W.D. Henshaw **Math 6800: Solutions for Problem Set 5**

1. Let $A, B \in \mathbb{C}^{m\times m}$ be nonsingular and let $\kappa(A)$ be the condition number of A. Let $\|\cdot\|$ denote any induced matrix norm.

(a) Show that for any induced norm where $\|I\| \geq 1$, then $\kappa(A) \geq 1$.
(b) Show that $\kappa(AB) \leq \kappa(A)\kappa(B)$ and $\kappa(\alpha A) = \kappa(A)$ for any scalar $\alpha \in \mathbb{C}$, $\alpha \neq 0$.
(c) Let $A \in \mathbb{C}^{m\times m}$ be nonsingular with SVD $A = U\Sigma V^*$. Show that if $Ax = b$ then

$$x = \sum_{i=1}^{m} \frac{u_i^* b}{\sigma_i} v_i,$$

where U has columns u_i and V has columns v_i. In which direction will perturbations in b be amplified the most?

Solution:

(a)

$$AA^{-1} = I \ \rightarrow 1 \leq \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A).$$

(a)

$$\kappa(AB) = \|AB\|\|(AB)^{-1}\| = \|AB\|\|B^{-1}A^{-1}\| \leq \|A\|\|B\|\|A^{-1}\|\|B^{-1}\| = \kappa(A)\kappa(B).$$

(b)

$$\kappa(\alpha A) = \|\alpha A\|\|(\alpha A)^{-1}\| = |\alpha||\alpha^{-1}|\kappa(A) = \kappa(A).$$

(c)

$$A = U\Sigma V^* \ \rightarrow A^{-1} = V\Sigma^{-1}U^* = \sum_{i=1}^{m} \sigma_i^{-1} v_i u_i^*$$

and thus

$$Ab = \sum_{i=1}^{m} \frac{u_i^* b}{\sigma_i} v_i.$$

A perturbation in b, $b \rightarrow b + \delta b$ is amplified by $u_i^* \delta b/\sigma_i$ in direction v_i. Thus perturbations along the direction u_m will be amplified the most since $\sigma_m = \min_{i=1}^{m} \sigma_i$. (If there are multiple singular values equal to σ_m then these directions will also be amplified to the same degree).

2. Let $A \in \mathbb{C}^{m\times m}$ and let $\|\cdot\|$ denote any induced matrix norm.
(a) Show that if $\|A\| < 1$ in any induced matrix norm then

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

This shows that when $\|A\| < 1$, $I - A$ is nonsingular. (Hint, consider $I - A$ times the partial sums $S_N = \sum_{k=0}^{N} A^k$ and let $N \to \infty$).

(b) Show that if $\|A\| < 1$, and $\|I\| = 1$ then

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

Solution:

(a) Let $S_N = \sum_{k=0}^{N} A^k$, then

$$(I - A)S_N = \sum_{k=0}^{N}(I - A)A^k = \sum_{k=0}^{N} A^k - A^{k+1} = I - A^{N+1}$$

But since $\|A\| < 1$ then

$$\|A^{N+1}\| \leq \|A\|^{N+1} \to 0 \text{ as } N \to \infty,$$

and thus

$$(I - A)^{-1} = \lim_{N\to\infty} \sum_{k=0}^{N} A^k = \sum_{k=0}^{\infty} A^k.$$

(b) Consider

$$(I - A)(I - A)^{-1} = I \;\to\; (I - A)^{-1} - A(I - A)^{-1} = I \;\to\; (I - A)^{-1} = I + A(I - A)^{-1}$$

Whence

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

and therefore

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

Alternate solution:

$$\|(I - A)^{-1}\| = \|\sum_{k=0}^{\infty} A^k\| \leq \sum_{k=0}^{\infty} \|A^k\| \leq \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|},$$

using the sum of a geometric series since $\|A\| < 1$.

3. (a) Show that if $|\epsilon_i| \leq \epsilon_{\text{machine}}$, $i = 1, 2, \ldots, n$ where $0 < \epsilon_{\text{machine}} < 1$ then

$$(1 - \epsilon_1)(1 - \epsilon_2) \cdots (1 - \epsilon_n) = (1 + \epsilon)^n$$

for some ϵ with $|\epsilon| \leq \epsilon_{\text{machine}}$.

