CAAM 454/554: Stationary Iterative Methods

Yin Zhang (draft)

CAAM, Rice University, Houston, TX 77005

2007, Revised 2010

Abstract

Stationary iterative methods for solving systems of linear equations are considered by some as out of date and out of favor, as compared to methods based on Krylov subspace iterations. However, these methods are still useful in many circumstances because they are easier to implement and, more importantly, can be used as pre-conditioners in combination with Krylov-subspace methods. In this note, we briefly introduce the fundamental ideas of stationary iterative methods.

1 Introduction

We consider solving a linear system of equations

$$Ax = b, \tag{1}$$

where we will always assume, unless specified otherwise, that A is n by n and real, i.e., $A \in \Re^{n \times n}$, nonsingular, and the right-hand side (RHS) $b \in \Re^n$ is nonzero.

For any nonsingular matrix $Q \in \Re^{n \times n}$, one can rewrite the system into an equivalent form:

$$Ax = b \iff x = Mx + c, \tag{2}$$

where,

$$M = I - Q^{-1}A, \quad c = Q^{-1}b, \tag{3}$$

The equivalence can be easily seen from

$$Mx + c = x + Q^{-1}(b - Ax). \tag{4}$$

Based on the equation $x = Mx + c$, we can derive a stationary iterative method of the form: given an initial guess $x^0 \in \Re^n$, for $k = 0, 1, \cdots$, do

$$x^{k+1} = Mx^k + c, \tag{5}$$

until some stopping criterion is met. Methods of this form are called stationary because we do exactly the same thing at every iteration, which is to multiply the iterate by M and add to it the c vector.

Let x^* be the solution to $Ax = b$ or equivalently, let it satisfy

$$x^* = Mx^* + c. \tag{6}$$

Subtracting (6) from (5), we have

$$x^{k+1} - x^* = M(x^k - x^*). \tag{7}$$

Therefore, for any vector norm $\| \cdot \|$,

$$\|x^{k+1} - x^*\| \leq \|M\|\|x^k - x^*\|, \tag{8}$$

where by convention the matrix norm is the one induced by the given vector norm. Recall an induced matrix norm is defined by

$$\|M\| := \max_{x \neq 0} \frac{\|Mx\|}{\|x\|}.$$

It is clear from (8) that after every iteration the error, as is measured by the given norm, is at least reduced by a fixed factor of $\|M\|$ whenever $\|M\|$ is less than one. Therefore, we have a sufficient condition for convergence.

2

Theorem 1. *Let x^* satisfy $x^* = Mx^* + c$. The stationary iterative method (5) converges, i.e., for any initial guess x^0,*

$$\lim_{k\to\infty} x^k = x^*,$$

if $\|M\| < 1$ for some induced matrix norm.

We note that $\|M\| < 1$ implies that $I - M = Q^{-1}A$ is nonsingular, thus so is A.

We still have the freedom to choose the nonsingular matrix Q, hence the *iteration matrix $M = I - Q^{-1}A$*. Noting that $Mx = x - Q^{-1}(Ax)$, to make the iterative method practical we need to choose Q so that solving the linear system $Qx = d$ is very inexpensive, much more so than solving the original system.

We have two central questions: (1) how to choose the matrix Q, and (2) for given A and Q, whether or not the method converges?

2   Jacobi method

From Wikipedia, the free encyclopedia, it reads

> "The Jacobi method is an algorithm in linear algebra for determining the solutions of a system of linear equations with largest absolute values in each row and column dominated by the diagonal element. Each diagonal element is solved for, and an approximate value plugged in. The process is then iterated until it converges. The method is named after German mathematician Carl Gustav Jakob Jacobi."

I would not call the above a clear description of the Jacobi method (in fact, some statements are technically wrong), though it at least informs us that Jacobi was a German mathematician.

In the Jacobi method, Q is chosen as the diagonal matrix formed by the diagonal of A; that is, $Q = D$ where

$$D_{ij} = \begin{cases} A_{ii}, & i = j, \\ 0, & i \neq j. \end{cases} \qquad (9)$$

Therefore, the Jacobi method can be written into the following scheme:

$$x \leftarrow x + D^{-1}(b - Ax). \qquad (10)$$

The idea behind the Jacobi method is simple. At each iteration, one solves the i-th equation in $Ax = b$ for a new value of x_i, the i-th variable in x, while fixing all the other variables at their values in the prior iteration.

