B Error Analysis

All numerical processes are subject to error. Errors may be of the following types:

1. Truncation errors that are inherent in the numerical algorithm
2. Rounding errors due to the necessity to work to a finite number of significant figures
3. Errors due to inaccurate input data
4. Simple human errors in coding, which should not happen but does!

Examples of the errors described in (1) can be found throughout this text—see, for example, Chapters 3, 4, and 5. Here we consider the implications of the errors described in (2) and (3). Errors of the type described in (4) are outside the scope of this text.

B.1 Introduction

Error analysis estimates the error in some computation caused by errors in some previous process. The previous process may be some experimentation, observation, or rounding in a calculation. Generally we require an upper estimate of the error that can arise when circumstances conspire to be at their worst! We now illustrate this by a specific example. Suppose that $a = 4 \pm 0.02$ (which implies an error of $\pm 0.5\%$) and $b = 2 \pm 0.03$ (which implies an error of $\pm 1.5\%$); then the highest value of a/b results when we divide 4.02 by 1.97 (to give 2.041) and the lowest value of a/b results when we divide 3.98 by 2.03 (to give 1.960). Thus compared with the nominal value of a/b $(= 2)$ we see that the extremes are 2.05% above and 2.0% below the nominal value.

A particular aspect of error analysis is to determine how sensitive a particular calculation is to an error in a specific parameter. Thus we deliberately modify the value of a parameter to determine how sensitive the final answer is to changes in that parameter. For example, consider the following equation:

$$a = 100\frac{\sin\theta}{x^3}$$

If a is evaluated for $\theta = 70^\circ$ and $x = 3$, then $a = 3.4803$. If θ is increased by 10%, then $a = 3.6088$, an increase of 3.69%. If θ is decreased by 10%, then $a = 3.3$, a decrease of 5.18%, Similarly, if we independently increase x by 10%, then $a = 2.6148$, which is a decrease of 24.8%. If we decrease x by 10%, then $a = 4.7741$, an increase of 37.17%. Clearly the value of a is much more sensitive to small changes in x than in θ.

B.2 Errors in Arithmetic Operations

More usually, each of the independent variables has a specified error and we wish to find the overall error in a calculation. We now consider how we can estimate the errors that arise from the standard arithmetic operations. Let x_a, y_a, and z_a be approximations to the exact values x, y, and z, respectively. Let the errors in x, y, and z be x_ε, y_ε, and z_ε, respectively. Then these are given by

$$x_\varepsilon = x - x_a,\;\; y_\varepsilon = y - y_a,\;\; z_\varepsilon = z - z_a$$

Thus

$$x = x_\varepsilon + x_a,\;\; y = y_\varepsilon + y_a,\;\; z = z_\varepsilon + z_a$$

If $z = x \pm y$, then

$$z = (x_a + x_\varepsilon) \pm (y_a + y_\varepsilon) = (x_a \pm y_a) + (x_\varepsilon \pm y_\varepsilon)$$

Now $z_a = x_a \pm y_a$ and hence from the preceding definitions $z_\varepsilon = x_\varepsilon \pm y_\varepsilon$. Normally we are concerned with the maximum possible error and since x_ε and y_ε may be positive or negative quantities, then

$$\max(|z_\varepsilon|) = |x_\varepsilon| + |y_\varepsilon|$$

Consider now the process of multiplication. If $z = xy$, then

$$z = (x_a + x_\varepsilon)(y_a + y_\varepsilon) = x_a y_a + x_\varepsilon y_a + y_\varepsilon x_a + x_\varepsilon y_\varepsilon \tag{B.1}$$

