Chapter 4: Roundoff and Truncation Errors

Outline

- Errors
  - Accuracy and Precision
  - Error Definitions
- Roundoff Errors
- Truncation Errors
  - The Taylor Series
  - The Remainder for the Taylor Series Expansion
  - Numerical Differentiation
  - Error Propagation
  - Stability and Condition
- Total Numerical Error
- Blunders, Model Errors, and Data Uncertainty
Approximations and Round-off Errors

- All numerical methods involve the approximation and round-off of numbers, and it will be extremely important for us to be able to quantify the effect these approximations have on the results of our calculations.
- The **significant digits** of a number are those that can be used with confidence, usually corresponding to a certain number of certain digits plus one estimated digit.

Approximations and Round-off Errors

- **Accuracy** refers to how closely a computed or measured value agrees with the true value.
- **Precision** is related to repeatability and refers to how closely individual computed or measured values agree with each other.
- Often we will have to make decisions to accept lower levels of accuracy in order to efficiently solve problems.
- The techniques we choose must be sufficiently accurate to meet the requirements of a particular problem and must be precise enough to meet the requirements of engineering design.
- We will often use the term **error** to both inaccuracy and imprecision.
Definitions of Error

- There are two basic types of errors inherent in numerical methods.

- **Truncation errors** are those which result when approximations are used to represent exact mathematical procedures.

- **Roundoff** errors result when numbers with limited significant figures are used to represent exact numbers.

Both types of error can be formulated as:

\[ \text{True value} = \text{approximation} + \text{error} \]

or

\[ E_t = \text{True value} - \text{approximation} \]

where \( E_t \) is the true error.

- The relative error is defined as

\[ \epsilon_t = \frac{\text{true error}}{\text{true value}} \]
Definitions of Error

In the application of numerical methods, exact values are known when we have an analytical solution to compare. Instead, we will normalize the error to the best available estimate of the true value, that is, the approximation itself:

$$\varepsilon_a = \frac{\text{approximate error}}{\text{approximation}}$$

In addition, we need to estimate the error without knowing the true value. Certain numerical methods involve iterative calculations, where a calculation is made based on the results of a previous calculation to compute better and better approximations:

$$\varepsilon_a = \frac{\text{current approximation} - \text{previous approximation}}{\text{current approximation}}$$
Definitions of Error

- Errors can be either positive or negative. Normally we are not concerned with the sign but whether it is smaller than some pre-specified percent tolerance, $\varepsilon_s$.
- Calculations are performed until the absolute value of the relative approximate error falls below the tolerance:

$$|\varepsilon_{\alpha}| < \varepsilon_s$$

- To find a result that is correct to at least $n$ significant digits:

$$\varepsilon_s = (0.5 \times 10^{2-n})\%$$

- See Example 4.1

Roundoff Errors

Without going into too much detail, computers (and calculators) retain only a fixed number of significant figures during calculations. This leads to what is known as roundoff error. The following aspects are always important to keep in mind when performing computer or calculator calculations:

- There is a limited number of quantities that may be represented. Only pre-defined types of numbers can be stored in computers, and each type has a defined range associated with it. Any attempt to use numbers outside of these ranges results in overflow errors.
Roundoff Errors

There are only a finite number of quantities that can be represented within the range. Due to the fact that computers use a fixed number of significant figures, the degree of precision is limited. Clearly, irrational numbers cannot be precisely represented but most rational numbers also cannot be represented precisely either. These types of errors are called quantizing errors.

The interval between numbers increases as the numbers grow in magnitude. Floating point representations preserve significant digits, but this feature also means that quantizing errors are proportional to the magnitude of the number being represented.

Even though round-off errors can be significant, most engineering calculations can be carried out with more than acceptable precision on most computers and calculators. When precision is of the utmost importance, the use of extended precision quantities can greatly mitigate the effects of roundoff errors and is recommended.
Truncation Errors and the Taylor Expansion: The Taylor Series

The Taylor Series is an important mathematical tool used to approximate the values of a function, and we will make extensive use of its properties. Recall from calculus that if a function \( f \) and its \( n+1 \) derivatives are continuous in the vicinity of a point \( x_p \) then its value at a nearby point \( x_{i+1} \) can be expressed as

\[
f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \frac{f^{(3)}(x_i)}{3!}(x_{i+1} - x_i)^3 + \frac{f^{(4)}(x_i)}{4!}(x_{i+1} - x_i)^4 + R_n
\]

where the remainder term accounts for the rest of the terms of the infinite series.

\[
R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x_{n+1} - x_n)^{(n+1)}
\]

The subscript \( n \) denotes that this is an \( nth \)-order approximation and \( \xi \) is a value that lies between point \( x_i \) and point \( x_{i+1} \). Note that this remainder is an exact representation of the error that would be a result of using a finite number of terms of the Taylor Series.
The Taylor Series

We will often simplify the Taylor series to be defining a step size $h = x_{i+1} - x_i$ and writing the previous Eq. as

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!} h^2 + \frac{f'''(x_i)}{3!} h^3 + \frac{f^{(4)}(x_i)}{4!} h^4 + R_n$$

where the remainder is now given as

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!} h^{n+1}$$

In general, $n$th-order Taylor expansions are exact for $n$th-order polynomials. For other differentiable and continuous functions, such as exponentials and transcendental functions, a finite number of terms does not exactly represent the function. Each additional term adds some improvement to the approximation. The assessment of how many terms to include in order to be “close enough” for a given problem is the essence of numerical analysis.
The Taylor Series

Even though the Taylor Series will be the basis of a number of techniques we will study, its use has two important drawbacks. First, the value of $\xi$ is not known exactly, only that it is between the values $x_i$ and $x_{i+1}$. This means that we will rarely be able to precisely know the error in our approximations. Secondly, since we will often not know the function that we are approximating (if we did, there would be no need to approximate it!), and its derivatives will also be unknown. However, the properties of the Taylor Series are very useful in estimating the errors associated with a numerical technique, particularly truncation errors.

