Convergence Theorems for Two Iterative Methods

A stationary iterative method for solving the linear system:

$$Ax = b \tag{1.1}$$

employs an iteration matrix B and constant vector c so that for a given starting estimate x^0 of x, for $k = 0,1,2,\ldots$

$$x^{k+1} = Bx^k + c\,. \tag{1.2}$$

For such an iteration to converge to the solution x it must be consistent with the original linear system and it must converge. To be consistent we simply need for x to be a fixed point – that is:

$$x = Bx + c\,. \tag{1.3}$$

Since that is equivalent to $(I - B)x = c$, the consistence condition can be stated independent of x by saying

$$A(I - B)^{-1}c = b. \tag{1.4}$$

The easiest way to develop a consistent stationary iterative method is to split the matrix A:

$$A = M + N \tag{1.5}$$

then rewrite $Ax = b$ as

$$Mx = -Nx + b\,. \tag{1.6}$$

The iteration will then be

$$Mx^{k+1} = -Nx^k + b\,. \tag{1.7}$$

Recasting this in the form above we have

$$B = -M^{-1}N \text{ and } c = M^{-1}b\,.$$

It is easy to show that this iteration is consistent for any splitting as long as M is nonsingular. Obviously, to be practical the matrix M must be selected so that the system $My = d$ is easily solved. Popular choices for M are diagonal matrices (as in the Jacobi method), lower triangular matrices (as in the Gauss-Seidel and SOR methods), and tridiagonal matrices.

Convergence:

Thus, constructing consistent iterations is easy – the difficult issue is constructing *convergent* consistent iterations. However, notice that if is equation (1.3) subtracted from equation (1.2) we obtain

$$e^{k+1} = Be^k\,, \tag{1.8}$$

where e^k is the error $x^k - x$.

Our first result on convergence follows immediately from this.

Theorem 1:

The stationary iterative method for solving the linear system:

$$x^{k+1} = Bx^k + c \;\; for \;\; k = 0,1,2,...$$

converges for any initial vecrtor x^0 *if* $\|B\| < 1$ *for some matrix norm that is consistent with a vector norm*

Proof:

Let $\|.\|$ be a matrix norm consistent with a vector norm $\|.\|$ and such that $\|B\| < 1$.
We then have

$$\|e^{k+1}\| = \|Be^k\| \le \|B\|\|e^k\| \tag{1.9}$$

and a simple inductive argument shows that in general

$$\|e^k\| \le \|B\|^k \|e^0\|. \tag{1.10}$$

Since $\|B\| < 1$, $\|e^k\|$ must converge to zero (and thus x^k converge to x) independent of e^0.

■

This theorem provides a sufficient condition for convergence. Without proof we offer this theorem that provides both necessary and sufficient conditions for convergence. It employs the *spectral radius* of a matrix:

$\rho(A) =$ the absolute value of the largest eigenvalue of A in absolute value.

Theorem 2:

The stationary iterative method for solving the linear system:

$$x^{k+1} = Bx^k + c \;\; for \;\; k = 0,1,2,...$$

converges for any initial vector x^0 *if and only if* $\rho(B) < 1$.

The easiest way to prove this uses the Jordan Normal Form of the matrix B. Notice that the theorem does not say that if $\rho(B) \ge 1$ the iteration **will not converge**. It says that if $\rho(B) \ge 1$ the iteration **will not converge for some initial vector** x^0. In practical terms though the difference is minor: the only way to have convergence with $\rho(B) \ge 1$ is to have an initial error e^0 having no component in any direction of an eigenvector of B corresponding to an eigenvalue at least one in absolute value. This is a probability zero event.

