Introduction to Numerical Analysis I
Handout 15

1 Numerical Linear Algebra (cont)

1.3 Iteration Methods

All the direct methods are we have learned require

1. $O(n^3)$ computational operations, and
2. $O(n^2)$ of memory, (in general) regardless the sparsity of the matrix. For example LU decomposition of a space matrix may be a full matrix.

We would like to find a solution to the linear system $Ax = b$ faster then $O(n^3)$. In order to do so we would to compromise on the accuracy, that is we will look for an approximation to the solution of $Ax = b$ instead of an exact (unto round-off error) solution. We will construct a sequence of approximations asymptotically convergent to the exact solution. Such methods are called iterative methods, since we compute the next term in the sequence, i.e. next approximation, until the error is small enough for our needs.

Iterative methods would often require less memory then any direct method, for example it could be implemented without keeping the matrix A in the memory. That is one can solve a very big linear system almost without limitations prescribed by the memory size.

For an iterative method one requires:

1. the computation of x_k from x_{k-1} to be (computationally) cheap, and also
2. the convergence of x_k to be fast.

However such requirements are little contradictive, for example $x_k = A^{-1}b$ will converge fast but not cheap, whenever $x_k = x_{k-1}$ is cheap but never converge.

In most of cases the cost of computing x_k will be matrix-vector multiplication dominated for some matrix B derived from the matrix A. The main idea is to describe the matrix A as a sum of matrices, e.g. $Ax = A_1x + A_2x = b$. Assuming A_1 is invertible gives $x = -A_1^{-1}A_2x + A_1^{-1}b$. Next, let $B = -A_1^{-1}A_2$ and $C = -A_1^{-1}b$ to get $x = Bx + C$. That is, x is a fixed point of iteration function $g(x) = Bx + C$. We denote B Iteration Matrix, and a Fixed Point Iteration is given by $x_k = Bx_{k-1} + C$.

Theorem 1.1 (Convergence of the iterative method). The convergence of an iterative method is understood as $\lim_{k\to\infty} \|e_k\| \to 0$ which gives

$$\|e_k\| = \|x_k - x\| = \|(Bx_{k-1} + C) - (Bx + C)\| = \|B(x_{k-1} - x)\| \leq$$
$$\leq \|B\| \|x_{k-1} - x\| = \|B\| \|e_{k-1}\| \leq \|B\|^2 \|e_{k-2}\| \leq \ldots \leq \|B\|^k \|e_0\|$$

That is, **the sufficient criteria** for the convergence would be $\|B\|^k \|e_0\| \to 0$, or more precisely $\|B\| < 1$ in some norm.

Note also the similarity to the fixed point theorem:

$$\|g'(x)\| = \left\|\tfrac{d}{dx}(Bx + C)\right\| = \|B\| < 1.$$

The words "in some norm" says that it is in genral possible that $\|B\|_* > 1$ in some another norm $\|\cdot\|_* > 1$. Since it is impossible to check all the norms one requires better criteria.

Theorem 1.2. The **necessary and sufficient criteria** for the iterative method to converge is $\rho(B) < 1$. In other words, if $\rho(B) < 1$ there exist a norm for which $\|B\| < 1$.

In the following discussion we consider the form of $A = L + D + U$, where L is lower triangular matrix, D is diagonal matrix and U is upper triangular matrix.

$$\begin{pmatrix} a_{1,1} & a_{1,2} & & a_{1,n} \\ a_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ a_{n,1} & & a_{n,n-1} & a_{n,n} \end{pmatrix} = \begin{pmatrix} a_{2,1} & & 0 \\ \vdots & \ddots & \\ a_{n,1} & & a_{n,n-1} \end{pmatrix} +$$
$$\begin{pmatrix} a_{1,1} & & 0 \\ & \ddots & \\ 0 & & a_{n,n} \end{pmatrix} + \begin{pmatrix} a_{1,2} & & a_{1,n} \\ & \ddots & \vdots \\ 0 & & a_{n-1,n} \end{pmatrix}$$

1.3.1 Jacobi Iteration Method

Jacobi method suggests to solve $Ax = (L+D+U)x = b$ for $Dx = -(L + U)x + b$. Assuming that the diagonal entries of A (that is of D) are all non zero. In this case $\left(D^{-1}\right)_{ii} = (D)_{ii}^{-1}$, that is it is significantly (computationally) cheap to compute the inverse matrix of the diagonal matrix D. Finally, $x = -D^{-1}(L + U)x + D^{-1}b = Bx + C$ which defines the iteration (for $i = 1, 2, \ldots n$)

$$x_i^{k+1} = D^{-1}(b - (A - D)x^k = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij}x_j^k\right),$$

Definition 1.3. A matrix is called row (column) Diagonal Dominant if the absolute value of the diagonal entry is greater then sum of absolute values of the other entries in the row (column). That is the row DD matrix satisfies $|a_{ii}| > \sum_{j\neq i=0}^{n} |a_{ij}|$ for every i, and the column DD matrix satisfies $|a_{jj}| > \sum_{i\neq j=0}^{n} |a_{ij}|$ for every j.

Theorem 1.4. Let A be diagonal dominant matrix (either row DD or column DD), then the Jacobi Iteration for A is convergent.

Proof hint: Show that $\|B\|_\infty < 1$ for row DD, and $\|B\|_1 < 1$ for column DD.

