Name: SOLUTIONS

Math 614 Numerical Linear Algebra (Bueler) | Wednesday 15 November 2023

Midterm Quiz 2

In-class or proctored. No book, notes, electronics, calculator, internet access, or communication with other people. 100 points possible.
65 minutes maximum!

1. (*10 pts*) Suppose $f : X \to Y$ is a problem, where X and Y are normed vector spaces. For $x \in X$, define the *relative condition number* of the problem:

$$\kappa(x) = \lim_{\|\delta x\| \to 0} \sup_{\delta x \neq 0} \frac{\|f(x+\delta x) - f(x)\|_Y / \|f(x)\|_Y}{\|\delta x\|_X / \|x\|_X}$$

2. (*15 pts*) Our textbook Trefethen & Bau defines an idealized floating point system $\boldsymbol{F}$, also written $\mathbb{F}$. Define/describe it. (*Hints.* A floating point system is scientific notation based on a base β and a precision t. Both β and t are integers; what are their ranges? Describe the allowed fractions and exponents, and where they appear.)

$$\mathbb{F} = \{0\} \cup \left\{ \pm \frac{m}{\beta^t} \beta^e \right\}$$

$\beta \geq 2$ integer

$t \geq 1$ integer

e any integer

example: (to check)

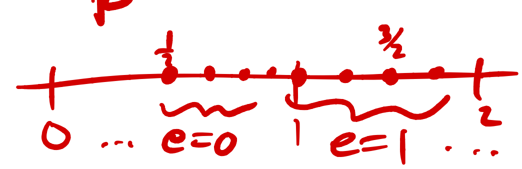
$\beta = 2$

$t = 3$

$\beta^{t-1} \leq m \leq \beta^t - 1$

$4 \leq m \leq 7$

$\frac{1}{2} \leq \frac{m}{\beta^t} < 1$



3. **(a)** (*5 pts*) State axiom (13.5).

if $x \in \mathbb{R}$ then $fl(x) \in \mathbb{F}$ satisfies

$fl(x) = x(1+\varepsilon)$ for some $\varepsilon \in \mathbb{R}$

so that $|\varepsilon| \leq \varepsilon_{machine}$

(b) (*5 pts*) State axiom (13.7).

if $x, y \in \mathbb{R}$ and $* = +, -, \times, \div$ then

$x \circledast y = (x * y)(1+\varepsilon)$ for some $\varepsilon \in \mathbb{R}$

so that $|\varepsilon| \leq \varepsilon_{machine}$

4. (*15 pts*) Suppose $f : X \to Y$ is a problem and $\tilde{f} : X \to Y$ is an algorithm to compute (approximate) that problem on a computer satisfying axioms (13.5) and (13.7). Define what it means for the algorithm $\tilde{f}$ to be *backward stable* for the input $x \in X$.

$\tilde{f}(x) = f(\tilde{x})$ for some $\tilde{x}$ so that

$$\frac{\|\tilde{x} - x\|_X}{\|x\|_X} = O(\varepsilon_{machine})$$

("exactly the right answer for nearly the right question")

5. (*7 pts*) Show that $(1+O(t))(1+O(t)) = 1+O(t)$ as $t \to 0$.

pf: Given $f_i(t)$ so that $|f_i(t)| \le c_i |t|$, $i=1,2$ for t sufficiently close to zero, we are to show

$$\left|(1+f_1(t))(1+f_2(t)) - 1\right| \le C|t|.$$

But

$$\begin{aligned}|(1+f_1(t))(1+f_2(t)) - 1| &= |\cancel{1} + f_1(t) + f_2(t) + f_1(t)f_2(t) \cancel{-1}| \\ &\le |f_1(t)| + |f_2(t)| + |f_1(t)|\,|f_2(t)| \\ &\le c_1|t| + c_2|t| + c_1|t|c_2|t| \quad \text{assume } |t|\le 1 \\ &\le (c_1+c_2+c_1c_2)|t|. \quad \text{Let } C = c_1+c_2+c_1c_2. \ \blacksquare\end{aligned}$$

6. (*7 pts*) Consider the problem (function) $f(x) = x^4$ on real numbers. Compute the absolute condition number $\hat{\kappa}(x)$ and the relative condition number $\kappa(x)$.

$$J(x) = [4x^3] \quad \text{1×1 matrix :)}$$

$$\hat{\kappa}(x) = |J(x)| = 4|x|^3$$

$$\kappa(x) = \frac{|J(x)|}{|f(x)|/|x|} = \frac{4|x|^3 \cdot |x|}{|x|^4} = 4$$

7. (*8 pts*) Suppose $A \in \mathbb{C}^{m\times m}$ is invertible, and that $b \in \mathbb{C}^m$. Explain, via major steps, how to use the QR factorization to solve the linear system $Ax = b$. How much work,[1] i.e. how many floating point operations, is required for each step?

