Computational methods-Lecture 7

Direct methods: LU decomposition, Cholesky decomposition

1 Finding the inverse of an invertible matrix

If A is a matrix of size $n \times n$ such that there is a matrix B of size $n \times n$ satisfying

$$AB = I,$$

then A is invertible and $B = A^{-1}$.

In fact, the above condition says that $Ax = e_i$ for any $i = 1, \cdots, n$ has a solution. Since $\{e_i\}$ forms a basis, then $Ax = y$ has a solution for any y. Hence, A must be of full rank, and thus invertible. Multiplying both sides on left with A^{-1} one gets $B = A^{-1}$.

According to this observation, to find the inverse, we only need to solve the linear equation

$$Ax = e_i$$

using Gauss elimination. The solution will be the ith column of B. We can do this in a compact form as follows. We set up a large augmented matrix

$$[A \vdots I].$$

The we apply Gauss elimination with partial pivoting to make A into upper triangular form. To solve the solutions, we use backward substitution. However, this is also equivalent to using the pivoting elements to kill the nonzeros above them as well (this is called the **Gauss-Jordan** elimination) and use row scalings to make A into I:

$$[I \vdots B].$$

Clearly, the columns in B will be the solutions for $Ax = e_i$ since these row operations will not change solutions.

Hence, we must have B to be the inverse matrix. Note that Gauss-Jordan elimination for the upper triangular part again costs $O(n^3)$, while backward substition costs $O(n^2)$. Hence, in solving linear systems, one does not use Gauss-Jordan. However, to find the inverse, there are n vectors, if you do backward substitution, the total cost is also $O(n^3)$. Thus, for finding inverse, the Gauss-Jordan is preferred.

Example Find the inverse matrix of

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix}$$

Then, we form the augmented matrix

$$\left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 2 & 0 & 0 & 1 \end{array}\right]$$

Next, we get

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

Next,

$$\left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -3 & 2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1/3 & 1/3 & 1/3 \end{array}\right]$$

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

Doing scaling

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 2/3 & 1/3 & 1/3 \\ 0 & 1 & 0 & 4/9 & -1/9 & 2/9 \\ 0 & 0 & 1 & -1/3 & 1/3 & 1/3 \end{array}\right]$$

Eventually,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -2/9 & 5/9 & -1/9 \\ 0 & 1 & 0 & 4/9 & -1/9 & 2/9 \\ 0 & 0 & 1 & -1/3 & 1/3 & 1/3 \end{array}\right]$$

In other viewpoint as we shall see, *doing row operations are equivalent to multiplying suitable matrices on the left.* Hence, we in fact have

$$[I \vdots B] = P[A \vdots I]$$

In other words,

$$PA = I,\ B = PI = P.$$

Hence, $BA = I$ and we must have $B = A^{-1}$ (note that if A is not square, there can also be B to make $BA = I$. In this case, B is not the inverse).

Remark 1. *During this process, if you ever did column operations to make A into I (apply similar column operations to the second matrix as well), then you in fact did $P_1AQ_1 = I$ and have $B = P_1Q_1$. In this case, B is in general not A^{-1}. However, if $P_1 = I$, Q_1 is the inverse. In other words, if you do pure column operations to make A into I, you can still obtain the inverse matrix, but mixture of both row and column operations will not work.*

2 The LU decomposition

In the Gauss elimination (without pivoting), we have applied row operations to reduced A into an upper triangular form U.

Recall how we performed Gauss elimination: we use the pivot elements $a_{ii}^{(i)}$ to kill the nonzero entries below it. There is one important observation in linear algebra: doing row operations is equivalent to multiplying certain fundamental matrices on the left. For operations of this type, we are multiplying matrices of the following form on the left:

$$L_i = \begin{pmatrix} 1 & & & & \\ & \cdots & & & \\ & & 1 & & \\ & & -m_{i+1,i} & 1 & \\ & & \vdots & & \vdots \\ & & -m_{mi} & & 1 \end{pmatrix}.$$

For example, you can check what happens if you multiply L_1 on the left. Hence, the Gauss elimination tells us that if all the pivot elements are nonzero $a_{ii}^{(i)} \neq 0$, then

$$L_nL_{n-1}\cdots L_1A = U,$$

where U is given as follows:

$$U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{n1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{n2}^{(2)} \\ & & \cdots & \cdots \\ & & & a_{nn}^{(n)} \end{pmatrix}.$$

Hence, we get

$$A = L_1^{-1}\cdots L_{n-1}^{-1}L_n^{-1}U.$$

Note that the inverse of L_i is just

$$L_i^{-1} = \begin{pmatrix} 1 & & & & \\ & \cdots & & & \\ & & 1 & & \\ & & m_{i+1,i} & 1 & \\ & & \vdots & & \vdots \\ & & m_{mi} & & 1 \end{pmatrix}.$$

This is easily understood because we can multiply the constants m_{ki} and add to the corresponding rows to recover the original rows.

