# Floating Point Errors and Norms

## Fundamental Axiom of Floating Point Arithmetic

For any of the four arithmetic operations (addition, subtraction, multiplication, division) we have the following error bound:

Such that . In other words, there exist finite bounds on floating point error for arithmetic.

Commonly u is (machine epsilon). [Looking at the literature, there seem to be a couple of alternative formulations.]

**Example** of the kind of problem we’d like to be able to solve

Suppose you have a matrix . We would like to show there is an upper bound on the errors in representation as . In other words

**Homework**:

* Show that the is equal to the maximum norm of the columns of A.
* Show that the is equal to the maximum norm of the rows of A.
* Show that .

## Matrix Norms

We might be interested in several different matrix norms. If we consider a matrix as a collection of vectors or points, as in the example above, we might be interested in the Frobenious (aka Euclidean) norm for measuring the error due to floating point representation. If we consider a matrix as a linear transformation (i.e. as the coefficients in a collection of linear equations) then we might be interested in measuring the error due to using the matrix as an operator. Vector subordinate matrix norms get at this concept. A generic subordinate matrix norm is defined as

, where .

For an norm we can ask what sort of vector x will maximize this quantity, and for any particular matrix A, what will the actual quantity be? Now, an norm is just a sum of the values of a vector so we can simplify our search somewhat by just considering vectors x whose elements sum to 1. Each element of Ax is a weighted sum along a row of A, i.e.

And the norm of Ax is

Rearranging the summations we have

Which is just , summed, or a weighted sum of the column vectors. This sum is maximized by giving all the weight ( to the column vector with the largest sum. In other words, the standard basis vector where j is the column with the greatest sum or norm. Then the norm itself is just the sum of the elements of column j.

For a we can go through a similar process. Here we are searching for the which has the maximum value. In this case we are interested in vectors x where the maximum value of all elements is 1 (because we are forming ). The normed x which will generate the largest values for each is therefore the x where every element is 1, the maximum value allowed. So we have

And the maximal value of will derive from the row of A which has the greatest sum, that is, the row of A with the greatest norm.

## Dot Product Errors

Let be vectors. We would like to calculate , the error in the floating point representation.

**Lemma**: Let be vectors, and . Then

Proof:

Let denote the sum s after p iterations, .

We start with .

At each iteration