2003/11/10	**M174 : Numerical Methods in Science and Engineering**
Steven Pav	SPAV@MATH.UCSD.EDU

§6.5 Iterative Solutions

Recall we are trying to solve

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

We saw that Gaussian Elimination requires around $\frac{2}{3}n^3$ operations just to find the LU factorization, then about n^2 operations to solve the system, when A is $n \times n$. When n is large, this may take too long to be practical, so we attempt to find an approximate solution iteratively. That is, we have some initial estimate of the solution, $\boldsymbol{x}^{(0)}$, and some means of calculating $\boldsymbol{x}^{(k+1)}$ from $\boldsymbol{x}^{(k)}$.

Our iterative update is defined implicitly. For some matrix Q, we will let

$$\mathsf{Q}\boldsymbol{x}^{(k+1)} = (\mathsf{Q} - \mathsf{A})\,\boldsymbol{x}^{(k)} + \boldsymbol{b}.$$

Now suppose that as $k \to \infty$, $\boldsymbol{x}^{(k)}$ converges to some vector $\boldsymbol{x}^*$. Then we have

$$\begin{aligned}
\mathsf{Q}\boldsymbol{x}^* &= (\mathsf{Q} - \mathsf{A})\,\boldsymbol{x}^* + \boldsymbol{b}, \\
\mathsf{Q}\boldsymbol{x}^* &= \mathsf{Q}\boldsymbol{x}^* - \mathsf{A}\boldsymbol{x}^* + \boldsymbol{b}, \\
\mathsf{A}\boldsymbol{x}^* &= \boldsymbol{b}.
\end{aligned}$$

We have some freedom in choosing Q, but there are two considerations we should keep in mind:

1. Choice of Q affects convergence and speed of convergence of the method.
2. Choice of Q affects ease of computing the update. That is, given

   $$\boldsymbol{z} = (\mathsf{Q} - \mathsf{A})\,\boldsymbol{x}^{(k)} + \boldsymbol{b},$$

   we should pick Q such that the equation

   $$\mathsf{Q}\boldsymbol{x}^{(k+1)} = \boldsymbol{z}$$

   is easy to solve exactly.

This last point is important. Remember that we chose an iterative method because we couldn't easily solve a linear equation.

We illustrate some choices of Q.

Richardson Iteration

The simplest iterative method is Richardson Iteration, which chooses Q to be the identity matrix. Solving the system

$$\mathsf{Q}\boldsymbol{x}^{(k+1)} = \boldsymbol{z}$$

is trivial: we just have $\boldsymbol{x}^{(k+1)} = \boldsymbol{z}$.

Example 1. Use Richardson Iteration to solve the system for

$$\mathsf{A} = \begin{bmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{bmatrix}, \qquad \mathsf{b} = \begin{bmatrix} 12 \\ 0 \\ 6 \end{bmatrix}.$$

We let

$$\mathsf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad (\mathsf{Q} - \mathsf{A}) = \begin{bmatrix} 5 & -1 & -1 \\ -2 & 3 & 0 \\ -1 & -2 & 5 \end{bmatrix}.$$

We start with an arbitrary $\boldsymbol{x}^{(0)}$, say $\boldsymbol{x}^{(0)} = [2\,2\,2]^\top$. We get $\boldsymbol{x}^{(1)} = [18\,2\,10]^\top$, and $\boldsymbol{x}^{(2)} = [90\ -30\,34]^\top$.

Note the real solution is $\boldsymbol{x} = [2\ -1\ 1]^\top$. The Richardson Iteration does not appear to converge for this example, unfortunately.

We can rethink the Richardson Iteration as

$$\boldsymbol{x}^{(k+1)} = (\mathsf{I} - \mathsf{A})\,\boldsymbol{x}^{(k)} + \boldsymbol{b} = \boldsymbol{x}^{(k)} + \left(\boldsymbol{b} - \mathsf{A}\boldsymbol{x}^{(k)}\right).$$

Thus at each step we are adding the *residual*, defined as $\boldsymbol{b} - \mathsf{A}\boldsymbol{x}^{(k)}$, to the iterate.

Jacobi Iteration

The Jacobi Iteration chooses Q to be the matrix consisting of the diagonal of A. Solving the system

$$\mathsf{Q}\boldsymbol{x}^{(k+1)} = \boldsymbol{z}$$

is trivial.

Example 2. Use Jacobi Iteration to solve the system for

$$\mathsf{A} = \begin{bmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{bmatrix}, \qquad \mathsf{b} = \begin{bmatrix} 12 \\ 0 \\ 6 \end{bmatrix}.$$

We let

$$\mathsf{Q} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}, \qquad (\mathsf{Q} - \mathsf{A}) = \begin{bmatrix} 0 & -1 & -1 \\ -2 & 0 & 0 \\ -1 & -2 & 0 \end{bmatrix}, \qquad \mathsf{Q}^{-1} = \begin{bmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix}.$$

We start with an arbitrary $\boldsymbol{x}^{(0)}$, say $\boldsymbol{x}^{(0)} = [2\,2\,2]^\top$. We get $\boldsymbol{x}^{(1)} = \left[\frac{4}{3}\ -1\ -1\right]^\top$. Then $\boldsymbol{x}^{(2)} = \left[\frac{7}{3}\ -\frac{2}{3}\,\frac{1}{9}\right]^\top$.