① $A = QR$ (by Householder or GS)

$O(m^3)$ work

② $y = Q^* b$ (by mat-vec)

$O(m^2)$ work

③ solve $Rx = y$ (by back-substitution)

$O(m^2)$ work

notes:

$Ax = b$

$Q(Rx) = b$

8. (*8 pts*) Suppose $A \in \mathbb{C}^{m\times n}$ is full rank, and that $m \geq n$. Suppose $b \in \mathbb{C}^m$. Explain, via major steps, how to use the reduced SVD factorization to solve the overdetermined system "$Ax = b$" by least squares. How much work is required for each step?

① $A = \hat{U}\hat{\Sigma}V^*$ (by $\boxed{?}$)

$O(mn^2)$ $\boxed{?}$ work

we will get to Lecture 31

② $y = \hat{U}^* b$ (by mat-vec)

$O(mn)$ work

③ $z = \hat{\Sigma}^{-1} y$ (n scalar divisions)

$O(n)$ work

④ $x = Vz$ (by mat-vec)

$O(n^2)$ work

notes:

"$Ax = b$"

$Ax = Pb$

$\cancel{\hat{U}}\hat{\Sigma}(V^* x) = \cancel{\hat{U}}(\hat{U}^* b)$

[1] For problems **7** and **8**, use big-O notation to communicate the amount of work at leading order in m and/or n, as they go to infinity. You do not need to prove your big-O usage.

9. Suppose $x \in \mathbb{R}^2$ and that $f(x) = x_1^2 + x_2^2$.

(a) (*4 pts*) Write the obvious floating point algorithm for computing f, using notation $\mathrm{f\ell}(\cdot)$, $\oplus$, $\otimes$:

$$\tilde{f}(x) = fl(x_1) \otimes fl(x_1) \oplus fl(x_2) \otimes fl(x_2)$$

(b) (*8 pts*) Assuming a computer satisfying axioms (13.5) and (13.7), show that the above algorithm is backward stable. You may assume here, without proof, that $(1 + O(t))(1 + O(t)) = 1 + O(t)$ and $\sqrt{1 + O(t)} = 1 + O(t)$ as $t \to 0$.

Pf:

$$\tilde{f}(x) \overset{(13.5)}{=} x_1(1+\varepsilon_1) \otimes x_1(1+\varepsilon_1) \oplus x_2(1+\varepsilon_2) \otimes x_2(1+\varepsilon_2)$$

$$\overset{(13.7)}{=} x_1^2(1+\varepsilon_1)^2(1+\varepsilon_3)(1+\varepsilon_5) + x_2^2(1+\varepsilon_2)^2(1+\varepsilon_4)(1+\varepsilon_5)$$

Let $\tilde{x}_1 = (1+\varepsilon_1)\sqrt{1+\varepsilon_3}\sqrt{1+\varepsilon_5}$,

$\tilde{x}_2 = (1+\varepsilon_2)\sqrt{1+\varepsilon_4}\sqrt{1+\varepsilon_5}$.

Then $\tilde{f}(x) = f(\tilde{x})$. And:

$$\frac{\|\tilde{x} - x\|_1}{\|x\|_1} = \frac{|x_1(1+\varepsilon_1)(1+O(\varepsilon_m))(1+O(\varepsilon_m)) - x_1| + |x_2(1+\varepsilon_2)(1+O(\varepsilon_m))(1+O(\varepsilon_m)) - x_2|}{|x_1| + |x_2|}$$

$$= \frac{|x_1(1+O(\varepsilon_m)) - x_1| + |x_2(1+O(\varepsilon_m)) - x_2|}{|x_1| + |x_2|} = \frac{|x_1| O(\varepsilon_m) + |x_2| O(\varepsilon_m)}{|x_1| + |x_2|}$$

$$\le \frac{|x_1| + |x_2|}{|x_1| + |x_2|} O(\varepsilon_m) = O(\varepsilon_m). \blacksquare$$

Extra Credit. (*2 pts*) Assuming that the result in **9 (b)** above can be extended to $x \in \mathbb{R}^m$ for any m, argue that the obvious algorithm for computing the 2-norm of a vector is backward stable. Along the way you will need to describe, and briefly justify, the expected stability properties of a fifth arithmetic operation.

