Stability in Numerical Programming: An Introduction

Robert P. Goddard

Applied Physics Laboratory
University of Washington
Seattle, WA

Northwest C++ User's Group
18 September 2013

Outline

Richard W. Hamming's Five Main Ideas

The purpose of computing is insight, not numbers.

R. W. Hamming, *Numerical Methods for Scientists and Engineers, Second Edition*, McGraw-Hill, New York etc. (1973), Chapter 1: "An Essay on Numerical Methods"

- 0. *Numbers*: Counting, fixed-point, floating-point (Hamming Chapter 2)
- 1. *Purpose*: Computing is intimately bound up with both the source of the problem and the use that is going to be made of the answers – it is not a step to be taken in isolation from reality.
- 2. *Generality*: It is necessary to study families and to relate one family to another when possible, and to avoid isolated formulas and isolated algorithms.

Hamming's Five Main Ideas
continued

- 3. *Roundoff error*: The greatest loss of significance in the numbers occurs when two numbers of about the same size are subtracted so that most of the leading digits cancel out.
- 4. *Truncation error*: Many of the processes of mathematics, such as differentiation and integration, imply the use of a limit which is an infinite process. The machine can only do a finite number of operations in a finite length of time.
- 5. *Feedback*: Numbers at one stage are fed back into the computer to be processed again and again. Feedback leads to the idea of *stability* of the feedback loop – will a small error grow or decay through the successive iterations?

Problems that involve feedback

- Difference equations, e.g. recurrence relations (example next)
- Indefinite integrals approximated by difference equations
- Differential equations approximated by difference equations
- Digital filters – specifically IIR (Infinite Impulse Response) filters
- Kalman filters, used for tracking objects based on multiple observations
- Linear algebra, e.g. eigenvalue-eigenvector computation
- Others: Audience?

Spherical Bessel Functions: Why?

- *Problem*: Compute sound scattered from a complicated object in a (locally) uniform medium with sound speed c.
- *Physics*: Helmholtz equation $\nabla^2 p + k^2 p = 0$ for the sound pressure p. (∇^2 is the *Laplacian*)
- *Observation*: Wavenumber k is uniform in a region near the scatterer but outside some sphere of radius R.
- *Approach*: Use a *multipole expansion* of the field in the uniform region.
- *Math*: Express the Helmholtz equation in spherical coordinates (range and two angles) instead of Cartesian coordinates (X, Y, Z). It is separable into range-dependent and direction-dependent parts.
- *Our interest here*: Compute the range-dependent parts: the *spherical Bessel functions*

Spherical Bessel Functions: Definitions

The range-dependent part is the *Spherical Bessel Equation*:

$$z^2 \frac{\mathrm{d}^2}{\mathrm{d}z^2} f_n(z) + \frac{\mathrm{d}}{\mathrm{d}z} f_n(z) + (z^2 - n(n+1)) f_n(z) = 0$$

where $z = kr$, $k = 2\pi c/f$ is the wavenumber, and n is an integer. Since this is a second-order homogeneous differential equation, any solution is a linear combination of two indepent solutions, of which a standard pair is

- $j_n(z)$: Spherical Bessel function of the first kind. Finite at $z = 0$.
- $y_n(z)$: Spherical Bessel function of the second kind. Pole (z^{n+1}) at $z = 0$.

Spherical Bessel Functions: Recurrence Relation

Our computational problem is:

- Compute both $j_n(z)$ and $y_n(z)$
- ... for a given z of order R/λ (which can be large)
- ... for a range of n from 0 to $n_{\text{max}} > z$ (even larger)

Fortunately, this *recurrence relation* holds for any kind of spherical Bessel function:

$$f_{n+1}(z) + f_{n-1}(z) = (2n+1)z^{-1} f_n(z)$$

We also have explicit formulas for the first two members:

$$j_0(z) = \frac{\sin z}{z} \qquad y_0(z) = -\frac{\cos z}{z}$$

$$j_1(z) = \frac{\sin z}{z^2} - \frac{\cos z}{z} \qquad y_1(z) = -\frac{\cos z}{z^2} - \frac{\sin z}{z}$$

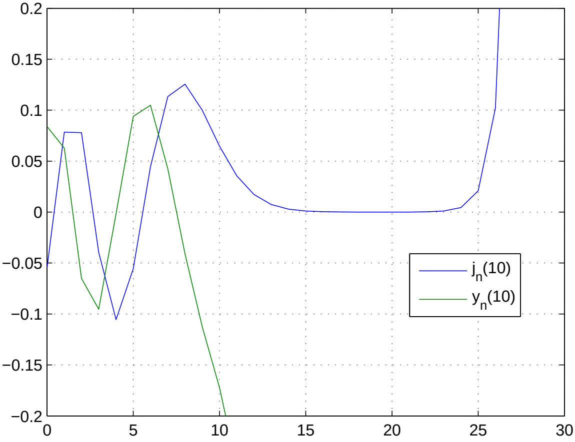
So the answer is easy – *right*?

