MATH3230A Numerical Analysis

Tutorial 3

1 Recall:

1. **Floating-point arithmetic**:

   (a) Floating-point representation of a binary number is:

   $$a = \pm q \times 2^m$$

   where $\pm q$ is a a real number and denoted as **significand** or **mantissa**, m is an integer and denoted as **exponent**.

   (b) IEEE floating-point arithmetic standard:
   Single precision floating-point representation (stored on 32 bits) is:

   $$a = (-1)^s (1.f_1 f_2 \dots f_{23})_2 \times 2^{(m_1 m_2 \dots m_8)_2 - 127}$$

   Double precision floating-point representation (stored on 64 bits) is:

   $$a = (-1)^s (1.f_1 f_2 \dots f_{52})_2 \times 2^{(m_1 m_2 \dots m_{11})_2 - 1023}$$

   A machine number is a real number which can be represented as the normalized floating-point form as above.
   In both representation above, values of m with $(00\dots00)_2$ and $(11\dots11)_2$ are reserved for ± 0 and $\pm\infty$.

   (c) Given a real number x, let $fl(x)$ be the floating point representation of x, which means

   $$\left|\frac{fl(x) - x}{x}\right| \leq 2^{-\beta} := \epsilon_m$$

   where ϵ_m is the machine precision/ machine unit roundoff error. Then we can write

   $$fl(x) = x(1 + \epsilon)$$

   with $|\epsilon| \leq \epsilon_m$.

2. **Solutions of linear systems of algebraic equations**

   (a) p-norm of vector is defined as:

   $$||x||_p = \begin{cases} (|x_1|^p + |x_2|^p \dots + |x_n|^p)^{1/p} & \text{for} \quad 1 \leq p < \infty \\ \max_{1\leq i \leq n} |x_i| & \text{for} \quad p = \infty \end{cases}$$

   And the corresponding matrix norm is $||A||_p := \max\limits_{||x||_p = 1} ||Ax||_p$ for $1 \leq p \leq \infty$, $p \in \mathbb{N}$.

   (b) Sensitivity of linear systems:
   Consider the linear system $Ax = b$, $b \neq 0$ and the pertubed system: $\tilde{A}\tilde{x} = b$. If we write $\tilde{A} = A + E$, then

   $$\frac{\|\tilde{x} - x\|}{\|\tilde{x}\|} \leq \|A^{-1}E\| = \|A^{-1}\tilde{A} - I\|$$

In addition, if $\|A^{-1}E\| < 1$, we have

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|}$$

The real number $\kappa(A)$ given by

$$\kappa(A) = \|A\|\|A^{-1}\|$$

is called the **condition number** of the matrix A. For $\kappa(A)$, we have:
If $\kappa(A) = 10^k$, one should expect to lose at least k digits of accuracy in solving the system $Ax = b$.

2 Exercises:

Please submit solutions of problems with star(*) before 6:30PM on Wednesday and finish the rest by yourself.

1. * Recall that most computers adopt the binary system. Numbers can be decoded as the following normalized floating-point representation:

$$a = (-1)^s q \times 2^{(-1)^p \cdot m}, \tag{1}$$

where $s, p = 0$ or 1, $q = (1.f_1 f_2 \cdots f_h)_2$ and $m = (m_1 m_2 \cdots m_k)_2$.

Remark: in this form of representation, we don't consider reserved values of m for 0 and ∞.

(a) Let $h = 9$, $k = 2$, find the smallest and second smallest positive numbers of the form (1).

(b) Let $h = 4$, $k = 8$, find the largest and second largest numbers of the form (1).

Solution. (a) Put $s = 0$, $p = 1$, $f = (\underbrace{00...00}_{9})_2$ and $m = (11)_2$, the smallest positive number is

$$2^{-3}$$

. Put $s = 0$, $p = 1$, $f = (\underbrace{00...0}_{8}1)_2$ and $m = (11)_2$, the second smallest positive number is:

$$(1.\underbrace{00...0}_{8}1)_2 \times 2^{-3} = (1 + 2^{-9}) \times 2^{-3}.$$

(b) Put $s = 0$, $p = 0$, $f = (1111)_2$ and $m = (\underbrace{11...11}_{8})_2$, the largest number is:

$$(1.1111)_2 \times 2^{2^8 - 1} = (2 - 2^{-4}) \times 2^{255}.$$

Put $s = 0$, $p = 0$, $f = (1110)_2$ and $m = (\underbrace{11...11}_{8})_2$, the second largest number is:

$$(1.1110)_2 \times 2^{255} = (2 - 2^{-3}) \times 2^{255}.$$

□

2. Estimate the approximation errors for the following floating point operations. You can use ϵ to represent the machine precision.

