Computational methods-Lecture 6

Direct methods for solving linear systems

We now move onto linear system of equations (the matrices), which will be the second main part of our course.

1 Linear system of equations and matrices

As we have known, the linear system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ &\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_n \end{aligned}$$

can be written into the matrix form

$$Ax = b,$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

There are several special types of matrices that are important in numerical linear algebra.

1.1 Positive definite matrices

Definition 1. *A real matrix A is positive definite if it is symmetric such that $x^T Ax > 0$ for all $x \in \mathbb{R}^n, x \neq 0$.*

Proposition 1. *If A is of size $n \times n$ and is positive definite, then*

- *A is nonsingular (invertible)*
- $a_{ii} > 0$
- $\max_{i,j} |a_{ij}| = \max_i |a_{ii}|$
- $a_{ij}^2 < a_{ii}a_{jj}$ *for* $i \neq j$

The first is obvious since $Ax = 0$ only has $x = 0$ solution. The second is true by taking $x = e_i$. The fourth holds by taking $x = \alpha e_i + e_j$, and for all $\alpha \in \mathbb{R}$. The fourth then implies the third.

The leading principal submatrix for $k \leq n$ is deifned by

$$A_k = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}.$$

The following is a characterization of the PD matrices

Theorem 1. *A symmetric matrix A is positive definite if and only if each of its leading principal submatrices has a positive determinant (leading principal minors are positive).*

There are other charaterizations

Theorem 2. *A symmetric matrix is positive definite if and only if all the eigenvalues are positve, if and only if it can be written as LL^T for some invertible lower triangular matrix L*

1.2 Strictly diagonally dominant matrices

Definition 2. *The $n \times n$ matrix A is said to be strictly diagonally dominant if*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

for all $i = 1, 2, \cdots n$.

We first introduce the Gershgorin Circle Theorem about the distribution of eigenvalues, to conclude that A is nonsingular. Other properties about strictly diagonally dominant matrices will be discussed after talking about Gauss elimination.

Theorem 3. *Suppose A is any complex $n \times n$ matrix. Let P be any invertible matrix. Consider*

$$PAP^{-1} = D + F,$$

where D is a diagonal matrix and F is a matrix with zero diagonals, then the eigenvalues are contained in

$$\cup_i D_i = \cup_i \Big\{ z : |z - d_i| \leq \sum_{j:j\neq i} |f_{ij}| \Big\}.$$

Hence, the eigenvalues must be in

$$\bigcap_P \left\{ \bigcup_i \left\{ z : |z - d_i| \leq \sum_{j:j\neq i} |f_{ij}| \right\} \right\}$$

Using this, it is easy to see that the strictly diagonally dominant matrices are invertible.

1.3 Band matrices

A matrix is called a band matrix if there exist p, q with $1 < p, q < n$, such that $a_{ij} = 0$ for $i + p \leq j$ or $j + q \leq i$. The bandwidth is $w = p + q - 1$. If $p = q = 2$, the matrix is called tridiagonal.

We will come back this this matrix later.

1.4 Sparse matrices

A sparse matrix is a matrix in which most of the elements are zero. Usually, the term "sparse" makes sense when the size of A is big.

2 Gauss elimination

2.1 The algorithm

We start with an example.

$$\begin{array}{rrrrl} x_1 & +x_2 & & +3x_4 & = 4, \\ 2x_1 & +x_2 & -x_3 & +x_4 & = 1, \\ 3x_1 & -x_2 & -x_3 & +2x_4 & = -3 \\ -x_1 & +2x_2 & +3x_3 & -x_4 & = 4. \end{array}$$

For a linear system, the following operators will not change the set of solutions

- Mulitplying an equation with a nonzero constant.
- Add one equation into another one.
- Interchange the order of two equations.

We will then use these three operations to change the system into another one that is easier to solve. We multiply the first equation with -2 and add it to the second equation; multiply the first equation with -3 and add to the third equation, etc:

$$\begin{array}{rrrrr} x_1 & +x_2 & & +3x_4 & = 4, \\ & -x_2 & -x_3 & -5x_4 & = -7, \\ & -4x_2 & -x_3 & -7x_4 & = -15 \\ & 3x_2 & +3x_3 & +2x_4 & = 8. \end{array}$$

Now, we focus on the last three equations. We do the same trick: multiply the second equation in the new system with suitable constants to kill other x_2's below it:

$$\begin{array}{rrrrr} x_1 & +x_2 & & +3x_4 & = 4, \\ & -x_2 & -x_3 & -5x_4 & = -7, \\ & & 3x_3 & +13x_4 & = 13 \\ & & & -13x_4 & = -13. \end{array}$$

Clearly, using this form, we can solve $x_4 = 1$ out and then substitute this back to the third equation and solve x_3. Then, solve x_2 and lastly x_1.

