4 Homework Solutions

18.335 - Fall 2004

4.1 Trefethen 11.1

First note that any $x \in \mathbb{C}^m$ can be writtten as $x = x_R + x_R^{\perp}$ where $x_R \in \mathcal{R}(A)$, $x_R^{\perp} \in \mathcal{R}(A)^{\perp}$. Now since:

$$\left(x_R^{\perp}\right)^* \underbrace{Ay}_{\in \mathcal{R}(A)} = 0,\ \forall y \in \mathbb{C}^n \Longrightarrow y^* \left(A^* x_R^{\perp}\right) = 0,\ \forall y \in \mathbb{C}^n \Longrightarrow A^* x_R^{\perp} = 0$$

we have by definition:

$$\begin{aligned}
\left\|A^+\right\| &= \max_{\substack{z \in \mathbb{C}^m \\ z \neq 0}} \frac{\left\|\left(A^*A\right)^{-1}\left(A^* x_R + A^* x_R^{\perp}\right)\right\|}{\|x\|} \\
&\leq \max_x \frac{\left\|\left(A^*A\right)^{-1} A^* x_R\right\|}{\|x_R\|} = \max_w \frac{\left\|\left(A^*A\right)^{-1} A^* A w\right\|}{\|Aw\|} = \max_w \frac{\|w\|}{\|Aw\|} \\
&\leq \max_w \frac{\|w\|}{\left\|\sqrt{\|A_1 w\|^2 + \|A_2 w\|^2}\right\|} \leq \max_w \frac{\|w\|}{\|A_1 w\|} = \max_w \frac{\left\|A_1^{-1} w\right\|}{\|w\|} = \left\|A_1^{-1}\right\|
\end{aligned}$$

4.2 Prove that (13.7) in Trefethen is valid for complex arithmetic (all four arithmetic operations) with ε now bounded by a modest multiple of $\varepsilon_{\text{machine}}$.

For addition and subtraction we have

$$(a + ib) \pm (c + id) := (a \pm c) + i(b \pm d).$$

Let δ_i be small numbers bounded in absolute value by ϵ. We have

$$\begin{aligned}
fl((a + ib) \pm (c + id)) &= (a \pm c + i(b \pm d)) \left(1 + \frac{(a \pm c)\delta_1 + i(b \pm d)\delta_2}{(a \pm c) + i(b \pm d)}\right) \\
&= (a \pm c + i(b \pm c))(1 + \delta)
\end{aligned}$$

where

$$|\delta|^2 = \frac{(a \pm c)^2 \delta_1^2 + (b \pm d)\delta_2^2}{(a \pm c)^2 + (b \pm d)^2} \leq 2\epsilon^2$$

so $|\delta| \leq \sqrt{2}\epsilon$. For multiplication we have:

$$(a + ib)(c + id) := (ac - bd) + i(ad + bc)$$

For some $|\delta_i| \leq 2\epsilon$ we have:

$$\begin{aligned} fl((a+ib)(c+id)) &= (ac(1+\delta_1) - bd(1+\delta_2)) + i(ad(1+\delta_3) + bc(1+\delta_4)) \\ &= [(ac-bd) + i(ad+bc)] + [(ac\delta_1 - bd\delta_2) + i(ad\delta_3 + bc\delta_4)] \end{aligned}$$

We will use the fact that $|u+iv| \leq |u| + |v| \leq \sqrt{2}|u+iv|$ to write

$$fl((a+ib)(c+id)) = (a+ib)(c+id)(1+\beta)$$

where

$$\beta = \frac{(ac\delta_1 - bd\delta_2) + i(ad\delta_3 + bc\delta_4)}{(a+ib)(c+id)}$$

$$\begin{aligned} |\beta| &\leq \frac{(|ac\delta_1| + |bd\delta_2|) + (|ad\delta_3| + |bc\delta_4|)}{\frac{1}{2}(|a|+|b|)(|c|+|d|)} \\ &\leq 4\epsilon \frac{(|a|+|b|)(|c|+|d|)}{(|a|+|b|)(|c|+|d|)} = 4\epsilon \end{aligned}$$

This result does not guarantee high relative accuracy in the individual components of the product. For example if we take two numbers whose product is nearly real, the imaginary part will be the result of cancellation and so be small but probably not accurate. But the real part will be large, so the bound holds. Another way to look at it is that the true product lies in a little ball in the complex plane centered at the true product p and with radius $4\epsilon|p|$. If this ball interects the real (or imaginary) axis, then we can't even guarantee the sign of the real (or imaginary) part.