Assuming the errors are small, we can neglect the product of errors in the preceding equation. It is convenient to work in terms of relative error, where the relative error in x, x_ε^R is given by

$$x_\varepsilon^R = x_\varepsilon / x \approx x_\varepsilon / x_a$$

Thus, dividing (B.1) by $z_a = x_a y_a$ we have

$$\frac{(z_a + z_\varepsilon)}{z_a} = 1 + \frac{x_\varepsilon}{x_a} + \frac{y_\varepsilon}{y_a}$$

or

$$\frac{z_\varepsilon}{z_a} = \frac{x_\varepsilon}{x_a} + \frac{y_\varepsilon}{y_a} \tag{B.2}$$

(B.2) can be written

$$z_\varepsilon^R = x_\varepsilon^R + y_\varepsilon^R$$

Again, we want to estimate the worst-case error in z and since the error in x and y may be positive or negative, we have

$$\max\left(\left|z_\varepsilon^R\right|\right) = \left|x_\varepsilon^R\right| + \left|y_\varepsilon^R\right| \tag{B.3}$$

It can easily be shown that if $z = x/y$, the maximum relative error in z is also given by (B.3). This proof is left as an exercise for the reader.

A more general approach to error analysis is to use a Taylor series. Thus if $y = f(x)$ and $y_a = f(x_a)$, then we can write

$$y = f(x) = f(x_a + x_\varepsilon) = f(x_a) + x_\varepsilon f'(x_a) + \cdots$$

Now

$$y_\varepsilon = y - y_a = f(x) \; - f(x_a)$$

Therefore

$$y_\varepsilon \approx x_\varepsilon f'(x_a)$$

For example, consider $y = \sin\theta$ where $\theta = \pi/3 \pm 0.08$. Thus $\theta_\varepsilon = \pm 0.08$. Hence

$$y_\varepsilon \approx \theta_\varepsilon \frac{d}{d\theta}\{\sin(\theta)\} = \theta_\varepsilon \cos\left(\pi/3\right) = 0.08 \times 0.5 = 0.04$$

B.3 Errors in the Solution of Linear Equation Systems

We now consider the problem of estimating the error in the solution of a set of linear equations, $\mathbf{Ax} = \mathbf{b}$. For this analysis we must introduce the concept of a matrix norm.

The formal definition of a matrix p-norm is

$$\|\mathbf{A}\|_p = \max \frac{\|\mathbf{Ax}\|_p}{\|\mathbf{x}\|_p} \quad \text{if} \quad x \neq 0$$

where $\|\mathbf{x}\|_p$ is the vector norm defined in Section A.10. In practice matrix norms are not computed using this definition directly. For example, the 1-norm, 2-norm, and the infinity-norm are computed as follows:

$||\mathbf{A}||_1 =$ maximum absolute column sum of $\mathbf{A}$

$||\mathbf{A}||_2 =$ maximum singular value of $\mathbf{A}$

$||\mathbf{A}||_\infty =$ maximum absolute row sum of $\mathbf{A}$

Having defined the matrix norm we now consider the solution of the equation system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Let the exact solution of this system be $\mathbf{x}$ and the computed solution be $\mathbf{x}_c$. Then we may define the error as

$$\mathbf{x}_e = \mathbf{x} - \mathbf{x}_c$$

We can also define the residual $\mathbf{r}$ as

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}_c$$

We note that large residuals are indicative of inaccuracies but small residuals do not guarantee accuracy. For example, consider the case where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 2+\varepsilon & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 3+\varepsilon \end{bmatrix}$$

The exact solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ (with a residual $\mathbf{r} = \mathbf{0}$) is

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

However, if we consider the very poor approximation

$$\mathbf{x}_c = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$$

then the residual is

$$\mathbf{b} - \mathbf{A}\mathbf{x}_c = \begin{bmatrix} 0 \\ -0.5\varepsilon \end{bmatrix}$$

If $\varepsilon = 0.00001$, then the residual is very small even though the solution is very inaccurate.

