TUTORIAL 5

October 24, 2024

1 Basic Linear Algebra Review

For your review, here are some basic linear algebra contents necessary for you to survive in this course.

Definition 1.1. Let $A \in \mathbb{C}^{n\times n}$, its adjoint A^* is defined by

$$A^*_{ij} = \overline{A_{ji}}$$

A is ***self adjoint / Hermitian*** if $A = A^*$, A is ***normal*** if $AA^* = A^*A$, and A is ***unitary*** if $AA^* = I$

Definition 1.2. Given a matrix $A \in \mathbb{C}^{n\times n}$. its characteristic polynomial is $p_A(\lambda) = \det(A-\lambda I)$. The roots λ of the characteristic polynomial are called the ***eigenvalues*** of A and the set of all eigenvalues are called ***spectrum*** of A and denoted by $\sigma(A)$. The maximum modulus of the eigenvalues is called ***spectral radius*** and is denoted by $\rho(A)$

$$\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda|$$

Proposition 1.3. (**Eigen Decomposition**)

1. $A \in \mathbb{C}^{n\times n}$ is **normal** if and only if there exists a unitary matrix U and a **complex** valued diagonal matrix Λ such that $A = U\Lambda U^*$
2. $A \in \mathbb{C}^{n\times n}$ is **self-adjoint** if and only if there exists a unitary matrix U and a **real** valued diagonal matrix Λ such that $A = U\Lambda U^*$
3. $A \in \mathbb{C}^{n\times n}$ is **unitary** if and only if there exists a unitary matrix U and a complex valued diagonal matrix Λ such that $A = U\Lambda U^*$, where each diagonal element of Λ is of modulus 1.

Note that each column vector in U is generally complex valued and distinct column vectors are orthonormal, i.e., $\langle e_i, e_j\rangle = e_i^T\overline{e_j} = \delta_{ij}$. But when the matrix is real symmetric, the case becomes easier.
$A \in \mathbb{R}^{n\times n}$ is **symmetric** if and only if there exists a real orthonormal matrix Q ($Q^TQ = I$) and a real valued diagonal matrix D such that $A = QDQ^T$.

Eigen Decomposition are not applicable to all matrices but Jordan Decomposition can apply to all.

Proposition 1.4. (**Jordan Block**) An $m \times m$ upper triangular matrix $B(\lambda, m)$ is called a *Jordan block* provided all m diagonal elements are the same eigenvalue λ and all super-diagonal elements are one:

$$B(\lambda, m) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \quad (m \times m \text{ matrix}).$$

(**Jordan Form**) Given an $n \times n$ matrix A, a *Jordan form* J for A is a block diagonal matrix

$$J = \mathbf{diag}(B(\lambda_1, m_1), B(\lambda_2, m_2), \ldots, B(\lambda_k, m_k)),$$

where $B(\lambda_i, m_i)$ is a Jordan block corresponding to the eigenvalue λ_i with size m_i.
(**Jordan Decomposition**) For any $A \in \mathbb{C}^{n\times n}$, there always exists a nonsigular $X \in \mathbb{C}^{n\times n}$ and a Jordan form J such that $A = XJX^{-1}$

2 Iterative Method

Given an $n \times n$ real matrix A and a real n-vector b, the problem considered here is to find $x \in \mathbb{R}^n$ such that $Ax = b$. Most of the methods covered in this chapter involve passing from one iterate to the next by modifying one or a few components of an approximate vector solution at a time.

We begin with the decomposition $A = D + L + U$ where D is the diagonal part of A, L its strict lower part and U its strict upper part.

Definition 2.1. 1. (**Jacobi iteration**) $Dx_{k+1} = (D - A)x_k + b$

2. (**Gauss-Seidel iteration**) $(D + L)x_{k+1} = -Ux_k + b$

3. (**Successive Over Relaxation iteration**)

$$(D + \omega L)x_{k+1} = [-\omega U + (1 - \omega)D]x_k + \omega b$$

Divide both sides by ω, we obtain the form presented in the class

$$(L + \frac{1}{\omega}D)x_{k+1} = [\frac{1}{\omega}D - (D + U)]x_k + b$$

Note that a hidden assumption of this iteration method is that all diagonal elements must be nonzero

Two core problems we care:

1. whether the iteration method converges for any initial guess?
2. If the iteration converges, what's its convergence factor?

2.1 Convergence

All above iteration methods introduced above define a sequence of iterates of the form

$$x_{k+1} = Gx_k + f, \tag{1}$$

in which G is a certain iteration matrix, where G is invertible.

