leq |x - a|$$

So sin is like a contraction, where mistakes going through a process will be 'redeemed' and 'forgotten', as long as our expressions have bounded derivatives.

Theorem 1.4. Taylor's Theorem (generalization of MVT)

$$\begin{aligned}f(x) &= f(a) + \int_a^x [1]f'(s)\ ds \\ &\qquad \text{(use int. by parts ; integrate the 1, differentiate the } d'(s)) \\ &= f(a) - \int_a^x \left[\frac{d}{ds}(x - s)\right] f'(s)\ ds \\ &= f(a) - (x - s)f'(s)\Big|_a^x - \left(\int_a^x -(x - s)f''(s)\ ds\right) \\ &= f(a) + (x - a)f'(a) + \int_a^x (x - s)f''(s)\ ds \\ &= f(a) + (x - a)f'(a) + \int_a^x \left[-\frac{d}{ds} \cdot \frac{(x - s)^1}{1!}\right] f''(s)\ ds \\ &= \frac{(x - a)^0}{0!}f(a) + \frac{(x - a)^1}{1!}f'(a) + \frac{(x - a)^2}{2!}f''(a) + \cdots\end{aligned}$$

Common Taylor Expansions:

$$\begin{aligned} f(x) = e^{bx} \quad f'(x) = be^{bx} = f^{(k)}(x) = b^k e^{bx} \\ f(0) = 1 \quad f'(0) = bf^{(k)}(0) = b^k \\ e^{bx} = \frac{x^0}{01}1 + \frac{x^1}{1!}b + \frac{x^2}{2!}b^2 + \frac{x^3}{3!}b^3 + \cdots \\ e^{bx} = \sum_{k=0}^{\infty} \frac{b^k}{k!}x^k \end{aligned}$$

Writing infinite series like that something $= \cdots$ means that the remainder is

$$\sum_{k=K}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k = R_K(x) \to 0_{\text{as } K \text{ goes to } \infty}$$

Also $f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots$, let's say we stop at some

$$\frac{(x-a)^{k-1}}{(k-1)!}f^{(K-1)}(a) + \underbrace{\int_a^x \frac{(x-s)^{k-1}}{(k-1)!}f^{(k)}(s)\ ds}_{R_K(x)}$$

with $f(s) > 0$. So we have, for the error of a Taylor Polynomial stopping BEFORE the k-th term:

$$\begin{aligned} R_k(x) &= f^{(k)}(c)\int_a^x \frac{(x-s)^{(k-1)}}{(k-1)!}\ ds \\ &= f^{(k)}(c)\left[\frac{-(x-s)^k}{k!}\right]_a^x \\ &= f^{(k)}(c)\frac{(x-a)^k}{k!} \end{aligned}$$

Hence, for our example b^{bx} above:

$$|R_k(x)| \le b^k e^{bx}\frac{(x-a)^k}{k!}$$

Stirling's Approximation:

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

Example:

$$\begin{aligned}
f(x) &= \frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \quad \text{, if } |x| < 1, \quad 0 < c < x \\
f'(x) &= \left((1-x)^{-1}\right)' = 1!(1-x)^{-2} \\
f'(0) &= 1! \\
f''(x) &= 2!(1-x)^{-3} \\
f''(0) &= 2!
\end{aligned}$$

So we check, with our formula

$$\begin{aligned}
R_k(x) &= f^{(k)}(c)\frac{(x-0)^k}{k!} = k!(1-c) \\
&= k!(1-c)^{-k-1}\frac{x^k}{k!} \\
&= \frac{1}{1-c}\left(\frac{1}{1-c}\right)^k \\
&\leq \frac{1}{1-x}\left(\frac{x}{1-x}\right)^k \\
&\leq \frac{2^{-k}}{1-x} \text{ if } x \leq \frac{1}{3}
\end{aligned}$$

So we have a bound for the relative error. Question: How big does k have to be until we don't care anymore? 10 is usually enough for mechanical engineering. 6-digit accuracy can be good enough for electric engineering. For bio-engineering and heart surgery, we'll probably want 9-digit accuracy.