The following theorem uses Theorem 1 to show the Jacobi iteration converges if the matrix is strictly row diagonally dominant. Recall that Jacobi iteration is

$$x_i^{k+1} = (b_i - \sum_{j \ne i} a_{i,j} x_i^k) / a_{i,i} \qquad \text{for } i = 1,2,...,n \tag{1.11}$$

and that strict row diagonal dominance says that

$$\sum_{j \ne i} |a_{i,j}| < |a_{i,i}| \qquad \text{for } i = 1,2,...,n. \tag{1.12}$$

The splitting for the Jacobi method is $A = D + (L+U)$, where $D, L,$ and U are the diagonal, strict lower triangle, and strict upper triangle of the matrix, respectively. Thus the iteration matrix is $-D^{-1}(L+U)$.

Theorem 3:

The Jacobi iterative method

$$x_i^{k+1} = (b_i - \sum_{j \neq i} a_{i,j} x_i^k)/a_{i,i} \qquad \text{for } i = 1, 2, \dots, n$$

for solving the linear system $Ax = b$ *converges for any initial vector* x^0 *if the matrix* A *is strictly row diagonally dominant.*

Proof:

Let $\| . \|_\infty$ indicate the infinity vector norm as well as its subordinate matrix norm. To prove the theorem it suffices to show $\left\| -D^{-1}(L+U) \right\|_\infty < 1$. To that end consider the row sums in absolute values of the matrix $-D^{-1}(L+U)$. These are $\sum_{j \neq i} \frac{|a_{i,j}|}{|a_{i,i}|}$, but property (1.12) guarantees that this is strictly less than one. The maximum of the row sums in absolute value is also strictly less than one, so $\left\| -D^{-1}(L+U) \right\|_\infty < 1$ as well. ■

The next theorem uses Theorem 2 to show the Gauss-Seidel iteration also converges if the matrix is strictly row diagonally dominant. Recall that Gauss-Seidel iteration is

$$x_i^{k+1} = (b_i - \sum_{j<i} a_{i,j} x_i^{k+1} - \sum_{j>i} a_{i,j} x_i^k)/a_{i,i} \qquad \text{for } i = 1, 2, \dots, n \tag{1.13}$$

The splitting for the Gauss-Seidel method is $A = (L+D) + U$, . Thus the iteration matrix is $-(L+D)^{-1}U$.

Theorem 4:

The Gauss-Seidel *iterative method*

$$x_i^{k+1} = (b_i - \sum_{j<i} a_{i,j} x_i^{k+1} - \sum_{j>i} a_{i,j} x_i^k)/a_{i,i} \qquad \text{for } i = 1, 2, \ldots, n$$

for solving the linear system $Ax = b$ *converges for any initial vector* x^0 *if the matrix* A *is strictly row diagonally dominant.*

Proof:

According to Theorem 2, it suffices to show $\rho(-(L+D)^{-1}U) < 1$. To that end let v be any eigenvector corresponding to an eigenvalue λ of $-(L+D)^{-1}U$ such $|\lambda| = \rho(-(L+D)^{-1}U)$. We shall show $|\lambda| < 1$ and thus $\rho(-(L+D)^{-1}U) < 1$. We have

$$Uv = -\lambda(L+D)v \tag{1.14}$$

so

$$-(L+D)^{-1}Uv = \lambda v\,. \tag{1.15}$$

In a component fashion, this says

$$\sum_{j>i} a_{i,j} v_j = -\lambda \sum_{j \le i} a_{i,j} v_j\,. \tag{1.16}$$

Let m denote an index of v corresponding to the largest component in absolute value. That is

$$|v_m| = \max_j \{|v_j|\} \tag{1.17}$$

so

$$\frac{|v_j|}{|v_m|} \le 1. \tag{1.18}$$

We also have for row m in particular

$$\begin{aligned}
\sum_{j>m} |a_{m,j}||v_j| &\ge \left|\sum_{j>m} a_{m,j} v_j\right| \\
&= |\lambda| \left|\sum_{j \le m} a_{m,j} v_j\right| \\
&= |\lambda| \left|a_{m,m} v_m + \sum_{j<m} a_{m,j} v_j\right| \\
&\ge |\lambda| \left(|a_{m,m} v_m| - \left|\sum_{j<m} a_{m,j} v_j\right| \right) \\
&\ge |\lambda| \left(|a_{m,m}||v_m| - \sum_{j<m} |a_{m,j}||v_j| \right)
\end{aligned}$$

