Solving Linear Systems by Iteration

We saw in chapter 2 that you can solve a root-finding problem

$$f(x) = 0$$

by restructuring the problem as a fixed point problem

$$g(x) = x$$

and solving by iteration. If certain conditions on $g(x)$ and its derivative are satisfied the sequence of iterates will converge to a fixed point of $g(x)$.

Our primary concern in chapters 6 and 7 is now solving systems of linear equations

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

Once again, we are going to try to solve the original problem by recasting it as a fixed point problem and seeking a solution by iteration. Specifically, we will seek to rewrite the system as a fixed point problem involving vectors

$$\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$$

for some appropriate matrix T and vector $\mathbf{c}$.

In a little while we will see how to recast the original system into this form. First, let us tackle the question of whether or not the fixed point problem can be solved by iteration.

The first thing to note is that

$$\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$$

can be rewritten

$$\mathbf{x} - T\,\mathbf{x} = \mathbf{c}$$

or

$$(I - T)\,\mathbf{x} = \mathbf{c}$$

or

$$\mathbf{x} = (I - T)^{-1}\,\mathbf{c}$$

Here is a lemma from the text that can help us to make sense of the right hand side.

Lemma 7.18 If the spectral radius of T satisfies $\rho(T) < 1$ then $(I - T)^{-1}$ exists and

$$(I - T)^{-1} = I + T + T^2 + T^3 + \cdots = \sum_{j=0}^{\infty} T^j$$

The following theorem uses this lemma to show that if $\rho(T) < 1$ then the iteration

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c}$$

converges.

Theorem If $\rho(T) < 1$ then for any $\mathbf{x}^{(0)} \in \mathbb{R}^n$ the sequence defined by

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c}$$

converges to the solution of the fixed point equation

$$\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$$

Proof Iterating the iteration a few times gives

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c}$$

$$= T\left(T\,\mathbf{x}^{(k-2)} + \mathbf{c}\right) + \mathbf{c} = T^2\,\mathbf{x}^{(k-2)} + T\,\mathbf{c} + \mathbf{c}$$

$$= T\left(T\left(T\,\mathbf{x}^{(k-3)} + \mathbf{c}\right) + \mathbf{c}\right) + \mathbf{c} = T^3\,\mathbf{x}^{(k-3)} + T^2\,\mathbf{c} + T\,\mathbf{c} + \mathbf{c}$$

$$\vdots$$

$$= T^k\,\mathbf{x}^{(0)} + \left(T^{k-1} + T^{k-2} + \cdots + T + I\right)\mathbf{c}$$

Since $\rho(T) < 1$ the term $T^k\,\mathbf{x}^{(0)}$ converges to the $\mathbf{0}$ vector for any $\mathbf{x}^{(0)}$. In the limit as k gets very large the other term converges to

$$\left(\sum_{j=0}^{\infty} T^j\right)\mathbf{c} = (I - T)^{-1}\,\mathbf{c} = \mathbf{x}$$

Two ways to convert a linear system to an iteration

Our hope now is to convert a linear system

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

into a fixed point problem

$$\mathbf{x} = T\,\mathbf{x} + c$$

in such a way that $\rho(T) < 1$ so we can use the theorem above to guarantee convergence of the fixed point interation from any starting $\mathbf{x}^{(0)}$.

The first technique for doing this is called the Jacobi method. This method begins by writing the matrix A in a special form:

$$A = (D - L - U)$$

where L is the negative of the portion of A below the main diagonal, D contains just the main diagonal of A and the rest 0s, and U is the negative of the portion of A above the main diagonal. (Note that the L and U here have nothing to do with the L and U of the LU decomposition. These L and U matrices are defined in a much simpler way.)

$$(D - L - U)\,\mathbf{x} = \mathbf{b}$$

$$D\,\mathbf{x} = (L + U)\,\mathbf{x} + \mathbf{b}$$

$$x = D^{-1}\left(L + U\right)\mathbf{x} + D^{-1}\,\mathbf{b} = T_j\,\mathbf{x} + \mathbf{c}_j$$

The second method, the Gauss-Seidel method uses a slightly different approach:

$$(D - L - U)\,\mathbf{x} = \mathbf{b}$$

$$(D - L)\,\mathbf{x} = \ U\,\mathbf{x} + \mathbf{b}$$

$$\mathbf{x} = \left(D - L\right)^{-1}\,U\,\mathbf{x} + \left(D - L\right)^{-1}\,\mathbf{b} = T_g\,x + c_g$$

Because D^{-1} is much easier to compute than $\left(D - L\right)^{-1}$ the Jacobi method is easier to implement. The Gauss-Seidel method has the advantage of converging more quickly.

Another fact about the Gauss-Seidel method is that in practice one can implement it without really having to compute $\left(D - L\right)^{-1}$. Here is the trick. What we ultimately want to do is to iterate the form $T\,\mathbf{x} + c$:

$$\mathbf{x}^{(k)} = \left(D - L\right)^{-1}\,U\,\mathbf{x}^{(k-1)} + \left(D - L\right)^{-1}\,\mathbf{b}$$

This can be written

$$(D - L)\,\mathbf{x}^{(k)} = U\,\mathbf{x}^{(k-1)} + \mathbf{b}$$

Note that if we know $\mathbf{x}^{(k-1)}$ we can compute $U\,\mathbf{x}^{(k-1)} + \mathbf{b} = \mathbf{v}$ and then solve

$$(D - L)\,\mathbf{x}^{(k)} = \mathbf{v}$$

by back-substitution. (Note that $D - L$ is a lower triangular matrix.) In fact, if you work out the details you can see that the result of the back substitution can be summarized in a set of formulas. Here are the first few steps in the calculation.

$$a_{1,1}\,\left(\mathbf{x}^{(k)}\right)_1 = \sum_{j=2}^{n} -a_{1,j}\,\left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_1$$

$$\left(\mathbf{x}^{(k)}\right)_1 = \frac{\sum_{j=2}^{n} -a_{1,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_1}{a_{1,1}}$$

$$a_{2,1} \left(\mathbf{x}^{(k)}\right)_1 + a_{2,2} \left(\mathbf{x}^{(k)}\right)_2 = \sum_{j=3}^{n} -a_{2,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_2$$

$$\left(\mathbf{x}^{(k)}\right)_2 = \frac{-a_{2,1} \left(\mathbf{x}^{(k)}\right)_1 - \sum_{j=3}^{n} a_{2,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_2}{a_{2,2}}$$

$$a_{3,1} \left(\mathbf{x}^{(k)}\right)_1 + a_{3,2} \left(\mathbf{x}^{(k)}\right)_2 + a_{3,3} \left(\mathbf{x}^{(k)}\right)_3 = \sum_{j=4}^{n} -a_{3,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_3$$

$$\left(\mathbf{x}^{(k)}\right)_3 = \frac{-\sum_{j=1}^{2} a_{3,j} \left(\mathbf{x}^{(k)}\right)_j - \sum_{j=4}^{n} a_{3,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_3}{a_{3,3}}$$

The general pattern that emerges here is

$$\left(\mathbf{x}^{(k)}\right)_m = \frac{-\sum_{j=1}^{m-1} a_{m,j} \left(\mathbf{x}^{(k)}\right)_j - \sum_{j=m+1}^{n} a_{m,j} \left(\mathbf{x}^{(k-1)}\right)_j + \mathbf{b}_m}{a_{m,m}}$$