This process can be performed in the form of a matrix. Let us construct the so-called augmented matrix:

$$\left[\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right]$$

Hence, the above algorithm is to use the nonzero entries to make the entries below them zeros.

$$\left[\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right] \Rightarrow \left[\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right]$$

The above method is called Gauss Elimination with Backward Substitution.

Gauss elimination

Given a matrix A of size $m \times n$. Let $s = \min(m-1, n)$.

For $i = 1, \cdots, s$ do the following:

(1) If $a_{ii} = 0$, stop, and output error message.

(2) For $k = i+1, \cdots, m$ do

(a) $m_{ki} = a_{ki}/a_{ii}$

(b) $a_{kj} \leftarrow a_{kj} - m_{ki}a_{ij}$ for $j = i+1, \cdots, n$

Note that in the above code, we should have a_{ki} is equal to zero for $k > i$. However, in practice, one often stores the m_{ki} at the place of a_{ki}. Hence, we do $a_{ki} \leftarrow a_{ki}/a_{ii}$. This is not only for convenience, but also has a practical significance in the LU decomposition as we shall soon.

Backward substitution

Consider $A_1 = [A|b]$ to be the augmented matrix after the Gauss elimination has been performed. Suppose A is of size $n \times n$. Then, $[A|b]$ is a upper triangular system.

For $i = n, \cdot, 1$ do the following

1. $x_i \leftarrow [b_i - \sum_{j=i+1}^{n} a_{ij}x_j]/a_{ii}$.

If i equals n, the summation over empty set is assumed to be zero. This algorithm clearly applies to any upper triangular system.

2.2 The complexity

In the ith iteration, we need to perform $(m-i)$ divisionss for step $2(a)$ and $(n-i)*(m-i)$ multiplications (note that for a_{kj} with $j = i$, the multiplication is not done since it is guaranteed to result in zero for a_{ki}, and the place is reserved for m_{ki}). The total number of divisions/multiplications is

$$\sum_{i=1}^{s}(m-i)(n-i+1) \sim \frac{1}{3}s^2 - \frac{1}{2}s^2(m+n) + mns$$

When $m = n$, the total number to the leading order is like $\frac{1}{3}n^3$.

The number of additions/subtractions can be similarly estimated. It is still $O(n^3)$. However, each addition/subtraction is much cheaper than division/multiplication.

3 Gauss elimination with partial pivoting

Let the element $a_{\ell k}^{(i)}$ be the elements in the ith step. Clearly, the elements $a_{ii}^{(i)}$'s are most important, since we use them to divide the numbers below it. These are called the **pivot elements**.

If $a_{ii}^{(i)} = 0$, then the above Gauss elimination fails. The following theorem guarantess when these are nonzero

Theorem 4. *The pivot elements are nonzero if and only if all the leading principal minors are nonzero.*

The proof relies on one fact: multiplying a row with a constant and adding to another row will not change the determinants. Here, we do not care this much, so we skip the details.

If we have $a_{ii}^{(i)} = 0$, we cannot continue Gauss elimination, but this does not mean the original system does not have solutions. One typical example is

$$\begin{aligned} x_2 &= 1, \\ x_1 &= 3 \end{aligned}$$

Even if $a_{ii}^{(i)}$ is not zero, but it can be small. This then will result in large roundoff errors in computing m_{ki}'s and back substitutions. Hence, we desire row interchanges to reduce errors. Often, we choose the **maximal column pivoting**, i.e., using the one with largest absolute value in the corresponding column below (and including) $a_{ii}^{(i)}$ for the pivot element. This is also called **partial pivoting**.

If we use the partial pivoting, as soon as the matrix is invertible, the Gauss elmination can be performed to the end.

Example

$$\begin{array}{ccccc} 0.001x_1 & +2x_2 & +3x_3 & = 1, \\ -x_1 & +3.712x_2 & +4.623x_3 & = 2, \\ -2x_1 & +1.072x_2 & +5.643x_3 & = 3 \end{array}$$

We write out the augmented matrix

$$\left[\begin{array}{ccc|c} 0.001 & 2 & 3 & 1 \\ -1 & 3.712 & 4.623 & 2 \\ -2 & 1.072 & 5.643 & 3 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} -2 & 1.072 & 5.643 & 3 \\ -1 & 3.712 & 4.623 & 2 \\ 0.001 & 2 & 3 & 1 \end{array}\right]$$

Then, we get

$$\left[\begin{array}{ccc|c} -2 & 1.072 & 5.643 & 3 \\ 0 & 3.176 & 1.801 & 0.5 \\ 0 & 2.001 & 3.003 & 1.002 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} -2 & 1.072 & 5.643 & 3 \\ 0 & 3.176 & 1.801 & 0.5 \\ 0 & 0 & 1.868 & 0.687 \end{array}\right]$$

The final answer is $(-0.49, -0.05113, 0.3678)$.

Remark 1. *Sometimes, the partial pivoting is not also quite good enough. People then want to combine it with scalings. This will result in Gauss elimination with scaled partial pivoting.*