Question 2.1. Given the initial guess x_0, show that

$$x_N = G^N x_0 + (I - G)^{-1}(I - G^N)f$$

What's the sufficient and necessary condition of G for x_N to converge, and whether the limit is independent of x_0? If converges, does the limit x satisfy that

$$x = Gx + f$$

and further

$$Ax = b$$

From this exercise, we can prove that

Theorem 2.2. Let G be a square matrix such that $\rho(G) < 1$. Then $I - G$ is nonsingular and the iteration converges for any f and x_0, with the limit being $(I - G)^{-1}f$. Conversely, if the iteration converges for any f and x_0, then $\rho(G) < 1$.

Therefore, a standard workflow to analyze the convergence of an iteration method is that usually we start with a form like 1, and compute the convergence factor $\rho(G)$ by eigendecomposition. For example, in the Jacobi iteration, $G = -D^{-1}(L+U)$, whose eigenvalues in some cases are easy to compute. If so, you may directly apply this strategy to analyze the convergence. Besides, we want to know how fast a method converges. What quantity can be used to measure the convergence speed?

Question 2.2. Assume that the iteration 1 converges, we suppose let x^* satisfy $x^* = Gx^* + f$, and denote the error $d_k = x_k - x^*$, show that

$$d_k = G^k d_0$$

(**challenging**) Use the Jordan Decomposition to prove that

$$\rho(G) = \lim_{k\to\infty} (\frac{\|d_k\|}{\|d_0\|})^{\frac{1}{k}}$$

Therefore, $\rho(G)$ is called *convergence factor* and used to measure how fast an iteration method converges. The smaller the factor is, the faster the method converges.

Question 2.3. Sometimes we can improve the efficiency of iteration schemes by *relaxation*. Specifically, instead of letting $\boldsymbol{x}^{(k+1)} = H\boldsymbol{x}^{(k)} + \boldsymbol{v}$, we let

$$\widehat{\boldsymbol{x}}^{(k+1)} = H\boldsymbol{x}^{(k)} + \boldsymbol{v}, \quad \text{and then} \quad \boldsymbol{x}^{(k+1)} = \omega\widehat{\boldsymbol{x}}^{(k+1)} + (1-\omega)\boldsymbol{x}^{(k)}$$

where ω is a real constant called the relaxation parameter. Note that $\omega = 1$ corresponds to the standard "unrelaxed" iteration. Good choice of ω leads to a smaller spectral radius of the iteration matrix compared with the "unrelaxed" method. Suppose we know the smallest and largest eigenvalues of H are α and β, respectively. Additionally, $-1 < \alpha < \beta < 1$, what is the optimal ω?

2.2 Special cases

However, usually it's not easy to compute $\rho(G)$ unless G is of some special form, like tridiagonal. In the following, we would like to discuss several cases in which we can easily know whether the certain iteration methods converge without the need to compute $\rho(G)$.

2.2.1 A is strictly diagonal dominant

Definition 2.3. A matrix A is

1. diagonally dominant by rows (resp. by columns) if

$$|a_{jj}| \geqslant \sum_{i\neq j}^{n} |a_{ji}|, \quad (\text{resp.} \quad |a_{jj}| \geqslant \sum_{i\neq j}^{n} |a_{ij}|) \quad j = 1, 2, \ldots, n$$

2. strictly diagonally dominant by rows (resp. by columns) if

$$|a_{jj}| > \sum_{i\neq j}^{n} |a_{ji}|, \quad (\text{resp.} \quad |a_{jj}| > \sum_{i\neq j}^{n} |a_{ij}|) \quad j = 1, 2, \ldots, n$$

An important tool is the Gershgorin Theorem.

Theorem 2.4. (**Gershgorin**) $\forall \lambda \in \sigma(A), \quad \exists i \quad \text{such that} \quad |\lambda - a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|$

Corollary 2.1. If A is strictly diagonally dominant (either by rows or columns), then it is non-singular.

Question 2.4.

1. Suppose A is a real symmetric matrix, if A is strictly diagonally dominant and its diagonal elements are positive, then A is symmetric positive definite.
2. Suppose A is a real symmetric matrix, if A is diagonally dominant and its diagonal elements are nonnegative, then A is symmetric positive semi-definite.
3. Using some spectrum-preserving operations, we may find more restrictions on the eigenvalues of A.
   (a) show that $\forall \lambda \in \sigma(A), \quad \exists i \quad \text{such that} \quad |\lambda - a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ji}|$
   (b) SAS^{-1} has the same spectrum as A, so given any positive numbers d_j, find a suitable S and show that $sp(A) \subseteq \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} \colon |z - a_{i,i}| \leqslant d_i \sum_{j=1, j\neq i}^{n} \frac{1}{d_j} |a_{i,j}| \right\}$
4. Suppose A is real symmetric and strictly diagonally dominant, show that

$$\rho(A^{-1}) \leqslant (\min_i \left\{ a_{ii} - \sum_{j\neq i} |a_{ij}| \right\})^{-1}$$

A technique used here is to write the matrix-vector product in the component-wise form.

