ESMI Applied Math - Summer 2019 13 Jun 2019

Floating Point Representation

Expected Skills.
Students can...

- *explain the how computers represent scalars in single and double float.*
- *identify and explain the important of nonnumeric values.*
- *identify and contrast the three sources of error with floating point arithmetic.*

Discussion Questions

(a) What is the difference between binary and floating point?

(b) What the three parts of a floating point representation and how are they used to represent a value?

(c) How do computers understand `Inf` and `NaN`?

(d) What is the difference between round-off, cancellation, and truncation error?

Floating Point Arithmetic

(a) What is the largest value that can be represented by a double float?

(b) How many real numbers can be represented by a double precision float?

(c) Addition and multiplication of real numbers are both commutative and associative, is this true for floating point values as well?

(d) What is the largest double you can add to 1 billion for MatLab to return 1 billion? Verify your answer.

(e) What is the value of e?

```
e = 1 - 3*(4/3 - 1)
```

What is happening here?

(f) What is the value of e?

```
e = sqrt(1e-16 + 1) - 1
```

What is happening here?

(g) How does truncation error play into our approximation of pi yesterday?