Note the real solution is $\boldsymbol{x} = [2\ -1\ 1]^\top$.

We see there is an easy way to describe the update performed by the Jacobi Iteration. Let $x_j^{(k)}$ be the j^{th} element of $\boldsymbol{x^{(k)}}$. Then we see easily that

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{i=1, i \neq j}^{n} a_{ji} x_i^{(k)} \right).$$

Thus an update takes less than $2n^2$ operations.

Gauss Seidel Iteration

The Gauss Seidel Iteration chooses Q to be lower triangular part of A, including the diagonal. In this case solving the system

$$\mathsf{Q}\boldsymbol{x^{(k+1)}} = \boldsymbol{z}$$

is performed by forward substitution.

Example 3. Use Gauss Seidel Iteration to again solve for

$$\mathsf{A} = \begin{bmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{bmatrix}, \qquad \mathsf{b} = \begin{bmatrix} 12 \\ 0 \\ 6 \end{bmatrix}.$$

We let

$$\mathsf{Q} = \begin{bmatrix} 6 & 0 & 0 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{bmatrix}, \qquad (\mathsf{Q} - \mathsf{A}) = \begin{bmatrix} 0 & -1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We start with an arbitrary $\boldsymbol{x^{(0)}}$, say $\boldsymbol{x^{(0)}} = [2\,2\,2]^\top$. We get $\boldsymbol{x^{(1)}} = \left[\frac{4}{3}\; -\frac{2}{3}\; 1\right]^\top$. Then $\boldsymbol{x^{(2)}} = \left[\frac{35}{18}\; -\frac{35}{36}\; 1\right]^\top$.

Already this is fairly close to the actual solution $\boldsymbol{x} = [2\; -1\; 1]^\top$.

There is an easier way to describe the Gauss Seidel Iteration. In this case we will keep a single vector $\boldsymbol{x}$ and overwrite it, element by element. Thus for $j = 1, 2, \ldots, n$, we set

$$x_j \leftarrow \frac{1}{a_{jj}} \left(b_j - \sum_{i=1, i \neq j}^{n} a_{ji} x_i \right).$$

This looks exactly like the Jacobi update. However, in the sum on the right there are some "old" values of x_i and some "new" values; the new values are those x_i for which $i < j$.

Again this takes less than $2n^2$ operations.

An alteration of the Gauss Seidel Iteration is to make successive "sweeps" of this redefinition, one for $j = 1, 2, \ldots, n$, the next for $j = n, n-1, \ldots, 2, 1$. This amounts to running Gauss Seidel once with Q the lower triangular part of A, then running it with Q the upper triangular part. This iterative method is known as "red-black Gauss Seidel."

Successive Overelaxation Iteration (SOR)

The Successive Overelaxation (SOR) Iteration chooses Q to be lower triangular part of A, but with the diagonal divided by a relaxation factor, ω. In this case solving the system

$$\mathsf{Q}\boldsymbol{x}^{(\boldsymbol{k}+\mathbf{1})} = \boldsymbol{z}$$

is performed by forward substitution.

Example 4. Use SOR Iteration with $\omega = 1.5$ to again solve for

$$\mathsf{A} = \begin{bmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{bmatrix}, \qquad \mathsf{b} = \begin{bmatrix} 12 \\ 0 \\ 6 \end{bmatrix}.$$

We let

$$\mathsf{Q} = \begin{bmatrix} 4 & 0 & 0 \\ 2 & \frac{8}{3} & 0 \\ 1 & 2 & 4 \end{bmatrix}, \qquad (\mathsf{Q} - \mathsf{A}) = \begin{bmatrix} -2 & -1 & -1 \\ 0 & -\frac{4}{3} & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

We start with an arbitrary $\boldsymbol{x}^{(\mathbf{0})}$, say $\boldsymbol{x}^{(\mathbf{0})} = [2\,2\,2]^\top$. We get $\boldsymbol{x}^{(\mathbf{1})} = \left[1\ -\frac{7}{4}\ \frac{9}{8}\right]^\top$.