In Matlab, (10) could be implemented as

$$\mathtt{r = b - A * x; \ x = x + invd .* r;}$$

where $\mathtt{r}$ is the current residual and $\mathtt{invd} = \mathtt{1./diag(A)}$.

A natural stopping criterion is that the relative residual is less than some appropriately chosen tolerance $\epsilon > 0$:

$$\frac{\|b - Ax\|}{1 + \|b\|} < \epsilon, \qquad (11)$$

where in the denominator one is added to the norm of b to guard against the case of an excessively small right-hand side b.

When is the Jacobi method convergent? Let us inspect the ℓ_∞ norm of its corresponding iteration matrix $M = I - D^{-1}A$ (which is the maximum row sum in absolute value). Since the diagonal of M is zero, we have

$$\|M\|_\infty = \max_{1 \leq i \leq n} \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|},$$

where in the summation j is running and i is fixed. Therefore,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \ \forall i \quad \Rightarrow \quad \|M\|_\infty < 1. \qquad (12)$$

4

A matrix satisfying the left inequality in (12) is called row strictly diagonally dominant. Similarly, a matrix is called column strictly diagonally dominant if for each column the absolute value of the diagonal element is great than the sum of the absolute values of all the off-diagonal elements in that column.

Proposition 1. *If A is row strictly diagonally dominant, then the Jacobi method converges from any initial guess.*

This theorem does not say that if a matrix is not strictly diagonally dominant, then the Jacobi method does not converge. It provides a sufficient, but not necessary, condition for convergence.

Now what if A is column but not row strictly diagonally dominant? In this case, we can measure the errors in the vector norm $\|\cdot\|_D$ defined by

$$\|x\|_D := \|Dx\|_1,$$

which induces the matrix norm $\|M\|_D \equiv \|D(\cdot)D^{-1}\|_1$ since

$$\frac{\|Mx\|_D}{\|x\|_D} = \frac{\|DMx\|_1}{\|Dx\|_1} = \frac{\|DMD^{-1}(Dx)\|_1}{\|Dx\|_1}.$$

Observe that

$$DMD^{-1} = D(I - D^{-1}A)D^{-1} = I - AD^{-1}.$$

Hence

$$\|M\|_D = \|I - AD^{-1}\|_1 = \max_{1 \le j \le n} \sum_{i \ne j} \frac{|a_{ij}|}{|a_{jj}|} < 1$$

whenever A is column strictly diagonally dominant. As a result, the theorem below follows.

Theorem 2. *If A is, either row or column, strictly diagonally dominant then the Jacobi method converges from any initial guess.*

3 Fundamental Theorem of Convergence

We already mentioned that Theorem 1 is only a sufficient condition. The following theorem gives a necessary and sufficient condition in terms of the *spectral radius* of M, defined to be

$$\rho(M) = \max\{|\lambda| : \det(M - \lambda I) = 0\}, \tag{13}$$

that is, the largest modulus over the set of eigenvalues of M.

Theorem 3. *The stationary iterative method (5) converges from any initial guess if and only if*

$$\rho(M) < 1. \tag{14}$$

To prove the theorem, we will make use of the following lemma.

Lemma 1. *For any $B \in \mathbb{C}^{n \times n}$,*

$$\rho(B) = \inf_{\|\cdot\|} \|B\|, \tag{15}$$

where the infimum is taken over all induced matrix norms.

Proof. Obviously, in any vector norm and for any eigen-pair (λ, x) of B,

$$Bx = \lambda x \quad \Rightarrow \quad |\lambda|\|x\| = \|Bx\| \leq \|B\|\|x\| \quad \Rightarrow \quad |\lambda| \leq \|B\| \quad \Rightarrow \quad \rho(B) \leq \|B\|.$$

Indeed $\rho(B)$ is a lower bound of $\|B\|$ over all induced norms. We prove (15) by showing that for any $\epsilon > 0$, there exists a nonsingular matrix S that defines an induced norm $\|\cdot\|_S := \|S(\cdot)S^{-1}\|_1$ such that $\|B\|_S \leq \rho(B) + \epsilon$.