Due to the meaning of multplying L_i^{-1} on the left, we can easily seen that

$$L_1^{-1} \cdots L_{n-1}^{-1} L_n^{-1} = \begin{pmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & \vdots & \vdots & \\ m_{n1} & m_{m2} & m_{n3} & \cdots & 1 \end{pmatrix}.$$

Hence, we conclude the following

Theorem 1. *If A is of size $n \times n$ and its leading principal minors D_i ($i \leq n-1$) are nonzero, then A can be decomposed as*

$$A = LU,$$

where L is a lower triangular matrix with diagonal elements being 1 and U is an upper triangular matrix. Moreover, such decomposition is unique.

The proof of the uniqueness is left as an exercise.

Since the elements in L are just the multipliers generated during Gauss elimination. As we recall, we store these elements already in place at a_{ki}. Hence, after the Gauss elimination is done. The LU decomposition is already completed: we just read out the elements in the lower triangular part!

These observations can be summarized into the following code for LU decomposition:

```
for k=1:n-1
    A(k+1:n, k)=A(k+1: n, k)/A(k, k);
    for j=k+1:n
        for i=k+1:n
            A(i,j)=A(i, j)-A(i, k)*A(k,j);
```

```
        end
    end
end
```

This can be written into a more compact form in MATLAB:

```
for k=1:n-1
    A(k+1:n, k)=A(k+1: n, k)/A(k, k);
    j=k+1:n
    i=k+1:n;
    A(i,j)=A(i, j)-A(i, k)*A(k,j);
end
```

There are other versions, which are essentially doing the same thing, but the orders of the operations are changed. Two typical examples are as following (the following two are not required for exam):

```
for j=1:n
    for k=1:j-1
        i=k+1:n;
        A(i,j)=A(i, j)-A(i, k)*A(k,j);
    end
    i=j+1:n;
    A(i,j)=A(i,j)/A(j,j);
end
```

```
for i=2:n
    for j=2:i
        A(i, j-1)=A(i, j-1)/A(j-1, j-1);
        k=1:j-1;
        A(i,j)=A(i,j)-A(i,k)*A(k,j);
    end
    k=1:i-1;
    j=i+1:n;
    A(i,j)=A(i, j)-A(i, k)*A(k,j);
end
```

Why do we care about LU decomposition?

Once decomposed, we can use this decomposition to solve the linear system for any b, instead of doing GEM once for every b.

One first solve:

$$Lz = b,$$

and solve

$$Ux = z.$$

The total cost is $O(n^2)$.

Example Find the LU decomposition of the following matrix.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & -1 \\ 2 & -2 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & -1 \\ 2 & -4 & -1 \end{bmatrix}$$

Then,

$$\Rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & -1 \\ 2 & -1 & -2 \end{bmatrix}.$$

3 With partial pivoting*(Not required)

The LU decomposition requires the pivot elements to be nonzero. For general invertible matrices, these elements can be zero. Hence, the rwo interchanges are necessary. In fact interchanging ith and jth rows corresponds to multiplying the matrix I_{ij} on the left.

$$I_{ij} = \begin{pmatrix} 1 & & & & & & & \\ & \cdots & & & & & & \\ & & 0 & \cdots & \cdots & 1 & \cdots & \\ & & & 1 & \vdots & & & \\ & & & \cdots & 1 & \cdots & & \\ & & 1 & \cdots & & 0 & & \\ & & \cdots & & & & \cdots & \\ & & \vdots & & \vdots & & & 1 \end{pmatrix}.$$

This matrix has the property that

$$I_{ij}^{-1} = I_{ij}^T = I_{ij}.$$

If we multiply this matrix on the right, we are switching the column i and column j.

If $i = j$, there is no exchange. Hence, the GEM with partial pivoting can be written as

$$L_{n-1}I_{n-1,i_{n-1}} \cdots L_2 I_{2,i_2} L_1 I_{i,i_1} A = U,$$

where $i_k \geq k$.

We now define the permutation matrix

$$P = I_{n-1,i_{n-1}} \cdots I_{2,i_2} I_{i,i_1}.$$

Note that this permutation matrix is not symmetic. Its inverse is its transpose, but not itself. This is like doing the row interchanges all at once. One question is like this: if we do row exchanges all at once, we can then apply the GEM to obtain some LU decomposition

$$L^{-1}PA = U. \Rightarrow PA = LU.$$

Hence, we conclude

Theorem 2. *For an invertible matrix, there is a permutation matrix P such that*

$$PA = LU,$$

where L is a lower triangular matrix with diagonal elements to be 1 and U is upper triangular.

