MTH303

Section 1.3: Error Analysis

R.Touma

These lecture slides are not enough to understand the topics of the course; they could be used along with the textbook

The numerical solution of a mathematical problem is an approximation of the analytical exact solution. The precision or the accuracy of the numerical solution can be diminished in several ways. It is important to understand and ....

Definition

Let $\hat{p}$ denote an approximation to p. The *absolute error* is $E_p = |p - \hat{p}|$, and the *relative error* is $R_p = \frac{|p-\hat{p}|}{|p|}$ provided that $p \neq 0$.

Example

Let $\hat{x} = 3.14$ denote an approximation to $x = 3.141592$.

The absolute error is:

$$E_x = |x - \hat{x}| = |3.141592 - 3.14| = 0.001592$$

The relative error is:

$$R_x = \frac{|x - \hat{x}|}{|x|} = \frac{|3.141592 - 3.14|}{|3.141592|} = 0.000507$$

In this case there is not too much difference between E_x and R_x, and either could be used to determine the accuracy of $\hat{x}$.

Let $y = 1,000,000$ and $\hat{y} = 999,996$.
The absolute error is:

$$E_y = |y - \hat{y}| = |1,000,000 - 999,996| = 4$$

The relative error is:

$$R_y = \frac{|y - \hat{y}|}{|y|} = \frac{|1,000,000 - 999,996|}{|1,000,000|} = 0.000004$$

In this case the error E_y is large while the relative error R_y is small; $\hat{y}$ would probably be considered a good approximation to y.
Let $z = 0.000012$ and $\hat{z} = 0.000009$.
The absolute error is:

$$E_z = |z - \hat{z}| = |0.000012 - 0.000009| = 0.000003$$

The relative error is:

$$R_z = \frac{|z - \hat{z}|}{|z|} = \frac{|0.000012 - 0.000009|}{|0.000012|} = 0.25$$

z is of magnitude of 10^{-6}. The error E_z is the smallest of all three cases but the relative error R_z is the largest (about 25%) and thus $\hat{z}$ is not a good approximation to z.

Definition

The number $\hat{p}$ is said to *approximate* p to d significant digits if d is the largest nonnegative integer for which

$$\frac{|p-\hat{p}|}{|p|} < \frac{10^{1-d}}{2}$$

Example

Determine the number of significant digits for the approximations $\hat{x}, \hat{y}$, and $\hat{z}$.

(a) $\hat{x} = 3.14$ and $x = 3.141592$, then $|x-\hat{x}|/|x| = 0.000507 < 0.005 = \frac{10^{-2}}{2}$. We set $1-d = -2$ or $d = 3$. Therefore $\hat{x}$ approximates x to three significant digits.

(b) If $y = 1,000,000$ and $\hat{y} = 999,996$, then $|y-\hat{y}|/|y| = 0.000004$ $< 0.000005 = \frac{10^{-5}}{2}$; We set $1-d = -5$, thus $d = 6$. Therefore $\hat{y}$ approximates y to six significant digits.

(c) If $z = 0.000012$ and $\hat{z} = 0.000009$, then $|z - \hat{z}|/|z| = 0.25 < \frac{10^{-0}}{2}$. We set $1 - d = 0$ or $d = 1$. Therefore $\hat{z}$ approximates z to one significant digit.

Truncation Error

The truncation error is the error introduced when a complicated mathematical expression is replaced with a more elementary formula. Usually complicated functions are replaced with a truncated Taylor series. For example, the infinite Taylor series

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^3}{3!} + \frac{x^8}{4!} + \frac{x^{10}}{5!} + \cdots + \frac{x^{2n}}{n!} + \cdots$$

might be replaced with just the first 5 terms $1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$. This is useful when approximating an integral numerically.

Example Given that $\int_0^{1/2} e^{x^2}\, dx = 0.544987104184 = p$. Determine the accuracy of the approximation obtained by replacing the the integrand

$f(x) = e^{x^2}$ with the truncated Taylor series $P_8(x) = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$.
Integrating term by term, we obtain:

$$\begin{aligned}
\int_0^{\frac{1}{2}} \left(1 + x^2 + \frac{x^4}{2!} + \frac{x^3}{3!} + \frac{x^8}{4!}\right) dx &= \left[x + \frac{x^3}{3} + \frac{x^5}{5(2!)} + \frac{x^7}{7(3!)} + \frac{x^9}{9(4!)}\right]_{x=0}^{x=\frac{1}{2}} \\
&= \frac{1}{2} + \frac{1}{24} + \frac{1}{320} + \frac{1}{5376} + \frac{1}{110,592} \\
&= 0.544986720817 = \hat{p}.
\end{aligned}$$

Now we compute $|p - \hat{p}|/|p|$, we obtain:

$$|p - \hat{p}|/|p| = 0.703443 \times 10^{-6} < 10^{-5}/2$$

Thus $1 - d = -5$, so $d = 6$, and $\hat{p}$ approximates the p to six significant digits.