The Taylor Series

In any numerical technique, we have control over certain parameters of the analysis. One of the most important is the step size $h$. We can look at the remainder Eq. as $R_n = O(h^n+1)$, which means that the truncation error is “of order” $h^{n+1}$, that is, it is proportional to $h^{n+1}$. This gives us a guideline with which to change an analysis to increase accuracy and reduce error.
The Taylor Series

As an example of how this is used in numerical analysis, let's estimate the first derivative of the function $f$ in the vicinity of $x_i$.

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + R_i$$

solving for the first derivative gives

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} + \frac{R_i}{h}$$

$$\frac{R_i}{h} = \frac{f^{(2)}(\xi)}{h(n+1)!} h^2 = O(h)$$

Therefore, the estimate of the first derivative has a truncation error of $O(h)$, which means that the approximation is proportional to the step size $h$. Consequently, if we halve the step size, we would expect to halve the truncation error. This particular approximation for the first derivative is called the first forward difference.
Numerical Differentiation

The Taylor expansion of a function can be represented by:

\[ f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f'''(x_i)}{3!}h^3 + \cdots \quad (*) \]

The Taylor Series can also be expanded backwards to calculate a previous value based on a current value.

\[ f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{f''(x_i)}{2!}h^2 - \frac{f'''(x_i)}{3!}h^3 + \cdots \quad (**) \]

Subtracting Eq. (**) from Eq. (*) and solving for the first derivative yields the centered difference approximation:

\[ f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + O(h^2) \]

Note that this error term is proportional to \( h^2 \), so that if we halve the step size, the error is quartered.
**Error Propagation**

Using the first-order Taylor Series, it can be shown the errors associated with common mathematical operations using inexact numbers \( \bar{u} \) and \( \bar{v} \) can be summarized as

<table>
<thead>
<tr>
<th>Operation</th>
<th>Estimated Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addition</td>
<td>( \Delta(\bar{u} + \bar{v}) )</td>
</tr>
<tr>
<td>Subtraction</td>
<td>( \Delta(\bar{u} - \bar{v}) )</td>
</tr>
<tr>
<td>Multiplication</td>
<td>( \Delta(\bar{u}\bar{v}) )</td>
</tr>
<tr>
<td>Division</td>
<td>( \Delta(\frac{\bar{u}}{\bar{v}}) )</td>
</tr>
</tbody>
</table>

**Stability and Condition**

The *condition* of a mathematical problem is a measure of its sensitivity to changes in its input values. A computation is *numerically unstable* if the uncertainty of the input values is grossly magnified by the numerical method.

Using a first-order Taylor Series,

\[
f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x})
\]

The relative error of \( f(x) \) is

\[
\frac{f(x) - f(\bar{x})}{f(\bar{x})} \geq \frac{f'(\bar{x})(x - \bar{x})}{f(\bar{x})}
\]
Stability and Condition

The relative error of \( x \) is \( \frac{x - \hat{x}}{\hat{x}} \)

The condition number is the ratio of these relative errors

\[
\text{Condition number} = \frac{\hat{x} f'(\hat{x})}{f(\hat{x})}
\]

and is a measure of how the uncertainty in \( x \) is magnified in \( f(x) \). A value of 1 means that the function’s relative error matches the relative error in \( x \). A value greater than 1 means that the relative error is amplified; values less than one mean that the error is attenuated. Functions with large condition numbers are said to be ill-conditioned.

Total Numerical Error

- The total numerical error is the summation of the truncation and round-off errors. In general, the way to minimize round-off errors is to increase the number of significant figures used in the calculation. Round-off error increases due to subtractive cancellation and as the number of calculations in an analysis grows. Truncation error can be reduced by decreasing the step size, which leads to increased numbers of calculations and an increase in round-off errors.
Total Numerical Error

Numerical errors can be controlled with certain general guidelines:

- Avoid subtracting two nearly-equal numbers to avoid subtractive cancellation
- Work with the smallest numbers first and progressively include the largest
- The use of numerical experiments and sensitivity analyses may provide insight to numerical strategies
- Performing the analysis with different modeling and numerical techniques may provide more confidence in the final results.

Blunders

- Gross errors, or *blunders*, are sometimes unavoidable. Sources of blunders are mainly based in human imperfection and include errors of the modeling process and incorrect programming. We will frequently mention techniques with which to avoid blunders and check our numerical calculations.
Formulation Errors

- Formulation or modeling errors are due to incomplete mathematical models. It is important to note that formulation errors cannot be resolved with increased numerical analysis. Poorly conceived models cannot yield useful results, no matter how sophisticated the analysis techniques.

Data Uncertainty

- All physical data is subject to variation and uncertainty, exhibiting both inaccuracy and imprecision. Where such variation is important to interpret the results to be meaningful, the analysis has to be carried out with data that is described in statistical terms of a central measure and the degree of spread about that central measure.