ASSIGNMENT 1 - Solution

Problem 1 - Solution

First it is important to realize that all input numbers have to be first converted to the chosen precision before any computation can be performed on them. Right after any computation, the result must also be rounded to the chosen precision before continuing with other computations. Notice that with a 4-digit or higher decimal precision, the numbers 4, 8, $r_0 = 6370$, and $r_1 = r_0 + 1 = 6371$ are all machine numbers, and therefore suffer no rounding. However rounding is expected to affect all other numbers and intermediate results.

Second because the associative property of various operations (in particular with multiplication here) may not hold in finite precision arithmetic especially at low precision, using the formula $A = 4 \cdot \pi \cdot r^2$ to compute the area may give slightly different results depending on how the factors are multiplied together. I think one is supposed to compute the area by first squaring the radius and then compute A from multiplying from left to right. One could also compute it using the formula $\tilde{A} = 4 \cdot \pi \cdot r \cdot r$, where all the factors are multiplied together from left to right. There are other ways also.

(a) Using 4-digit decimal precision we have $\pi = 3.142e0$ and therefore $4\pi = 1.257e1$. For r_0^2 we have $4.058e7$. The computed value for A is $5.101e8$. On the other hand, the value for $4 \cdot \pi \cdot r$ is $8.007e5$, and the result for $\tilde{A}$ is $5.100e8$. As we will see later, this second result $\tilde{A}$ is slightly more accurate.

(b) Next we repeat the above computation for $r_1 = 6371$. The result for A is $5.102e8$, and is the same for $\tilde{A}$. The difference in the area A is therefore $1.000e5$, and for $\tilde{A}$ is $2.000e5$. The fact that these two results differ by so much tell us that neither one is accurate even for the leading digit. We can see that in a different way in parts (c), (d) below.

(c) From calculus we know that for an infinitesimal change in the radius dr, the change in the area is $dA = 8\pi r dr$. If we let $dr = h$ we can derive this result together with an error term as follows.

$$\begin{aligned} A(r+h) - A(r) &= 4\pi(r+h)^2 - 4\pi r^2 = 4\pi(r^2 + 2rh + h^2 - r^2) \qquad (1) \\ &= 8\pi r h(1 + \frac{h}{2r}). \end{aligned}$$

The second term inside the parentheses gives the relative error due to the fact that h is not infinitesimally small. For $h = 1$ and $r = 6370$ this error can at most affect the least significant digit by plus or minus 1. When we compute the change in the area using the formula $8\pi r h$ for $r = r_0$ and $h = 1$ with 4-digit decimal precision,

we obtain $1.601e5$. This result is expected to be rather accurate. Thus we see that the result for the difference in the areas computed in part (b) has roughly a 50% error.

(d) We repeat the above computations using a 6-digit decimal precision. The change in the surface area obtained by computing separately the two areas and then subtracting the two results (method 1)is $1.59000e5$. This time exactly the same results are obtained no matter how the factors are grouped in the computation of the areas. The result when we use the $8\pi rh$ formula (method 2) is $1.60095e5$.

(e) Method 1 obtains the change in the surface area by subtracting two nearly equal numbers and therefore suffers from huge cancellations. The computed result is therefore not accurate compared with the one computed using the formula from calculus which does not have cancellation problems. Of course the cancellation becomes less problematic when higher precision is used. That explains the results obtained in part (d).

(f) We write a MATLAB program to compute the changes in the surface area for a given h using methods 1 and 2. All computations are carried out using double-precision (16-digit decimal precision). The computation is performed initially for $h = 1$. Its value is then reduced by one half at each step. We see that when h is not too small, both methods give essentially the same results. The small discrepancies come mainly from the fact that h is not infinitesimally small. Cancellation begins to affect the accuracy of the results from method 1 when h is smaller than about 2e-12. For h less than about 4e-13, the difference in the area computed from method 1 gives a result of zero. We can easily see why that is so because

$$A(r+h) = 4\pi(r+h)^2 \approx 4\pi(r^2+2rh) = 4\pi r^2\left(1+\frac{2h}{r}\right).$$

When the fraction inside the parentheses is less than eps, $A(r+h)$ is numerically indistinguishable from $A(r)$ and so the difference in the area is zero. We see that

$$\frac{2h}{r} = \frac{2\times 4\times 10^{-13}}{6370} \approx 1.3\times 10^{-16},$$

which is indeed very close to eps.