Round-off Error

As we previously saw, usually, computers don't store the exact value of a given mathematical quantity, but instead, an approximation is being stored. This representation of real numbers is limited to the fixed precision of the mantissa. This is called the round-off error. For example when the number $1/10 = 0.0\overline{0011}_{two}$ is stored in the computer, its binary representation must be truncated; the actual number stored may undergo chopping or rounding of the last digit.

Chopping Off versus Rounding Off

Let p denote any real number expressed in *normalized decimal form*:

$$p = \pm 0.d_1 d_2 d_3 \cdots d_k d_{k+1} \cdots \times 10^n$$

where $1 \leq d_1 \leq 9$ and $j \leq d_j \leq 9$ for $j > 1$.
If k is maximum number of decimal digits carried in the floating-point

computation of a computer; then p is represented by $fl_{chop}(p)$ and is given by:

$$fl_{chop}(p) = \pm 0.d_1 d_2 d_3 \cdots d_k \times 10^n$$

The number $fl_{chop}(p)$ is called the *chopped floating-point representation* of p. An alternative $k-$digits representation is the *rounded floating-point representation* $fl_{round}(p)$ which is given by:

$$fl_{chop}(p) = \pm 0.d_1 d_2 d_3 \cdots r_k \times 10^n$$

The digit r_k is obtained by rounding the the number $d_k d_{k+1} d_{k+2} \cdots$ to the nearest integer.

Example The real number $p = \frac{22}{7} = 3.14285714285714285 \cdots$ has the following six digits representations:

$$\begin{aligned} fl_{chop}(p) &= 0.314285 \times 10^1 \\ fl_{round}(p) &= 0.314286 \times 10^1 \end{aligned}$$

Usually computers use some form of rounded floating point representation

method.

$O(h^n)$ Order of Approximation

Definition The function $f(h)$ is said to be *big Oh* of $g(h)$, denoted by $f(h) = \mathcal{O}(g(h))$, if there exist constants C and c such that

$$|f(h)| \leq |Cg(h)|, \quad \text{whenever } h \leq c. \tag{1}$$

Definition 0.1 *The function $f(h)$ is said to be big Oh of $g(h)$, denoted $f(h) = \mathcal{O}(g(h))$, if there exist constants C and c such that*

$$|f(h)| \leq C|g(h)|, \quad \textit{whenever } h \geq c. \tag{2}$$

Example 0.1 *Consider the function $f(x) = x^2 + 1$ and $g(x) = x^3$. Since $x^2 \leq x^3$ and $1 \leq x^3$ for $x \geq 1$, it follows that $x^2 + 1 \leq 2x^3$ for $x \geq 1$. Therefore, $f(x) = \mathcal{O}(g(x))$*

Definition 0.2 *The sequence $\{x_n\}_{n=1}^{\infty}$ is said to be of order big Oh of $\{y_n\}_{n=0}^{\infty}$ (denoted $x_n = \mathcal{O}(y_n)$), if there exist constants C and N such that*

$$|x_n| \leq C|y_n|, \quad \textit{whenever } n \geq N. \tag{3}$$

Example 0.2 *Consider the sequence $x_n = \frac{n^2-1}{n^3}$.*
x_n is of the order $\mathcal{O}(\frac{1}{n})$, since $\frac{n^2-1}{n^3} \leq \frac{n^2}{n^3} = \frac{1}{n}$ whenever $n \geq 1$

Remark 0.1 *When a function $f(h)$ is approximated by the function $p(h)$, the error bounds is known to be $M|h^n|$.*

Definition 0.3 *We say that the function $p(h)$ approximates the function $f(h)$ with order of approximation $\mathcal{O}(h^n)$ and we write:*

$$f(h) = p(h) + \mathcal{O}(h^n),$$

if there exist a constant $M > 0$ and a positive integer n so that

$$\frac{|f(h) - p(h)}{|h|^n} \leq M \quad \textit{for sufficiently small h.}$$