# CMPEN 331 Exam 2 Review

Will Bochnowicz

March 14, 2023

## Contents



# 1 Floating Point

The standard of floating point numbers are described in IEEE 754. This was created because of a divergence in early computing where different manufacturers were using different means of storing floating point numbers, making program portability more difficult for developers.

## 1.1 Floating Point Representation

The representation of a floating point number is split into three parts:

| Precision | Sign Bit | Exponent | Fraction |
|---|---|---|---|
| Single | 1 bit | 8 bits | 23 bits |
| Double | 1 bit | 11 bits | 52 bits |

Where the number represented is calculated as $(-1)^S \times (1 + F) \times 2^{E-B}$, $B$ stands for Bias. The bias is used to allow representations of both positive and negative exponent values. In order to calculate it, use the following equation: $2^{(\text{Exponent Bit Count}-1)} - 1$.

This system is not perfect. When doing operations on floating point numbers, it is possible to have errors with your handling of the exponent field. Overflow is when the exponent field after an operation is too great to fit in the allotted number of bits, while underflow occurs when the negative exponent is too large to fit in the exponent field.

#### 1.1.1 Floating Point Problems

How would you represent the value 4.125 in binary? (1)

What is the following single-precision float in decimal? 10101101001100011011000110110001 (2)

### 1.2 Floating Point Operations

## 2 Solutions

1:

2:

# Index