Supplementary Notes 1

Tiago Salvador (tiago.saldanhasalvador@mail.mcgill.ca)

Abstract

We discuss round-off errors and their propagation in computer arithmetic and review some important results from calculus. We discuss as well the bisection method.

Contents

1 Round-off errors and Computer Arithmetic

Errors occur during calculations when using a computer. Let's denote with $*$ any operation, be it $+$, $-$,$\times$ or $/$. Denote the floating point approximation for any real number x with $fl(x)$. The main thing to remember is that in general

$$fl(fl(x) * fl(y)) \neq fl(x * y) \neq fl(x) * fl(y).$$

Arithmetic operations in the computer are a cause of error and the way the error propagates is not always obvious.

Example 1.1 (An example of error propagation)**.** *Suppose we are working with a computer that rounds to five significant digits and we want to subtract the rational numbers*

$$x = \frac{301}{2000} \approx 0.15050000\ldots \quad \textit{and} \quad y = \frac{301}{2001} \approx 0.150424787\ldots.$$

The exact result $x - y$ is approximately

$$0.0000752139\ldots.$$

The computer rounds x and y to 0.15050 *and* 0.15042 *respectively. Subtracting these two numbers we get* 0.00008*. This has no significant digits in common with the exact result. Alternatively, we can choose to calculate the difference by writing*

$$\begin{aligned}\frac{301}{2000} - \frac{301}{2001} &= \frac{301 \times 2001 - 301 \times 2000}{2000 \times 2001} \\ &= \frac{602301 - 602000}{4002000} \\ &\approx \frac{6.0230 \times 10^5 - 6.0200 \times 10^5}{4.0020 \times 10^6} \\ &= \frac{3.0000 \times 10^2}{4.0020 \times 10^6} \\ &\approx 0.000074963.\end{aligned}$$

The way we computed things now gets us some significant digits. What this means is that floating point computation does not satisfy things we know to be true for arithmetic operations, like associative and distributive laws.

2 Review of Calculus

We start with some notation. Let $[a, b]$ be a closed interval. For $n \in \mathbb{N}$,

$$C^n[a, b] = \{f : f \text{ is an n-times continuously differentiable function on } [a, b]\}$$

Specifically,

- $f \in C^n[a, b]$ means $f(x)$ is n-times continuously differentiable function on $[a, b]$,
- $f \in C^0[a, b]$ or $f \in C[a, b]$ means $f(x)$ is a continuous function on $[a, b]$.

Example 2.1. *Let $f(x) = e^{-|x|}$. We have $f \in C[-1, 1]$ but $f \notin C^1[-1, 1]$.*

We now review some important theorems in Calculus.

Theorem 2.1 (Extreme value theorem)**.** *If $f \in C[a, b]$ then there are $c, d \in [a, b]$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in [a, b]$. In addition, if f is differentiable on (a, b), then the numbers c and d occur either at the endpoints of $[a, b]$ or where f' is zero.*

Theorem 2.2 (Intermediate value theorem)**.** *If $f \in C[a, b]$ and K is any number between $f(a)$ and $f(b)$, then there is $\xi \in (a, b)$ for which $f(\xi) = k$.*

Remark 2.3. *The intermediate value theorem only guarantees the existence of ξ. In general, we don't know where ξ is, besides the fact that it belongs to the interval (a, b). Moreover, such ξ may not be unique.*

Theorem 2.4 (Rolle's theorem)**.** *Suppose $f \in C[a, b]$ and f is differentiable on (a, b). If $f(a) = f(b)$, then there is $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

Theorem 2.5 (Mean value theorem)**.** *If $f \in C[a, b]$ and f is differentiable on (a, b), then there is $\xi \in (a, b)$ such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Exercise 2.1. *Find the number of roots and some intervals containing them for the functions $f(x) = x - 3^{-x}$ and $g(x) = x^3 - 2x^2 - 4x + 2$.*

Solution: We have $f \in C[0, 1]$ and $f'(x) = 1 + 3^{-x} \log(3)$. We have that f' is strictly positive. Hence f is strictly increasing and can have at most one zero. Now observe that $f(0) = -1$ and $f(1) = 2/3$ and so by the Intermediate Value Theorem there is $\xi \in [0, 1]$ such that $f(\xi) = 0$.

