Chapter 3

Systems of Linear Equations - Iterative Approach

Many linear systems arised in real-world applications are large and sparse. Because of the large number of equations and unknowns, storage becomes a serious concern. When it is possible, Gaussian elimination remains a very economical, accurate, and useful algorithm. Elimination is possible as long as there is space to store all the nonzero elements of the triangular matrices associated with the elimination and when the coding necessary to locate these elements can be programmed. Techniques along this line can be found, for example, in SPARSPAK by George, A., and Liu, J.W.H., "Computer Solution of Large Sparse Positive Definite Systems".

There are cases where the orders n are so large that is is impossible to store the fill-in resulted from the Gaussian elimination. It is, therefore, desirable to solve such linear systems $Ax = b$ by methods that never alter the matrix A and never require storing more than a few vectors of length n. Iterative methods are especially suitable for this purpose.

In an iterative method, beginning with an initial vector $x^{(0)}$, we generate a sequence of vectors $x^{(1)} \to x^{(2)} \to \cdots$ according to the iteration scheme. We hope that as $k \to \infty$, $x^{(k)}$ will converge to the exact solution. The computational effort in each individual step $x^{(i)} \to x^{(i+1)}$, generally, is comparable to the multiplications of A with a vector. This is a very modest amount when A is sparse.system.

A iterative method may be motivated from the following consideration. Given a linear system

$$Ax = b \tag{3.1}$$

and an approximate solution $\tilde{x}$, the residual corresponding to $\tilde{x}$ is defined by

$$r := b - A\tilde{x}. \tag{3.2}$$

It follows that the error $e := x - \tilde{x}$ satisfies the equation

$$Ae = r. \tag{3.3}$$

If we could solve (3.3) exactly, then $x := \tilde{x} + e$ would be the solution to (3.1). In iterative methods, instead of solving (3.3) for the correction e, we solve

$$Se = r \tag{3.4}$$

where S is an approximation to A. The difference between A and S here is that (3.4) is much easier to be solved than (3.3). Now adding approximate correction to the approximate $\tilde{x}$ gives what we hope is a better approximate to the true solution x. This procedure can be summarized as follows:

(a) x^{old}: = The current approximation to x;

(b) Compute the residual $r := b - Ax^{\text{old}}$;

(c) Solve $Se = r$ for the unknown e;

(d) Set

$$x^{\text{new}} := x^{\text{old}} + e; \tag{3.5}$$

(e) Go back to (a).

Multiplying (3.5) by S yields

$$\begin{aligned} Sx^{new} &= Sx^{old} + Se \\ &= Sx^{old} + b - Ax^{old} \\ &= (S - A)x^{old} + b := Tx^{old} + b \end{aligned} \tag{3.6}$$

where

$$T := S - A \tag{3.7}$$

is called the splitting of the matrix A. Note that if the iterates converges to a limit x, then $x = x^{old} = x^{new}$ and, by (3.6), we see that $Ax = b$. In other words, a limit point of the iteration scheme (3.6) is a solution of the system (3.1).

The choice of S gives rise to different iterative schemes. For instance, if the matrix A is split as

$$A = D - L - U \tag{3.8}$$

where D, $-L$ and $-U$ are, respectively, the diagonal, the strictly lower triangular and the strictly upper triangular matrices of A. Then we may describe three classical iterative schemes as follows:

(1) <u>The Jacobi Method.</u>

$$S = D; T = L + U; \tag{3.9}$$

$$Dx^{new} = (L + U)x^{old} + b; \tag{3.10}$$

$$x_i^{new} = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{old} - \sum_{j=i+1}^{n} a_{ij}x_j^{old}). \tag{3.11}$$

(2) The Gauss-Seidel method.

$$\begin{aligned} S &= D - L;\ T = U; && (3.12)\\ (D-L)x^{new} &= Ux^{old} + b; && (3.13)\\ x_i^{new} &= \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{new} - \sum_{j=i+1}^{n} a_{ij}x_j^{old}). && (3.14) \end{aligned}$$

(3) The SOR method.

$$\begin{aligned} S &= \sigma D - L;\ T = (\sigma-1)D + U; && (3.15)\\ (D-\omega L)x^{new} &= (1-\omega)Dx^{old} + \omega U x^{old} + \omega b; && (3.16)\\ & \omega = \frac{1}{\sigma};\\ \hat{x}_i^{new} &= \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{new} - \sum_{j=i+1}^{n} a_{ij}x_j^{old}); && (3.17)\\ x_i^{new} &= (1-\omega)x_i^{old} + \omega\hat{x}_i^{new}. && (3.18) \end{aligned}$$

The main concerns raised in any iterative method are that

1. Will the sequence $(x^{(k)})$ every converge? Does the limit point depend upon the starting point $x^{(0)}$?
2. If the sequence $(x^{(k)})$ does converge, how fast? Is there a way to accelerate the convergence?

3.1 General Consideration

We realize that most of the iterative schemes are of the form

$$x^{new} = Hx^{old} + d \tag{3.19}$$

for some constant square matrix H and constant vector d. We first prove an important theorem concerning the estimate of the spectral radius of H.

