Computational methods-Lecture 8

Error analysis for solving linear systems; condition number

1 LU decomposition for special types of matrices

1.1 Positive definite matrices and Cholesky

For positive definite matrices, all the pivot elements will be positive in GEM. Hence, we will have

$$A = LU.$$

One again, one can rewrite

$$U = DU_1,$$

where U_1 is upper triangular with diagonal elements to be 1. Hence,

$$A = LDU_1.$$

By symmetry

$$LDU_1 = U_1^T DL^T.$$

Due to the uniqueness of LU decomposition, $L = U_1^T$. Hence,

$$A = LDL^T.$$

Moreover, if we define

$$L_1 = L\sqrt{D},$$

then L_1 is also lower triangular, thus leading to the **Cholesky decomposition**

Theorem 1. *If A is positive definite, there is a unique L which lower triangular, with positive diagonal entries, such that*

$$A = LL^T.$$

Both the LDL^T and Cholesky decomposition can be done in a similar fashion as in the Doolittle's or Crout's method. For example, the Cholesky can be done as

For $j = 1 : n$

1. $l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$
2. $l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk})/l_{jj}$, $i \geq j + 1$.

1.2 Diagonally dominant

The Gauss elimination can be performed for diagonally dominant matrices without row interchanges. In fact, in the Gauss elimination, the submatrices are preserved to be diagonally dominant.

To prove this, by induction, one only needs to show that $A^{(2)}$ is diagonally dominant. In fact,

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}},\ i \geq 2, j \geq 2.$$

Then,

$$\begin{aligned}
\sum_{j=2, j\neq i}^{n} |a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}}| &\leq \sum_{j=2, j\neq i}^{n} |a_{ij}^{(1)}| + \sum_{j=2, j\neq i}^{n} \frac{|a_{1j}^{(1)}||a_{i1}^{(1)}|}{|a_{11}^{(1)}|} \\
&\leq |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} \sum_{j=2, j\neq i}^{n} |a_{1j}^{(1)}| \\
&\leq |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} (|a_{11}^{(1)}| - |a_{1i}^{(1)}|) \\
&= |a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \leq |a_{ii}^{(2)}|
\end{aligned}$$

1.3 Tridiagonal matrices

For tridiagonal matrices, the Doolitle's method or Crout's method can be done in $O(n)$ time. Let us consider the Crout's decomposition briefly.

Clearly, in LU decomposition, the nonzeros of L and U matrices will again be in the three diagonals. Direct computation shows that

$$a_{i,i-1} = \ell_{i,i-1},\ i = 2, 3, \cdots, n.$$

$$a_{ii} = \ell_{i,i-1} u_{i-1,i} + \ell_{ii},\ i = 1, \cdots, n$$

here if $i = 1$, $a_{11} = \ell_{11}$.

Lastly,

$$a_{i,i+1} = \ell_{ii} u_{i,i+1}.$$

We fist use the first relation to find all $\ell_{i,i-1}$ elements. Then, use the second equation to find ℓ_{11}, which gives u_{12}. Using u_{12} and the second, we find ℓ_{22}, which together with the third gives u_{23}, and so on.

Alternatively,

$$u_{i,i+1} = \frac{a_{i,i+1}}{\ell_{ii}} = \frac{a_{i,i+1}}{a_{ii} - a_{i,i-1}u_{i-1,i}}$$

As soon as we have the LU decomposition, the solution can be solved easily in linear time.

Theorem 2. *If* $|a_{11}| > |a_{12}| > 0$, $|a_{ii}| \geq |a_{i,i-1}| + |a_{i,i+1}|$, *and* $|a_{nn}| > |a_{n,n-1}| > 0$, *then* $|u_{i,i+1}| \in (0,1)$, *and* $|a_{ii}| - |a_{i,i-1}| < |\ell_{ii}| < |a_{ii}| + |a_{i,i-1}|$.

2 Norms of vectors and matrices

2.1 Norms of vectors

Suppose the vectors x in $\mathbb{R}^d$ ($\mathbb{C}^d$) are equipped with some norm $\|\cdot\|$, which satisfies three properties:

- $\|x\| \geq 0$ and the equality holds if and only if $x = 0$
- $\|\alpha x\| = |\alpha|\|x\|$, $\alpha \in \mathbb{R}$ ($\mathbb{C}$),
- $\|x + y\| \leq \|x\| + \|y\|$.

