COL726 Homework 1

Lovish Madaan
2015CS50286

1. Computing $Ax = b$ is equivalent to computing $x = A^{-1}b$. So, we'll first calculate A^{-1}. For any matrix, A^{-1} is given by $A^{-1} = \frac{1}{det(A)} adj(A)$. We have,

$$A = \begin{bmatrix} 1 & 2.01 \\ 1.01 & 2.03 \end{bmatrix}$$

$$det(A) = 1 \times 2.03 - 2.01 \times 1.01 = -10^{-4}$$

$$adj(A) = \begin{bmatrix} 2.03 & -2.01 \\ -1.01 & 1 \end{bmatrix}$$

Therefore,

$$A^{-1} = \begin{bmatrix} -20300 & 20100 \\ 10100 & -10000 \end{bmatrix}$$

Now, the function f defined in the question can be interpreted as

$$x = f(b) = A^{-1}b$$

(a) For $b = \begin{bmatrix} 1.01 \\ 1.02 \end{bmatrix}$, $x = f(b) = A^{-1}b = \begin{bmatrix} -20300 & 20100 \\ 10100 & -10000 \end{bmatrix} \begin{bmatrix} 1.01 \\ 1.02 \end{bmatrix} = \begin{bmatrix} -203 \times 101 + 201 \times 102 \\ 101 \times 101 - 100 \times 102 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$$f(b + \Delta b) = A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}\Delta b = f(b) + f(\Delta b)$$

$$f(\Delta b) = \begin{bmatrix} -20300 & 20100 \\ 10100 & -10000 \end{bmatrix} \begin{bmatrix} 0 \\ 0.0001 \end{bmatrix} = \begin{bmatrix} 2.01 \\ -1 \end{bmatrix}$$

Therefore,

$$x + \Delta x = f(b + \Delta b) = f(b) + f(\Delta b) = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2.01 \\ -1 \end{bmatrix} = \begin{bmatrix} 1.01 \\ 0 \end{bmatrix}$$

(b) Since $\kappa = \max_{\Delta b} \frac{||\Delta x||/||\Delta b||}{||x||/||b||}$ for input b and output x, we have for a given Δb and $||.|| = ||.||_\infty$

$$\kappa \geq \frac{||\Delta x||_\infty / ||\Delta b||_\infty}{||x||_\infty / ||b||_\infty}$$

Since $||x||_\infty = \max_i |x_i|$, we have

$$\kappa \geq \frac{2.01/0.0001}{1/1.02}$$

$$\kappa \geq 2.0502 \times 10^4$$

κ is large, therefore the problem instance f is ill-conditioned.

2. We have

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad a, b, c > 0$$

(a) Since b^2 and $4ac$ have the same sign, the calculation $b^2 - 4ac$ is ill conditioned when b^2 and $4ac$ are close to each other. Thus it will lead to loss in accuracy. This situation cannot be avoided in computation of either root. Another situation which can lead to loss in accuracy is in the computation of the numerator in root x_2 since $\sqrt{b^2 - 4ac}$ and b are of the same sign and when they are close to each other, the computation becomes ill conditioned. The second situation of accuracy loss only affects root x_2.

(b) Multiplying the numerator and denominator of roots x_1 and x_2 with $-b + \sqrt{b^2 - 4ac}$ and $-b - \sqrt{b^2 - 4ac}$ respectively, we get

$$x_1 = \frac{(-b - \sqrt{b^2 - 4ac})(-b + \sqrt{b^2 - 4ac})}{2a(-b + \sqrt{b^2 - 4ac})} = \frac{(-b)^2 - (\sqrt{b^2 - 4ac})^2}{2a(-b + \sqrt{b^2 - 4ac})} = \frac{2c}{-b + \sqrt{b^2 - 4ac}}$$

Similarly,

$$x_2 = \frac{2c}{-b - \sqrt{b^2 - 4ac}}$$

Now, the first situation of accuracy loss still affects both roots but the second situation now affects root x_1 instead of x_2 because the signs are reversed. So, the following formulas should be used for the computation of roots.