Lemma 3.1.1. For any square matrix H and any $\epsilon > 0$, there exist an induced matrix norm such that

$$\rho(H) \le \|H\| \le \rho(H) + \epsilon \tag{3.20}$$

(pf): The first inequality is true for all induced norm. We prove the second inequality by construction. Given H, there exist a nonsingular matrix P such that

$$P^{-1}HP = \Lambda + U$$

where $\Lambda :=$ a diagonal matrix with $\lambda_1(H)$ as elements, and $U :=$ an upper triangular matrix with zero diagonal. (See, Linear Algebra by Lane, p184). With $\delta > 0$, define $D := \text{diag}\{1, \delta, \delta^2, \ldots, \delta^{n-1}\}$. Then $D^{-1} = \text{diag}\{1, \delta^{-1}, \ldots, \delta^{1-n}\}$. Now

$$D^{-1}P^{-1}HPD = D^{-1}(\Lambda + U)D = \Lambda + D^{-1}UD := C.$$

Note that

$$D^{-1}UD = \begin{bmatrix} 0 & x & x & x & x \\ & 0 & x & x & x \\ & & 0 & x & x \\ & & & 0 & x \\ 0 & & & & 0 \end{bmatrix} \begin{array}{l} \leftarrow \text{ elements have factor } \delta^{n-1} \\ \\ \leftarrow \text{ elements have factor } \delta^2 \\ \leftarrow \text{ elements have factor } \delta \\ \\ \end{array}$$

Now we define a vector norm $\|\cdot\|$ by

$$\|x\| := \|D^{-1}P^{-1}x\|_2. \text{ (Show that this is a norm!)}$$

Then the induced matrix norm for H is

$$\begin{aligned} \|H\| &= \sup_{\|x\|=1} \|Hx\| = \sup_{\|x\|=1} \|D^{-1}P^{-1}Hx\|_2 = \sup_{\|x\|=1} \|CD^{-1}P^{-1}x\|_2 \\ &= \sup_{\|z\|_2=1} \|Cz\|_2 \text{ where } z := D^{-1}P^{-1}x, \text{ (So, } \|x\| = 1 \Rightarrow \|z\|_2 = 1) \\ &= \|C\|_2. \end{aligned}$$

Observe that $\|C\|_2$ depends continuously on δ. When $\delta = 0$, it is obvious that $\|C\|_2 = \rho(C) = \rho(H)$ since $C = \Lambda$. Therefore, as $\delta \to 0$, $\|C\|_2 \to \rho(H)$. Now given ϵ, we may choose δ small enough such that $\|C\|_2 \le \rho(H) + \epsilon$. $\oplus$

Lemma 3.1.2. The following three statements are equivalent:

(1) $\lim_{n\to\infty} H^n = 0$,

(2) $\lim_{n\to\infty} \|H^n\| = 0$ for <u>some</u> norm,

(3) $\rho(H) < 1$.

(pf): This is a homework problem.

We now apply these results to our iterative method.

Theorem 3.1.1 *Suppose $x = Hx + d$ has a unique solution x^*. Then the sequence $\{x^{(k)}\}$ computed from (3.19) with any starting point $x^{(0)}$ converges to x^* if and only $\rho(H) < 1$.*

(pf): Observe first that $x^{(k+1)} - x^* = H(x^{(k)} - x^*) = \ldots = H^{k+1}(x^{(0)} - x^*)$. Thus
($\Rightarrow$) If $\rho(H) < 1$, then $\|x^{(k+1)} - x^*\| \le \|H^{k+1}\|\|x^{(0)} - x^*\| \to 0$ by Lemma 3.1.2 Thus $x^{(k)} \to x^*$¿
($\Leftarrow$) Suppose $\|x^{(k)} - x^*\| \to 0$ for every $x^{(0)}$. Take $x^{(0)} = x^* + e_i$. Then $x^{(k)} - x^* = H^k e_i =$ The i-th column of H^k. Use the $\|\cdot\|_1$ norm, then we have

$\|$ The i-th column of $H^k\|_1 \to 0$ as $k \to \infty$. Since i is arbitrary, it follows that $\|H^k\|_1 \to 0$. By Lemma 3.1.2 again, $\rho(H) < 1$. $\oplus$.

Remark. Suppose $\rho(H) < 1$. Consider a sequence $\{x^{(k)}\}$ generated from the scheme (3.1.1). The estimate $\|x^{(k)} - x^*\|^{1/k}$ represents the geometric average error improvement in k iterations. The sequence $\{\|x^{(k)} - x^*\|^{1/k}\}$ generally may not converge. Thus, instead, we consider the number $\beta := \inf_k \sup_{n \geq k} \|x^{(n)} - x^*\|^{1/n}$ to be the rate of convergence of the given sequence, that is, for every $\epsilon > 0$, there is a k such that for every $n \geq K$, $\|x^{(n)} - x^(\| \leq (\beta + \epsilon)^n$. Note that β depends on the starting vector $x^{(0)}$.