By the well-know Schur's theorem, any square matrix is similar to a upper triangular matrix. Namely, $PBP^{-1} = D + U$ where D is diagonal and U is strictly upper triangular. Furthermore, let T be the diagonal matrix with

$$T_{ii} = 1/t^i, \ i = 1, 2, \cdots, n,$$

for any arbitrary $t \neq 0$. Then

$$TPBP^{-1}T^{-1} = D + TUT^{-1}.$$

The strict upper triangular matrix $\hat{U} := TUT^{-1}$ has elements

$$\hat{u}_{ij} = \begin{cases} u_{ij}t^{j-i}, & j > i, \\ 0, & j \le i, \end{cases}$$

where the nonzero elements can be made arbitrarily small by choosing t arbitrarily small. Therefore, letting $S = TP$, we have shown that

$$SBS^{-1} = D + \hat{U}, \quad \|\hat{U}\|_1 \le \epsilon$$

for some t value chosen sufficiently small (however, $\hat{U} \ne 0$ unless B is diagonalizable). Noting that D is a diagonal matrix with the eigenvalues of B on the diagonal, we see that

$$\|B\|_S := \|SBS^{-1}\|_1 \le \|D\|_1 + \|\hat{U}\|_1 \le \rho(B) + \epsilon.$$

Since $\| \cdot \|_S$ is an induced norm (please verify), we have proved the lemma. $\qquad \square$

Now we prove Theorem 3. Since $\rho(M) < 1$ implies $\|M\| < 1$ for some induced norm $\| \cdot \|$, the sufficiency of condition (14) for convergence follows directly from Theorem 1. For necessity, let us assume that there exists an eigen-pair (λ, d) of M such that $Md = \lambda d$ and $|\lambda| \ge 1$. Let $x^0 = x^* + d$, where x^* is the solution, so that $e^0 = x^0 - x^* = d$. Then

$$\|e^k\| = \|M^k e^0\| = |\lambda|^k \|e^0\| \ge \|d\|,$$

implying non-convergence. This establishes the necessity of condition (14) for convergence from any initial point.

However, it should be clear from the proof that convergence from some initial guesses is still possible even when $\rho(M) \ge 1$.

4   Gauss-Seidel Method

Another popular stationary iterative method is the Gauss-Seidel (GS) method where Q is chosen to be the lower triangular part, including the diagonal, of A. If one

7

partitions A into three parts:

$$A = D - L - U$$

where D is the diagonal and $-L$ ($-U$) is the strictly lower (upper) triangular part of A. Then for the GS method

$$Q = D - L$$

is a lower triangular matrix (hence $Qx = r$ is easy to solve), and the corresponding iteration matrix M is

$$M = I - (D - L)^{-1}A \equiv (D - L)^{-1}U. \tag{16}$$

Both the Jacobi and the GS method solves one equation (the i-th) for one variable (the i-th) at a time. The difference is that while the Jacobi method fixes other variables at their prior iteration values, the GS method immediately uses new values once they become available. Therefore, the GS method generally converges faster.

Like the Jacobi method, the GS method has guaranteed convergence for strictly diagonally dominant matrices.

Theorem 4. *If A is, either row or column, strictly diagonally dominant then the Gauss-Seisel method converges from any initial guess.*

Proof. (i) Row case: In this case,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \Rightarrow \quad |a_{ii}| - \sum_{j < i} |a_{ij}| > \sum_{j > i} |a_{ij}|. \tag{17}$$

We will show $\rho(M) < 1$. Let (λ, x) be an eigen-pair of M where x is scaled so that $|x_j| \leq 1$ for $j = 1, 2. \cdots, n$ and $|x_i| = 1$ for some index i. From the definition of M,

$$Mx = \lambda x \quad \Rightarrow \quad \lambda(Dx - Lx) = Ux.$$

Consider the i-th equation on the right,

$$\lambda \left(a_{ii}x_i + \sum_{j < i} a_{ij}x_j \right) = - \sum_{j > i} a_{ij}x_j.$$