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{2c}{-b - \sqrt{b^2 - 4ac}}$$

3. 3 kinds of errors will be introduced in the computation of $f(x) = (x-1)^2$: input error, subtraction error and multiplication error. Therefore, we'll have

$$\widetilde{f}(x) = (fl(x) \ominus 1) \otimes (fl(x) \ominus 1)$$

because we can represent 1 exactly. For positive x, we have:

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

$$fl(x) - 1 \leq x + x\epsilon_m - 1$$

$$fl(x) \ominus 1 = (fl(x) - 1)(1 + \epsilon) \leq (fl(x) - 1)(1 + \epsilon_m)$$

Now,

$$\widetilde{f}(x) = (fl(x) \ominus 1) \otimes (fl(x) \ominus 1) = (fl(x) \ominus 1)^2[1 + \epsilon] \leq (fl(x) \ominus 1)^2(1 + \epsilon_m) \leq (x + x\epsilon_m - 1)^2(1 + \epsilon_m)^3$$

Now, we want to collect the coefficients of ϵ_m in $|\widetilde{f}(x) - f(x)|$ which on some expansion can be written as $|(x^2 + x^2\epsilon_m^2 + 1 + 2x^2\epsilon_m - 2x\epsilon_m - 2x)(1 + \epsilon_m^3 + 3\epsilon_m^2 + 3\epsilon_m) - (x-1)^2|$, which are basically the collection of zero order terms from bracket 1 and first order terms from bracket 2 and vice versa w.r.t. ϵ_m. Therefore,

$$c\epsilon_m = |(x^2 + 1 - 2x)(3\epsilon_m) + (2x^2\epsilon_m - 2x\epsilon_m)(1)|$$

$$c = |3(x^2 - 2x + 1) + 2x^2 - 2x| = |5x^2 - 8x + 3|$$

For negative x, we'll have

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

but,

$$|x(1 + \epsilon)| \leq |x(1 + \epsilon_m)|$$

therefore, it will give the same value of c because we will multiply 2 negative values to get a positive value.

4. We'll represent any $m \times n$ matrix X as $[x_1\ x_2\ \cdots\ x_n]$, where x_i is the m dimensional i^{th} column of X.

   (a) We can write $C = AB$ in column interpretation as follows:

   $$c_j = Ab_j \Rightarrow c_j = \sum_{k=1}^{n} b_{kj} a_k$$

   where c_j and b_j are the j^{th} columns of C and B respectively and a_k is the k^{th} column of A. So we can say that each column of C is a linear combination of columns of A. So we have

   $$c_j \subset colspace(A)\ \forall\ j = 1, 2, \cdots, p$$

   which implies

   $$colspace(C) \subset colspace(A) \Rightarrow rank(C) \leq rank(A) \leq \min(m, n) = n$$

   (b) Since A is full rank, using Trefethen & Bau's Theorem 1.2, we have that A does not map 2 distinct vectors to the same vector. We have that $A\vec{0} = \vec{0}$, so no other vector other than $\vec{0}$ will be mapped to $\vec{0}$. Hence, all nonzero vectors will be mapped to nonzero vectors.
   Now, suppose we have n linearly independent vectors $v_1, v_2, \cdots, v_n$. We will prove by contradiction that A will map linearly independent vectors to linearly independent vectors. Assuming the mapped vectors are not linearly independent, there exists a vector v_j such that,

   $$Av_j = \sum_{i=1}^{j-1} k_i A v_i + \sum_{i=j+1}^{n} k_i A v_i \quad \text{where } k_i \text{ is a scalar}$$

   This can be interpreted as,

   $$Av_j = \sum_{i=1}^{j-1} A k_i v_i + \sum_{i=j+1}^{n} A k_i v_i$$

   $$Av_j = A(\sum_{i=1}^{j-1} k_i v_i + \sum_{i=j+1}^{n} k_i v_i)$$

   Since A is full rank,

   $$v_j = \sum_{i=1}^{j-1} k_i v_i + \sum_{i=j+1}^{n} k_i v_i$$

   which is a contradiction to the fact that vectors $v_1, v_2, \cdots, v_n$ are linearly independent.
   Hence A maps linearly independent vectors to linearly independent vectors.