Definition 3.1.1 *The asymptotic convergence factor α of an iterative scheme (3.1.1) is defined to be*

$$\alpha := \sup_{x^{(0)} \neq 0} \inf_k \sup_{n \geq k} \|x^{(n)} - x^*\|^{1/n}$$

Remark. By the norm equivalence theorem and the fact that $\lim_{n \to \infty} c^{1/n} = 1$ for any nonzero constant c, it follows that β (and hence, α) is norm independent.

Theorem 3.1.2 *Suppose $\rho(H) < 1$. Then the iterative scheme has asymptotic convergence factor $\alpha = \rho(H)$.*

(pf): Since $\rho(H) < 1$, we may choose a norm $\|\cdot\|$ such that $\|H\| \leq \rho(H) + \epsilon < 1$. We have already seen that $\|x^{(k)} - x^*\| \leq \|H\|^k |x^{(0)} - x^0\|$. It follows that $\beta \leq \rho(H) + \epsilon$. Since ϵ is arbitrary, it follows that $\alpha \leq \rho(H)$. To show equality, we construct a sequence $\{x^{(k)}\}$ such that the equality holds. Toward this, we consider two cases:

(i) Suppose λ is a real eigenvalue of H such that $\|\lambda\| = \rho(H)$. Let u be the associated real unit eigenvector of λ. We choose $x^{(0)} := x^* + u$. Then $x^{(k)} - x^* = H^k u = \lambda^k u$.

For this sequence $\beta = |\lambda| = \rho(H)$.

(ii) Suppose $\rho(H)$ corresponds to a pair of complex conjugate eigenvalues λ and $\overline{\lambda}$. Let u and $\overline{u}$ be the corresponding eigenvector. We may select a basis $\{u_i\}$ for $\mathbf{C^n}$ such that $u_1 = u$ and $u_2 = \overline{u}$. Any vector $y \in \mathbf{C^n}$ may be expressed as $y = \sum_{i=1}^n c_i u_i$. We may take $\|y\| := \Sigma |c_i|$ as norm for y. Now we choose $x^{(0)} := +\frac{1}{2}(u + \overline{u})$. Then $x^{(k)} - x^* = H^k \frac{1}{2}(u + \overline{u}) = \frac{1}{2}(\lambda^k u + \overline{\lambda}^k \overline{u})$. Using the norm just defined, we have $\|x^{(k)} - x^*| = \frac{1}{2}(|\lambda|^k + |\overline{\lambda}|^k) = \rho(H)^k$. It follows that $\beta = \rho(H)$. $\oplus$.

Recall that the splitting of the matrix A

$$A = S - T$$

induces the iterative scheme

$$Sx^{new} = Tx^{old} + b$$

for the system $Ax = b$. Thus it is imperative to find conditions such that $\rho(S^{-1}T) < 1$. We discuss below several possible sufficient conditions that has been established in the literature (cf: R. S. Varga, Matrix Iterative Analysis).

Definition 3.1.2 *Let $A \in \mathbf{R^{n\times n}}$. Then $A = S - T$ is said to be a regular splitting if $S^{-1} \geq 0$ and $T \geq 0$.*

Theorem 3.1.3 *If $A \in \mathbf{R^{n\times n}}$, $\mathbf{A^{-1} \geq 0}$ and $A = S - T$ is a regular splitting, then $\rho(S^{-1}T) < 1$.*

(pf): Let $H := S^{-1}T$. Then $H \geq 0$ and $S^{-1}A = I - H$. Note that $(I + H + \ldots + H^m)(I - H) = I - H^{m+1}$. It follows that $0 \leq (I + H + \ldots + H^m)S^{-1} = (I - H^{m+1})A^{-1} \leq A^{-1}$. Let $D^{(m)} := I + H + \ldots + H^m$, $F := D^{(m)}S^{-1}$. Denote $A^{-1} = (\alpha_{ij})$, $S^{-1} = (\beta_{ij})$. Observe that $f_{ik} = \sum_{j=1}^{n} d_{jk}^{(m)}$ = linear combinations of $d_{ij}^{(m)}$ with nonnegative coefficients. Since S^{-1} is nonsingular, no row of S^{-1} can be identically zero. It follows that each $d_{ij}^{(m)}$ must involve at least once with one of f_{i*}. For each j, there exists a column index $k(j)$ such that $\beta_{jk(j)} > 0$. Then $f_{ik(j)} = \Sigma d_{ij}^{(m)}\beta_{jk(j)} \leq \alpha_{ik(j)}$ implies that $d_{ij}^{(m)}\beta_{jk(j)} \leq \alpha_{ik(j)}$. Thus $d_{ij}^{(m)}$ is always bounded above. Since $\{d_{ij}^{(m)}\}$ is a monotone increasing sequence as $m \to \infty$. It follows that $\sum_{m=0}^{\infty} H^m$ converges. So $\lim_{m\to\infty} H^m = 0$. $\oplus$

Definition 3.1.3 *A nonsingular matrix $A \in \mathbf{R^{n\times n}}$ is said to be an M-matrix if $a_{ij} \leq 0$ for $i \neq j$, and if $A^{-1} \geq 0$.*

Remark. Suppose A is an M-matrix. Then $a_{ii} > 0$.