We have $g'(x) = 3x^2 - 4x - 4$ and so the zeros of g' are $x_1 = -\frac{2}{3}$ and $x_2 = 2$. In addition

	$-\infty$	x_1		x_2	∞
$g'(x)$	$+$	0	$-$	0	$+$
$g(x)$	$\nearrow$	$\frac{94}{27}$	$\searrow$	-6	$\nearrow$

Table 1: Behavior of g.

In addition,

$$\lim_{x \to -\infty} g(x) = -\infty \quad \text{and} \quad \lim_{x \to \infty} g(x) = \infty.$$

Therefore g has 3 solutions. Now we observe that $g(-2) = -6$ and $g(4) = 18$. Hence by the Intermediate Value Theorem there is a solution in each of the following intervals: $(-2, -\frac{2}{3})$, $(-\frac{2}{3}, 2)$ and $(2, 4)$.

Exercise 2.2. *Let $f(x) = (x-1)\tan(x) + x\sin(\pi x)$. Show that f' has at least one zero in $[0,1]$.*

Solution: Note that $f \in C[0,1]$ and f is differentiable on $(0,1)$. Also $f(0) = 0$ and $f(1) = 0$. Hence by Rolle's theorem there is $\xi \in [0,1]$ such that $f'(\xi) = 0$.

3 Taylor's Theorem

The Taylor's theorem we present next is one of the most important Theorems in class and it will be used constantly throughout the course.

Theorem 3.1 (Taylor's theorem)**.** *Suppose $f \in C^n[a,b]$, $f^{(n+1)} \in C(a,b)$ and that $x_0 \in [a,b]$. For every $x \in [a,b]$, there exists a number $\xi(x)$ between x_0 and x with*

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

where

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \end{aligned}$$

and

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}.$$

Here $P_n(x)$ is called the n-th Taylor polynomial for f about x_0 and $R_n(x)$ is called the remainder term (or truncation error) associated with $P_n(x)$.

Remark 3.2. *In general, $P_n(x)$ is a good approximation to $f(x)$ only for $x \approx x_0$. The actual error is usually smaller than the upper bound.*

Exercise 3.1. *Let $f(x) = e^x \cos(x)$.*

a) Find the second Taylor polynomial $P_2(x)$ for f about $x_0 = 0$.

Solution: *We first compute f' and f''. We have*

$$f'(x) = e^x\cos(x) - e^x\sin(x) \quad \text{and} \quad f''(x) = -2e^x\sin(x).$$

Hence

$$P_2(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f'(x_0)}{2!}(x-x_0)^2 = 1 + x.$$

b) Find an upper bound for the error if $x \in [0,1]$.

Solution: *We need to first compute the remainder term. We have*

$$f^{(3)}(x) = -2e^x \sin x - 2e^x \cos x = 2e^x(\sin x + \cos x).$$

and so

$$R_2(x) = -\frac{1}{3}e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))x^3.$$