For example, in the proof of Gershgorin Theorem, a key equality is

$$(\lambda - a_{mm})\varsigma_m = -\sum_{j\neq m}^{n} a_{mj}\varsigma_j$$

By some assumptions on ς_j, we can prove the theorem. Using this technique, we can also prove that

Theorem 2.5. If A is strictly diagonally dominant either by rows or columns, then the associated Jacobi and Gauss-Seidel iterations converge for any x_0.

Question 2.5. Define a new norm of a squared matrix by

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

Prove that $\rho(A) \leqslant \|A\|_\infty$

Question 2.6. This question is going to prove an lower bound of the determinant of a strictly diagonally dominant matrix. Suppose A is strictly diagonally dominant matrix,

1. show that the system of linear equations

$$a_{i1} + \sum_{j=2}^{n} a_{ij}x_j = 0, \quad i = 2, 3, \ldots, n$$

has a unique solution and the solution $x = (x_2, \ldots, x_n)$ satisfies that

$$\max_i \{x_i\} \leqslant 1$$

2. using the Gauss elimination, show that

$$\det(A) = \left(a_{11} + \sum_{j=2}^{n} a_{1j}x_j\right) \det \begin{pmatrix} a_{22} \cdots a_{2n} \\ \vdots \ddots \vdots \\ a_{n2} \cdots a_{nn} \end{pmatrix}$$

3. show that

$$|\det(A)| \geqslant \prod_{i=1}^{n} (|a_{ii}| - \sum_{j\neq i} |a_{ij}|)$$

Question 2.7. Here is a property of the eigenvector of a strictly diagonally dominant matrix. Suppose A is a strictly diagonally dominant squared matrix, and $\vec{u}$ is an eigenvector of A. Let $\alpha = \max_i |u_i|$, prove that it is impossible that the absolute values of all u_i's are α.

2.2.2 A is symmetric positive definite

Theorem 2.6. (**Householder-John**) If A and B are real matrices such that both A and $A - B - B^T$ are symmetric positive definite, then the spectral radius of $H = -(A-B)^{-1}B$ is strictly less than one.

Corollary 2.2. 1. If A is symmetric positive definite, then the Gauss-Seidel method converge.

2. If both A and $2D - A$ are symmetric positive definite, then the Jacobi method converge.

2.2.3 A is irreducible diagonally dominant

2.2.4 $A = M - N$ is a regular splitting

2.3 Successive Over-Relaxation(SOR) method

Let us discuss the iteration method in a component-wise form. What we have after the $k-1$-th iteration?

1. latest updated $x_i^{(k-1)}, \quad i = 1, 2, \ldots, n$
2. residual of each component $r_i^{(k)} = b_i - \sum\limits_{j=1} a_{ij} x_j^{(k-1)}$

Usually, our aim of updating x_i is to eliminate the residual induced by $x_i^{(k-1)}$. The Jacobi method **updates all $x_i^{(k)}$ in parallel**, i.e., $x_i^{(k)}$ is to eliminate the residual $b_i - \sum\limits_{j\neq i} a_{ij} x_j^{(k)}$,

$$x_i^{(k)} a_{ii} = b_i - \sum_{j\neq i} a_{ij} x_j^{(k)}$$

The Gauss-Seidel method **updates $x_i^{(k)}$ component by component, i.e., $x_i^{(k)}$ can be updated only after all $x_j^{(k)}, j < i$ have been updated**, and $x_i^{(k)}$ is to eliminate the residual induced by un-updated $x_j^{(k-1)} (j \geqslant i)$ and updated $x_j^{(k)} (j < i)$,

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

A common characteristic of these two methods is that they don't explicitly consider the affect of $x_i^{(k-1)}$ when updating $x_i^{(k)}$. That's how the SOR method comes. We consider a weighted average of $x_i^{(k-1)}$ and the newly updated value by Gauss-Seidel method.

$$x_i^{(k)} = (1-\omega) x_i^{(k-1)} + \omega \frac{1}{a_{ii}} (b_i - \sum_{j<i} a_{ij} x_j^{(k)} - \sum_{j>i} a_{ij} x_j^{(k-1)}), \quad \text{for} \quad i = 1, 2, \ldots, n$$

Question 2.8. Let $A = D+L+U$, prove that above updated scheme can be written as

$$(L + \frac{1}{\omega}D)x^{(k)} = [\frac{1}{\omega}D - (D + U)]x^{(k-1)} + f, \quad \text{for some vector } f$$

and equivalently

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

Define $G_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U]$ and $f_\omega = \omega(D + \omega L)^{-1}b$, to analyze the convergence of SOR method, we need to analyze the $\rho(G_\omega)$. Here is a necessary condition on ω for the convergence.