(b) Let $\tilde{f}(x_1, x_2, \ldots, x_n)$ denote the algorithm to sum n floating point numbers $x_i \in \mathbb{F}$, evaluated, in floating point arithmetic from left to right, that is

$$\tilde{f}(x_1, x_2, \ldots, x_n) = (\ldots((x_1 \oplus x_2) \oplus x_3) \oplus \ldots \oplus x_n).$$

Show that

$$\tilde{f}(x_1, x_2, \ldots, x_n) = x_1(1+\epsilon_1)^{n-1} + x_2(1+\epsilon_2)^{n-1} + +x_3(1+\epsilon_3)^{n-2} + \ldots + x_n(1+\epsilon_n),$$

for some ϵ_i with $|\epsilon_i| \leq \epsilon_{\text{machine}}$.

Solution:

(a) Let $\epsilon_m = \min_{i=1}^n \epsilon_i$ and $\epsilon_M = \max_{i=1}^n \epsilon_i$. Then

$$(1-\epsilon_M)^n \leq (1-\epsilon_1)(1-\epsilon_2)\cdots(1-\epsilon_n) = (1+\epsilon)^n \leq (1-\epsilon_m)^n.$$

Since $f(\epsilon) = (1+\epsilon)^n$ is a continuous function of ϵ it will take on all values between $f(-\epsilon_M) = (1-\epsilon_M)^n$ and $f(-\epsilon_m) = (1-\epsilon_m)^n$ and thus there exists an ϵ with $|\epsilon| \leq \epsilon_{\text{machine}}$ such that

$$(1-\epsilon_1)(1-\epsilon_2)\cdots(1-\epsilon_n) = (1+\epsilon)^n$$

(b) By the floating point axioms,

$$\begin{aligned}
\tilde{f}(x_1, x_2, \ldots, x_n) &= (\ldots((x_1 \oplus x_2) \oplus x_3) \oplus \ldots \oplus x_n), \\
&= (\ldots((x_1 + x_2)(1+\tilde{\epsilon}_2) + x_3)(1+\tilde{\epsilon}_3) + x_4)(1+\tilde{\epsilon}_4) + \ldots x_n)(1+\tilde{\epsilon}_n), \\
&= x_1(1+\tilde{\epsilon}_2)(1+\tilde{\epsilon}_2)\cdots(1+\tilde{\epsilon}_n) \\
&\quad + x_2(1+\tilde{\epsilon}_2)(1+\tilde{\epsilon}_2)\cdots(1+\tilde{\epsilon}_n) \\
&\quad + x_3(1+\tilde{\epsilon}_3)(1+\tilde{\epsilon}_3)\cdots(1+\tilde{\epsilon}_n) \\
&\quad + \ldots \\
&\quad + x_n(1+\tilde{\epsilon}_n)
\end{aligned}$$

where $\|\tilde{\epsilon}_i| \leq \epsilon_{\text{machine}}$. Using the results of (a) it follows that

$$\tilde{f}(x_1, x_2, \ldots, x_n) = x_1(1+\epsilon_1)^{n-1} + x_2(1+\epsilon_2)^{n-1} + +x_3(1+\epsilon_3)^{n-2} + \ldots + x_n(1+\epsilon_n),$$

for some ϵ_i with $|\epsilon_i| \leq \epsilon_{\text{machine}}$.

4. NLA 15.1 : *Each of the following problems describes an algorithm implemented on a computer satisfying axioms (13.5) and (13.7). For each ...*

Solution:

Let ϵ_i be numbers with $|\epsilon_i| \leq \epsilon_{\text{machine}}$.

(a) Data: $x \in \mathbb{C}$. Solution: $2x$ computed as $x \oplus x$.
$f(x) = 2x$ and

$$\begin{aligned}
\tilde{f}(x) &= \text{fl}(x) \oplus \text{fl}(x) = (x(1+\epsilon_1) + x(1+\epsilon_1))(1+\epsilon_2), \\
&= x(1+\epsilon_1)(1+\epsilon_2) + (1+\epsilon_1)(1+\epsilon_2)x, \\
&= 2x(1+\tilde{\epsilon}) \\
&= f(x(1+\tilde{\epsilon}))
\end{aligned}$$

where $|\tilde{\epsilon}| \leq 2\epsilon_{\text{machine}} + O(\epsilon_{\text{machine}}^2)$. Here we have assumed that $\text{fl}(x)$ always returns the same value. This algorithm is thus backward-stable with $\tilde{x} = x(1+\tilde{\epsilon})$, and therefore also stable.

