Numerical Analysis **Answers to Assignment #2**

1. (i) Here, x^n is computed by

$$\begin{aligned} p_1 &= x \\ p_k &= \mathrm{fl}(xp_{k-1}), \quad k = 2, 3, \ldots, n, \\ p_n &= \mathrm{fl}(x^n). \end{aligned}$$

Since $p_2 = x^2(1+\varepsilon_2)$, $p_3 = x^3(1+\varepsilon_2)(1+\varepsilon_3), \ldots$, with $|\varepsilon_k| \leq$ eps, we have

$$p_n = x^n(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n),$$

hence

$$\begin{aligned} \left|\frac{p_n - x^n}{x^n}\right| &= |(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n) - 1| \\ &\approx |\varepsilon_2 + \varepsilon_3 + \cdots + \varepsilon_n| \leq (n-1) \text{ eps}. \end{aligned}$$

(ii) Here,

$$\mathrm{fl}(x^n) = e^{n[\ln x](1+\varepsilon_1)(1+\varepsilon_2)}(1+\varepsilon_3), \quad |\varepsilon_i| \leq \text{eps}, \quad i = 1, 2, 3,$$

where ε_1, ε_2, ε_3 are the relative errors committed in computing respectively the logarithm, the product by n, and the exponential. Thus,

$$\begin{aligned} \mathrm{fl}(x^n) &\approx e^{n[\ln x](1+\varepsilon_1+\varepsilon_2)}(1+\varepsilon_3) = e^{n[\ln x]}e^{(\varepsilon_1+\varepsilon_2)n[\ln x]}(1+\varepsilon_3) \\ &\approx x^n(1 + (\varepsilon_1+\varepsilon_2)n\ln x + \varepsilon_3), \end{aligned}$$

and

$$\left|\frac{\mathrm{fl}(x^n) - x^n}{x^n}\right| \approx |(\varepsilon_1+\varepsilon_2)n\ln x + \varepsilon_3| \leq (2n|\ln x| + 1) \text{ eps}.$$

It follows that (i) is more accurate than (ii) if

$$n - 1 \leq 2n|\ln x| + 1,$$

that is, if

$$n(1 - 2|\ln x|) \leq 2.$$

This is always true if $|\ln x| \geq \frac{1}{2}$, that is, if $0 < x \leq e^{-\frac{1}{2}}$ or $x \leq e^{\frac{1}{2}}$, and on the interval $e^{-\frac{1}{2}} < x < e^{\frac{1}{2}}$ it is true if

$$n \leq \frac{2}{1 - 2|\ln x|}.$$

2.
$$(x+y)^{1/4}-y^{1/4}=\frac{(x+y)^{1/2}-y^{1/2}}{(x+y)^{1/4}+y^{1/4}}=\frac{x}{[(x+y)^{1/4}+y^{1/4}][(x+y)^{1/2}+y^{1/2}]}$$

3. (a) $(\text{cond}\, f)(x)=\left|\dfrac{x\cdot\frac{1}{n}x^{\frac{1}{n}-1}}{x^{\frac{1}{n}}}\right|=\dfrac{1}{n}\leq 1$

Evidently, f is very *well* conditioned for all $x>0$.

(b) $(\text{cond}\, f)=\left|\dfrac{x(1-\frac{x}{\sqrt{x^2-1}})}{x-\sqrt{x^2-1}}\right| = \left|\dfrac{x(\sqrt{x^2-1}-x)}{\sqrt{x^2-1}(x-\sqrt{x^2-1})}\right| = \dfrac{x}{\sqrt{x^2-1}}$

Here, f is *ill* conditioned near $x=1$, *well* conditioned as $x\to\infty$.

