Midterm Solutions 553.481/681, February 25, 2025

Problem 1. *For each of the numbers x given below, answer the following three questions:*

(i) Write the number in the form $x = (1+F)\times 2^E$ where the fraction $0 \le F < 1$ should be given its base-2 representation and the exponent E is an integer.

(ii) Is x a machine number in IEEE standard double precision arithmetic? Explain why or why not, using the results of (i).

(iii) If x is not a machine number, what machine number $\hat{x}$ is used to represent it in IEEE standard double precision arithmetic if using the "round-to-even" rule? Give the result as $\hat{x} = (1+\hat{F})\times 2^{\hat{E}}$ with $\hat{F}$ in base-2 for normalized numbers, as $\hat{x} = \hat{F}\times 2^{\hat{E}}$ and $\hat{F}$ in base-2 for denormalized numbers, or else as `INF`, `NAN`, *etc.*

$$(a)\quad x = 2^{52}+1 \qquad\qquad (b)\quad x = 2^{53}+1$$

3

Solution:

(a) $x = 2^{52}+1 = (1+2^{-52})\times 2^{52} = (1.\underbrace{00\cdots01}_{52})_2 \times 2^{52}$ so that $E = 52$ and

3 $F = (0.\underbrace{00\cdots01}_{52})_2$. Yes, since $E \le 1023$ and F has only 52 bits, this is an IEEE standard double-precision number. 3

3

(b) $x = 2^{53}+1 = (1+2^{-53})\times 2^{53} = (1.\underbrace{00\cdots01}_{53})_2 \times 2^{53}$ so that $E = 53$ and

3 $F = (0.\underbrace{00\cdots01}_{53})_2$. NO, this is not an IEEE double precision number, since still $E \le 1023$ but now F has 53 bits. 3

2+2

Using the round-to-even rule

$$F = (0.\underbrace{00\cdots01}_{53})_2 \longrightarrow \hat{F} = (0.\underbrace{00\cdots0}_{52})_2 = 0$$

3 and $\hat{E} = 53$, so that $\hat{x} = (1+0)\times 2^{53} = 2^{53}$.

In fact, $x = 2^{53}+1$ is the smallest integer that cannot be exactly represented in IEEE standard double-precision arithmetic!

Problem 2. *This problem concerns the symmetric-difference approximation to the second-derivative*

$$D_h^2 f(x) \equiv \frac{f(x+h)+f(x-h)-2f(x)}{h^2} \doteq f''(x), \quad h>0.$$

By means of the Taylor expansion with remainder, the truncation error of this approximation can be proved to be of the form:

$$D_h^2 f(x) = f''(x) + \frac{1}{12} f^{(4)}(\xi)h^2, \quad \xi \in [x-h, x+h]. \qquad (*)$$

(a) If this symmetric-difference approximation is implemented in IEEE standard double precision, derive an estimate for the optimal choice h_* to minimize total error. Your estimate should be given in terms of the numbers $M = \max_x |f(x)|$, $M_4 = \max_x |f^{(4)}(x)|$, and eps, *the machine epsilon.*

(b) Give a corresponding estimate for the minimum error. What is the expected accuracy of this derivative approximation evaluated in double-precision arithmetic?

Solution. (a) The error in the machine-arithmetic approximation

$$\widehat{D_h^2 f(x)} \equiv \frac{\widehat{f(x+h)} + \widehat{f(x-h)} - 2\widehat{f(x)}}{h^2}$$

must also include the round-off error. Assuming that $Rel(\widehat{f(x)}) \simeq eps$, we see that this round-off error is at most $4M \cdot eps/h^2$.

Thus, the total error satisfies

6 5

$$Err \leq \frac{4\, eps\, M}{h^2} + \frac{M_4 h^2}{12}.$$

Minimizing the upper bound by $0 = \frac{d}{dh}\left(\frac{4\, eps\, M}{h^2} + \frac{M_4 h^2}{12}\right) = -\frac{8\, eps\, M}{h^3} + \frac{M_4 h}{6}$ gives the optimal value $h_* = (48M\, eps/M_4)^{1/4}$. 6

(b) Substituting h_* from (a), the optimal upper bound is $Err_* = (eps\, M M_4/3)^{1/2} + (eps\, M M_4/3)^{1/2} = 2(eps\, M M_4/3)^{1/2}$. Since $eps \doteq 10^{-16}$ in double-precision arithmetic, the error should be about $eps^{1/2} \doteq 10^{-8}$, which is single-precision accuracy.

6

2

Problem 3. *Consider the problem of finding the set of all of the real roots (up to multiplicity) of the function*

$$f(x) = 1 + \cos x.$$

(a) If ChatGPT is asked whether this problem is well-posed, it's response is as follows:

> The problem is not well-posed in the strict sense because uniqueness fails due to the infinitely many solutions. However, if we restrict the problem to a finite interval (e.g., finding the smallest positive root), it can be made well-posed.

