✍1 *Non associativity in the presence of round-off.*

Solution: *This is done in a class demo and the diary should be posted. Here are the commands.*

```
n = 10000;
a = randn(n,1);  b = randn(n,1);  c = randn(n,1);
t = ((a+b)+c == a+(b+c));
sum(t)
```

Right-hand side in 3rd line returns 1 for each instance when the two numbers are the same. ☐

✍2 *Find machine epsilon in matlab.*

Solution:

```
u = 1;
```

```
for i=0:999
    fprintf(1,' i =  %d , u = %e \n',i,u)
    if (1.0 +u == 1.0) break, end
    u = u/2;
end
u = u*2
```

□

✍4 *Proof of Lemma: If* $|\delta_i| \le \underline{\mathbf{u}}$ *and* $n\underline{\mathbf{u}} < 1$ *then*

$$\Pi_{i=1}^{n}(1+\delta_i) = 1 + \theta_n \quad \text{where} \quad |\theta_n| \le \frac{n\underline{\mathbf{u}}}{1 - n\underline{\mathbf{u}}}$$

Solution:

The proof is by induction on n.

1) Basis of induction. When $n = 1$ *then the product reduces to* $1 + \delta_i$ *and so we can take*

$\theta_n = \delta_n$ *and we know that* $|\delta_n| \leq \underline{\mathbf{u}}$ *from the assumptions and so*

$$|\theta_n| \leq \underline{\mathbf{u}} \leq \frac{\underline{\mathbf{u}}}{1 - \underline{\mathbf{u}}},$$

as desired.

2) Induction step. Assume now that the result as stated is true for n *and consider a product with* $n + 1$ *terms:* $\Pi_{i=1}^{n+1}(1 + \delta_i)$*. We can write this as* $(1 + \delta_{n+1})\Pi_{i=1}^{n}(1 + \delta_i)$ *and from the induction hypothesis we get:*

$$\Pi_{i=1}^{n+1}(1 + \delta_i) = (1 + \theta_n)(1 + \delta_{n+1}) = 1 + \theta_n + \delta_{n+1} + \theta_n\delta_{n+1}$$

with θ_n *satisfying the inequality* $\theta_n \leq (n\underline{\mathbf{u}})/(1 - n\underline{\mathbf{u}})$*. We call* θ_{n+1} *the quantity* $\theta_{n+1} = \theta_n + \delta_{n+1} + \theta_n\delta_{n+1}$*, and we have*

$$\begin{aligned}
|\theta_{n+1}| &= |\theta_n + \delta_{n+1} + \theta_n\delta_{n+1}| \\
&\leq \frac{n\underline{\mathbf{u}}}{1 - n\underline{\mathbf{u}}} + \underline{\mathbf{u}} + \frac{n\underline{\mathbf{u}}}{1 - n\underline{\mathbf{u}}} \times \underline{\mathbf{u}} = \frac{n\underline{\mathbf{u}} + \underline{\mathbf{u}}(1 - n\underline{\mathbf{u}}) + n\underline{\mathbf{u}}^2}{1 - n\underline{\mathbf{u}}} = \frac{(n+1)\underline{\mathbf{u}}}{1 - n\underline{\mathbf{u}}} \\
&\leq \frac{(n+1)\underline{\mathbf{u}}}{1 - (n+1)\underline{\mathbf{u}})}
\end{aligned}$$

This establishes the result with n replaced by $n+1$ as wanted and completes the proof. $\square$

✍ 5 *Assume you use single precision for which you have $\underline{\mathbf{u}} = 2. \times 10^{-6}$. What is the largest n for which $n\underline{\mathbf{u}} \leq 0.01$ holds? Any conclusions for the use of single precision arithmetic?*

Solution: *We need $n \leq 0.01/(2.0 \times 10^{-4})$ which gives $n \leq 5,000$. Hence, single precision is inadequate for computations involving long inner products.*

✍ 6 *What does the main result on inner products imply for the case when $y = x$? [Contrast the relative accuracy you get in this case vs. the general case when $y \neq x$]* $\square$

Solution: *In this case we have*

$$|fl(x^T x) - (x^T x)| \leq \gamma_n x^T x$$

which implies that we will always have a small relative error. Not true for the general case because

➤ *This leads to the final result (forward form)*

$$\left|fl(y^T x) - (y^T x)\right| \leq \gamma_n |y|^T |x|$$

does not imply a small relative error which would mean $|fl(y^T x) - (y^T x)| \leq \epsilon |y^T x|$ *where* ϵ *is small.* ☐

✍ 7 *Show for any* x, y, *there exist* $\Delta x, \Delta y$ *such that*

$$
\begin{aligned}
fl(x^T y) &= (x + \Delta x)^T y, \quad \text{with} \quad |\Delta x| \leq \gamma_n |x| \\
fl(x^T y) &= x^T (y + \Delta y), \quad \text{with} \quad |\Delta y| \leq \gamma_n |y|
\end{aligned}
$$

Solution:

The main result we proved is that

$$
fl(y^T x) = \sum_{i=1}^{n} x_i y_i (1 + \theta_i) \qquad \text{where} \quad |\theta_i| \leq \gamma_n
$$

The first relation comes from just attaching each $(1 + \theta_i)$ *to* x_i *so* x_i *is replaced by* $x_i + \theta_i x_i$ *... Similarly for the second relation.* ☐

✍ 8 *(Continuation) Let* A *an* $m \times n$ *matrix,* x *an* n*-vector, and* $y = Ax$. *Show that there* *exist a matrix ΔA such*

$$fl(y) = (A + \Delta A)x, \quad \text{with} \quad |\Delta A| \le \gamma_n |A|$$

Solution: *The result comes from applying the result on inner products to each entry y_i of y – which is the inner product of row i with y. We use the first of the two results above:*

$$fl(y_i) = (a_{i,:} + \Delta a_{i,:})^T y \quad \text{with} \quad |\Delta a_{i,:}| \le \gamma_n |a_{i,:}|$$

the result follows from expressing this in matrix form. □

✍ 9 *(Continuation) From the above derive a result about a column of the product of two matrices A and B. Does a similar result hold for the product AB as a whole?*

Solution: *We can have a result each column since this is just a matrix-vector product. How this does not translate into a result for AB because the ΔA we get for each column will depend on the column. Specifically, for the j-th column of B you will have a certain matrix $(\Delta A)_j$ such that $fl(AB(:,j)) = (A + (\Delta A)_j)B(:,j)$ with certain conditions as set in previous exercise. However this $(\Delta A)_j$ is *NOT* the same matrix for each column. So you cannot say*

$\boldsymbol{fl}(\boldsymbol{A}) = (\boldsymbol{A} + \boldsymbol{\Delta A})\boldsymbol{B}$, ... $\square$