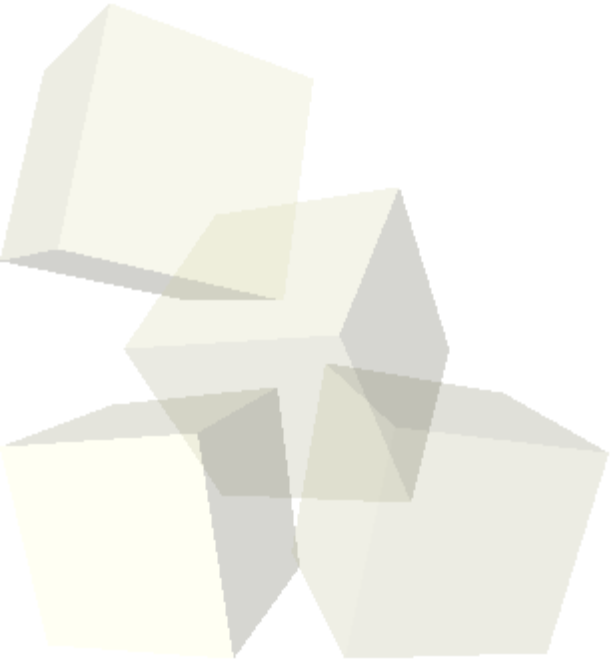
Really Doing Number Systems

8-30-2006

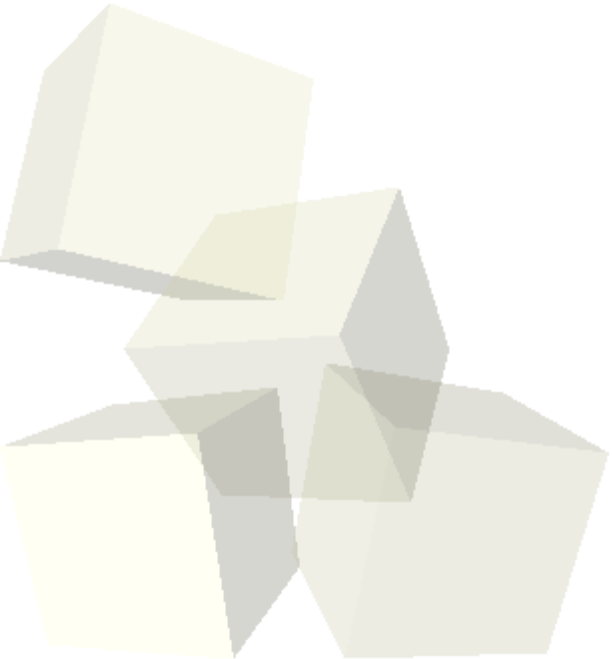
Opening Discussion

- What did we talk about last class?



Closing out Mandelbrot

- Let's real quick finish off what we did last time and give it a bit more functionality. We will also make it more object oriented.



Matlab

- We have Matlab installed on these machines. Why don't you go ahead and start up Matlab and we will do the rest of the day in it, even though what we are going to do could be done in C/C++.

Floating Point Numbers

- Computers represent all numbers in binary (base 2). For fractional numbers we typically us a notation called floating point. This is very much like scientific notation using binary.
- The standard for floating point is IEEE 754. This comes in both single and double precision. Both styles have one byte for the sign followed by bits for an exponent and a fractional part.
- For single precision the exponent gets 8 bits (with a bias of 127) and the fraction/mantissa gets the remaining 23.
- Double precision gets 11 for the exponent (1023 biased) and 52 for the fraction.
- Both assume a leading 1.

Advantages of IEEE 754

- There are some aspects of the floating point representation that are interesting to note. In particular, the differences between integer and floating point representation.
  - The mantissa has a sign bit instead of using 2s complement.
  - The exponent has a bias instead of being 2s complement.
- Combined with the ordering of sign, exponent, mantissa you get much faster comparisons.
- The leading 1 makes 0 a special form. Some implementations have unnormalized forms for numbers below the range of the standard.

What it means to you

- The details of IEEE 754 don't matter to most people. What does matter is that these numbers are not the “real” numbers from math.
- Having limited precision means some properties, like associativity, don't always hold.
- The real key is that not all numbers can be represented perfectly. This is similar to the fact that we can't represent numbers like 1/3rd in decimal with a finite number of digits.
- Similarly, there are some types of operations that should be avoided.
  - Don't subtract two numbers that differ by an amount much smaller than their magnitude.
  - Don't divide by very small numbers.

Advantages of Tools

- These rules of thumb are part of the reason many scientists will use tools like Matlab or existing libraries. Often these libraries will use special alternate formulas in situations where the “normal” one is poorly behaved.
- It is often difficult enough just to get the math right without worrying too much about how the numerics will behave.

Closing

- Do you feel clear on how floating point numbers work on computers?
- Read chapters 4 & 5 for next class. We'll just proceed with Matlab until we think of something better to do.
- Also remember that assignment #1 is due on Friday.