MTH 373/573:
Scientific Computing

Solutions to Quiz 1

IIIT-Delhi, Monsoon 2022
Kaushik Kalyanaraman

Problem 1: Problem and Algorithm

Which of the following statements is *False*?

A. A stable algorithm produces a result which is relatively insensitive to perturbations due to approximations made during a computation.

B. A stable algorithm will always produce an accurate solution to an underlying problem.

C. The computed solution to a well conditioned problem is relatively insensitive to perturbations in the input.

D. Computation of an irrational number like π is always ill conditioned because of finite precision in the IEEE double precision floating point number system.

Option A is *True* because that is a definition of stability of the algorithm – the computational errors are minimum. Option C is likewise *True* because that is the definition of a well conditioned problem.

Option B is *False* without the qualification of the underlying problem being well conditioned while Option D is *False* because it is almost word salad – conditioning of a smooth problem has nothing to do with floating point representation which is a feature of the discrete problem.

Problem 2: Properties of Floating Point

Which of the following statements is *True* for floating point multiplication and addition?

(*Hint:* Consider sources of significant error in floating point arithmetic such as cancellation, overflow, and underflow.)

A. Floating point addition is associative.

B. Floating point addition is commutative.

C. Floating point multiplication is associative.

D. Floating point multiplication is commutative.

Counterexamples to floating point addition not being associative are $\left(1+\frac{\varepsilon_M}{2}\right)+\frac{\varepsilon_M}{2} \neq 1+\left(\frac{\varepsilon_M}{2}\right)+\frac{\varepsilon_M}{2}$ and $\left(10^{308}+10^{308}\right)-9\times10^{307} \neq 10^{308}+\left(10^{308}-9\times10^{307}\right)$. Similar examples can be concocted for the non associativity of floating point multiplication.

Problem 3: Subnormal Numbers

Which of the following is not a valid reason for allowing subnormal numbers in a floating-point system?

A. To reduce the unit roundoff ε_{M}

B. To reduce the underflow level UFL

C. To fill in the gap around zero due to normalization

D. To extend the range of representable numbers

The ε_{M} has to do with the precision in the representation of a floating point number whilst subnormal numbers have to do with what values are allowed in the mantissa and the exponent below the normalization.

Problem 4: Linear Algebra

Which of the following statements about a matrix $A \in \mathbb{R}^{n\times n}$ is/are *not* mathematically equivalent to the others? (Recall that statements p and q are equivalent if $p \implies q$ and $q \implies p$.)

A. $\det(A) \neq 0$.

B. There is no vector $x \neq 0$ such that $A\,x = 0$.

C. The column rank of A and row rank of A are equal.

D. There exists a unique $n \times n$ matrix B such that $A\,B = B\,A = \mathbb{I}$.

The column rank of a matrix being equal to its row rank is *always* true. To make it equivalent to the matrix being nonsingular, this rank should also be *full*, that is equal to the dimension of the space in which the image of the matrix is a subspace of.

Problem 5: Catastrophic Cancellation

For a quadratic equation $ax^2 + bx + c$, one of the roots can be evaluated using either:

$$x = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \qquad \text{or} \qquad x = \frac{2c}{-b + \sqrt{b^2 - 4ac}}.$$

In this context, which of the following statements are *True*?

A. For the equation $0.0501x^2 - 100x + 5.015$, it is better to use the first/left expression to compute the root in finite precision.

B. In infinite precision, both expressions are identical.

C. In finite precision, both expressions are identical.

D. For the equation $x^2 + x + 10^{-10}$, it is better to use the first/left expression to compute the root in finite precision.

Option A is *False* because $b^2 - 4ac = (-100)^2 - 4\times 0.0501\times 5.015 \approx 100^2 - 4\times 0.05\times 5 = 100^2 - 1 \approx 100^2$. Hence, for this quadratic, $-b$ and $\sqrt{b^2 - 4ac}$ are nearly of the same magnitude but different signs leading to cancellation. Option C is *False* precisely because of floating point idiosyncrasies like cancellation, overflow or underflow.

In infinite precision, of course, both expressions for the root of the quadratic are equivalent and for the quadratic in Option D, there will $b^2 - 4ac = 1^2 - 4\times 1\times 10^{-10} \approx 1$. Hence, $-b - \sqrt{b^2 - 4ac} \approx -2$, leading to the better estimation of this root than with the second expression which will incur cancellation.

Problem 6: System of Linear Equations

Which of the following statements is *never True* for a system of linear equations?

A. There can be no solution.

B. There can be exactly one solution.

C. There can be exactly two solutions.

D. There can be infinitely many solutions.

Problem 7: Matrix 1- and ∞-norms

Which of the following properties is/are *True* for the vector induced matrix 1-norm and ∞-norm? (*Hint:* Observe that the maximizing vector in $\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty$ can always be chosen to have $|x_i| = 1$ for all i. Similarly for $\|A\|_1 = \max_{\|x\|_1 = 1} \|Ax\|_1$, the maximizing vector can be chosen to satisfy $|x_i| = 1$ for one i and $x_j = 0$ for all other $j \neq i$.)

A. $\|A^{-1}\|_1 = \|A\|_\infty$

B. $\|A\|_1$ is the maximum 1-norm of any column of A

C. $\|A\|_\infty$ is the maximum 1-norm of any row of A

D. $\|A^T\|_1 = \|A\|_\infty$

Problem 8: Ill Conditioning

Which of the following nonsingular matrices is/are *ill conditioned*?

