MATH 320: CLASS NOTES

1. Error - Chapter 4

An algorithm is a set of rules that gives instructions for a sequence of operations. Often takes some input and performs a calculation.

An iterative algorithm is an algorithm that iterates (or repeats), providing a better answer at each step.

Definition 1.1. Suppose an algorithm yields an approximate answer $\hat{x}$ for some unknown x.

(1) Error is defined as $E_t = x - \hat{x}$.

(2) Relative error is $\epsilon_t = \dfrac{x - \hat{x}}{x}$.

Definition 1.2. Suppose an iterative algorithm yields a sequence of approximations $(\hat{x}_k)_{k=1,2,\ldots,n}$. Then we define:

(1) Approximate error is $E_a = x_n - x_{n-1}$

(2) Approximate relative error: $\epsilon_a = \dfrac{x_n - x_{n-1}}{x_n}$.

We can take a stopping criterion ϵ_s such that the algorithm stops when $|\epsilon_a| < \epsilon_s$.

Error can also arise from roundoff errors in computation.

1.1. How are numbers stored in computers.

Definition 1.3. A base-n number system is a system of representing rational numbers as a sum of powers of n.

The base-10 number system, or decimal system, uses powers of 10.

Example 1.4. 134.57 is equal to

$$1 * 10^2 + 3 * 10^1 + 4 * 10^0 + 5 * 10^{-1} + 7 * 10^{-2}$$

Definition 1.5. The signed magnitude method sets aside the first bit to denote the sign, and the remaining bits denote the number. 1 means a negative number. Since ± 0 is redundant, the number -0 is taken to be the number -2^{n-1}.

Definition 1.6. Double precision floating point representation looks like

$$\pm s \times b^e.$$

s is called the significand (or mantissa), b is the base of the number system, and e is the exponent. s is written with one (nonzero) digit to the left of the decimal point.

When $b = 10$, this is called scientific notation. Computers use $b = 2$ with 11 bits set aside for a signed exponent, and 52 bits for a mantissa.

Definition 1.7. An overflow error occurs when numbers are larger than the allotted maximum. An underflow error occurs if it is smaller than the minimum.

Definition 1.8. Roundoff error happens in computations when the true answer is rounded off losing some small ϵ.

Machine epsilon is defined as the smallest value ϵ such that $1 + \epsilon \neq 1$, i.e. ϵ is the smallest value within our precision.

If you were to add $.1 + .01$ in a number system with only one significant bit in the mantissa, the second term would be rounded off. Furthermore if a number does not have a finite binary expansion, even simple operations can yield small roundoff errors that accumulate.

Example 1.9. Consider $\frac{1}{10}$ in binary. Suppose we round off to 5 digits. Perform the addition.

Definition 1.10. An infinite series is the limit of n-th partial sums of a sequence as $n \to \infty$.

For example, geometric series.

Definition 1.11. Truncation error is the error made by approximating an infinite sum by a finite sum.

Theorem 1.12 (Taylor's theorem)**.** *Let $k \geq 1$ be an integer, and let $f : \mathbb{R} \to \mathbb{R}$ be k-times differentiable at a. Then*

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k$$

with $h_k(x) \to 0$ as $x \to a$.

Proof. Use L'Hopital's rule k times. □

Explicit formula for remainder:

Theorem 1.13. *Suppose f is $k+1$ times differentiable on the open interval (x, a) and continuous on the closed interval, then the remainder term*

$$R_k = \frac{f^{(k+1)}(x^*)}{(k+1)!}(x-a)^{k+1}.$$

where x^* is a number in the open interval (x_0, x).

Definition 1.14. Let f and g be two functions of a real variable. Then, $f(x) = O(g(x))$ as $x \to a$ if there are positive numbers M, δ such that $|f(x)| \leq M|g(x)|$ for all x such that $|x - a| \leq \delta$. Read aloud as 'f is big-O of g'.

$R_k(x)$ is big-O of $(x-a)^{k+1}$.

Corollary 1.15. *Suppose we want an error bound for an approximation $f(x)$ using a Taylor polynomial of degree k. If we have a uniform bound on $f^{(k+1)}$, this gives a bound on R_k.*

Example 1.16.

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

Let us approximate $\log(1.5)$ using the first three terms of the the Taylor series. $.5 - .125 + .0417 = .4167$. The remainder term is

$$R_3 = \frac{f^{(4)}(x^*)}{4!}(.5)^4$$

The fourth derivative is strictly decreasing:

$$f^{(4)}(x) = (-1)^4 4!(1+x)^{-4} \leq 4!. \Rightarrow R_3 \leq .0625.$$

Therefore, $\log(1.5) \in (.3542, .4792)$. True answer $\approx .4055$.