Error Analysis

Suppose that $\boldsymbol{x}$ is the solution to equation 1. Define the error vector:

$$\boldsymbol{e}^{(\boldsymbol{k})} = \boldsymbol{x}^{(\boldsymbol{k})} - \boldsymbol{x}.$$

Now notice that

$$\begin{aligned}
\boldsymbol{x}^{(\boldsymbol{k}+\mathbf{1})} &= \mathsf{Q}^{-1}(\mathsf{Q} - \mathsf{A})\,\boldsymbol{x}^{(\boldsymbol{k})} + \mathsf{Q}^{-1}\boldsymbol{b}, \\
\boldsymbol{x}^{(\boldsymbol{k}+\mathbf{1})} &= \mathsf{Q}^{-1}\mathsf{Q}\boldsymbol{x}^{(\boldsymbol{k})} - \mathsf{Q}^{-1}\mathsf{A}\boldsymbol{x}^{(\boldsymbol{k})} + \mathsf{Q}^{-1}\mathsf{A}\boldsymbol{x}, \\
\boldsymbol{x}^{(\boldsymbol{k}+\mathbf{1})} &= \boldsymbol{x}^{(\boldsymbol{k})} - \mathsf{Q}^{-1}\mathsf{A}\left(\boldsymbol{x}^{(\boldsymbol{k})} - \boldsymbol{x}\right), \\
\boldsymbol{x}^{(\boldsymbol{k}+\mathbf{1})} - \boldsymbol{x} &= \boldsymbol{x}^{(\boldsymbol{k})} - \boldsymbol{x} - \mathsf{Q}^{-1}\mathsf{A}\left(\boldsymbol{x}^{(\boldsymbol{k})} - \boldsymbol{x}\right), \\
\boldsymbol{e}^{(\boldsymbol{k}+\mathbf{1})} &= \boldsymbol{e}^{(\boldsymbol{k})} - \mathsf{Q}^{-1}\mathsf{A}\boldsymbol{e}^{(\boldsymbol{k})}, \\
\boldsymbol{e}^{(\boldsymbol{k}+\mathbf{1})} &= \left(\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}\right)\boldsymbol{e}^{(\boldsymbol{k})}.
\end{aligned}$$

We want to ensure that $\boldsymbol{e}^{(\boldsymbol{k}+\mathbf{1})}$ is "smaller" than $\boldsymbol{e}^{(\boldsymbol{k})}$. To do this we take a detour into matrix and vector norms.

Matrix Norms

First we define a vector norm. For $\boldsymbol{x}$, we define its two-norm as

$$||\boldsymbol{x}||_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}.$$

Norms have many interesting properties. We will list only a few of them here:

1. The norm obeys the triangle inequality:

$$||\boldsymbol{x}+\boldsymbol{y}||_2 \leq ||\boldsymbol{x}||_2 + ||\boldsymbol{y}||_2 \, .$$

2. The norm scales positively:

$$||\alpha \boldsymbol{x}||_2 = |\alpha| \, ||\boldsymbol{x}||_2 \, .$$

3. Only the zero vector has zero norm.
4. If $||\boldsymbol{x}||_2 = r$, then $\boldsymbol{x}$ is on a sphere centered at the origin of radius r, in R^n.

Now we can define a matrix norm. We define

$$||\mathsf{A}||_2 = \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{||\mathsf{A}\boldsymbol{x}||_2}{||\boldsymbol{x}||_2}.$$

We will need the following properties of matrix norms:

1. Because the matrix norm is defined as a max, then if $\boldsymbol{x} \neq \mathbf{0}$, we have

$$||\mathsf{A}\boldsymbol{x}||_2 \leq ||\mathsf{A}||_2 \, ||\boldsymbol{x}||_2 \, .$$

2. The norm of a matrix is equal to its largest eigenvalue, in absolute value.

Now we note that since

$$\boldsymbol{e}^{(\boldsymbol{k}+1)} = \left(\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}\right) \boldsymbol{e}^{(\boldsymbol{k})},$$

then

$$\left|\left|\boldsymbol{e}^{(\boldsymbol{k}+1)}\right|\right|_2 = \left|\left|\left(\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}\right) \boldsymbol{e}^{(\boldsymbol{k})}\right|\right|_2 \leq \left|\left|\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}\right|\right|_2 \left|\left|\boldsymbol{e}^{(\boldsymbol{k})}\right|\right|_2 .$$

Thus our iteration converges if

$$\left|\left|\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}\right|\right|_2 < 1.$$

This gives the theorem:

Theorem 5. An iterative solution scheme converges for any starting $\boldsymbol{x}^{(\mathbf{0})}$ if and only if all eigenvalues of $\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}$ are less than 1 in absolute value.

Another way of saying this is "the spectral radius of $\mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}$ is less than 1."

An easier condition to check is *diagonal dominance*. The matrix A is said to be diagonally dominant if for $j = 1, 2, \ldots, n$,

$$|a_{jj}| > \sum_{i=1, i\neq j}^{n} |a_{ji}| \, .$$

Thus in a diagonally dominant matrix, a diagonal element is larger, in absolute value, than the sum of the absolute values of the rest of the row. A diagonally dominant matrix plus the right choice of Q gives the right eigenvalues, as the following theorem asserts:

Theorem 6. If the matrix A is diagonally dominant, then the Jacobi and Gauss Seidel iterative solution schemes converge for any starting $\boldsymbol{x}^{(\mathbf{0})}$.