Spherical Bessel Functions: Forward Recurrence Code

```
/// Compute spherical Bessel functions using forward recurrence relation
/**
 * Results are written into the vector f, which must already contain
 * the first two members of the set.
 *
 * The recurrence relation used is Abramowitz & Stegun Eq. 10.1.19.
 */
template <typename T>
void SpherBesselRecur_1( T z, std::vector<T>& f )
{
    assert( z > 0.0 );
    size_t nmax = f.size() - 1;
    assert( nmax > 1 );

    for ( size_t n = 1; n < nmax; ++n ) {
        T b = T(n+n+1)/z;
        f[n+1] = b*f[n] - f[n-1];
    }
}
```

Forward Recurrence: Result



Forward Recurrence: What Happened?

- The recurrence relation (a second-order difference equation) has two solutions, $j_n(z)$ and $y_n(z)$.
- For $n > z$, $y_n(z)$ grows with n, whereas $j_n(z)$ shrinks with n.
- Inevitably, an error comes in.
- Thereafter, the solution is a linear combination of $j_n(z)$ and $y_n(z)$, not one or the other.
- Sooner or later, the growing solution dominates the shrinking one.

Solution:

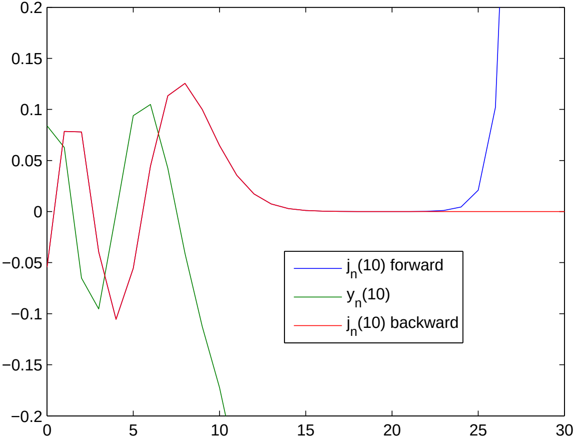
- Apply the recurrence relation backward, for decreasing n.
- The starting value doesn't matter much after a few iterations because now the solution you want is dominant.
- Use the known $j_0(z)$ to renormalize the results.
- (Better: Use a sum rule, more accurate if $j_0(z)$ is near zero.)

Spherical Bessel Functions: Backward Recurrence Code

```
/// Compute spherical Bessel functions using backward recurrence
/**
 * Results are written into the vector f, which must already contain
 * the first member of the set in f[0]..
 *
 * The recurrence relation used is Abramowitz & Stegun Eq. 10.1.19.
 */
template <typename T>
void SpherBesselRecur_2( T z, std::vector<T>& f )
{
    assert( z > 0.0 );
    size_t nmax = f.size() - 1;
    assert( nmax > 1 );

    f[nmax] = 0;
    f[nmax-1] = T(1.0e-10);
    T f0 = f[0];
    for ( size_t n = nmax-1; n > 0; --n ) {
        T b = T(n+n+1)/z;
        f[n-1] = b*f[n] - f[n+1];
    }
    T ratio = f0/f[0];
    for ( size_t n = 0; n <= nmax; ++n )
        f[n] = ratio*f[n];
}
```

Backward Recurrence: Result



Generalizing: Properties of Difference Equations

$$f_{n+1}(z) + f_{n-1}(z) = (2n+1)z^{-1}f_n(z)$$

- Second order difference equation: 2 starting values, 2 independent solutions. Generalize: Order N, N starting values, N independent solutions.
- If any solution grows relative to the others as the equation is iterated, that one will dominate.
- You can use the equation in that directon *only* to compute the dominant solution. For the other solutions, it is unstable.
- Sometimes you can reverse direction to select a different solution, dominant in that direction.
- For Bessel functions, we know the general properties of the solutions, so the behavior is no surprise.
- *Problem*: How can we generalize this idea for other difference equations?

Characteristic Equation

$$f_{n+1}(z) - (2n+1)z^{-1} f_n(z) + f_{n-1}(z) = 0$$

Observation: For large n, the factor $(2n+1)z^{-1}$ doesn't change much in a few steps. So, we can learn about local behavior by treating it as constant. Generalize and simplify notation:

$$x_{n+1} + Bx_n + Cx_{n-1} = 0$$

Guess: Try solutions of the form $x_n = a^n$ for some constant a. Subsitute:

$$a^2 + Ba + C = 0 \qquad \text{Characteristic Equation}$$
$$a = -B/2 \pm \sqrt{(B/2)^2 - C} \qquad \text{Solutions}$$

Now the local behavior is clear:

- If $|a| > 1$, the solution a^n grows exponentially in size.
- If $|a| < 1$, the solution a^n shrinks exponentially in size.
- If $(B/2)^2 - C < 0$, a is complex, so the solution a^n is oscillatory.