   (c) Since A & B are full rank, $rank(A) = min(m, n) = n$ and $rank(B) = min(n, p) = n$. Equivalently, n columns of B out of p are linearly independent. Now, since A is full rank and we can write $C = AB$ as

   $$c_j = Ab_j$$

   i.e. A maps j^{th} column of B to j^{th} column of C, we have that the n linearly independent columns of B will be mapped to n linearly independent columns of C. Therefore, atleast n out of p columns of C are linearly independent. Hence, $rank(C) \geq n$. Using the result from part (a) that $rank(C) \leq n$, we have that $rank(C) = n$.

5. Given 2 vectors x and y, the triangle inequality states that $||x+y|| \leq ||x|| + ||y||$.

   (a) We can write $||x||$ as $||(x+y)-y||$. Using triangle inequality, we have

   $$||x|| = ||(x+y)-y|| \leq ||x+y|| + ||-y|| = ||x+y|| + ||y|| \tag{1}$$

   Similarly,

   $$||y|| \leq ||x+y|| + ||x|| \tag{2}$$

   Using (1) and (2),

   $$||x+y|| \geq ||x|| - ||y|| \quad \text{and} \quad ||x+y|| \geq ||y|| - ||x||$$

   $$||x+y|| \geq \max\{||x|| - ||y||, ||y|| - ||x||\} = \Big|\,||x|| - ||y||\,\Big|$$

   (b) Suppose $I + \epsilon A$ is invertible. Then, $I + \epsilon A$ will not map 2 distinct vectors to the same vector(using Trefethen & Bau's Theorem 1.2). Since it already maps $\vec{0}$ to 0, for $x \neq \vec{0}$, we have $(I+\epsilon A)x \neq \vec{0}$, which further implies

   $$||(I+\epsilon A)x|| > 0 \quad \text{for} \quad x \neq \vec{0}$$

   Now, using the inequality derived in part (a) and assuming $\epsilon > 0$,

   $$||(I+\epsilon A)x|| = ||x + \epsilon Ax|| \geq \Big|\,||x|| - ||\epsilon Ax||\,\Big| \geq ||x|| - ||\epsilon Ax|| = ||x|| - \epsilon||Ax||$$

   Now, we know that $||Ax|| \leq ||A||\,||x||$, therefore,

   $$||(I+\epsilon A)x|| \geq ||x|| - \epsilon||Ax|| \geq ||x|| - \epsilon||A||\,||x|| = (1 - \epsilon||A||)||x|| \tag{3}$$

   We want $||(I+\epsilon A)x||$ to be strictly greater than 0 and since $x \neq \vec{0}$, we have $||x|| > 0$. From equation (3), this can be ensured by keeping the value of ϵ satisfying the following inequality

   $$1 - \epsilon||A|| > 0$$

   which implies,

   $$\epsilon < \frac{1}{||A||} = \delta$$

6. **Disclaimer**: Used wikipedia as reference for kahaan summation. Kahaan summation works because it does not accumulate errors in the whole summation by keeping a different variable for accumulating the errors and then subtracting.
   I have taken the input array as

   $$data1 = 10^7 + 0.001 * [1\ 2\ \cdots\ 10]$$

   Now,

   $$var(data1) = var(0.001 * [1\ 2\ \cdots\ 10]) = 10^{-6} var([1\ 2\ \cdots\ 10]) = 8.2500 \times 10^{-6}$$

   Using the 2 different methods, we have

   $$var1(data1) = 0.0$$

   $$var2(data1) = 8.2499 \times 10^{-6}$$

   Now,

   $$data2 = [10^{-10}\ 10^{-10}\ \cdots\ 10^{-10}]$$

   $$var(data2) = 0.0$$

   $$var2(data2) = 1.6704 \times 10^{-52}$$

   $$var3(data2) = 0.0$$