CENG 371 - Scientific Computing
Fall 2023
Homework 1

Adıgüzel, Gürhan İlhan
e2448025@metu.edu.tr

October 30, 2023

Answer 1

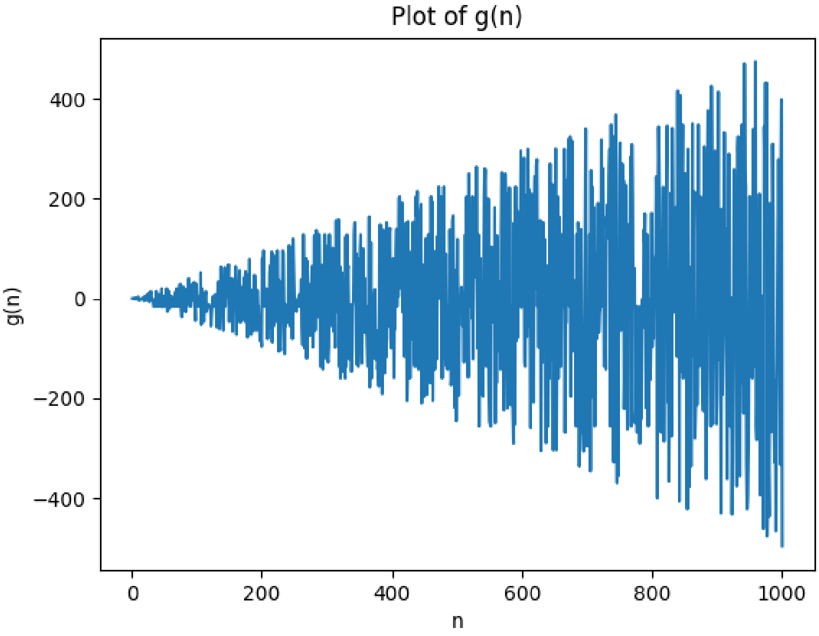

Figure 1: g(n)

(a)

(b) The values of n that satisfy g(n) = 0 are 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512.

(c) $g(n)$ is very large for most values of n because ϵ is very small, which is approximately 2.220446049250313e-16. This epsilon represents the smallest possible relative error in representing real numbers within the chosen floating-point format. In other words, it's the smallest discrepancy that can occur when approximating real numbers with the finite precision of floating-point representation. Floating-point arithmetic, which employs a fixed number of binary digits to represent real numbers, introduces rounding errors due to this limited precision. As a result, not all real numbers can be exactly represented, and they must be rounded to the nearest available value in the floating-point format. Consequently, the cumulative effect of these small rounding errors and precision losses can cause $g(n)$ to deviate from zero for the majority of n values..

(d) The non-zeros of $g(n)$ grow in size as n increases because the term $\frac{n}{\epsilon}$ dominates the other terms in the expression for $g(n)$. As n increases, the value of $\frac{n}{\epsilon}$ increases exponentially, while the other terms grow at a slower rate.

Answer 2

(a) Sum of an arithmetic series :

$$S = \frac{n}{2} * (firstTerm + lastTerm)$$

$$\frac{10^6}{2} * (1.1 + 1.00000001)$$

Therefore, the theoretical result for Theoretical Summation : 1005000.005000.

(b) Pairwise summation is an algorithm used to reduce the loss of precision in the summation of a large set of numbers. It works by splitting the numbers into pairs and adding them together iteratively, reducing the error that can accumulate in the final result when adding a large number of floating-point values.

(c) i. Naive Summation in single precision: 1002466.7
Naive Summation in double precision: 1005000.0049999995

ii. Compensated Summation in single precision: 1005000.0
Compensated Summation in double precision: 1005000.005

iii. Pairwise Summation in single precision: 1005000.0
Pairwise Summation in double precision: 1005000.0049999999

METHOD	ERROR	RUN TIME
Naive summation	Large	2.335 seconds
Compensated summation	Small	2.774 seconds
Pairwise summation	Small	2.696 seconds

(d) **Error:** The pairwise and compensated summation method has the smallest error, followed by the the naive summation method. This means that the pairwise and compensated summation method produces the most accurate results, while the naive summation method produces the least accurate results.

Run Time: The naive summation method is the fastest, followed by the pairwise summation method and the compensated summation method. This means that the naive summation method produces results the quickest, while the compensated summation method produces results the slowest.

(e) Overall, the best method to use depends on the specific needs of the problem we are trying to solve. If accuracy is the most important factor, then the pairwise summation method is the best choice. If speed is more important, then the naive summation method may be a better choice. If you need a good compromise between accuracy and speed, then the compensated summation method is a good option.

Possible Improvements:

- The pairwise summation method can be parallelized with threads, which can significantly improve its performance.