Corollary:

$$|\|x\| - \|y\|| \leq \|x - y\|$$

The most common norm is the Euclidean norm, induced by the inner product:

$$\langle x, y\rangle = \sum_{i=1}^{d} x_i \bar{y}_i.$$

This is also called the 2-norm

$$\|x\|_2 := \sqrt{\langle x, x\rangle} = (\sum_{i=1}^{n} |x_i|^2)^{1/2}.$$

This inner product is special in the sense that the Cauchy-Schwartz inequality holds

$$|\langle x, y\rangle| \leq \|x\|_2 \|y\|_2.$$

The reason that the above is called 2-norm is that one can define the general p-norm for $p \geq 1$:

$$\|x\|_p := (\sum_{i=1}^{d} |x_i|^p)^{1/p}.$$

This is a norm because of the Minkowski inequality.

Taking $p \to \infty$, one can have the following ∞-norm (exercise):

$$\|x\|_\infty = \max_{1 \le i \le d} |x_i|.$$

Claim 1 Every norm can induce a distance, called the metric, defined by

$$d(x, y) := \|x - y\|.$$

Claim 2 For finite dimensional space, all norms are equivalent. That means for any two norms $\|\cdot\|$ and $\|\cdot\|'$, there exists two constants $C_1 > 0, C_2 > 0$ such that

$$C_1\|x\|' \le \|x\| \le C_2\|x\|',$$

where C_1, C_2 could depend on the dimension d.

Read the proof in the reference book. For example,

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{d}\|x\|_\infty.$$

2.2 Norms of matrices

A norm for matrices satisfy the following

- $\|A\| \ge 0$ and the equality holds if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$
- $\|A + B\| \le \|A\| + \|B\|$.
- $\|AB\| \le \|A\|\|B\|$

Compared with the norms for vectors, the norms for matrices have one more requirement: it is consistent with matrices multiplication. This is because the set of matrices forms the so-called **algebra**.

One typical example is the Frobenius norm:

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}.$$

This is in fact the 2-norm by regarding the matrix as a vector in $\mathbb{R}^{n^2}$. Of course, you need to verify that the extra requirement is satisfied.

Proposition 1. *If for some norm $\|\cdot\|$, one has $\|B\| < 1$, then*

$$\|(I - B)^{-1}\| \le \frac{1}{1 - \|B\|}$$

Proof. In fact, if $\|B\| < 1$, one can show that the following series converges:

$$I + B + B^2 + \cdots = C.$$

Moreover, it is easy to verify

$$\|(I - B)(I + \sum_{n=1}^{N} B^n) - I\| \to 0.$$

By the continuity of norms, the inside of the left hand side converges to $(I - B)C - I$ and 0. Hence, $I - B$ is invertible and

$$(I - B)^{-1} = I + \sum_{n=1}^{\infty} B^n.$$

Thus,

$$\|(I - B)^{-1}\| \leq 1 + \sum_{n=1}^{\infty} \|B\|^n = \frac{1}{1 - \|B\|}.$$

□

The above proof works for any matrix norm.

Among the norms for matrices, there is a class of norms that are very important, namely those induced by vector norms, or the **operator norms**:

$$\|A\|_v := \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v}.$$

In previous homework, we have seen that

Lemma 1. *if $\|\cdot\|$ is some operator norm for the matrix A consistent with some norm for vectors, then*

$$\rho(A) \leq \|A\|,$$

where $\rho(A) = \max_i |\lambda_i|$ is the largest absolute value of eigenvalues, called the spectral radius of A.

Note that if we consider other norms of matrices that cannot be induced by vectors, then this may not be true.

The verification that this is a matrix norm is left as an exercise. There are some important such norms:

Theorem 3. *Denote $\|A\|_p$ be the matrix norms induced by $\|\cdot\|_p$ norms for vectors. Let A be a real matrix of size $m \times n$. Then,*

(a) $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$.

(b) $\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}|$

(c) $\|A\|_2 = \sqrt{\rho(A^TA)} = \sqrt{\rho(AA^T)}$. *Hence, A and A^T have the same 2-norm.*

Proof. Here, we prove the third as an example. You may read the book for the other two.

First of all

$$\langle Ax, Ax\rangle = \langle A^TAx, x\rangle.$$

Since A^TA is real symmetric and positive semi-definite, it has n real eigenvalues:

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0.$$

Moreover, its n eigenvectors are perpendicular to each other. Hence, we can make them orthonormal. We write

$$x = \sum_{i=1}^n c_i v_i.$$

$$\langle A^TAx, x\rangle = \langle \sum_{i=1}^n \lambda_i c_i v_i, \sum_{i=1}^n c_i v_i\rangle = \sum_{i=1}^n \lambda_i c_i^2 \le \lambda_1 \sum_{i=1}^n c_i^2 = \lambda_1 \|x\|_2^2.$$

Hence,

$$\|A\|_2^2 \le \lambda_1.$$

However, if we choose $x = v_1$, the equality can be achieved. This means that we in fact have

$$\|A\|_2 = \sqrt{\lambda_1}.$$

□

Note that AA^T and A^TA can have different sizes. However, the claim is that their eigenvalue sets are the same (though multiplicity may be different). This will be clearer after we know singular value decomposition (SVD). At this point, you may show that

$$\det(\lambda I - AA^T) = \det(\lambda I - A^TA)$$

for invertible matrices (this will be left as an exercise).