(c) *detailed analysis*:

$$(\text{cond}\, f)(x_1)=\frac{x_1\cdot\frac{1}{2}(x_1^2+x_2^2)^{-1/2}\cdot 2x_1}{(x_1^2+x_2^2)^{1/2}} = \frac{x_1^2}{x_1^2+x_2^2} < 1$$
$$(\text{cond}\, f)(x_2)=\frac{x_2^2}{x_1^2+x_2^2} < 1$$

$$(\text{cond}\, f)(x_1,x_2)=(\text{cond}\, f)(x_1)+(\text{cond}\, f)(x_2)=1$$

Again, f is *well* conditioned for arbitrary x_1, x_2 with $|x_1|+|x_2|>0$.

global approach (based on Euclidean norms):

$\parallel x\parallel_2 := \sqrt{x_1^2+x_2^2}$, $\parallel [a_1,a_2]\parallel_2 := \sqrt{a_1^2+a_2^2}=\parallel a^T\parallel_2$

$$(\text{cond}\, f)(x)=\frac{\parallel x\parallel_2\parallel f'(x)\parallel_2}{|f(x)|}$$
$$\frac{\parallel x\parallel_2\parallel [x_1(x_1^2+x_2^2)^{-1/2}, x_2(x_1^2+x_2^2)^{-1/2}]\parallel_2}{\sqrt{x_1^2+x_2^2}}=\frac{\parallel x\parallel_2\cdot 1}{\parallel x\parallel_2}=1$$

(d) *detailed analysis*:

$$(\text{cond}\, f)(x_1)=\left|\frac{x_1}{x_1+x_2}\right|,\ (\text{cond}\, f)(x_2)=\left|\frac{x_2}{x_1+x_2}\right|$$

$$(\mathrm{cond}\, f)(x_1, x_2) = (\mathrm{cond}\, f)(x_1) + (\mathrm{cond}\, f)(x_2) = \frac{|x_1| + |x_2|}{|x_1 + x_2|}$$

As expected, f is *ill* conditioned when $|x_1 + x_2|$ is very small, but $|x_1|$ (and hence $|x_2|$) is not (cancellation!).

global approach (based on L_1-norm):

$$\| x \|_1 := |x_1| + |x_2|$$

$$(\mathrm{cond}\, f)(x) = \frac{\| x \|_1 \| f'(x) \|_1}{|f(x)|} = \frac{\| x \|_1 \| [1, 1] \|_1}{|x_1 + x_2|}$$

$$= 2\, \frac{|x_1| + |x_2|}{|x_1 + x_2|}$$

4. We have

$$\begin{aligned}(\mathrm{cond}\, f \cdot g)(x) &= \left| \frac{x(fg)'(x)}{(fg)(x)} \right| = \left| \frac{xf'(x)g(x)}{f(x)g(x)} + \frac{xf(x)g'(x)}{f(x)g(x)} \right| \\ &\le \left| \frac{xf'(x)}{f(x)} \right| + \left| \frac{xg'(x)}{g(x)} \right| = (\mathrm{cond}\, f)(x) + (\mathrm{cond}\, g)(x).\end{aligned}$$

5. PROGRAMS

```
c The series in (a)
c
      n=0
      term=1.
      sum1=1.
   10 sum=sum1
      n=n+1
      if(n.gt.100) goto 20
      term=-term/float(n*n)
      sum1=sum+term
      if(sum1.ne.sum) goto 10
      write(*,2) n,sum
    2 format(1x,'n=',i1,'    sum(a) =',f10.7)
```

```
      stop
   20 write(*,3)
    3 format(1x,'no convergence')
      stop
      end


c The series in (b)
c
      n=0
      term=1.
      sum1=1.
   10 sum=sum1
      n=n+1
      if(n.gt.100) goto 20
      term=term/float(n*n)
      sum1=sum+term
      if(sum1.ne.sum) goto 10
      write(*,2) n,sum
    2 format(1x,'n=',i1,'    sum(b) =',f9.6)
      stop
   20 write(*,3)
    3 format(1x,'no convergence')
      stop
      end
```

```
    RESULTS

 n=8    sum(a) = 0.2238908
 n=7    sum(b) = 2.279585
```

The exact answers are $J_0(2) = .22389077914...$, $I_0(2) = 2.279585302...$ (see, e.g., Abramowitz & Stegun, Tables 9.4 and 9.11). The accuracy is perfect to 7 significant digits, there being now cancellation of terms, not even in the alternating series since the terms decrease rapidly in absolute value.