Theorem 3.1.4 *If $A \in \mathbf{R^{n\times n}}$ is an M-matrix. Then both the Jacobi splitting (3.10) and the Gauss-Seidel splitting are regular. In this case, both the Jacobi method and the Gauss-Seidel method are convergent.*

(pf): In the Jacobi method, $S = D$ and $T = L + U$. Since $D > 0$, $S^{-1} = D^{-1} > 0$. Now $T = L + U \geq 0$ by the definition of M-matrix. This shows the Jacobi splitting is regular. Convergence follows from Theorem 3.1.2.

In the Gauss-Seidel method, $S = D - L$ and $T = U$. Obviously S is nonsingular and $T \geq 0$. To show regular splitting, we need to show $S^{-1} \geq 0$. Now $S^{-1} = (D - L)^{-1} = (I - D^{-1}L)-1D^{-1}$. Note that $D^{-1}L \geq 0$. Since $D^{-1}L$ is a strictly lower triangular matrix, $(D^{-1}L)^{m+1} = 0$ whenever $m + 1 \geq n$. It follows that $(I - D^{-1}L)(I + D^{-1}L + \ldots + D^{-1}L)^m) = I - (D^{-1}L)^{m+1} = I$. But then $(I - D^{-1}L)^{-1} = I + D^{-1}L + \ldots + (D^{-1}L)^m \geq 0$. This shows that the Gauss-Seidel splitting is regular. Convergence follows from Theorem 3.1.2. $\oplus$

Definition 3.1.4 *A matrix $A \in \mathbf{R^{n\times n}}$ is said to be (strictly, if inequality holds) row-wise diagonally dominant if $|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ for all i.*

Theorem 3.1.5 *Both the Jacobi method and the Gauss-Seidel method converge if A is strictly diagonally dominant.*

(pf): In the Jacobi method, $J := H := D^{-1}(L+U)$. Taking the L_∞-norm, we have $\|J\|_\infty = \max_i \frac{1}{\|a_{ii}\|}\sum_{k\neq i}|a_{ik}| < 1$. The convergence follows from Lemma 3.1.1 and Theorem 3.1.1. In the Gauss-Seidel method, $G := H := (D-L)^{-1}U = (I-D^{-1}L)^{-1}D^{-1}U$. Note that $|J_e| \leq \|J\|_\infty e$. Thus $|D^{-1}U|e \leq (\|J\|_\infty I - |D^{-1}L|)e$. Note also that $0, \leq |(I-D^{-1}L)^{-1}| = |I + D^{-1}L + \ldots + (D^{-1}L)^{n-1}| \leq (I-|D^{-1}L|)^{-1}$. Thus $|G|e \leq (I-|D^{-1}L|)^{-1}(I-|D^{-1}L| + (\|J\|_\infty - I)I)e = (\|J\|_\infty - 1)(I-|D^{-1}L|)^{-1}e \leq (I+(\|J\|_\infty-1)I)e = \|J\|_\infty e$. It follows that $\|G\|_\infty \leq \|J\|_\infty < 1$. $\oplus$

Remark. In Theorem 3.1.5 we have actually proved a stronger result $\|G\|_\infty \leq \|J\|_\infty$. But it is not necessarily true that $\rho(G) \leq \rho(J)$. That is, it is not true, in general, that the Gauss-Seidel method converges at least as fast as the Jacobi method, although intuitively it seems this should be so. (cf: Stein and Rosenberg Theorem in Varga).

Theorem 3.1.6 *If A is symmetric and positive definite, then the Gauss-Seidel method converges.*

(pf): In the Gauss-Seidel method, $G := (D-L)^{-1}L^T$. Consider $G_1 := D^{1/2}GD^{-1/2} = (I-L_1)^{-1}L_1^T$ with $L_1 := D^{-1/2}LD^{-1/2}$.

Since G and G_1 are similar, G and G_1 have the same eigenvalues. Suppose $G_1x = \lambda x$ with $x^*x = 1$ (Note that x may be in $\mathbf{C^n}$). Then $L_1^Tx = \lambda(I-L_1)x$. It follows that $x^*L_1x = \lambda(1-x^*L_1x)$. Let $X^*L_1x = a+ib$. Then $\frac{|\lambda|^2=|a-ib}{1-a-ib|^2=\frac{a^2+b^2}{1-2a+a^2+b^2}}$. Note that $D^{-1/2}AD^{-1/2} = I - L_1 - L_1^T$ is still positive definite. Thus $1 - x^*L_1x - x^*L_1^Tx = 1-2a > 0$. It follows that $|\lambda| < 1$. $\oplus$

3.2 Relaxation Method

In the relaxation methods, we consider classes of matrices H that depend on certain parameters. The main idea is to vary these parameters so that the corresponding asymptotic convergence factor $\rho(H)$ becomes as small as possible.