Taking the moduli of both sides and in view of the fact that $|x_i| = 1$ and $|x_j| \leq 1$, we have

$$|\lambda| \left(|a_{ii}| - \sum_{j<i} |a_{ij}| \right) \leq \sum_{j>i} |a_{ij}|,$$

Hence, in view of (17),

$$|\lambda| \leq \sum_{j>i} |a_{ij}| / \left(|a_{ii}| - \sum_{j<i} |a_{ij}| \right) < 1,$$

for all eigenvalues of M, confirming $\rho(M) < 1$.

(ii) Column case: In this case, consider $\rho(M) = \rho(M^T)$ and apply a similarity transformation $(D - L)^{-T} M^T (D - L)^T$ first. Details are left as an exercise. $\square$

Unlike the Jacobi method, the Gauss-Seidel method has guaranteed convergence for another class of matrices.

Theorem 5. *The Gauss-Seidel method converges from any initial guess if A is symmetric positive definite.*

We sketch a proof as follows (the details are left as an exercise). Consider an eigen-pair of M, (λ, x), and rearrange the equation $Mx = \lambda x$. We have

$$Ax = (1 - \lambda)Qx \quad \Rightarrow \quad x^* Ax = (1 - \lambda)x^* Qx = (1 - \bar{\lambda})x^* Q^* x > 0,$$

from which there holds

$$\frac{x^* Dx}{x^* Ax} = \frac{x^*(Q + Q^* - A)x}{x^* Ax} = \frac{1 - |\lambda|^2}{|1 - \lambda|^2} > 0.$$

Therefore, $|\lambda| < 1$, implying $\rho(M) < 1$.

5 SOR Method

SOR stands for Successive Over-Relaxation. It is an extension to the GS method. For SOR, the diagonal is split into two parts and distributed to both the left and the

right hand sides. That is,

$$Q = \frac{1}{\omega}D - L, \quad Q - A = U - \left(1 - \frac{1}{\omega}\right)D.$$

Therefore,

$$M = \left(\frac{1}{\omega}D - L\right)^{-1}\left(U - \left(1 - \frac{1}{\omega}\right)D\right),$$

or equivalently,

$$M(\omega) = (D - \omega L)^{-1}(\omega U + (1 - \omega)D). \tag{18}$$

Obviously, $\omega = 1$ gives the GS method. The extra degree of freedom in ω, if appropriately chosen, can generally help reduce the spectral radius of $\rho(M(\omega))$ from the value of $\rho(M(1))$ for the GS method, though an optimal ω value is generally impossible or impractical to calculate.

As an extension to the GS method, it is not surprising that the SOR method converges for symmetric positive definite matrices.

Theorem 6. *The SOR method converges from any initial guess if A is symmetric positive definite and $\omega \in (0, 2)$.*

The proof follows from a similar argument as for the GS method and is left as an exercise. In addition, the condition $\omega \in (0, 2)$ is always necessary for convergence from any initial guess.

Exercises

1. Prove that for any nonsingular matrix $S \in \Re^{n \times n}$, $\|S(\cdot)S^{-1}\|_p$ is an induced matrix norm in $\Re^{n \times n}$ for $p \geq 1$ (where $\|\cdot\|_p$ is the matrix norm induced by the vector p-norm).

2. Prove the convergence of the GS method for column strictly diagonally dominant matrices.

3. Prove Theorem 5 in details following the given sketch.

4. Prove Theorem 6.

5. Prove that a necessary condition for SOR to converge is $\omega \in (0, 2)$. (Hint: First show $\det(M(\omega)) = (1 - \omega)^n$.)

References

[1] James M. Ortega Numerical Analysis, A Second Course. Academic Press, 1972.

[2] Richard S. Varga. Matrix Iterative Analysis. Prentice-Hall, Englewood, NJ, 1962.

[3] David M. Young. Iterative Solution of Large Linear Systems. Academic Press, New York, 1971.