(a) * a^n, where a is a positive machine number and n is a positive integer

(b) $(a + b)(a - b)$

(c) * $(a^2 + b^2 - c)d$

(d) $a^2 b^2 c$

Solution. (a) Note $fl(a^2) \approx a^2(1 + \epsilon)$, then $a^n \to fl((a)fl((a)\cdots))) \approx a(1 + (n-1)\epsilon)$

(b) Note $fl(fl(a)+fl(b)) \approx (a+b)(1+2\epsilon)$, then $(a+b)(a-b) \to fl(fl(fl(a)+fl(b)) \times fl(fl(a)-fl(b))) \approx (a+b)(a-b)(1+5\epsilon)$

(c) Note $fl(fl(a)\times fl(a)) \approx a^2(1+3\epsilon)$, then $(a^2+b^2-c)d \to fl(fl(fl(fl(fl(a)^2)+fl(fl(b)^2))-fl(c))fl(d)) \approx (a^2+b^2)d(1+7\epsilon) - cd(1+4\epsilon)$

(d) $a^2b^2c \to fl(fl(fl(fl(a)fl(a))fl(fl(b)fl(b)))fl(c)) \approx a^2b^2c(1+9\epsilon)$

□

3. * Given an invertible $n \times n$ matrix A. Let $b, b^\delta, x, x^\delta \in \mathbb{R}^n\backslash\{0\}$ be four non-zero vectors such that $Ax = b$ and $Ax^\delta = b^\delta$.

(a) Show that there exists $\kappa(A) > 0$ such that

$$\frac{1}{\kappa(A)}\frac{\|b-b^\delta\|}{\|b^\delta\|} \leq \frac{\|x-x^\delta\|}{\|x^\delta\|} \leq \kappa(A)\frac{\|b-b^\delta\|}{\|b^\delta\|},$$

where $\|\cdot\|$ is a given norm.

(b) Let

$$A = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}.$$

Find $\kappa(A)$ where the norm is

i. 1-norm.

ii. sup-norm

Solution. (a) Let us first recall that

$$\|Ax\| \leq \|A\|\|x\| \quad \forall x \in \mathbb{R}^n.$$

Using this inquality and the fact that

$$b = Ax \quad b^\delta = Ax^\delta,$$

we have

i. $\|b-b^\delta\| \leq \|A\|\|\|x-x^\delta\|$

ii. $\|x^\delta\| \leq \|A^{-1}\|\|b^\delta\|$

iii. $\|x-x^\delta\| \leq \|A^{-1}\|\|b-b^\delta\|$

iv. $\|b^\delta\| \leq \|A\|\|x^\delta\|$

Using (i) and (ii), we have

$$\|b-b^\delta\| \cdot \frac{1}{\|b^\delta\|}\frac{1}{\|A\|\|A^{-1}\|} \leq \|x-x^\delta\| \cdot \frac{1}{\|x^\delta\|}$$

which is the first inequality required.

Using (iii) and (iv), we have

$$\|x-x^\delta\| \cdot \frac{1}{\|x^\delta\|} \leq \|b-b^\delta\| \cdot \frac{1}{\|b^\delta\|}\|A\|\|A^{-1}\|$$

which is the second inequality required.

(b) i. $\kappa(A) = \|A\|_1\|A^{-1}\|_1 = 5 \times 5 = 25$.

ii. $\kappa(A) = \|A\|_\infty\|A^{-1}\|_\infty = 5 \times 5 = 25$.

□

4. * Consider a matrix $C \in \mathbb{R}^{n\times n}$ such that $\|C\| < 1$.

(a) Show that

$$\lim_{n\to\infty} C^n = \mathbf{0},$$

where $\mathbf{0}$ is a zero matrix.

(b) Show that $I - C$ is invertible and

$$(I - C)^{-1} = I + C + C^2 + \cdots$$

Solution.

(a) We have:

$$\|C^n\| = \|C(C^{n-1})\| \leq \|C\|\|C^{n-1}\| \leq \|C\|^2\|C^{n-2}\| \leq \|C\|^n$$

.

Thus,

$$\lim_{n\to\infty} \|C^n\| = 0.$$

Therefore,

$$\lim_{n\to\infty} C^n = 0.$$

(b) A direct computation yields

$$(I - C)(I + C + C^2 + \cdots C^n) = (I + C + C^2 + \cdots C^n) - (C + C^2 \cdots + C^{n+1}) = I - C^{n+1}$$

In view of the results above,

$$I = I - \lim_{n\to\infty} C^n = \lim_{n\to\infty}(I - C^n) = \lim_{n\to\infty}(I - C)(1 + C + C^2 + \cdots C^{n-1}) = (I - C)(I + C + C^2 + \cdots)$$

$\square$