One of the most popular relaxation methods is the SOR method (3.16) where

$$H(\omega) = (D-\omega L)^{-1}[(1-\omega)D + \omega U)]$$

Theorem 3.2.1 *Suppose $A \in \mathbf{R^{n\times n}}$ has nonzero diagonal elements. Then $\rho(H(\omega)) \geq |1-\omega|$. So for convergence of SOR, it is necessary to have $0 < \omega < 2$.*

(pf): We first observe that $\det(H(\omega)) = \det(D-\omega L)^{-1}\det[(1-\omega)D+\omega U] = \det(D^{-1})\det((1-\omega)D = (1-\omega)^n$. On the other hand, $\det(H(\omega)) = \prod_{i=1}^{n}\lambda_i$ with λ_i eigenvalues of $H(\omega)$. It follows that $|(1-\omega)^n| = \prod_{i=1}^{n}|\lambda_i| \leq (\rho(H(\omega)))^n$. The assertion follows. $\oplus$

Theorem 3.2.2 *(Ostrowski and Reich Theorem) Let A be real, symmetric and positive definite. Then the SOR method converges if and only if $0 < \omega < 2$.*

(pf): The SOR method comes from the splitting $A = S - T$ where

$$S := \omega^{-1}D - L \text{ and } T := (\omega^{-1} - 1)D + U.$$

Obviously S is nonsingular. Let

$$Q := A - (S^{-1}T)^T A(S^{-1}T).$$

We claim that both $S + T$ and Q are positive definite. Suppose these claims are true. Let λ be any eigenvalues of $H = S^{-1}T$, and y the corresponding eigenvector. Then $0 < y^*Qy = y^*AY - \overline{\lambda}y^*A\lambda y = (1 - |\lambda|^2)y^*Ay$. It follows that $\|\lambda| < 1$, and hence $\rho(H) < 1$.

Now we prove the claims. Recall that any given matrix M can be written as $M = \frac{1}{2}(M + M^T) + \frac{1}{2}(M - M^T) := M_s + M_k$ where M_s is symmetric and M_k is skew-symmetric. Note also that $x^TMx = x^TM_sx$. So it suffices to check the symmetric part of $S + T$ for positive definiteness. Now $(S+T)_s = \frac{1}{2}\{S+S^T+T+T^T\} = \frac{1}{2}\{(\omega^{-1}D-L)+(\omega^{-1}D-U)+((\omega^{-1}-1)D+U)+((\omega^{-1}-1)D+L)\} = D(\omega^{-1}(2-\omega))$ which obviously is positive definite. To check the matrix Q for positive definiteness, we first observe that $Q = A - H^TAH = A-(I-S^{-1}A)^TA(I-S^{-1}A) = A-\{A-(S^{-1}A)^TA-A(S^{-1}A)^TA(S^{-1}A)\} = (S^{-1}A)^T\{S + S^T - A\}(S^{-1}A) = (S^{-1}A)^T\{S^T + T\}(S^{-1})$. Now $x^TQx = x^T(S^{-1}A)^T\{S^T + T\}(S^{-1}A)x = y^T(S^T + T)y = y^T(S + T)y > 0$ for all $x \neq 0$. So Q is positive definite. $\oplus$

Remark. It is often possible to choose the parameter ω so that the SOR method converges rapidly; much more rapidly than the Jacobi method or the Gauss-Seidel method. Normally, such an optimum value for ω can be prescribed if the coefficient matrix A, relative to the partitioning imposed, has the so called property A and is the so called consistently order (cf: Hageman and Young, Applied Iterative Methods, Chapter 9). In practice, the estimates for the SOR ω is obtained by an adaptive procedure.

3.3 Acceleration Methods

In this section we discuss a general procedure for accelerating the rates of convergence of basic iterative methods. The procedure involves the formation of a new vector sequence from linear combinations of the iterates obtained from the basic method.

Let $\{x^{(k)}\}$ be the sequence of iterates generated by a basic method (3.19. That is, $\{x^{(k)}\}$ is formed by

$$x^{(k)} = Hx^{(k-1)} + d. \tag{3.21}$$

Then the error vector $e^{(k)} := x^{(k)} - x^*$ satisfies

$$e^{(k)} = H^k e^{(0)}. \tag{3.22}$$

We consider a new vector sequence $\{u^{(k)}\}$ determined by the linear combination

$$u^{(k)} := \sum_{i=0}^{k} \alpha_{k,i} x^{(i)}, k = 0, 1, \ldots \tag{3.23}$$

where the real numbers $\alpha_{k,i}$ are required to satisfy the consistency condition

$$\sum_{i=0}^{k} \alpha_{k,i} = 1, \quad k = 0, 1, \ldots \tag{3.24}$$

Let $\epsilon^{(k)} := u^{(k)} - x^*$. Then we have

$$\begin{array}{rcl} \epsilon^{(k)} & = & \sum_{i=0}^{k} \alpha_{k,i} x^{(i)} - x^* = \sum_{i=0}^{k} \alpha_{k,i} e^{(i)} \\ \epsilon^{(k)} & = & \sum_{i=0}^{k} \alpha_{k,i} x^{(i)} = \sum_{i=0}^{k} \alpha_{k,i} e^{(i)} \\ & = & (\sum_{i=0}^{k} \alpha_{k,i} H^i) e^{(0)} = (\sum_{i=0}^{k} \alpha_{k,i} H^i) \epsilon^{(0)} \\ & := & Q_k(H) \epsilon^{(0)} \end{array}$$