For division we have the following algorithm for computing $(a+bi)/(c+di)$:

- $\alpha = \max(|c|, |d|)$
- $c_1 = c/\alpha$
- $d_1 = d/\alpha$, ... therefore $c_1 + d_1 i = \frac{1}{\alpha}(c+di)$
- $s = \alpha(c_1^2 + d_1^2)$, ... same as $\alpha(c_1 + d_1 i)(c_1 - d_1 i)$
- $w = (a+bi)(c_1 - d_1 i)$
- $z = w/s$, ... same as $z = \frac{1}{s} \cdot w$
- return z

This clearly produces the right answer in exact arithmetic. These operations can be interpreted as complex multiplications and forming inverses of real numbers. Each of those operation preserves the relative accuracy and the overall error bound is a product of all $(1+\delta)$ terms from each complex multiply. Over all we get a relative error bounded by 22ϵ.

4.3 Prove that in IEEE binary floating point arithmetic $\sqrt{x}$ returns x exactly.

Recall that any IEEE number, x, can be written as

$$x = 2^a \left(1 + \frac{m}{2^{53}}\right), \text{ with } 0 \le m < 2^{53}$$

Note that here we assumed double precision, even though this is not necessary. Since we are not concerned with overflow or underflow, no limits were placed on a. Then we have

$$x^2 = 2^{2a} \left(1 + \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}}\right)$$

To show that $\text{fl}(\sqrt{x^2}) = x$ we need to verify that:

$$2^a \left(1 + \frac{m - \frac{1}{2}}{2^{53}}\right) \le \sqrt{2^{2a} \left(1 + \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}}\right)} < 2^a \left(1 + \frac{m + \frac{1}{2}}{2^{53}}\right) \tag{1}$$

In order to do that we have to distinguish 2 cases

- $\frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}} < 1$

  This implies that for

  $$-\frac{1}{2} + \frac{m^2}{2^{53}} \le y \le \frac{1}{2} + \frac{m^2}{2^{53}} \tag{2}$$

  we get that

  $$\text{fl}(x^2) = 2^{2a} \left(1 + \frac{2m + y}{2^{53}}\right)$$

  Thus (1) becomes

  $$1 + \frac{m - \frac{1}{2}}{2^{53}} \le \sqrt{1 + \frac{2m + y}{2^{53}}} < 1 + \frac{m + \frac{1}{2}}{2^{53}}$$
  $$\Rightarrow -1 + \frac{\left(m - \frac{1}{2}\right)^2}{2^{53}} \le y < 1 + \frac{\left(m + \frac{1}{2}\right)^2}{2^{53}}$$

  which is obiously true because of (2) .

- $1 < \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}} < 3$

  In this case we have that for some $0 \le k < 2^{52}$

  $$1 + \frac{2k - \frac{1}{2}}{2^{53}} \le \frac{2m}{2^{53}} + \frac{m^2}{2^{106}} < 1 + \frac{2k + \frac{1}{2}}{2^{53}} \tag{3}$$

and therefore

$$\text{fl}(x^2) = 2^{2a}\left(1 + 1 + \frac{2k}{2^{53}}\right) = 2^{2a+1}\left(1 + \frac{k}{2^{53}}\right)$$

To show that $\text{fl}(\sqrt{x^2}) = x$ we need to verify from (1) that

$$1 + \frac{m - \frac{1}{2}}{2^{53}} \leq \sqrt{2\left(1 + \frac{k}{2^{53}}\right)} < 1 + \frac{m + \frac{1}{2}}{2^{53}}$$

This follows directly from (3) since by re-arranging terms in (3) we get:

$$\left(1 + \frac{m}{2^{53}}\right)^2 - \frac{1}{2^{54}} \leq 2\left(1 + \frac{k}{2^{53}}\right) < \left(1 + \frac{m}{2^{53}}\right)^2 + \frac{1}{2^{54}}$$

Combining these 2 cases we complete the proof.

4.4 Let a and b be positive IEEE binary floating point numbers such that $a < b < 2a$. Prove that $\text{fl}(b - a) = b - a$ exactly.

Proof: Assume $a = 1.a_1a_2...a_n \times 2^k, b = 1.b_1b_2...b_n \times 2^r$ $(a_i, b_i \in \{0, 1\})$. Also we may assume $k = 0$. $b \geq a$ implies $r \leq 0$ and $2a \geq b$ implies $r + 1 \geq 0$, so we have either $r = 0$ or $r = -1$.

For $r = 0$ we have $b - a = 1.b_1b_2...b_n - 1.a_1a_2...a_n = 0.c_1c_2...c_n$, which is an exact floating point number, since it has less than $n + 1$ fraction bits.

For $r = -1$ we have $b - a = 1.b_1b_2...b_n - 1.a_1a_2...a_n \times 2^{-1} = 1.b_1b_2...b_n0 - 0.1a_1a_2...a_n = c_{-1}.c_0c_1c_2...c_n$. Since $2a \geq b$ we have $c_{-1} = 0$. So the result has at most $n + 1$ fraction bits which is again an exact floating point number.