Shawn Hinnebusch

Parallel Computing for Engineers

Oct 9, 2020

Professor: Dr. Senocak

Homework #4: Infinite Series Calculated with Single and Double Precision

1)

a. Largest positive integer is: 18446744073709551615
b. - d.

Table 1: Results of the summation 1/n for single and double precision

	Single (SP)	Double (DP)
Epsilon	1.192093e-07	2.220446e-16
Iterations	603975	1500000000000 *
Time (s)	0.003364	3667.204075 *
Sum (1/n)	13.889009	28.613702447488990*

* 1.5e12 was chosen as a stopping criterion to keep the run to about an hour. It never hit the convergence criterion to converge and would have run longer if the additional constraint was set higher. The residual got down to 2.329886e-14 which is still above the machine epsilon for double precision. It maxed out the 1-hour time before fully converging.

e. How long does it take to run in single precision vs. double precision to get the largest number in each case? Comment on your findings.

Single precision converged in a fraction of a second whereas a double took over an hour and still never reached the convergence criterion because it was stopped due to the time constraint for this homework assignment.

f. Now consider $\sum_{n=1}^{\infty} \frac{1}{n^2}$ What is the largest number you can obtain?

Table 2: Results of the summation $1/n^2$

	Single	Double
Sum ($1/n^2$)	1.644494	1.644934047831179

2) Write a C program to compute the exponential function using the infinite series

a) Summing in the natural order, what stopping criterion should you use to calculate the infinite series?

Originally, when developing this series, the stopping criteria was the factorial as the largest value a single precision can hold is $3.4x10^{38}$ which will hold a factorial of 34 max before going over the max value. The max value of a double is $1.8x10^{308}$ which can hold a factorial of 170. For a SP, the factorial goes over the max value it can hold before it meets the convergence criteria on several occasions so two stopping criterions need to be used.

Stopping criteria

Single: n > 34 && residual < MACHINE_PRECISION

Double: n > 170 && residual < MACHINE_PRECISION

b) Can you use the series in this form to obtain accurate results for x < 0?

To keep the exact formula as is, the series will not give accurate results when x is negative. By saying $1/e^x$, will give more accurate results, but by regrouping the series terms as discussed in the next section will give the best results to this accuracy problem.

c) Can you rearrange the series or regroup the series terms in any way to obtain more accurate results for single precision calculations?

Yes, by regrouping the series terms as in equation 1, you can get much more accurate results in these calculations for SP.

$$e^x = 1 + \left(\frac{x}{1}\right)(1 + \left(\frac{x}{2}\right)(1 + \left(\frac{x}{3}\right)(\ldots))) \quad (1)$$

This assumes x > 0. In the case where x is negative, e^x needs calculated first, then an additional calculation of $1/e^x$ is added in equation 2.

$$e^x = 1 + \left(\frac{(-1)x}{1}\right)(1 + \left(\frac{(-1)x}{2}\right)(1 + \left(\frac{(-1)x}{3}\right)(\ldots))) \quad (2)$$

Once this value is calculated, $1/e^x$ calculation is completed to give much more accurate answers than the original series. The values were very close to the math function of exp(x).