see last page

for $A \in \mathbb{C}^{m\times m}$ invertible and $b \in \mathbb{C}^m$

10. (*8 pts*) Suppose I invent a new way of solving linear systems which is even more stable than the Householder reflection QR method. The Bueler algorithm solves $Ax = b$ in a backward stablle manner, with numerical result $\tilde{x} \in \mathbb{C}^m$ satisfying $(A + \delta A)\tilde{x} = b$ where $\|\delta A\|_2/\|A\|_2 \le 30\log_{10}(m)\epsilon_{\text{machine}}$.[2] On a computer with $\epsilon_{\text{machine}} = 10^{-16}$, I apply the Bueler algorithm to solve a linear system for a certain matrix $A \in \mathbb{C}^{1000\times 1000}$ for which I know that the 2-norm condition number is $\kappa_2(A) = 10^9$. How many digits of accuracy will I have in the answer $\tilde{x}$? (*Hints.* Start by being clear on what is the problem. Apply big ideas precisely, but avoid little algebra.)

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

we know about the Bueler algorithm that

$$\tilde{x} = \tilde{f}(A) = f(\tilde{A}) \quad \text{where} \quad \tilde{A} = A + \delta A$$

$$\text{and} \quad \frac{\|\tilde{A} - A\|_2}{\|A\|_2} = \frac{\|\delta A\|_2}{\|A\|_2} \le 30\log_{10}(m)\cdot\varepsilon_{\text{machine}}$$

by Thm 15.1 = FTSC:

key idea here

$$\frac{\|\tilde{f}(A) - f(A)\|_2}{\|f(A)\|_2} \le \underbrace{\kappa(A)}_{\kappa(A) = \|A\|_2\|A^{-1}\|_2} \cdot \overbrace{O(\varepsilon_{\text{machine}})}^{= 30\log_{10}(m)\,\varepsilon_{\text{machine}} \text{ in this case}}$$

$\kappa(A) = \|A\|_2\|A^{-1}\|_2$ is rel. cond. # of problem of solving linear system (Thm 12.2)

So: using $m = 10^3$, $\kappa(A) = 10^9$, $\varepsilon_{\text{machine}} = 10^{-16}$

we get

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \le 10^9 \cdot 30\log_{10}(10^3)\cdot 10^{-16} = 10^9 \cdot 30 \cdot 3 \cdot 10^{-16}$$

$$= 3^2 \cdot 10^{-6} \approx 10^{-5}$$

So we expect about 5 decimal digits of accuracy

[2] That is, $\|\delta A\|_2/\|A\|_2 = O(\epsilon_{\text{machine}})$ where the constant is $C = 30\log_{10}(m)$. This is a much smaller constant than the one for Householder QR.

Extra Credit:

The problem is $f(x) = \sqrt{\sum_{i=1}^{m} x_i^2} = \|x\|_2$.

We assume that $\sqrt{\cdot}$ is backward stable:

$$\widetilde{\sqrt{x}} = \sqrt{\tilde{x}} \quad \text{where} \quad \frac{|\tilde{x}-x|}{|x|} = O(\varepsilon_{machine}).$$

This is reasonable because (e.g.) Newton's method should get almost all bits of $\sqrt{x}$ correct for $x \geq 0$. Then

$$\tilde{f}(x) = \widetilde{\sqrt{\sum_{i=1}^{m} x_i \otimes x_i}}$$

A proof by induction, extending 9(b), shows $\tilde{g}(x) = \sum_{i=1}^{m} x_i \otimes x_i$ is backward stable for $g(x) = \sum_{i=1}^{m} x_i^2$. Then $\tilde{f}(x) = \widetilde{\sqrt{\tilde{g}(x)}}$.

So $\tilde{f}(x) = \widetilde{\sqrt{\tilde{g}(x)}} = \sqrt{\tilde{g}(x) + \delta g}$ $\left[\frac{|\delta g|}{|g|} = O(\varepsilon_m)\right]$

$= \sqrt{g(\tilde{x}) + \delta g} = \sqrt{g(\tilde{\tilde{x}})}$ [$\tilde{\tilde{x}}$ has entries multiplied by ≈ 1 since $|\delta g| = O(\varepsilon_m)|g|$]

$= f(\tilde{\tilde{x}})$.

I think one can show $\|\tilde{\tilde{x}} - x\| / \|x\| = O(\varepsilon_{machine})$.