One natural question: We can do

- if we do Gauss Elimination Method with partial pivoting, we also store the m_{ik} elements in the lower half of the matrix and do row exchanges when we do pivoting.
- We do all row exhanges at the very beginning, and then apply GEM to obtain the LU decomposition.

Will the final lower triangular half in the first way be the same as in the second way, or as in the theorem? The answer is **yes**.

To see this, let us take $n = 4$ as the example.

$$L_3 I_{3,i_3} L_2 I_{2,i_2} L_1 I_{1,i_1} = L_3 (I_{3,i_3} L_2 I_{3,i_3})(I_{3,i_3} I_{2,i_2} L_1 I_{2,i_2} I_{3,i_3}) P =: L_3 \tilde{L}_2 \tilde{L}_1 P.$$

By the uniqueness of LU decomposition, we have

$$L_3 \tilde{L}_2 \tilde{L}_1 = L.$$

In fact, this can also be understood using the following observation:

Lemma 1. *Let P_k be permuation that only operates on indicies bigger than k, and P_k generates permutation $\sigma(\cdot)$: $\sigma(i)$ means the new ith row is the original $\sigma(i)$th row. Then,*

$$P_k L_k P_k^T = \begin{pmatrix} 1 & & & & \\ & \cdots & & & \\ & & 1 & & \\ & & -m_{\sigma(i+1),i} & 1 & \\ & & \vdots & & \vdots \\ & & -m_{\sigma(m),i} & & 1 \end{pmatrix}.$$

It is then clear that the new $\tilde{L}_k$ matrix is exactly that acting on the permuted rows.

Hence, we need to very that $L_3\tilde{L}_2\tilde{L}_1$ corresponds to the lower triangular half matrices generated by the GEM with partial pivoting. Let us think about what happens if we do GEM with partial pivoting. At the kth iteration, you do

$$L_k = \begin{pmatrix} 1 & & & & \\ & \cdots & & & \\ & & 1 & & \\ & & -m_{i+1,i} & 1 & \\ & & \vdots & & \vdots \\ & & -m_{m,i} & & 1 \end{pmatrix}.$$

and this generates the part

$$\begin{pmatrix} & & & \\ \cdots & & & \\ & & & \\ & m_{i+1,i} & & \\ & \vdots & \vdots & \\ & m_{m,i} & & \end{pmatrix}.$$

In the lower triangular half. Later, this column will only be changed by the row exchanges after the kth step, which is exactly what P_k is doing, which

will make this to be

$$\begin{pmatrix} & \cdots & & \\ & & & \\ & & m_{\sigma(i+1),i} & \\ & & \vdots & \vdots \\ & & m_{\sigma(m),i} & \end{pmatrix},$$

agreeing with the above. Hence, the conclusion follows.

4 Compact form of decomposition: Doolittle and Crout

As we have seen, as soon as we have LU decomposition, we can solve the linear systems in $O(n^2)$ time. Hence, one question is whether we can obtain the LU decomposition directly without doing the Gauss Elimination step by step. Here we consider A which can be decomposed into LU directly. For the ones with P, you may read the book.

With $A = LU$, one has

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir} u_{rj}.$$

There are n^2 equations with $n^2 + n$ unknowns.

There are often two choices to determine these unknowns. We can fix $l_{ii} = 1$, which leads to the same LU as in the Gauss Elimination, the resulted method will be called the Doolittle's method. If we impose $u_{ii} = 1$, one will have the Crout's method.

Let us consider the Doolittle's method first. Assume that the first $k-1$ ($k = 1, \cdots, n$) rows of U and the first $k-1$ columns of L are known already. Then, we determine the kth row of U and kth column of L:

$$a_{kj} = \sum_{r=1}^{k-1} l_{kr} u_{rj} + u_{kj},\ j \geq k.$$

Hence, the kth row of U can be determined by

$$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj},\ j \geq k.$$

Then, using these computed values, we can determined l_{ik}:

$$a_{ik} = \sum_{r=1}^{k-1} l_{ir}u_{rk} + l_{ik}u_{kk},\ i \geq k+1,$$

we have

$$l_{ik} = \frac{1}{u_{kk}}(a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk}),\ i \geq k+1.$$

This method is called the Doolittle's method, and it costs also like $n^3/3$ multiplications/divisions, comparable to Gauss elimination.

The Crount's method is similar, where we impose $u_{rr} = 1$:

$$l_{ik} = a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk},\ i = k, \cdots, n,$$

$$u_{kj} = \frac{1}{l_{kk}}(a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj}),\ j = k+1, \cdots, n.$$