To obtain a formula that provides bounds on the relative error of the computed value, $\mathbf{x}_c$, we proceed as follows:

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}_c = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_c = \mathbf{A}\mathbf{x}_\varepsilon \tag{B.4}$$

From (B.4) we have

$$\mathbf{x}_\varepsilon = \mathbf{A}^{-1}\mathbf{r}$$

Taking the norms of this equation we have

$$\|\mathbf{x}_\varepsilon\| = \left\|\mathbf{A}^{-1}\mathbf{r}\right\| \tag{B.5}$$

We can choose to use any p-norm and in the analysis that follows the subscript p is omitted. A property of norms is that $||\mathbf{AB}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$. Thus we have, from (B.5),

$$\|\mathbf{x}_\varepsilon\| \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{r}\| \tag{B.6}$$

But $\mathbf{r} = \mathbf{A}\mathbf{x}_\varepsilon$ and so

$$\|\mathbf{r}\| \leq \|\mathbf{A}\|\ \|\mathbf{x}_\varepsilon\|$$

Therefore

$$\frac{\|\mathbf{r}\|}{\|\mathbf{A}\|} \leq \|\mathbf{x}_\varepsilon\|$$

Combining this equation with (B.6) we have

$$\frac{\|\mathbf{r}\|}{\|\mathbf{A}\|} \leq \|\mathbf{x}_\varepsilon\| \leq \left\|\mathbf{A}^{-1}\right\|\ \|\mathbf{r}\| \tag{B.7}$$

Now since $\mathbf{x}\ = \mathbf{A}^{-1}\mathbf{b}$ we have, similarly,

$$\frac{\left\|\mathbf{b}\right\|}{\|\mathbf{A}\|} \leq \|\mathbf{x}\| \leq \left\|\mathbf{A}^{-1}\right\| \left\|\mathbf{b}\right\| \tag{B.8}$$

If none of the terms in the preceding equation are zero, we can take reciprocals to give

$$\frac{1}{\left\|\mathbf{A}^{-1}\right\| \left\|\mathbf{b}\right\|} \leq \frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\left\|\mathbf{b}\right\|} \tag{B.9}$$

Multiplying the corresponding terms in (B.7) and (B.9) gives

$$\frac{1}{\|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|} \frac{\|\mathbf{r}\|}{\left\|\mathbf{b}\right\|} \leq \frac{\|\mathbf{x}_\varepsilon\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{r}\|}{\left\|\mathbf{b}\right\|} \tag{B.10}$$

This equation gives error bounds for the relative error in the computation that are directly computable. The condition number of $\mathbf{A}$ is given by $\text{cond}(\mathbf{A}, p) = ||\mathbf{A}||_p||\mathbf{A}^{-1}||_p$. Hence (B.10) can be rewritten in terms of $\text{cond}(\mathbf{A}, p)$. When $p = 2$, $\text{cond}(\mathbf{A})$ is the ratio of the largest singular value of $\mathbf{A}$ to the smallest.

We now show how (B.10) can be used to estimate the relative error in the solution of $\mathbf{Ax} = \mathbf{b}$ when $\mathbf{A}$ is the Hilbert matrix. We have chosen the Hilbert matrix because its condition number is large and its inverse is known and so we can compute the actual error in the computation of $\mathbf{x}$. The following MATLAB script evaluates (B.10) for a specific Hilbert matrix using the 2-norm.

```
n = 6, format long
a = hilb(n); b = ones(n,1);
xc = a\b;
x = invhilb(n)*b;
```

```
exact_x = x';
err = abs((xc-x)./x);
nrm_err = norm(xc-x)/norm(x)
r = b-a*xc;
L_Lim = (1/cond(a))*norm(r)/norm(b)
U_Lim = cond(a)*norm(r)/norm(b)
```

Running this script gives

```
n =
    6

nrm_err =
3.316798106133016e-11

L_Lim =
3.351828310510846e-21

U_Lim =
7.492481073232495e-07
```

We see that the norm of the actual relative error, 3.316×10^{-11}, lies between the bounds 3.35×10^{-21} and 7.49×10^{-7}.