3.2 Numerical Errors

Several different ways to approximate a continuous derivative by a discrete one were presented in the previous section. In these approximations, the function was only evaluated at discrete points and derivatives were reduced to differences between these discrete values. Hence, a differential calculus problem was transformed into an algebraic problem. Finite-difference equations are generated when the continuous derivatives in partial differential equations are approximated by their discrete counterparts. However, this transformation of a PDE into a FDE introduces different types of error in the solution. These errors are discussed in the present section.

3.2.1 Norms

Computers store the discrete version of continuous functions as arrays or vectors. Hence, it is possible to evaluate the error of any approximation at each grid point. However, too much unnecessary information is generated. The alternative is to include this information at all grid points into one single number. Norms, such as the L_p-norm, achieve that very purpose and have become the standard in numerical analysis.

The absolute value $|x|$ measures the size of scalar x. This scalar measurement has the following well-known properties

$$\begin{aligned} &|x| \geq 0 \quad \text{where} \quad |x| = 0 \quad \text{if and only if} \quad x = 0 \ , \\ &|\alpha\, x| = |\alpha|\,|x| \quad \text{for any scalar} \quad \alpha \ , \\ &|x + y| \leq |x| + |y| \quad \text{for any scalars} \quad x \quad \text{and} \quad y \ . \end{aligned} \tag{3.32}$$

Similarly, the norm $||\mathbf{x}||$ measures the size of vector $\mathbf{x}$. This is also a scalar measurement and has the following properties

$$\begin{aligned} &||\mathbf{x}|| \geq 0 \quad \text{where} \quad ||\mathbf{x}|| = 0 \quad \text{if and only if} \quad \mathbf{x} = 0 \ , \\ &||\alpha\, \mathbf{x}|| = |\alpha|\,||\mathbf{x}|| \quad \text{for any scalar} \quad \alpha \ , \\ &||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}|| \quad \text{for any vectors} \quad \mathbf{x} \quad \text{and} \quad \mathbf{y} \ . \end{aligned} \tag{3.33}$$

Hence, norms are the natural extensions of absolute value for scalars, but applied to vectors. However, there are infinite types of norm but only one absolute value. The most popular norm is the L_p-norm defined as

$$||\mathbf{x}||_p \,=\, \sqrt[p]{|x_1|^p \,+\, |x_2|^p \,+\, \ldots \,+\, |x_{N-1}|^p \,+\, |x_N|^p} \ , \tag{3.34}$$

where x_1 to x_N are the components of vector $\mathbf{x}$ in any orthogonal basis and p is any integer. The most used L_p-norms are the L_1-norm

$$||\mathbf{x}||_1 \,=\, |x_1| \,+\, |x_2| \,+\, \ldots \,+\, |x_{N-1}| \,+\, |x_N| \ , \tag{3.35}$$

the L_2-norm or Euclidean norm

$$||\mathbf{x}||_2 \;=\; \sqrt{x_1^2 \,+\, x_2^2 \,+\; \ldots \;+\, x_{N-1}^2 \,+\, x_N^2}\ , \tag{3.36}$$

and the L_∞-norm or maximum norm

$$||\mathbf{x}||_\infty \;=\; \max\Big[\,|x_1|\,,\,|x_2|\,,\;\ldots\;,\,|x_{N-1}|\,,\,|x_N|\,\Big]\ . \tag{3.37}$$

It is also possible to use norms to measure the absolute different $||\mathbf{x}-\mathbf{y}||$ and the relative or percentage difference $||\mathbf{x}-\mathbf{y}||/||\mathbf{x}||$ between vectors $\mathbf{x}$ and $\mathbf{y}$. Furthermore, the choice between different norms depends on the application. L_1- and L_2-norms represent a measure of average error and are more useful in minimization processes. On the other hand, the L_∞-norm is the maximum error and is more useful when a more conservative error analysis of numerical solutions is necessary.

3.2.2 Round-off Errors

Most computers today store information using packages of 32 bits. A binary digit or bit is a digit in a base-two number system, being either 0 or 1. Hence, these 32-bit packages can represent 2^{32} different types of information, which include characters, integers, real numbers, etc.

Virtually all numerical methods perform most of their calculations using real numbers. Theoretically, there is an infinite continuum of real numbers. However, due to the finite amount of information stored in 32-bit packages, there must be a maximum computer number and a minimum computer number. Furthermore, there must be a finite relative spacing ϵ between real computer numbers, called machine precision. Usually, computers use a single 32-bit package to store information. Hence, the maximum, minimum and machine precision computer numbers in a base-ten number system are approximately around 10^{39}, 10^{-39} and 10^{-8}, respectively. Computers can also use two, four or more 32-bit packages to store information. Double precision calculations modify the three previous computer numbers to approximately 10^{309}, 10^{-309} and 10^{-16}, respectively.