where

$$Q_k(H) := \alpha_{k,0} I + \alpha_{k,1} H + \ldots + \alpha_{k,k} H^k$$

is a matrix polynomial. The idea is to choose the polynomials $\{Q_k\}$ so that $\{u^{(k)}\}$ converges to x^* faster than $\{x^{(k)}\}$. Generally speaking, it requires a high arithmetic cost and a large amount of storage in using (3.23) to obtain $u^{(k)}$. Alternatively, we usually consider only the important family of polynomials satisfying the recurrence relation:

$$\begin{array}{rcl} Q_0(x) & = & 1 \\ Q_1(x) & = & \gamma_1 x - \gamma_1 + 1 \\ Q_{k+1}(x) & = & \rho_{k+1}(\gamma_{k+1} x + 1 - \gamma_{k+1}) Q_k(x) + (1 - \rho_{k+1}) Q_{k-1}(x), \text{ for } k \geq 1 \end{array} \tag{3.25}$$

where $\gamma_1, \rho_2, \gamma_2, \ldots$ are real numbers to be determined. We note that the consistency condition (3.24) is satisfied automatically for all $k \geq 0$.

Theorem 3.3.1 *If the polynomial sequence $\{Q_k\}$ in (3.25) is used, then the iterates $\{u^{(k)}\}$ of (3.23) may be obtained using the three-term relation:*

$$\begin{array}{rcl} u^{(1)} & = & \gamma_1 (H u^{(0)} + d) + (1 - \gamma_1) u^{(0)}, \\ u^{(k+1)} & = & \rho_{k+1} \{\gamma_{k+1} (H u^{(k)} + d) + (1 -, \gamma_{k+1}) u^{(k)} + (1 - \rho_{k+1}) u^{(k-1)}. \end{array} \tag{3.26}$$

(pf): This is a homework problem. $\oplus$ The polynomial $\{Q_k\}$ may be chosen to fulfill one of two purposes:

(1) (Conjugate Gradient Acceleration) From (3.3), we have for any norm that

$$\|\epsilon^{(k)}\| \leq \|Q_k(H) \epsilon^{(0)}\|. \tag{3.27}$$

So the polynomial sequence $\{Q_k\}$ may be chosen to minimize $\|Q_k(H)\epsilon^{(0)}\|$. (cf: Hageman and Young, Chapter 7).

(2) (Chebyshev Acceleration) This is motivated by the fact that

$$\|\epsilon^{(k)}\| \leq \|Q_k(H)\|\|\epsilon^{(0)}\| \leq (\rho(Q_k(H)) + \epsilon)\|\epsilon^{(0)}\| \tag{3.28}$$

for a certain norm. We note that

$$\rho(Q_k(H)) = \max_{1\leq i\leq n} \|Q_k(\lambda_i)|. \tag{3.29}$$

Let $M(H)$ and $m(H)$ denote, respectively, the algebraically largest and smallest eigenvalues of H. So the polynomial $\{Q_k\}$ is chosen such that the virtual spectral radius of $Q_k(H)$ defined by

$$\overline{\rho}(Q_k(H)) := \max_{m(H)\leq x\leq M(H)} |Q_k(x)| \tag{3.30}$$

is minimized. (cf: Hageman and Young, Chapter 4-6).

3.4 Conjugate Gradient Method

The conjugate gradient method is a very useful technique in many areas of numerical computation. In this section, we shall study how it can be applied to solve the linear system (3.1) when A is symmetric and positive definite.

Consider the quadratic functional $F : \mathbf{R^n} \to \mathbf{R}$ where

$$F(x) = \frac{1}{2}x^T Ax - x^T b. \tag{3.31}$$

Suppose $\overline{x}$ is the solution to (3.1). Then

$$F(x) = F(\overline{x}) + \frac{1}{2}(x - \overline{x})^T A(x - \overline{x}). \tag{3.32}$$

It follows that the problem of solving $Ax - b$ is equivalent to the problem of minimizing $F(x)$. Moreover, the gradient of $F(x)$ is given by

$$\bigtriangledown F(x) = Au - b. \tag{3.33}$$

The direction of the vector $\bigtriangledown F(x)$ is the direction for which the functional $F(x)$ at the point x changes most rapidly. Suppose $x^{(k)}$ is an approximation to $\overline{x}$, then in the direction of steepest descent $r_k := - \bigtriangledown F(x^{(k)})$ we should obtain an improved approximation

$$x^{(k+1)} := x^{(k)} - \alpha_k r_k \tag{3.34}$$

if α_k is chosen to minimize $F(x^{(k)} + \alpha r_k))$. Using (3.31), we can easily calculate the number a_k. Thus we have derived

Algorithm 3.4.1. (The Steepest Descent Method)

Given $x^{(0)}$ arbitrary
For $k = 0, 1, \ldots,$
  $r_k := b - Ax^{(k)}$
  If $r_k = 0$
    then stop
  Else

$$\begin{aligned} \alpha_k &:= \frac{r_k^T r_k}{r_k^T A r_k}; (\text{Why?}) \\ x^{(k+1)} &:= x^{(k)} + \alpha_k r_k. \end{aligned}$$

When the condition number $k_2 = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ is large, the level curves of F are very elongated hyperellipsoids and minimization corresponds to finding the lowest point on a relatively flat, steep-side valley. In steepest descent, we are forced to traverse back and forth across the valley rather than down the valley. This is a slow process. We would like to choose certain descent directions $\{p_k\}$ other than $\{r_k\}$.

