18.330: Introduction to Numerical Analysis

Problem Set #1: Fundamentals of Numerical Analysis

due September 26^{th} @ 9:30 am

Instructions: There are eight problems listed below, six are theoretical and two are computational, denoted by an (M) and (C) respectively. You may complete as many problems as you want, but please **submit only four problems** for grading. Particularly challenging problems have been marked with a star next to them. For computational problems, please include a copy of your source code and output. You may collaborate with classmates and reference outside material, but you must write your own solutions and note your collaborators and sources. You may either submit your problem set solutions via Canvas, or hand them in at the beginning of class. **Late submissions will not be accepted** – to be granted an exception (due to illness or extenuating circumstance) please contact Student Support Services and have them contact me before the assignment deadline.

1 Accumulating Errors (M)

To simplify expressions in the analysis of floating point arithmetic, the following result is quite useful:

Claim. *If* $|\delta_i| \leq u$, $\rho_i = \pm 1$ *for* $i = 1, ..., n$, *and* $nu < 1$, *then*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \theta_n, \qquad \textit{where } |\theta_n| \leq \frac{nu}{1-nu}.$$

Please provide a proof of this fact.

2 Relative Error for Sample Variance (M)*

The sample variance $V(x) : \mathbb{R}^n \to \mathbb{R}$ of a set of n numbers can be computed in a number of different ways. One popular technique is the two-pass formula

$$V(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \text{where} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

Prove that the computed sample variance $\widehat{V}(x) = \text{fl}(V(x))$ using the above two-pass formula satisfies

$$\frac{|V - \widehat{V}|}{V} \leq (n+3)u + O(u^2),$$

where u is the unit roundoff.

3 Norms and Matrices (M)

A matrix norm $\|\cdot\|$ is said to be *natural* if $\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. Prove the following:

Claim. *Let $A \in \mathbb{R}^{n\times n}$. If $\|A\| < 1$ for some natural norm $\|\cdot\|$, then $I - A$ is non-singular and*

$$\frac{1}{1+\|A\|} \leq \|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

4 Avoiding Loss of Significance (M)

Each of the below expressions suffers from loss of significance due to cancellation. Show how to evaluate the following expressions in a numerically stable fashion using an alternate representation:

(a)	$\frac{1}{1+2x} - \frac{1-x}{1+x}$	for $\quad \lvert x\rvert \ll 1$
(b)	$\sqrt{x+\frac{1}{x}} - \sqrt{x-\frac{1}{x}}$	for $\quad x \gg 1$
(c)	$\frac{1-\cos x}{x}$	for $\quad \lvert x\rvert \ll 1,\ x \neq 0$
(d)	$\sin x - \sin y$	for $\quad x \approx y$
(e)	$(a^2+b^2-2ab\cos\theta)^{1/2}$	for $\quad a \approx b,\ \lvert\theta\rvert \ll 1$

5 Condition Numbers for Sample Variance (M)*

A condition number for the sample variance $V(x) : \mathbb{R}^n \to \mathbb{R}$ (see Equation 1 above) can be defined by

$$\kappa_C := \lim_{\epsilon\to 0} \sup \left\{ \frac{|V(x+\Delta x) - V(x)|}{\epsilon V(x)} \,\middle|\, |\Delta x_i| \leq \epsilon |x_i|,\ i = 1, ..., n \right\}.$$

Show that

$$\kappa_C = 2\frac{\sum_{i=1}^n |x_i - \bar{x}||x_i|}{(n-1)V(x)}.$$

κ_C measures perturbations in x componentwise. An analogous condition number for the 2-norm is given by

$$\kappa_N := \lim_{\epsilon\to 0} \sup \left\{ \frac{|V(x+\Delta x) - V(x)|}{\epsilon V(x)} \,\middle|\, \|\Delta x\|_2 \leq \epsilon \|x\|_2 \right\}.$$

Show that

$$\kappa_N = 2\frac{\|x\|_2}{\sqrt{(n-1)V(x)}} = 2\left(1 + \frac{n}{n-1}\frac{\bar{x}^2}{V(x)}\right)^{1/2}.$$

Which condition number is bigger?

6 IEEE Rounding in Double Precision (E)

(a) Do the following sums by hand in IEEE double precision computer arithmetic, using the Rounding to Nearest Rule (e.g., the rounding rule introduced in class):

$$(1) \qquad \left(1 + (2^{-51} + 2^{-53})\right) - 1$$
$$(2) \qquad \left(1 + (2^{-51} + 2^{-52} + 2^{-53})\right) - 1$$
$$(3) \qquad \left(1 + (2^{-51} + 2^{-52} + 2^{-54})\right) - 1$$
$$(4) \qquad \left(1 + (2^{-51} + 2^{-52} + 2^{-60})\right) - 1$$

Verify your answers using a computer.

(b) Is $1/3 + 2/3$ exactly equal to 1 in double precision floating point arithmetic using the IEEE Rounding to Nearest Rule? Explain why. Verify your answers using a computer. What would the result be if chopping (simply removing extra digits without rounding) was used instead?

7 Cancellation of Roundoff Error (E)

Consider the quantity $(e^x - 1)/x = \sum_{i=0}^{\infty} x^i/(i+1)!$. Compute $(e^x - 1)/x$ for $x = 10^{-k}$, $k = 5, 6, ..., 16$, using the following two different techniques:

- Technique 1: evaluate $(e^x - 1)/x$ directly
- Technique 2: compute $y = e^x$, then evaluate $(y - 1)/\log y$

Report the output using each technique. Which technique gives the more accurate answer, and why is that the case? Please justify your answer with some error analysis.

8 The Matrix Infinity Norm (M)

Let A be an $n \times n$ matrix, and consider the natural norm $\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$. In class, we saw that

$$\|A\|_\infty = \max_{i=1,...,n} \sum_{j=1}^{n} |A_{i,j}|.$$

Please evaluate the infinity norm of the following matrices and, in each case, give a vector that achieves the norm (e.g., find an x such that $\|Ax\|_\infty = \|A\|_\infty \|x\|_\infty$):

$$(a) \qquad A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$
$$(b) \qquad A = \begin{pmatrix} 1 & 5 & 1 \\ -1 & 2 & -3 \\ 1 & -7 & 0 \end{pmatrix}$$

Student Feedback

Please let me know how you're finding the course and the first problem set. What are you hoping to get out of the class? How is the pace of lecture? Please rate the difficulty and volume of the first problem set.