DS284: Numerical Linear Algebra
Assignment 2

Instructor: Dr. Phani Motamarri
TA: Karan Jeswani
Office Hours: Sat 2 PM to 4 PM

Total: 100 Marks
Posted on Oct 28, 2020 and due on Nov 9, 2020

Problem 1 [10 marks]

Let $\|.\|$ be any vector induced matrix norm and k a positive integer. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ Prove that

$$[\rho(\mathbf{A})]^k = \rho(\mathbf{A}^k) \leq \|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k$$

And therefore, $\rho(\mathbf{A}) \leq (\|\mathbf{A}^k\|)^{1/k} \leq \|\mathbf{A}\|$

Here ρ denotes spectral radius i.e. largest absolute eigen value $|\lambda|$ of $\mathbf{A}$.

Problem 2 [20 marks]

If $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$, then show the following.

(a) $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$

(b) $\|\mathbf{x}\|_2 \leq \sqrt{m}\|\mathbf{x}\|_\infty$

(c) $\|\mathbf{A}\|_\infty \leq \sqrt{n}\|\mathbf{A}\|_2$

(d) $\|\mathbf{A}\|_2 \leq \sqrt{m}\|\mathbf{A}\|_\infty$

Give an example of a non-zero vector or matrix for which equality is achieved in the above inequalities.

Problem 3 [15 marks]

For matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, show that

(a) $\|\mathbf{A}\|_F = \sqrt{tr(\mathbf{A}^T\mathbf{A})}$

(b) $\frac{1}{\sqrt{m}}\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_2 \leq \sqrt{n}\|\mathbf{A}\|_1$

(c) $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty}$

Problem 4 [10 marks]

Induced matrix norm is defined as $\|\mathbf{A}\|^{(m,n)} = \max_{\mathbf{x}} \|\mathbf{Ax}\|^{(m)}$, where $\mathbf{x} \in \mathbb{R}^n$ and is a unit vector. $\|.\|$ corresponds to p-norm ($1 \leq p < \infty$). For this exercise, let us consider p to be a natural number.

Using MATLAB or Octave programming environment, create a matrix using "$\mathbf{A} = randn(100, 2)$". Subsequently, create random unit vectors $\mathbf{x}$ using "$temp = randn(2, 1)$" and normalize $\mathbf{x}$ using "$\mathbf{x} = temp/norm(temp)$". Check for multiple random vectors $\mathbf{x}$ (use a loop, and check for about 1000 random vectors $\mathbf{x}$) using "$norm\_of\_Ax = norm(\mathbf{Ax}, p)$" for $p = 1, 2, 3, 4, 5, 6, \infty$. What is the maximum value of p-norm for the vector $\mathbf{Ax}$? Now calculate p-norm of $\mathbf{A}$ using "$norm\_of\_A = norm(\mathbf{A}, p)$" for $p = 1, 2, \infty$. Verify the equality $\|\mathbf{A}\|^{(m,n)} = \sup\|\mathbf{Ax}\|^{(m)}$ for $p = 1, 2, \infty$. Note that this equality is true for other values of p as well but you are restricting to $p = 1, 2, \infty$ in this exercise.

Problem 5 [20 marks]

Recall in IEEE single precision binary floating point representation, we use 32 bits to represent numbers 1 bit is for sign, 8 bits for the exponent and 23 bits for the mantissa. Using normalized binary scientific notation a floating point number in IEEE single precision can be represented as

$$(-1)^s \times (1.f)_2 \times 2^{(exponent-127)}$$

Here $s = 0$ for positive numbers and $s = 1$ for negative numbers. f represents the bits in mantissa. Note the digit 1 in $1.f$ and is explicitly shown for clarity and all binary representations are normalized to take the form $1.f$. The subscript 2 in the above $1.f$ denotes that we are representing the digits in base 2.

Now, in this exercise we will construct a dummy floating point number system where we use 5 bits of precision to represent numbers. In this simplified floating point number system, let us assume that we are representing numbers such that the exponent field admits values -1, 0 and 1 only. Imagine only positive numbers are represented in this system. Then the normalized binary scientific notation in this toy system would be

$$(1.f)_2 \times 2^{(exponent-1)}$$

Note here we use a biased representation in the exponent field instead of using a separate sign bit for exponent. The value of this bias is 1. Recall our toy floating point system admits -1, 0, 1 in the exponent field. Thus a value of -1 in the exponent field means $exponent = 0$, value of 0 in exponent field means $exponent = 1$, value of 1 in exponent field means $exponent = 2$. In this normalized binary scientific notation, 3 bits are used to store f, 2 bits are used to store the $exponent$.

In this backdrop answer the following questions for the toy floating point system we constructed above:

(a) How many numbers can this toy system describe?

(b) Create a table with 3 columns. First column should contain normalized binary scientific notation of the form

$$(1.f)_2 \times 2^{(exponent-1)}$$

of all the above numbers. (Make sure the numbers you are representing here just use 5 bits to store them). Second column should contain the usual binary representation. Third column should contain decimal representation (base 10 representation). Arrange the numbers in increasing order in the base 10 representation.

(c) From the table above, what is the minimum real number and maximum real number you can represent using our toy floating point number system.

(d) What can you say about absolute gaps between the numbers? Are they constant or do they change with the magnitude of the number you are representing?

(e) What can you say about machine epsilon for our toy floating point system?

(Hint: Pick $\mathbf{x} \in \mathbb{R}$, there exists $\mathbf{x}' \in \mathbb{F}$ such that $\frac{|\mathbf{x}-\mathbf{x}'|}{|\mathbf{x}|} \leq \epsilon_{machine}$)

Problem 6 [10 marks]

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be invertible matrix and consider the solution of the problem $\mathbf{Ax} = \mathbf{b}$ for some given non-zero $\mathbf{b} \in \mathbb{R}^n$.

(a) Derive the relative condition number of the problem of computing $\mathbf{x}$ given $\mathbf{b}$ with respect to perturbations in $\mathbf{b}$.

(b) Find the value of the tight lower bound of the relative condition number obtained in (a).

Problem 7 [15 marks]

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be invertible matrix and suppose that $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$ for some given non-zero $\mathbf{b}$. Consider perturbations $\Delta\mathbf{A} \in \mathbb{R}^{n\times n}$ of $\mathbf{A}$ satisfying the following in some given matrix norm induced by the vector norm $\|.\|$,

$$K(\mathbf{A})\|\Delta\mathbf{A}\| < \|\mathbf{A}\|$$

where $K(\mathbf{A})$ is the condition number of $\mathbf{A}$ in the given norm. Consider also some perturbation $\Delta\mathbf{b} \in \mathbb{R}^n$ of $\mathbf{b}$ and let $\mathbf{x} + \Delta\mathbf{x}$ solve

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$$

Prove that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left[\frac{K(\mathbf{A})}{1 - K(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}}\left[\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right]\right]$$