Definition: Taylor Polynomial, Remainder (BFB) -

Let $P_n(x)$ be the nth **Taylor polynomial** for f about x_0, and $R_n(x)$ be the **remainder term** or **truncation error** associated with $P_n(x)$. Suppose $f \in C^n[a,b], f^{(n+1)}$ exists on $[a,b]$, and $x_0 \in [a,b]$. For every $x \in [a,b]$, there exists a number $\xi(x)$ between x_0 and x with:

$$f(x) = P_n(x) + R_n(x),$$

$$\begin{aligned}
P_n(x) &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \\
R_n(x) &= \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}
\end{aligned}$$

Alternative to Taylor expansion:

$$
\begin{aligned}
f(1) &= f(0) + \int_0^1 \underbrace{1}_{-\frac{d}{ds}(1-s)} f'(s)\, ds \\
&= f(0) + \int_0^1 \underbrace{1}_{-\frac{d}{ds}(\frac{1}{2}-s)} f'(s)\, ds \\
&= f(0) + \underbrace{\left(-(\frac{1}{2}f'(s)\big|_0^1)\right)}_{-\frac{1}{2}f'(1)+\frac{1}{2}f'(0)} - \int_0^1 - \left[\frac{1}{2} - s\right] f''(s)\, ds \\
&= f(0) - \frac{1}{2}f'(1) + \frac{1}{2}f'(0) - \int_0^1 - \left[-\frac{d}{ds}\frac{(\frac{1}{2}-s)^2}{2!}\right] f''(s)\, ds \\
&= f(0) - \frac{1}{2}\left[f'(1) - f'(0)\right] - \frac{1}{2!}\left[\left(\frac{1}{2} - s\right)^2 f''(s)\right]_0^1 \\
&\qquad + \int_0^1 -\frac{d}{ds}\frac{(\frac{1}{2}-s)^3}{3!} f''(s)\, ds \\
\implies f(1) - f(0) &\approx -\frac{1}{2}\left[f'(1) - f'(0)\right] - \frac{1}{8}\left[f''(1) - f''(0)\right]
\end{aligned}
$$

2 Computer Arithmetic

Floating Point arithmetic $\approx$ real number arithmetic

64-bit IEEE Standard Arithmetic:

$$\underbrace{(-1)^s}_{\text{(sign, 1 bit)}} \quad \underbrace{2^{c-1023}}_{\text{(Characteristic, 11-bit exponent)}} \quad (1 + \quad \underbrace{f}_{\text{(mantissa, 52-bit binary fraction)}} \quad)$$

1023 is the bias applied to c.

$$[(-1)^s]\ [2^{e-1023}]\ [\ \underbrace{1.}_{\text{implied}}\ b_1 \dots b_{52}]$$

So the mantissa gives the numbers between 1 and 2 (not 0 1).

Definition: machine epsilon -

$$\varepsilon := 2^{-52}$$

$$\begin{gathered} \dots \quad |, \\ \frac{1}{2} - \frac{\varepsilon}{4} \quad |, \\ \frac{1}{2}, \dots, 1 - \frac{3\varepsilon}{2}, 1 - \frac{\varepsilon}{2} \quad |, \\ 1, 1+\varepsilon, \dots, 2-\varepsilon \quad |, \\ 2, 2+2\varepsilon, \dots, 4-2\varepsilon \quad |, \\ 4, 4+4\varepsilon, \dots \end{gathered}$$

So there are 2^{11} logarithmic intervals $[2^{+(k-1)}, 2^{+k}]$. On the interval $[2^{k-1}, 2^k)$, there are 2^{52} floating point numbers ("FP#s").

Example: Consider the number:

$$0 \quad 10000000011 \quad 1011100100010 \cdots 0$$

First bit is 0, so $s = 0$ and the number is positive.
The next 11 bits gives the characteristic, equal to:

$$c = 1 \times 2^{10} + 1 \times 2^1 + 1 \times 2^0 = 1024 + 2 + 1 = 1027$$

So the exponent part is $2^{1027-1023} = 2^4 = 16$.
The last 52 bits specify that the mantissa is

$$f = 1\left[\frac{1}{2}\right]^1 + 1\left[\frac{1}{2}\right]^3 + 1\left[\frac{1}{2}\right]^4 + 1\left[\frac{1}{2}\right]^5 + 1\left[\frac{1}{2}\right]^8 + 1\left[\frac{1}{2}\right]^{12}.$$

So in total, the machine number equals:

$$\begin{aligned} (-1)^s 2^{c-1023}(1+f) &= (-1)^0 2^{1027-1023}\left[1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right)\right] \\ &= 1 \times 16 \times (\cdots) \\ &= 27.56640625. \end{aligned}$$

Definition: Rounding -

$$fl(x) := \text{ nearest FP \# to } x$$

$$fl : \mathbb{R} \to FP\#S$$

$$2.718 \to 2.72$$

Round to "even" (last digit 0) if there is a tie. For example,

Example of tie: How do we round $1.111 \to ?$
$2 - \frac{1}{8} = 10.0$ last bit 0, is the stable choice
As opposed to $2 - \frac{1}{4} = 1.11$, which we don't choose.

Remark: END LECTURE 1 HERE