Round-off error can now be defined as the approximation made by computers when storing real numbers using a finite number of digits. Numbers greater than the maximum computer number or smaller than the minimum computer number cause overflow or underflow, respectively. The three main mechanisms that amplify round-off errors are known as systematic operation, catastrophic cancelation and operation sensitivity.

Systematic round-off error generation occurs with the repetitive addition or subtraction of numbers that are $1/\epsilon$ orders of magnitude different. For instance, a computer will say that $1+\epsilon=1$ according to the definition of machine precision. Hence, the final result will still be 1 even if this operation is repeated $1/\epsilon$ times, which means a relative error of 100%. The best way to minimize this type of error is to always add or subtract numbers from smallest to largest. In this case, the error would be reduced to $\epsilon\,\%$.

Catastrophic cancelation is the magnification of round-off errors through the subtraction of nearly equal numbers. For instance, consider the operation $\sqrt{x+1}-\sqrt{x}$ when $x = 1984$. A computer performing single precision calculations will say $\sqrt{x+1}-\sqrt{x} = 44.553339-44.542115 = 1.1224\times 10^{-2}$, which means three accuracy digits were lost. The best way to minimize this type of error is to rewrite the operation in an alternative way to avoid subtracting similar numbers. In this case, $\sqrt{x+1}-\sqrt{x} = 1/(\sqrt{x+1}+\sqrt{x}) = 1/(44.553339+44.542115) = 1.1223917\times 10^{-2}$, and the relative error is now reduced to machine precision again.

Operation sensitivity round-off errors occur in problems where small changes in the model parameters lead to large changes in the model solution. This hypersensitivity can be due to numerical and/or physical instability. An algebraic system of equations whose matrix is ill-conditioned often has operation sensitivity and is an example of the former. In such cases, the matrix determinant is close to zero. This occurs when two or more equations are close to being linearly dependent or when the matrix eigenvalues are orders of magnitude different. Furthermore, ill-posed problems also cause numerical instability due to the lack of model consistency. Physical instability, also known as thermal and/or hydrodynamic instability in the context of transport phenomena, also leads to operation sensitivity. Nonlinearity in PDEs is responsible for this unstable physical behavior.

3.2.3 Discretization Errors

Round-off errors arise from the computer representation of numbers. On the other hand, discretization errors arise from the computer representation of functions and functional operators. A functional operator will be continuous or discrete if the function it operates is continuous or discrete, respectively. Hence, discretization is the process of replacing continuous functions and functional operators by discrete functions and functional operators.

However, computers can only represent functions by finite sequences of numbers because they can only perform the four basic arithmetic operations: addition, subtraction, multiplication and division. For instance, the logarithmic function is defined as

$$\ln[x] = \int_1^x \frac{1}{\zeta}\, d\zeta \ , \tag{3.38}$$

which means a numerical integration procedure must be implemented using only the four basic arithmetic operations. The alternative is to represent this function as a truncated version of the infinite sequence

$$\ln[x] = \frac{2\,x-1}{x+1} + \frac{2}{3}\left(\frac{x-1}{x+1}\right)^2 + \frac{2}{5}\left(\frac{x-1}{x+1}\right)^5 + \ldots \ , \tag{3.39}$$

introducing discretization error in the process.

Several continuous functions can be represented as discrete functions using Taylor-series expansions. In general, one may write

$$f(x) = f(x_0) + \left.\frac{df}{dx}\right|_{x_0} + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_0} + \frac{1}{6}\left.\frac{d^3f}{dx^3}\right|_{x_0} + \dots\ , \tag{3.40}$$

which can be calculated using the basic four arithmetic operations if the same can be said about the function derivatives.

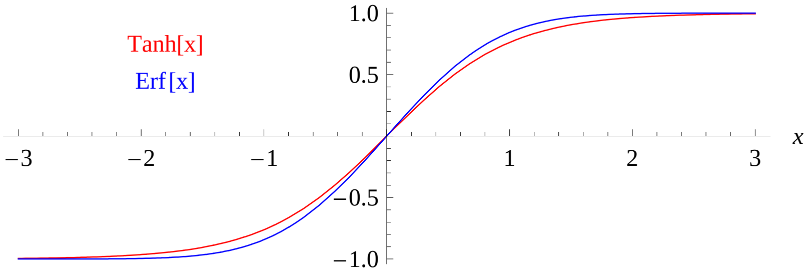

Figure 3.2: Hyperbolic tangent and error functions.