Definition 3.4.1 *Given a symmetric and positive definite matrix A, two vectors d_1, d_2 are said to be A-conjugate if and only $d_1^T A d_2 = 0$. A finite set of vectors $d_0, \ldots, d_k$ is called an A-conjugate set if $d_i^T A d_j = 0$ for all $i \neq j$.*

Lemma 3.4.1. If $\{d_0, \ldots, d_{n-1}\}$ is an A-conjugate set, then $d_0, \ldots, d_{n-1}$ are linearly independent.

Lemma 3.4.2. If $d_0, \ldots, d_{n-1}$ are A-conjugate, then the solution x^* to (3.1) may be written as

$$x^* = \sum_{i=0}^{n-1} (\frac{d_i^T b}{d_i^T A d_i}) d_i. \tag{3.35}$$

(pf): By Lemma 3.4.1, $d_0, \ldots, d_{n-1}$ form a basis of $\mathbf{R^n}$. The solution x^* to (3.1) has a unique representation $x^* = \sum_{j=0}^{n-1} \gamma_j d_j$. Also, we have $b = \sum_{j=0}^{n-1} \gamma_j A d_j$. Taking inner product of B and d_i, $\gamma_i = \frac{d_i^T b}{d_i^T A d_i}$. $\oplus$

Theorem 3.4.1 *(Conjugate Direction Theorem) Let $\{d_0, \ldots, d_{n-1}\}$ be a set of nonzero A-conjugate vectors. For any $x^{(0)} \in \mathbf{R^n}$, the sequence $\{x^{(k)}\}$ generated by*

$$x^{(k+1)} = x^{(k)} + \alpha_k d_k, \quad k \geq 0 \tag{3.36}$$

with

$$\alpha_k = \frac{r_k^T d_k}{d_k^T A d_k} \tag{3.37}$$

$$r_k = b - Ax^{(k)} \tag{3.38}$$

converges to the unique solution x^* of $Ax = b$ after n steps, i.e., $x^* = x^{(0)} + \sum_{j=0}^{n-1} \alpha_j d_j$.

(pf); Suppose $x^* - x^{(0)} = \sum_{j=0}^{n-1} \delta_j d_j$. Then

$$\delta_i = \frac{d_i^T(b - Ax^{(0)}}{d_i^T A d_i}) = \frac{d_i^T(r_k + Ax^{(k)} - Ax^{(0)})}{d_i^T A d_i}$$

for all $k \geq 0$. Observe from (3.36), $x^{(k)} - x^{(0)} = \sum_{j=0}^{k-1} \alpha_j d_j$. So $d_i^T(Ax^{(k)} - Ax^{(0)}) = 0$ for all $i \geq k$. That is, we have shown that $\delta_i = \frac{d_i^T r_i}{d_i^T A d_i} = \alpha_i$. $\oplus$

Theorem 3.4.2 *(Expanding Space Theorem) Let $(d_0, \ldots, d_{n-1})$ be a set of A-conjugate vectors. Let $S_k := [d_0, \ldots, d_{k-1}]$ denote the n-dimensional subspace spanned by the vectors $d_0, \ldots, d_{k-1}$. For any $x^{(0)} \in \mathbf{R^n}$, the sequence $\{x^{(k)}\}$ generated from the scheme (3.36) and (3.37) has the property that*

$$F(x^{(k)}) = \min_{\alpha} F(x^{(k-1)} + \alpha d_{k-1}). \tag{3.39}$$

In fact,

$$F(x^{(k)}) = \min_{x \in X^{(0)} + S_k} F(x) \tag{3.40}$$

(pf): Define $g(\alpha) = F(x^{(k-1)} + \alpha d_{k-1})$. Then $g'(\alpha) = d_{k-1}^T \bigtriangledown F(x^{(k-1)} + \alpha d_{k-1}) = d_{k-1}^T(A(x^{(k-1)} + \alpha d_{k-1}) - b) = d_{k-1}^T(\alpha A d_{k-1} - r_{k-1})$. It follows that the optimum value of α is given by (3.37).

To show that $x^{(k)}$ is a minimizer over the linear variety $x^{(0)} + S_k$, it suffices to show that $\bigtriangledown F(x^{(k)}) = r_k$ is perpendicular to S_k. Now for $k = 1$, we have $d_0^T r_1 = d_0^T(b - A(x^{(0)} + \alpha_0 d_0)) = 0$ by the definition of α_0. For $k = 2$, we have $d_1^T r_2 = d_1^T(b - A(x^{(1)} + \alpha_1 d_1)) = 0$ by the definition of α_1, and $d_0^T r_2 = d_0^T(r_1 - \alpha_1 A d_1) = 0$. The assertion follows from induction. $\oplus$

Remark. In the above theorem, we have actually proved the fact that

$$d_i \perp r_k$$

for all $i < k$.