Thus the maximum error can be written as

$$\begin{aligned}
&\max_{x\in[0,1]} |f(x) - P_2(x)| \\
=&\max_{x\in[0,1]} |R_2(x)| \\
=&\max_{x\in[0,1]} \left|-\frac{1}{3}e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))x^3\right| \\
\leq&\max_{x\in[0,1],\xi\in[0,x]} \left|-\frac{1}{3}e^{\xi}(\sin\xi + \cos\xi)x^3\right| \\
\leq&\max_{x\in[0,1],\xi\in[0,1]} \left|-\frac{1}{3}e^{\xi}(\sin\xi + \cos\xi)x^3\right| \\
\leq&\max_{\xi\in[0,1]} \left|-\frac{1}{3}e^{\xi}(\sin\xi + \cos\xi)\right| \max_{x\in[0,1]} \left|x^3\right| \\
=&\max_{\xi\in[0,1]} \left|\frac{1}{3}e^{\xi}(\sin\xi + \cos\xi)\right| \\
\leq&\frac{1}{3}\max_{\xi\in[0,1]} e^{\xi}\left(|\sin\xi| + |\cos\xi|\right) \\
\leq&\frac{1}{3}\max_{\xi\in[0,1]} e^{\xi}\left(\max_{\xi\in[0,1]}|\sin\xi| + \max_{\xi\in[0,1]}|\cos\xi|\right) \\
\leq&\frac{e}{3}(1+\sin(1)) \\
\approx&1.669
\end{aligned}$$

c) Approximate $\int_0^1 f(x)dx$ using $P_2(x)$ and find an upper bound for the error using $\int_0^1 |R_2(x)|dx$.

Solution: *We have*

$$\int_0^1 f(x)\,dx \approx \int_0^1 P_2(x)\,dx = \int_0^1 (1+x)\,dx = \frac{3}{2}.$$

We now show that $|R_2(x)| \leq \frac{1+\sin(1)}{3}ex^3$ for $x \in [0,1]$ by a similar reasoning to the one made

in b). We have that for $x \in [0,1]$

$$\begin{aligned}
&|R_2(x)| \\
&= \left| -\frac{1}{3} e^{\xi(x)} (\sin(\xi(x)) + \cos(\xi(x))) x^3 \right| \\
&\leq \max_{\xi \in [0,x]} \left| -\frac{1}{3} e^{\xi} (\sin\xi + \cos\xi) x^3 \right| \\
&\leq \max_{\xi \in [0,1]} \left| -\frac{1}{3} e^{\xi} (\sin\xi + \cos\xi) x^3 \right| \\
&\leq \max_{\xi \in [0,1]} \left| -\frac{1}{3} e^{\xi} (\sin\xi + \cos\xi) \right| |x^3| \\
&= x^3 \max_{\xi \in [0,1]} \left| \frac{1}{3} e^{\xi} (\sin\xi + \cos\xi) \right| \\
&\leq x^3 \frac{1}{3} \max_{\xi \in [0,1]} e^{\xi} \left(|\sin\xi| + |\cos\xi| \right) \\
&\leq x^3 \frac{1}{3} \max_{\xi \in [0,1]} e^{\xi} \left(\max_{\xi \in [0,1]} |\sin\xi| + \max_{\xi \in [0,1]} |\cos\xi| \right) \\
&\leq x^3 \frac{e}{3} (1 + \sin(1))
\end{aligned}$$

Hence we have

$$\int_0^1 |R_2(x)|\, dx \leq \int_0^1 \frac{1+\sin(1)}{3} e x^3\, dx = \frac{1+\sin(1)}{12} e \approx 0.4171.$$

We could have used the upper bound for $|R_2(x)|$ *deduced in b). However by keeping the* x *in the estimate we are able to get a tighter bound.*

Exercise 3.2. *Use the error term of a Taylor polynomial to estimate the error involved in using* $\sin(x) \approx x$ *to approximate* $\sin(1^o)$.

Solution: To approximate $\sin(x)$ by x means that we are approximating by the Taylor polynomial about $x_0 = 0$ with $n = 1$. The error term is then given by

$$R_1(x) = -\frac{\sin(\xi(x))}{2!} x^2.$$

Recalling that $1^o = \frac{\pi}{180}$, the estimate is then

$$\left| R_1\left(\frac{\pi}{180}\right) \right| \leq \frac{\sin(\frac{\pi}{180})}{2!} \left(\frac{\pi}{180}\right)^2 \approx 2.685 \times 10^{-6},$$

when the actual error is actually -8.861×10^{-7}.