Characteristic Equation for Spherical Bessel Recurrence

$$f_{n+1}(z) - 2Rf_n(z) + f_{n-1}(z) = 0 \qquad \text{Recurrence Relation}$$

$$\text{where} \quad R \approx (n+1/2)/z$$

$$a^2 - 2Ra + 1 = 0 \qquad \text{Characteristic Equation}$$

$$a = R \pm \sqrt{R^2 - 1} \qquad \text{Solutions for } a$$

If $R > 1$, a is real, $a_1 > 1$ and $a_2 < 1$. Hence $(a_1)^n$ grows exponentially and $(a_2)^n$ shrinks exponentially.
If $R < 1$, a is complex. Substitute: $R = \cos\phi$. Then

$$\begin{aligned} a &= \cos\phi \pm \sqrt{cos^2\phi - 1} \\ &= \cos\phi \pm \mathrm{i}\sin\phi \\ &= \mathrm{e}^{\mathrm{i}\phi}, \mathrm{e}^{-\mathrm{i}\phi} \end{aligned}$$

so both solutions are oscillatory, and neither is dominant.

First-Order ODE with Initial Condition

General Problem:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \mathrm{f}(x, y) \qquad\qquad y(0) = y_0$$

Solution Approach: Fourth-order Predictor-Corrector

$$\begin{aligned} y_{n+1} &= a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + a_3 y_{n-3} \\ &+ h^2 (b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}) \\ &+ E_5 \frac{h^5 y^{(5)}}{5!}(\theta) \\ \text{Predictor:} \quad & b_{-1} = 0 \\ \text{Corrector:} \quad & a_3 = 0 \text{ and } y''_{n+1} \text{ from Predictor} \end{aligned}$$

Coefficients for Fourth Order Convergence

Require: Exact for polynomials through degree 4: $y = 1, x, x^2, x^3, x^4$. That's 5 equations in 7 unknowns, leaving 2 free variables. For the corrector ($a_3 = 0$), that works out to:

$$\begin{aligned} a_0 &= 1 - a_1 - a_2 & b_0 &= (19 + 13a_1 + 8a_2)/24 \\ a_1 &= a_1 & b_1 &= (-5 + 13a_1 + 32a_2)/24 \\ a_2 &= a_2 & b_2 &= (1 - a_1 + 8a_2)/24 \\ b_{-1} &= (9 - a_1)/24 & E_5 &= (-19 + 11a_1 - 8a_2)/6 \end{aligned}$$

Characteristic Equation, Part 1

Let z_n be the true solution: $z' = f(x, z)$, and let y_n be the computed solution at $x = x_n$. If y_n is put into the differential equation, it will fit exactly, since the computed y' value was found from the equation. Thus (neglecting roundoff), $y_n' = f(x, y_n)$. Set

$$\begin{aligned} \epsilon_n &= z_n - y_n \\ \epsilon_n' &= z_n' - y_n' = f(x, z_n) - f(x, y_n) \\ &= \frac{\partial f(x, \theta)}{\partial y} \epsilon_n = A \epsilon_n \end{aligned}$$

where θ lies between y_n and z_n (mean value theorem).
The true solution z_n does not generally satisfy the difference equation:

$$\begin{aligned} z_{n+1} &= a_0 z_n + a_1 z_{n-1} + a_2 z_{n-2} + a_3 z_{n-3} \\ &+ h^2 (b_{-1} z_{n+1}' + b_0 z_n' + b_1 z_{n-1}' + b_2 z_{n-2}') \\ &+ e_n \end{aligned}$$

Characteristic Equation, Part 2

Subtract the difference equation with y from the version with z, subsititute $\epsilon'_n = A\epsilon_n$, and rearrange:

$$\begin{aligned}(1 - b_{-1}Ah)\epsilon_{n+1} &= (a_0 - b_0 Ah)\epsilon_n + (a_1 - b_1 Ah)\epsilon_{n-1} \\ &\quad +(a_2 - b_2 Ah)\epsilon_{n-2} + a_3\epsilon_{n-3} + e_n\end{aligned}$$

We are interested in *local* behavior. For that purpose, we assume e_n and $\partial f/\partial y = A$ are constants. Result: Difference equation in ϵ_n with constant coefficients.
Substitute $\epsilon_n = \rho^n$: That is the characteristic equation:

$$Ah = \frac{\rho^3 - a_0\rho^2 - a^1\rho - a_2}{b_{-1}\rho^3 + b_0\rho^2 + b_1\rho + b_2}$$

Characteristic Equation Roots

The roots (3 of them for the corrector) determine stability: The formula is unstable if $|\rho| > 1$ or if $|\rho| = 1$ and the root is a multiple root. For $Ah = 0$ (i.e. if $f(x, y)$ is effectively independent of y), the region of stability is a triangle in the (a_1, a_2) plane:

$$1 + a_1 + 2a_2 \geq 0 \qquad a_1 \leq 1 \qquad a_2 \leq 1$$

As $|ah|$ increases, that triangle shrinks and becomes distorted, but it remains within that outer triangle (Hamming Fig. 23.3.1). Within the stability region, other factors are used to choose (a_1, a_2) values:

- Minimize the the truncation error $E_5 = (-19 + 11a_1 - 8a_2)/6$
- Minimize the noise amplification $N_c = 1/(1 + a_1 + 2a_2)$
- Reduce the uncorrelated noise $N_a = (a_0^2 + a_1^2 + a^2)^{1/2}$