(b) Data: $x \in \mathbb{C}$. Solution: $f(x) = x^2$ computed as $\tilde{f}(x) = \text{fl}(x) \otimes \text{fl}(x)$.

$$\begin{aligned}\tilde{f}(x) &= \text{fl}(x) \otimes \text{fl}(x) = (x(1+\epsilon_1) \times x(1+\epsilon_1))(1+\epsilon_2),\\ &= \tilde{x} \times \tilde{x}\end{aligned}$$

where $\tilde{x} = (1+\epsilon_1)\sqrt{(1+\epsilon_2)} \leq \frac{3}{2}\epsilon_{\text{machine}} + O(\epsilon^2_{\text{machine}})$. This algorithm is thus backward-stable and therefore also stable.

(c) Data: $x \in \mathbb{C} \setminus \{0\}$. Solution: $f(x) = 1$ computed as $\tilde{f}(x) = \text{fl}(x) \oslash \text{fl}(x)$.

$$\begin{aligned}\tilde{f}(x) &= \text{fl}(x) \oslash \text{fl}(x) = (x(1+\epsilon_1)/x(1+\epsilon_1))(1+\epsilon_2),\\ &= (1+\epsilon_2)\end{aligned}$$

Therefore this algorithm is stable but NOT backward stable.

(d) Data: $x \in \mathbb{C}$. Solution: $f(x) = x - x$ computed as $\tilde{f}(x) = \text{fl}(x) \ominus \text{fl}(x)$.

$$\begin{aligned}\tilde{f}(x) &= \text{fl}(x) \ominus \text{fl}(x) = (x(1+\epsilon_1) - x(1+\epsilon_1))(1+\epsilon_2),\\ &= \tilde{x} - \tilde{x}\end{aligned}$$

where $\tilde{x} = x(1+\epsilon_1)(1+\epsilon_2) = x(1+\tilde{\epsilon})$, where $|\tilde{\epsilon}| \leq 2\epsilon_{\text{machine}} + O(\epsilon^2_{\text{machine}})$. The algorithm is backward stable and stable.

(e) Data none. Solution: $f(x) = e$ computed from as

$$f(x) = \sum_{k=0}^{N} \frac{1}{k!}$$

summed from left to right and stopping at the value of N when the summand reaches a magnitude less than $\epsilon_{\text{machine}}$.

Solution: Here we assume that $k \in \mathbb{F}$ (this does not change the conclusion, just the constants).

We stop the sum when approximately (the exact value depends on our definition of $\epsilon_{\text{machine}}$),

$$\frac{1}{(N+1)!} \approx \epsilon_{\text{machine}} \sum_{k=0}^{N} \frac{1}{k!} < \frac{1}{2}\epsilon_{\text{machine}}\ e$$

Since $N! \sim N^N\sqrt{2\pi N}e^{-N}$ (Stirling's forumula) as $N \to \infty$ we see that we will need to approximately choose N so that

$$\begin{aligned}&N^N\sqrt{2\pi N}e^{-N} \sim \frac{1}{\epsilon_{\text{machine}}},\\ \to &N\log(N) + \log(2\pi N) - N \sim \log(1/\epsilon_{\text{machine}}),\\ \to &N\log(N) \sim \log(1/\epsilon_{\text{machine}}),\\ \to &\log(N) + \log(\log(N)) \sim \log(\log(1/\epsilon_{\text{machine}})),\\ \to &\log(N) \sim \log(\log(1/\epsilon_{\text{machine}})),\end{aligned}$$

which implies that N grows as $\epsilon_{\text{machine}} \to 0$ but *very* slowly: when $\epsilon_{\text{machine}} \approx 10^{-16}$, $N \approx 18$ in the actual sum.

Let $g(k) = \frac{1}{k!}$ and $\tilde{g}(k)$ denote the machine implementation for computing $g(k)$. Then $\tilde{g}(k)$ is stable with

$$\tilde{g}(k) = \frac{1}{k!}(1+\epsilon_k)^{k+1}$$

for some ϵ_k with $|\epsilon_k| \le \epsilon_{\text{machine}}$. Then summing from left to right gives

$$\begin{aligned}
\tilde{f}(x) &= (((\ldots(\ (\tilde{g}(1) \oplus \tilde{g}(2)) \oplus \tilde{g}(3)) \oplus \ldots \oplus \tilde{g}(N-1)) \oplus \tilde{g}(N), \\
&= \sum_{k=0}^{N} \frac{1}{k!}(1+\epsilon_k)^{k+1}(1+\epsilon'_k)^{N-k}, \\
&= \sum_{k=0}^{N} \frac{1}{k!}(1+\tilde{\epsilon}_k)^{N+1}
\end{aligned}$$

Thus

$$\tilde{f}(x) - e = \sum_{k=0}^{N} \frac{1}{k!}\big[(1+\tilde{\epsilon}_k)^{N+1} - 1\big] - \sum_{k=N+1}^{\infty} \frac{1}{k!},$$