Algorithm 3.4.2. (The Conjugate Gradient Method)

Given $x^{(0)} \in \mathbf{R^n}$ arbitrary

$$d_0 := r_0 := b - Ax^{(0)}$$

For $k = 0, 1, \ldots, n-1$

If $r_k = 0$

Stop

Else

$$\alpha_k : = \frac{r_k^T d_k}{d_k^T A d_k} \text{ (Can be replaced by (3.46).)} \tag{3.41}$$

$$x^{(k+1)} := x^{(k)} + \alpha_k d_k \tag{3.42}$$

$$r_{k+1} := b - Ax^{(k+1)} \tag{3.43}$$

$$\beta_k := -\frac{r_{k+1}^T A d_k}{d_k^T A d_k} \text{ (Can be replaced by (3.47).)} \tag{3.44}$$

$$d_{k+1} : = r_{k+1} + \beta_k d_k. \tag{3.45}$$

Theorem 3.4.3 *(Conjugate Gradient Theorem) The conjugate gradient algorithm is a conjugate direction method. If it does not terminate at $x^{(k)}$, then*

1. $\text{span}\{r_0, r_1, \ldots, r_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$;
2. $\text{span}\{d_0, d_1, \ldots, d_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$;
3. $d_k^T A d_i = 0$ for $i < k$;
4. $$\alpha_k = \frac{r_k^T r_k}{d_k^T A d_k}; \tag{3.46}$$
5. $$\beta_k = +\frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}. \tag{3.47}$$

(pf): All proofs should be completed by induction.

(a) When $k = 0$, the case is trivial. Suppose the statement (a) is true for k. Want to show $\text{span}\{r_0, r_1, \ldots, r_{k+1}\} = \text{span}\{r_0, Ar_0, \ldots, A^{k+1} r_0\}$. Note

$$r_{k+1} = b - Ax^{(k+1)} = r_k - \alpha_k A d_k.$$

Note also

$$r_k \in \text{span}\{r_0, r_1, \ldots, r_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\} \subset \text{span}\{r_0, Ar_0, \ldots, A^{k+1} r_0\}.$$

But by construction, $d_k \in \text{span}\{r_0, r_1, \ldots, r_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$. Therefore, $r_{k+1} \in \text{span}\{r_0, Ar_0, \ldots, A^{k+1} r_0\}$. This shows that

$$\text{span}\{r_0, r_1, \ldots, r_{k+1}\} \subset \text{span}\{r_0, Ar_0, \ldots, A^{k+1} r_0\}.$$

Now we need to show

$$A^{k+1} r_0 \in \text{span}\{r_0, r_1, \ldots, r_{k+1}\}.$$

Note that $r_{k+1} = \sum_{i=0}^{k+1} \gamma_i A^i r_0$. Since

$$r_{k+1} \notin \text{span}\{r_0, r_1, \ldots, r_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\},$$

$\gamma_{k+1} \neq 0$. So $A^{k+1} r_0$ can be written as a linear combination of $r_0, \ldots, r_{k+1}$.

(b) The proof is similar to (a).

(c) Assume $d_k^T A d_i = 0$ for $i < k$. Want to show $d_{k+1}^T A d_i = 0$ for $i < k+1$. By construction, $d_{k+1}^T A d_i = r_{k+1}^T A d_i + \beta_k d_k^T A d_i$. If $i = k$, then $d_{k+1}^T A d_i = 0$ by the definition of β_k. If $i < k$, then $d_{k+1}^T A d_i = r_{k+1}^T A d_i = 0$ by (a) and (b) since $Ad_i \in \text{span}\{d_0, \ldots, d_{i+1}\}$ and $r_{k+1} \perp \text{span}\{d_0, \ldots, d_{i+1}\}$.

(d) By definition, $\alpha_k = \frac{r_k^T d_k}{d_k^T A d_k}$ and $d_k = r_k + \beta_{k-1} d_{k-1}$. So $\alpha_k = \frac{r_k^T r_k + \beta_{k-1} r_k^T d_{k-1}}{d_k^T A d_k} = \frac{r_k^T r_k}{d_k^T A d_k}$ since $r_k^T d_{k-1} = 0$.

(e) By definition, $\beta_k := -\frac{r_{k+1}^T A d_k}{d_k^T A d_k} = -\frac{r_{k+1}^T A d_k}{r_k^T r_k/\alpha_k} = -\frac{r_{k+1}^T (r_k - r_{k+1}/\alpha_k}{r_k^T r_k/\alpha_k} = +\frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$ because by (3.45) $r_k = d_k - \beta_{k-1} d_{k-1}$.

Remark. In exact arithmetic, the conjugate gradient method would have reached the solution in exactly n iterations. Because of the effect of floating-point arithmetic, the computed r_n generally is different from zero. In practice, therefore the method is simply continued until r_k is found sufficiently small.