Dividing by the necessarily positive values $|a_{m,m}|$ and $|v_m|$, we have

$$\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|} \ge \sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}\frac{|v_j|}{|v_m|} \ge |\lambda|\left(1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}\frac{|v_j|}{|v_m|}\right) \ge |\lambda|\left(1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}\right) \tag{1.19}$$

so

$$|\lambda| \le \frac{\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}}{1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}}. \tag{1.20}$$

But since $\sum_{j\ne m}\frac{|a_{m,j}|}{|a_{m,m}|} < 1$, it follows that

$$1 > \sum_{j\ne m}\frac{|a_{m,j}|}{|a_{m,m}|} = \sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|} + \sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}$$

and

$$|\lambda| \le \frac{\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}}{1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}} < 1. \quad \blacksquare$$

It is easy to show that $\frac{\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}}{1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}} < \max_i\left\{\sum_{j\ne i}\frac{|a_{i,j}|}{|a_{i,i}|}\right\}$ so the bound on the spectral radius iteration matrix of the Gauss-Seidel method is strictly less than the bound of the infinity norm of the iteration matrix of the Jacobi method. That does not guarantee that the Gauss-Seidel iteration always converges faster than the Jacobi iteration. However, it is often observed in practice that Gauss-Seidel iteration converges about twice as fast as the Jacobi iteration. To see this, imagine that $\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|} \approx \sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}$. Call this quantity $\frac{1}{2}-\theta$.

We have $\theta > 0$ and, if θ is small, then $\frac{\sum_{j>m}\frac{|a_{m,j}|}{|a_{m,m}|}}{1-\sum_{j<m}\frac{|a_{m,j}|}{|a_{m,m}|}} \approx 1-4\theta$. Yet

$\sum_{j\ne m}\frac{|a_{m,j}|}{|a_{m,m}|} \approx (\frac{1}{2}-\theta)+(\frac{1}{2}-\theta) = 1-2\theta$, and if we imagine for $\sum_{j\ne m}\frac{|a_{m,j}|}{|a_{m,m}|} \approx \max_i\left\{\sum_{j\ne i}\frac{|a_{i,j}|}{|a_{i,i}|}\right\}$, then our bound for the norm of the Jacobi iteration matrix is $1-2\theta$ while our bound on the spectral radius iteration matrix of the Gauss-Seidel method is $1-4\theta$.

Notice that if the iteration converges as $\frac{\|e^k\|}{\|e^0\|} \approx \sigma^k$, for some factor σ, then to reduce $\frac{\|e^k\|}{\|e^0\|}$ to some tolerance ε requires a value of k of about $\frac{\ln \varepsilon}{\ln \sigma}$. If $\sigma \approx 1$, then $\ln \sigma \approx -(1-\sigma)$ so we estimate about $\frac{-\ln \varepsilon}{(1-\sigma)}$ steps. With Jacobi we have $\frac{-\ln \varepsilon}{1-\sigma} \approx \frac{-\ln \varepsilon}{2\theta}$ but with Gauss-Seidel we have $\frac{-\ln \varepsilon}{1-\sigma} \approx \frac{-\ln \varepsilon}{4\theta}$ which justifies the claim that Jacobi converges twice as fast.

Lastly, without proof we state another theorem for convergence of the Gauss-Seidel iteration.

Theorem 5:

The Gauss-Seidel *iterative method*

$$x_i^{k+1} = (b_i - \sum_{j<i} a_{i,j} x_i^{k+1} - \sum_{j>i} a_{i,j} x_i^k) / a_{i,i} \qquad \text{for } i = 1, 2, \ldots, n$$

for solving the linear system $Ax = b$ converges for any initial vector x^0 if the matrix A is symmetric and positive definite.