Remark 1. *As another comment, the first two seems interersting: $\|A^T\|_1 = \|A\|_\infty$. This is by no means an coincidence. In fact, A^T can be viewed as the dual operator for A. The ℓ^∞ is the dual of ℓ^1.*

3 Error analysis for linear systems: condition number

In this section, all norms for the matrices will be **operator norms**.

3.1 Error bound by condition number

Consider the linear system $Ax = b$. Let $\tilde{x}$ be some approximation solution and the residue vector is

$$r = b - A\tilde{x}.$$

One question is: whether the smallness of r will imply the smallness of the error $x - \tilde{x}$?

Look at the example

$$A = \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}.$$

Consider

$$b = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix}.$$

The solution is $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. However, if we insert $\tilde{x} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$, the residue vector is still very small. The reason is that A is near singular: if 1.0001 is 1, the matrix is not invertible. In other words, A^{-1} could be very large in norms.

To describe this more accurately, let us first prove the following result.

Theorem 4. *Suppose $Ax = b$, $b \neq 0$ has an approximation solution $\tilde{x}$, then*

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|},$$

where

$$r = b - A\tilde{x} = b - \tilde{b} = \delta b.$$

Proof. Since

$$Ax - A\tilde{x} = b - A\tilde{x} = r,$$

one has

$$x - \tilde{x} = A^{-1}r \Rightarrow \|x - \tilde{x}\| \leq \|A^{-1}\|\|r\|.$$

However,

$$b = Ax \Rightarrow \|b\| \leq \|A\|\|x\| \Rightarrow \frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|}.$$

Hence, the claim

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|},$$

follows. □

Note that this error estimate is again in the form of **posterior** type, i.e., we use computed values to control the error for the solutions. The bound tells us that the relative error in b is amplied to the relative error in x by a factor

$$\kappa(A) := \|A\|\|A^{-1}\|.$$

This number measures whether the matrix A is near singular or not. The bigger it is, the more the matrix is close to being singular. This number is called the **condition number**.

Some properties of the condition number:

(a) For any nonsingular matrix A:

$$\kappa(A) \geq \|A^{-1}A\| = 1.$$

(b) For any constant

$$\kappa(cA) = \kappa(A)$$

(c) If the norm is chosen as the spectrum norm: $\|\cdot\| = \|\cdot\|_2$, then

$$\kappa(A) = \|A\|_2\|A^{-1}\|_2 = \sqrt{\frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)}}$$

These properties are very easy to verify.

If $\kappa(A) \approx 1$, then the matrix is away from being singular, and the system is said to be well-conditioned. Otherwise if $\kappa(A) \gg 1$, the system is said to be **ill-conditioned**.

3.2 Perturbation in b and A

In practice, there are always errors in determining A and b. We would like to know how these errors affect the computed solutions.

Theorem 5. *If* $\|\delta A\| < \frac{1}{\|A^{-1}\|}$, *then*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\kappa(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)$$

Proof. We have

$$(A+\delta A)(x+\delta x) = b+\delta b.$$

Using $Ax = b$, one has

$$(A+\delta A)\delta x = \delta b - \delta A x.$$

Hence,

$$\begin{aligned}\|\delta x\| &\leq \|A^{-1}(I+A^{-1}\delta A)^{-1}(\delta b - \delta A x)\| \\ &\leq \|A^{-1}\|\|(I+A^{-1}\delta A)^{-1}\|(\|\delta b\|+\|\delta A\|\|x\|) \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|}(\|\delta b\|+\|\delta A\|\|x\|)\end{aligned}$$

Hence,

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|x\|}+\|\delta A\|\right)$$

Moreover, since

$$\frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|},$$

one has

$$\begin{aligned}\frac{\|\delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|A\|\|\delta b\|}{\|b\|}+\|\delta A\|\right) \\ &= \frac{\kappa}{1-\kappa\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\end{aligned}$$

□

If $\delta A = 0$, this reduces to what we have seen $\kappa\frac{\|\delta b\|}{\|b\|}$. From this bound, it is clear that if the condition number is big, the error bound is big and potentially, the relative error can be very large.

Example Consider the *Hilbert matrix* we have seen in the least square chapter.

4 Iterative refinement*(Not required. This is left as free reading)