Ph.D. Coursework Tutorial Report

Error in Floating Point Computations

Ishwar Singh

Submitted to: **Prof. Awadesh Prasad**

November 9, 2020

Department of Physics and Astrophysics
University of Delhi
New Delhi, India

Contents

1 Introduction 1

2 Representation of Numbers in Computers 1

List of Figures

1 Introduction

Recently, a new category, computational physics, has been added to the traditional physics classification scheme. This new category acts as a bridge between the traditional theoretical and experimental physics. Computer simulations have become an intergral part of not only every branch of physics but also in other disciplines of science. With each advancement of computer technology, the usage of comupters for scientific computation has also increased substantially.

Although, computers are very powerful, but they have limitations. The representation of real numbers in computers is one of the most import limitations of computers. The computer is a finite system. This inability of the computers i.e. to store 'exact' real numbers, gives rise to the errors in floating point computations.

These errors, although small ($\epsilon \approx 10^{-7}$ for single precision and $\epsilon \approx 10^{-15}$ for double precision), propagates after every usage, might lead to unexpected results (known as garbage values). Therefore, it is imperative to study the errors involved in scientific computations before actually jumping into the field which requires computations.

2 Representation of Numbers in Computers

Computers are binary machines i.e. they understands sequences of binary integers (known as bits) i.e. zeros (0's) and ones (1's). All computations are also done using these arrays of zeros (0's) and ones (1's). These long arrays of binary digits are good only for computers, but it is not user friendly. A user enters his data in decimal numbers and expects his answer in decimal numbers only. The bits are treated together to make a word length. Different computers might have different word lenghts, but this length is generally expressed in bytes, with

$$1 \text{ byte} \equiv 1\text{B} = 8 \text{ bits}.$$

This seems a right time to introduce the single precision and double precision numbers. A single precision number is stored in an array of 32-bits i.e. 4 bytes while the double precision number is stored in an array of 64-bits are 8 bytes. These two terms will reoccur many a times in this report.