1.3.2 Gauss Seidel

The Jacobi Iteration can be improved very easily. Note that when we compute the component $i > 1$ of x^{k+1}, the values of x_j^{k+1} for $j < i$ are already known, but we still use the old x_j^k. Gauss Seidel method uses the newest values of x as soon as they are known. This can be formulated as

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$$

Which is a *forward substitution* of

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

that is $B = -(L + D)^{-1}U$ and $C = (L + D)^{-1}b$, since $L + D$ is lower triangular matrix. One expresses it as

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

Theorem 1.5. Let A be diagonal dominant matrix (either row DD or column DD), then the Gauss Seidel Iteration for A is convergent.

1.3.3 Relaxation Techniques

It is known that error often have a high oscillatory behavior. Relaxation techniques are often related to the methods that smoothes the the oscillatory behavior of the error.

Definition 1.6. Residual Vector Let $\tilde{x}$ be an approximation to the solution of the linear system $Ax = b$ then the residual vector reads for $r = b - A\tilde{x}$

Note that residual vector r is not the error vector which is defined as $e = x - \tilde{x}$. The relationship between the relationship between the error and the residue is given by $r = b - A\tilde{x} = b - A(x - e) = Ae$. Furthermore, for an ill-conditioned matrix the residual can be small while the error is big(why?).

The i'th entry of the residual vector is given by $r_i = b_i - \sum_{j=1}^n a_{i,j}x_j$ rewrite it as $r_i + a_{ii}x_i = b_i - \sum_{i\neq j=1}^n a_{i,j}x_j$. This give the following formulation of Jacobi Iteration:

$$x_i^{k+1} = \frac{r_i^k + a_{ii}x_i^k}{a_{ii}} = x_i^k + \frac{r_i^k}{a_{ii}}.$$

This has a general form of $x_{k+1} = x_k + Pr_k$, the error is given by $\|e_{k+1}| = \|x^{k+1} - x\| = \|x_k + Pr_k - x\| \leq \|x_k - x\| + \|Pr_k\| \leq e_0 + \|P\| \sum_{j=0}^k r_j$ which hints that to reducing the residual part of the error accelerates the convergence.

Jacobi Over Relaxation(JOR) : JOR defined by

$$x_i^{k+1} = x_i^k + \omega \frac{r_i^k}{a_{ii}} = x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^n a_{i,j}x_j \right),$$

where ω is an acceleration parameter, or in other words, a relaxation factor. One formulate it as

$$x_{k+1} = (\omega B + (1 - \omega)I)x_k + \omega C = \tilde{B}x_k + \tilde{C},$$

where $B = D^{-1}(L + U)$ and $C = D^{-1}$, that is B and C of the Jacobi method. Furthermore if$\omega = 1$ this is exactly Jacobi Iteration.

One may obtain the same formula from

$$x = Bx + C$$
$$\omega x + x - x = \omega(Bx + C)$$

Theorem 1.7. If Jacobi convergent, then JOR is convergent for any $0 < \omega \leq 1$.

Successive Over Relaxation (SOR) One applies the similar approach on a Gauss Seidel iteration to obtain a method known as SOR. That is $x_{k+1} = \omega(Bx^k + C) + (1 - \omega)x$ with $B = -(L + D)^{-1}U$ and $C = (L + D)^{-1}$. The commonly used form is given by

$$(\omega L + D)\mathbf{x} = \omega\mathbf{b} - [\omega U + (\omega - 1)D]\mathbf{x}$$

and the iteration is given by (for $i = 1, 2, \ldots, n$)

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$$

In order to write Gauss Seidel and SOR in the form $x^{k+1} = x^k + Pr$, one defines a residual of $x_i^{k+1} = (x_1^{k+1}, x_2^{k+1}, \ldots, x_i^k \ldots x_n^k)$ (note the paper index switches at i) to be $r_i^{k+1} = (r_{1i}^{k+1}, \ldots, r_{ni}^{k+1})$ which m'th component reads for

$$r_{mi}^{k+1} = b_m - \sum_{j=1}^{i-1} a_{mj}x_j^{k+1} - \sum_{j=i}^{n} a_{mj}x_j^k$$

Now Gauss Seidel and SOR reads for $x_i^{k+1} = x_i^k + \frac{r_{ii}^k}{a_{ii}}$ and $x_i^{k+1} = x_i^k + \omega\frac{r_{ii}^k}{a_{ii}}$, respectively.

Theorem 1.8. Let A be a matrix with no zeros in diagonal, then SOR may converge for $0 < \omega < 2$ is over relaxation.

In general $0 < \omega < 1$ is under relaxation, and $1 < \omega < 2$

Theorem 1.9. Let A be a positive definite matrix and $0 < \omega < 2$, then SOR converge from any initial guess.

1.3.4 QR via Least Squares and

One may try to approximate the linear system $Ax = b$ using minimization of the residual vector in some norm, e.g. $\|Ax - b\|$. Least squares is the most common method, that is

$$\min_x \|Ax - b\|_2 = \min_x \sqrt{\sum_{j=1}^n (A_{j\to}x - b_j)^2}$$

To see relationship with QR, let $A = QR$, where Q is an orthogonal basis of the columns of A and R is upper triangular matrix.

$$\|Ax - b\|^2 = \|Q^T(Ax - b)\|^2 = \|Q^TAx - Q^Tb)\|^2 =$$
$$\|(Q^TA)x - Q^Tb)\|^2 = \|Rx - Q^Tb)\|^2$$

Thus,

$$\min_x \|Ax - b\|_2 = \min_x \|Rx - Q^Tb)\|.$$

which has a solution $x = R^{-1}Q^Tb = A^{-1}b$.