Remark 3.3. *Note that the estimates for the error are a worst case situation. In general, the actual error will be smaller.*

Exercise 3.3. *Compute* $P_2(x)$ *of* $f(x) = \frac{e^x - 1}{x}$ *about* $x_0 = 0$.

Solution: We could compute $P_2(x)$ using Taylor's formula, but an easier way is to use Taylor series expansion of e^x. We know that

$$e^x = 1 + x + \frac{x^2}{2} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Hence,

$$\begin{aligned} f(x) &= \frac{\left(1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots\right) - 1}{x} \\ &= \frac{x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots}{x} \\ &= 1 + \frac{x}{2} + \frac{x^2}{6} + \ldots \end{aligned}$$

Then $P_2(x) = 1 + \frac{x}{2} + \frac{x^2}{6}$.

4 Bisection Method

Theorem 4.1 (Bisection Method)**.** *Suppose that $f \in C[a,b]$ and $f(a) \cdot f(b) < 0$. The bisection method generates a sequence $\{x_k\}_{k=0}^{\infty}$ approximating a zero x^* of f with*

$$|x_k - x^*| \leq \frac{b-a}{2^{k+1}} \quad \text{when } k \geq 0.$$

Remark 4.2. *The estimate on the error is independent of the function f. It only depends on the length of the starting interval.*

Exercise 4.1. *How many iterations are necessary a priori to guarantee that we find the root of $f(x) = x - \tan(x)$ with accuracy 10^{-3} on the interval $[4, 4.5]$?*

Solution: We need to find k such that

$$|x_k - x^*| \leq \frac{b-a}{2^{k+1}} = \frac{1}{2^{k+2}} < 10^{-3}.$$

We then want

$$\begin{aligned} \frac{1}{2^{k+2}} &< 10^{-3} \\ \frac{1}{10^{-3}} &< 2^{k+2} \\ \log\left(\frac{1}{10^{-3}}\right) &< \log\left(2^{k+2}\right) \\ k &> \frac{\log\left(\frac{1}{10^{-3}}\right)}{\log(2)} - 2 \approx 7.9658. \end{aligned}$$

Thus we need 8 iterations to guarantee an accuracy of 10^{-3}.

Exercise 4.2. *Explain how you would use the bisection method to find an approximation to $\sqrt{3}$ correct to within 10^{-4}.*

Solution: We need to consider a function which has $\sqrt{3}$ as a root, e.g., the function $f(x) = x^2 - 3$. Taking $a = 0$ and $b = 2$, we observe that we have $f(a)f(b) < 0$ since $f(0) = -3$ and $f(2) = 1$. It's also clear that $\sqrt{3}$ is the only root of f in $[0, 2]$. Thus by Theorem 4.1, the bisection method will produce a convergent sequence $\{x_k\}_{k=0}^{\infty}$ to $\sqrt{3}$ such that

$$|x_k - \sqrt{3}| \leq \frac{1}{2^k} \quad \text{when } k \geq 0.$$

By a similar reasoning to the previous exercise, we get

$$k > \frac{\log\left(\frac{1}{10^{-4}}\right)}{\log(2)} \approx 13.2877.$$

So $k = 14$ is sufficient.

Remark 4.3. *It is important to keep in mind that the error analysis gives only a bound for the number of iterations and in many cases is much larger than the actual number required. Also, the sequence of errors in the bisection method is not, in general, strictly decreasing, although, it always goes to 0.*

Remark 4.4. *The slow reduction of the error suggests that we use the bisection method as an "approaching" technique to the root. Indeed, with a few bisection steps, we can get a reasonable approximation to x^* to which a higher order method can then be applied for a rapid convergence to the solution within the fixed tolerance.*

References

[1] R. L. Burden and J. D. Faires. *Numerical Analysis*. 9^{th} edition. Brookes/Cole, 2004.

[2] A. Quarteroni, R. Sacco and F. Saleri. *Numerical Mathematics*. 2^{nd} edition. Springer, 2006.

[3] Math 317 Tutorial notes written by Jan Feys