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.

\[
\begin{align*}
n &= 10000; \\
a &= \text{randn}(n,1); \quad b = \text{randn}(n,1); \quad c = \text{randn}(n,1); \\
t &= ((a+b)+c == a+(b+c)); \\
\text{sum}(t)
\end{align*}
\]

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

Find machine epsilon in matlab.

Solution:

\[
\begin{align*}
u &= 1;
\end{align*}
\]
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

\[ \prod_{i=1}^{n}(1 + \delta_i) = 1 + \theta_n \text{ where } |\theta_n| \leq \frac{nu}{1 - nu} \]

\[ \textbf{Solution:} \]

\[ \text{The proof is by induction on } n. \]

\[ 1) \text{ Basis of induction. When } n = 1 \text{ then the product reduces to } 1 + \delta_i \text{ and so we can take } \]
\( \theta_n = \delta_n \) and we know that \( |\delta_n| \leq u \) from the assumptions and so

\[
|\theta_n| \leq u \leq \frac{u}{1-u},
\]
as desired.

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

\[
\prod_{i=1}^{n+1} (1 + \delta_i) = (1 + \theta_n)(1 + \delta_{n+1}) = 1 + \theta_n + \delta_{n+1} + \theta_n \delta_{n+1}
\]

with \( \theta_n \) satisfying the inequality \( \theta_n \leq (nu)/(1-nu) \). We call \( \theta_{n+1} \) the quantity \( \theta_{n+1} = \theta_n + \delta_{n+1} + \theta_n \delta_{n+1} \), and we have

\[
|\theta_{n+1}| = |\theta_n + \delta_{n+1} + \theta_n \delta_{n+1}|
\leq \frac{nu}{1-nu} + u + \frac{nu}{1-nu} \times u = \frac{nu + u(1-nu) + nu^2}{1-nu} = \frac{(n+1)u}{1-nu}
\]

\[
\leq \frac{(n+1)u}{1-(n+1)u}
\]
This establishes the result with \( n \) replaced by \( n + 1 \) as wanted and completes the proof. ■

\textcolor{red}{\textbf{\#5}} Assume you use single precision for which you have \( u = 2.0 \times 10^{-6} \). What is the largest \( n \) for which \( nu \leq 0.01 \) 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.

\textcolor{red}{\textbf{\#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 \( y \neq x \)] ■

**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

\( \diamond \) This leads to the final result (forward form)

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

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

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

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

**Solution:**

The main result we proved is that

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

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.

Let $A$ an $m \times n$ matrix, $x$ an $n$-vector, and $y = Ax$. Show that there
exist a matrix $\Delta A$ such

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

**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:

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

the result follows from expressing this in matrix form. □

(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 not into a result for $AB$ because the $\Delta A$ we get for each column will depend on the column. □