Theorem 2.7. (Kahan) If $a_{ii} \neq 0$, for each $i = 1, 2, \ldots .n$, then $\det(G_\omega) = (\omega - 1)^n$ and $\rho(T_\omega) \geqslant |\omega - 1|$. This implies that the SOR method can converge only if $0 < \omega < 2$.

Two questions naturally arise.

1. When A has some properties, can the range of ω be further shrinked?
2. Under what conditions on A, can the above necessary condition be also sufficient?

and here are some answers.

Theorem 2.8. When A is **strictly diagonally dominant**, SOR method converge if $0 < \omega < 1$

Note that this is a sufficient condition on ω, **it doesn't mean that SOR method won't converge if** $1 \leqslant \omega < 2$, instead, it just told you that if $0 < \omega < 1$, SOR method must converge.

Theorem 2.9. When A is **symmetric positively definite**, SOR method converges **if and only if** $0 < \omega < 2$.

The proof of theorem 2.9 is based on the Householder-John theorem. Besides, I would like to remind you one key factor in the above analysis.

Although $G_\omega = (D+\omega L)^{-1}[(1-\omega)D-\omega U]$ is an elegant way to represent the iteration matrix, it's more useful to write $G_\omega = (I+\omega D^{-1}L)^{-1}[(1-\omega)I-\omega D^{-1}U]$ in the analysis of $\det(G_\omega)$ and $\rho(G_\omega)$, or equivalently, $\det(G_\omega - \lambda I)$, because $\det((I + \omega D^{-1}L) = 1$. Next, it is a very **interesting and surprising** result of the consistently ordered matrix. And again its proof heavily depends on this key technique.

Theorem 2.10. First of all, recall the concept of a consistently ordered matrix. Suppose $A = D + L + U$ where L and U are strictly lower and upper triangular part, respectively, if eigenvalues of $\alpha D^{-1}L + \frac{1}{\alpha}D^{-1}U (\alpha \neq 0)$ are independent of α, then the matrix is said to be consistently ordered.

Let A be a consistently ordered matrix such that $a_{ii} \neq 0$ for all i, and let $\omega \neq 0$. Then if λ is a nonzero eigenvalue of the SOR iteration matrix G_ω, any scalar μ such that

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$$

is an eigenvalue of the Jacobi iteration matrix G_J

To use this theorem, the first step is to check whether A is consistently ordered. If so, we have two important conclusions.

Corollary 2.3. Suppose A is consistently ordered, we have

1. Let $\omega = 1$, the SOR method is exactly the Gauss Seidel method. Hence $\rho(G_{GS}) = \rho(G_J)^2$ and in such case it is said that Gauss-Seidel method converge twice faster that Jacobi method.

2. Further if we assume the Jacobi iteration matrix contains only real-valued eigenvalues, then given a fixed $\omega \in (0, 2)$, for each $\mu \in \sigma(G_J)$, we can obtain two eigenvalues of G_ω, thus

   $$\rho(G_\omega) = f(\omega, G_J)$$

   for some function f. Young's theorem told us that the optimal ω which results in the minimal $\rho(G_\omega)$ is exactly

   $$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(G_J^2)}}$$

Question 2.9. Consider an $n \times n$ tridiagonal matrix of the form

$$T_\alpha = \begin{pmatrix} \alpha & -1 & & & \\ -1 & \alpha & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & \alpha \end{pmatrix},$$

where α is a real parameter.

1. Verify that the eigenvalues of T_α are given by

$$\lambda_j = \alpha - 2\cos(j\theta), \quad j = 1, \ldots, n,$$

where

$$\theta = \frac{\pi}{n+1},$$

and that an eigenvector associated with each λ_j is

$$q_j = [\sin(j\theta), \sin(2j\theta), \ldots, \sin(nj\theta)]^T.$$

Under what condition on α does this matrix become positive definite?

2. Let $\alpha = 2$.

   (a) Will the Gauss-Seidel iteration converge for this matrix? If so, what will its convergence factor be?

   (b) For which values of ω will the SOR iteration converge? and what's the optimal value of ω_{opt}?