A. $\begin{bmatrix} 10^{20} & 1 \\ 1 & 10^{20} \end{bmatrix}$ B. $\begin{bmatrix} 10^{20} & 0 \\ 0 & 10^{-20} \end{bmatrix}$ C. $\begin{bmatrix} 10^{-20} & 1 \\ 1 & 10^{-20} \end{bmatrix}$ D. $\begin{bmatrix} 1.0000 & 2.0001 \\ 2.0000 & 4.0001 \end{bmatrix}$

There are many similar ways to argue this. For instance, $10^{20} \gg 1$, $1 \gg 10^{-20}$ and $\begin{bmatrix}1 & 2\end{bmatrix}^T \approx 2\begin{bmatrix}1 & 2\end{bmatrix}^T$ without referencing floating point representations directly.

However, notice also that in IEEE double precision floating point system, $\mathrm{fl}(1 + 10^{-20}) = 1$. Hence, the matrix in Option A is nearly the diagonal matrix with the same diagonal entry of 10^{20} while the matrix in Option C is nearly the identity matrix. Finally, the matrix in Option D is an approximate floating point representation of an exactly singular matrix, hence it will have a large condition number.

Problem 9: Perturbed Linear System

You just solved a linear system $Ax = b$. Unfortunately, the right hand side b that you solved it with was off by $\|\Delta b\|$. Worried, you compute $\frac{\|\Delta b\|}{\|b\|}$ and determine it to be $\approx 10^{-12}$. The condition number of your matrix is about 10^6. In this context, which of the following statements is/are *True*?

A. The expected worst-case relative error in the solution x due to the incorrect right hand side vector b is about 10^{-6}.

B. In IEEE double precision floating point system, this incorrect solution will still have about 10 to 11 correct digits.

C. Nothing can be stated about the correctness of the computed solution without knowing the algorithm used for its computation.

D. The linear system cannot have a unique solution since the condition number of A implies that it is a nearly singular matrix.

Option B "appears" to be *True* but is not because an error of 10^{-6} means that there are about 6 correct digits and not 6 digits off the precision.

Problem 10: Condition Number

What is the condition number of the matrix:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

A. 2 B. 3 [✓ C. 6] D. 3/2

Problem 11: Operator Conditioning

Consider the problem of evaluating $f(x)$ given any $x > 1$. For which of the following functions is the problem *ill-posed* using the more restrictive definition that a well-posed problem must have a finite condition number?

[✓ A. $f(x) = 3 - x$] B. $f(x) = \sqrt{x}$ C. $f(x) = 3x$ D. $f(x) = 3/x$

Using that the condition number for function evaluation is $\left|\frac{xf'(x)}{f(x)}\right|$, we have the condition numbers for the various choices to be:

A. $\kappa(x) = \left|\frac{x}{x-3}\right|$ B. $\kappa(x) = \frac{1}{2}$ C. $\kappa(x) = 1$ D. $\kappa(x) = 1$

Clearly, the functions in Options B, C and D are well posed because not only are their condition numbers finite but also constant. For $f(x) = 3 - x$, $\lim_{x\to\infty} \kappa(x) = 1$ and indeed this problem is well conditioned for large x. However, since the problem states *for any* $x > 1$, at $x = 3$, the condition number for this problem is ∞. This agrees with our understanding that subtraction is ill conditioned for two numbers of the same magnitude.

Problem 12: Absolute Backward Error

Consider the infinite series:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots,$$

which is valid for $|x| < 1$. If we truncate the series after the second term, we obtain the approximation:

$$\frac{1}{1-x} \approx 1 + x.$$

What is the magnitude of the *absolute backward error* in using this approximation to compute $1/(1-x)$ for $x = 1/2$?

A. 1/3 [✓ B. 1/6] C. 1/2 D. 3/2

Using the definitions and our notation, we have that $f(x) = 1/(1-x)$ and $\hat{f}(x) = 1 + x$. Given that $x = 1/2$, to compute the backward error, we first need to find $\hat{x}$ such that:

$$f(\hat{x}) = \hat{f}(x) \implies \hat{f}(x = 1/2) = 1{+}1/2 = 3/2 \implies f(\hat{x}) = 3/2 \implies 1/(1{-}\hat{x}) = 3/2 \implies \hat{x} = 1/3.$$

Thus, the absolute backward error is $|x - \hat{x}| = |1/2 - 1/3| = 1/6$.

Problem 13: Elementary Row Operations 1.5 points

Which of the following matrices *cannot* be obtained from the following matrix using only elementary row operations?

$$\begin{bmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 3 & 4 \\ 1 & 2 & 0 & 4 \end{bmatrix}$$

A. $\begin{bmatrix} 2 & 4 & 0 & 8 \\ 0 & 1 & 3 & 4 \\ 2 & 1 & 3 & 1 \end{bmatrix}$ B. $\begin{bmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 3 & 4 \\ 1 & 3 & 3 & 8 \end{bmatrix}$ C. $\begin{bmatrix} 1 & 2 & 3 & 1 \\ 1 & 0 & 3 & 4 \\ 2 & 1 & 0 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 1 & 3 & 4 \\ 1 & 2 & 0 & 4 \\ 2 & 1 & 3 & 1 \end{bmatrix}$

Option (A) results from row $3 \longleftarrow 2 \times \text{row } 3$ followed by swapping rows 1 and 3. Option (B) results from row $3 \longleftarrow$ row $2 +$ row 3. Option (C) involves a *column* swap, namely, columns 1 and 2. Option (D) is row $1 \longleftrightarrow$ row 2 followed by row $2 \longleftrightarrow$ row 3.

Solution Ready Reference

Question	Answer Choice(s)
1	B, D
2	B, D
3	A
4	C
5	B, D
6	C
7	B, C, D
8	B, D
9	A
10	D
11	A
12	B
13	C