Note that $(1+\tilde{\epsilon}_k)^{N+1} = 1 + NO(\epsilon_{\text{machine}})$, so that

$$\sum_{k=0}^{N} \frac{1}{k!}(1+\tilde{\epsilon}_k)^{N+1} = (1+NO(\epsilon_{\text{machine}}))\sum_{k=0}^{N} \frac{1}{k!} = 1 + NO(\epsilon_{\text{machine}}),$$

and thus

$$\frac{|\tilde{f}(x) - e|}{e} = NO(\epsilon_{\text{machine}})$$

Formally the algorithm is *not stable* due to the growth factor $N = N(\epsilon_{\text{machine}})$ which grows as $\epsilon_{\text{machine}}$ becomes smaller. (In practice N grows so slowly as to be effectively stable). The algorithm is not backward stable since it cannot ever compute "e" which is irrational.

(f) Similar to (e) but sum from right to left.

Summing from right to left gives

$$\begin{aligned}
\tilde{f}(x) &= \sum_{k=0}^{N} \frac{1}{k!}(1+\epsilon_k)^{k+1}(1+\epsilon'_k)^{k}, \\
&= \sum_{k=0}^{N} \frac{1}{k!}(1+\tilde{\epsilon}_k)^{2k+1}
\end{aligned}$$

Whence

$$\begin{aligned}
\tilde{f}(x) - e &= \sum_{k=0}^{N} \frac{1}{k!}\left[(1+\tilde{\epsilon}_k)^{2k+1} - 1\right] - \sum_{k=N+1}^{\infty} \frac{1}{k!}, \\
&= \sum_{k=0}^{N} \frac{1}{k!}\left[1 + (2k+1)O(\epsilon_{\text{machine}}) - 1\right] - \sum_{k=N+1}^{\infty} \frac{1}{k!}, \\
&= \sum_{k=0}^{N} \frac{1}{k!}(2k+1)O(\epsilon_{\text{machine}}) + O(\epsilon_{\text{machine}}), \\
&= O(\epsilon_{\text{machine}})
\end{aligned}$$

where we have used

$$\begin{aligned}
&\sum_{k=0}^{N} \frac{1}{k!} \leq e, \\
&\sum_{k=0}^{N} \frac{2k}{k!} = 2\sum_{k=1}^{N} \frac{1}{(k-1)!} = 2\sum_{k=0}^{N} \frac{1}{k!} \leq 2e.
\end{aligned}$$

Thus

$$\frac{|\tilde{f}(x) - e|}{e} = O(\epsilon_{\text{machine}})$$

In this case the algorithm is stable. It is not backward stable for the same reason as in (g).

(g) Data: none. $f(x) = \pi$. $\tilde{f}(x)$ computed by doing an exhaustive search...
For this problem let ϵ_i and ϵ_i', $i = 1, 2, \ldots$, denote values that are $O(\epsilon_{\text{machine}})$.
If $s(x)$ is a stable algorithm to compute $\sin(x)$ then for any $x \in \mathbb{F}$

$$s(x) = \sin(\tilde{x})(1 + \epsilon_1),$$

for some $\tilde{x} = x(1 + \epsilon_2)$.

We note that $s(3) = \sin(3(1+\epsilon_3))(1+\epsilon_4) > 0$ for $\epsilon_{\text{machine}}$ sufficiently small. Similarly $s(4) < 0$. Since $s(x)$ changes sign there must be a smallest $x \in \mathbb{F}$, $x \in [3, 4]$, such that $s(x) > 0$ and $s(x') \leq 0$, where $x' = x(1 + \epsilon_5)$ is the next consecutive floating point number. Now

$$\begin{aligned}
s(x) &= \sin(\tilde{x})(1 + \epsilon_1) > 0, \\
s(x') &= \sin(\tilde{x}')(1 + \epsilon_1') \leq 0
\end{aligned}$$

means $s(\tilde{x}) > 0$ and $s(\tilde{x}') \leq 0$ which implies (since $\sin(\pi) = 0$ is the only zero of $\sin(x)$ on $[3, 4]$)

$$\tilde{x} < \pi \leq \tilde{x}'$$

But since $\tilde{x}' = \tilde{x}(1 + \epsilon_5)$ it follows that

$$x = \tilde{x}(1 + \epsilon_1) = \pi(1 + \epsilon_6)$$

The algorithm is thus a stable way to compute π. It cannot be backward stable since π is irrational, $\pi \notin \mathbb{F}$.