Session 12

Xuanxi Zhang

1 Floating-Point

1.1 Binary Representation

7=111 0.375=0.011

1.2 Fixed-Point

several bits for integer part, several bits for fractional part. decimal point is fixed.

1.3 Floating-Point

$\pm 1.\text{mantissa} \times 2^{\text{exponent}}$

1.4 Machine Precision

The gap between 1 and the next larger floating-point. which is decided by the number of bits in the mantissa.

1.5 IEEE arithmetic

It's too complicated and impractical, and there's too much to remember. Ask Nour if it is necessary to remember all the details.

2 numerical eigen value problem

The eigenvalue and eigenvector of a matrix A are defined as the scalar λ and the vector v that satisfy $Av = \lambda v$

2.1 Gershgorin circle theorem

definition: Let $A \in \mathbb{C}^{n\times n}$ Gershgorin discs D_i are defined as $D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|\}$
Theorem: The eigenvalues of A are contained in the union of the Gershgorin discs.
Theorem: If $\{D_i\}$ can be divided into dijoint two set, the first has q disks and the other has p disks. Then there will be q eigenvalues in the first set and p eigenvalues in the second set.

2.2 Power method

A real symmtreic matrix A can be diagonalized as $A = Q\Lambda Q^T$.
similarity transform does not change the eigenvalues.
The power method is to find the largest eigenvalue of A and its corresponding eigenvector.
iteration scheme $x_{k+1} = Ax_k/\|Ax_k\|$
theorem: if A has a simple dominant eigenvalue, and the starting vector x_0 is not orthogonal to the eigenvector corresponding to the dominant eigenvalue, then the power method will converge to the dominant eigenvalue.

2.3 inverse power method

The inverse power method is to find the eigenvalue of A that is closest to a given μ.
scheme: $x_{k+1} = (A - \mu I)^{-1}x_k/\|(A - \mu I)^{-1}x_k\|$

2.4 find all eigenvalues

First step is to reduce the matrix to tridiagonal form. Then use QR algorithm to find all eigenvalues.

2.5 SVD

The singular value decomposition of a matrix A is $A = U\Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix with non-negative entries. It can be Representation as $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$.

we have $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$. To find the SVD of A, we can find the eigenvectors of AA^T and $A^T A$.

Solution to the low rank approximation problem: $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank-k approximation to A in the sense of Frobenius norm and 2-norm.

connection with matrix norm: $\|A\|_2 = \sigma_1$, $\|A\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$, $\kappa_2(A) = \sigma_1/\sigma_r$.