Is this ChatGPT response correct? In your answer, give the definition of a well-posed problem and explain why the definition is satisfied here or not.

(b) If ChatGPT is asked whether this problem is well-conditioned, it's response is as follows:*

> The problem is ill-conditioned because the roots are located at points where $\sin x = 0$, meaning that a tiny change in $f(x)$ results in a large shift in the root locations.

Is this ChatGPT response correct? If not, explain why. If the response is correct, also explain why and then explain approximately how many digits of accuracy would be lost in IEEE double precision arithmetic.

48 | 68 |

4 **Solution:** (a) The set of all real roots is easily checked to be $\{x_k = (2k+1)\pi | \, k \in \mathbb{Z}\}$. 3
4 A problem is said to be well-posed if the solution exists, is unique, and continuous in 3
4+4 the data. The above set exists and is unique. ChatGPT has correctly noted that the 2+2
set of roots has infinite cardinality, but the set itself is unique!

For a small change δy in $f(x)$, the small change δx_k in x_k can be obtained from

5 $$1 + \cos(x_k + \delta x_k) + \delta y = 0. \quad (*)$$ 3

Chat GPT wrong

For $\delta y > 0$ there is no solution, so that the entire set of roots becomes the empty set!

Hence, ChatGPT is correct that the problem is ill-posed, but its explanation is completely wrong. The set of roots is unique but is not stable under any small perturbation of $f(x)$ with $\delta y > 0$. 3

4

Note that this problem is in fact well-posed if one restricts to small perturbations with $\delta y < 0$. Each individual double root splits into a pair of simple roots, but these are both close to the original double root.

(b*) Taylor-expanding the condition (*) in δx_k

$$1 + \cos(x_k) - \sin(x_k)\delta x_k - \frac{1}{2}\cos(x_k)(\delta x_k)^2 \doteq -\delta y$$

so that using $\cos(x_k) = -1$, $\sin(x_k) = 0$ gives $\frac{1}{2}(\delta x_k)^2 \doteq -\delta y$. For the case that $\delta y < 0$ when this problem is well-posed, then $\delta x_k \doteq \pm(-2\delta y)^{1/2}$. Thus, $\delta x_k \to 0$ as $\delta y \to 0$, but far slower than δy itself. This problem is very ill-conditioned, with absolute condition number $\delta x_k/\delta y = \pm(-2/\delta y)^{1/2}$ diverging as $\delta y \to 0$. We see that ChatGPT's response is correct! 3

3

If one takes $\delta y \doteq eps \doteq 10^{-16}$ for double-precision arithmetic, then $\delta x_k \doteq \pm(-2\delta y)^{1/2}$ $\doteq 10^{-8}$. Hence, about 8 digits of accuracy will be lost in double-precision arithmetic. 3

Problem 4. *(a*) Show that if $f(x)$ is a function with a root x_* of multiplicity $p > 1$, then the function*

$$F(x) = f(x)/f'(x)$$

has x_* as a simple root.

(b) Write down in terms of the original function $f(x)$ the Newton iteration to determine the root x_* of $F(x)$.

(c) What will be the rate of convergence of the iteration in (b)?

4561 **Solution:** (a*) If $f(x) = (x-x_*)^p h(x)$ for $h(x_*) \neq 0$, then $f'(x) = p(x-x_*)^{p-1}h(x) + (x - x_*)^p h'(x)$. Hence, 681

$$F(x) = \frac{(x - x_*)^p h(x)}{p(x - x_*)^{p-1}h(x) + (x - x_*)^p h'(x)} = (x - x_*)H(x)$$

9

with

$$H(x) = \frac{h(x)}{ph(x) + (x - x_*)h'(x)}.$$

Since $H(x_*) = h(x_*)/ph(x_*) = 1/p \neq 0$, we see that $F(x)$ has x_* as a simple root.

(b) Using the quotient rule

$$F'(x) = \frac{d}{dx}\left[\frac{f(x)}{f'(x)}\right] = \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2}$$

so that

$$\frac{F(x)}{F'(x)} = \frac{f(x)f'(x)}{(f'(x))^2 - f(x)f''(x)}.$$

Thus, the Newton iteration for $F(x)$ becomes

13

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} = x_n - \frac{f(x_n)f'(x_n)}{(f'(x_n))^2 - f(x_n)f''(x_n)}.$$

8

12 (c) Because x_* is a simple root of $F(x)$, then according to the general theorems the convergence will be quadratic, at least when F is twice-continuously differentiable (or f thrice-continuously differentiable) near x_*. 8

Note, however, that expressed in terms of $f(x)$, the problem remains ill-conditioned and accuracy to working precision will not usually be obtained.