The hyperbolic tangent and error functions, commonly used in many *ad hoc* flow models, have a similar behavior as shown in Figure 3.2. Their Taylor-series expansions around $x = 0$ are given by

$$\tanh[x] = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 + O(x^7) \quad \text{and} \tag{3.41}$$

$$\operatorname{erf}[x] = \frac{2}{\sqrt{\pi}}x - \frac{2}{3\sqrt{\pi}}x^3 + \frac{1}{5\sqrt{\pi}}x^5 + O(x^7)\ . \tag{3.42}$$

The relative error associated with their difference at $x = 0.1$ is 3.3603423×10^{-3}, 1.5807875×10^{-5} and 2.1052435×10^{-7} with one, two and three terms in the series expansion, respectively. Two details must be mentioned. First, increasing the number of terms in expansions above will not reduce the relative error below machine precision. The computer user must remember this fact in order to avoid excessive algebra in the implementation of their codes. Second and most important, high accuracy is not the same as high order of accuracy. The expressions above will produce inaccurate results if evaluated with $|x| \gg 1$, even though they have high orders of accuracy.

The previous examples illustrate one important difference between the two types of error discussed so far. Contrary to round-off errors, discretization errors can be controlled by the user since truncation order can be tuned in order to generate the necessary accuracy. Furthermore, one can define truncation error as the error introduced when infinite sequences or series are replaced by finite sequences or series. Hence, truncation error is a type of discretization error. This is the most studied type of error in computational transport phenomena because it appears explicitly when PDEs are transformed into FDEs, i.e., when PDEs are discretized.

3.2.4 Truncation Errors

This section is divided in two major parts. First, the truncation errors associated with all finite-difference approximations to continuous derivative shown so far are analyzed. Then, this concept is extended to include the truncation errors generated when constructing finite-difference representations of partial differential equations.

Taylor-Series Expansion

The truncation error of forward-difference approximation (3.5) is obtained by subtracting it from Taylor-series expansion (3.4), which yields

$$T.E. = -\left.\frac{d^2f}{dx^2}\right|_i \frac{\Delta x}{2} - \left.\frac{d^3f}{dx^3}\right|_i \frac{(\Delta x)^2}{6} - \left.\frac{d^4f}{dx^4}\right|_i \frac{(\Delta x)^3}{24} - \cdots \,, \tag{3.43}$$

and represents the error introduced by truncating infinite series (3.4). Since, Δx is a small number, expression (3.43) indicates that approximation (3.5) is $O(\Delta x)$ accurate. As mentioned previously, this is not a statement about the accuracy of approximation (3.5), but about its order of accuracy. This shows that its $T.E.$ will decrease by half if Δx is decreased by half as well or, in other words, if the number of grid points N is doubled.

Similarly, the truncation error of backward-difference approximation (3.8) is obtained by subtracting it from Taylor-series expansion (3.7), which yields

$$T.E. = \left.\frac{d^2f}{dx^2}\right|_i \frac{\Delta x}{2} - \left.\frac{d^3f}{dx^3}\right|_i \frac{(\Delta x)^2}{6} + \left.\frac{d^4f}{dx^4}\right|_i \frac{(\Delta x)^3}{24} - \cdots \,, \tag{3.44}$$

and represents the error introduced by truncating infinite series (3.7), which is $O(\Delta x)$ as the previous truncation error.

Applying the same procedure to central-difference approximation (3.10) to generate its truncation error yields

$$T.E. = -\left.\frac{d^3f}{dx^3}\right|_i \frac{(\Delta x)^2}{6} - \left.\frac{d^5f}{dx^5}\right|_i \frac{(\Delta x)^4}{120} - \left.\frac{d^7f}{dx^7}\right|_i \frac{(\Delta x)^6}{5040} - \cdots \,. \tag{3.45}$$

There are three importante differences between this approximation and the previous two. First, it is $O\left((\Delta x)^2\right)$ accurate. Hence, doubling the number of grip points N will decrease the $T.E.$ four times instead of two. Second, expression (3.45) only has odd derivatives. Even derivatives are associated with diffusion effects whereas odd derivatives are associated with advection effects. Within the context of truncation errors, the absence of even derivatives indicates this approximation introduces no dissipative errors and has only dispersive errors. These two types of error will be discussed in more detail when analyzing the $T.E.$ of FDEs. Finally, even though equation (3.10) is an approximation to a continuous derivative at point x_0, it does not use use any information about the function at point x_0. Approximations with this characteristic tend to introduce an error known as odd-even decoupling, where the solution at odd and even grid points do not interact as needed.

The truncation error of central-difference approximation (3.13) is obtained by subtracting it from Taylor-series expansion (3.12), which yields

$$T.E. = \left.\frac{d^4 f}{dx^4}\right|_i \frac{(\Delta x)^2}{12} + \left.\frac{d^6 f}{dx^6}\right|_i \frac{(\Delta x)^4}{360} + \left.\frac{d^8 f}{dx^8}\right|_i \frac{(\Delta x)^6}{20160} + \cdots \ , \tag{3.46}$$

one possible approximation to a continuous second derivative. Two important remarks can be made regarding this error. First, discrete approximations to the second derivative are at least $O\left((\Delta x)^2\right)$ accurate. This is a direct consequence of the minimum number of terms needed to construct such an expression. Second, this T.E. only has even derivatives.