Numerical analysis, week 11: Linear systems of equations

Dr. Dmitry Batenkov

June 7, 2021

1 Matrix norms

Definition. V vector space over a field $\mathbb{F} \subseteq \mathbb{C}$, $v \in V$. A **norm** is a function $\|\cdot\| : V \to \mathbb{R}^{\geq 0}$ satisfying

1. $\|v\| \geq 0$ for all $v \in V$, and $\|v\| = 0$ iff $v = 0$
2. $\|\alpha v\| = |\alpha| \, \|v\|$ for all $v \in V$
3. $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in V$

Let $\underline{v} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$. Define

$$\|\underline{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \quad \|\underline{x}\|_1 = \sum_{i=1}^{n} |x_i| \quad , \quad \|\underline{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| \, .$$

It can be readily checked that these are indeed norms.

Definition. Norms $\|x\|, [\![x]\!]$ are **equivalent** if $\exists m, M > 0$ s.t. for all x we have

$$m[\![x]\!] \leq \|x\| \leq M[\![x]\!].$$

Claim 1. Every two norms on $\mathbb{R}^n$ are equivalent.

Check that

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty \quad , \quad \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty.$$

For matrices we can use the above norms as well when $\mathbb{R}^{n \times n}$ is isomorphic to $\mathbb{R}^{n^2}$, but it will be easier to use the so-called *induced norms*.

Definition 1. Given a vector norm $\|\underline{x}\|$ on $\mathbb{R}^n$, the induced norm on $\mathbb{R}^{n\times n}$ is defined as

$$\|A\| = \max_{\underline{x}\neq 0} \frac{\|A\underline{x}\|}{\|\underline{x}\|}.$$

Claim 2. For any induced norm we have

$$\|A\| = \max_{\|\underline{v}\|=1} \|A\underline{v}\|.$$

Corollary 1. *For any vector norm and the corresponding induced norm we have*

$$\|Ax\| \leq \|A\|\|x\|.$$

Claim 3. For any induced norm we have

$$\|AB\| \leq \|A\|\|B\|.$$

Claim 4. $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$.

Claim 5. $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$.

Definition. The **spectral radius** of a matrix A is

$$\rho(A) := \max\{|\lambda| : \ \lambda \text{ is an eigenvalue of } A\}.$$

Theorem 1. *For any induced norm,* $\|A\| \geq \rho(A)$.

Proof. Let $\lambda, \underline{v}$ be an eigenpair, then

$$\|A\| = \max_{\underline{x}\neq 0} \frac{\|A\underline{x}\|}{\|\underline{x}\|} \geq \frac{\|A\underline{v}\|}{\|\underline{v}\|} = |\lambda|.$$

Therefore $\|A\| \geq \rho(A)$ by definition. □

Claim 6. For any A and $\varepsilon > 0$ there exists an induced norm $\|\cdot\|_{A,\varepsilon}$ such that

$$\|A\|_{A,\varepsilon} < \rho(A) + \varepsilon.$$

Theorem 2. *The following are equivalent:*

1. $\lim_{n\to\infty} A^n = 0$.
2. $\lim_{n\to\infty} \|A^n\| = 0$ *for some induced norm* $\|\cdot\|$.

3. $\rho(A) < 1$.

4. *For any* $\underline{v}$,$\lim_{n\to\infty} A^n \underline{v} = 0$.

Proof. The implications are proved as follows:

- $1 \Rightarrow 2$: since the norm is continuous with respect to the matrix entries.
- $2 \Rightarrow 1$: By equivalence of norms, $\|A^n\|_\infty \leq M\|A^n\| \to 0$ and so we must have $A^n \to 0$.
- $2 \Rightarrow 3$: $[\rho(A)]^n = \rho(A^n) \leq \|A^n\| \to 0$ therefore we must have $\rho(A) < 1$.
- $3 \Rightarrow 2$: Since $\rho(A) < 1$, choose $\|\cdot\|$ such that $\|A\| < 1$, and then $\|A^n\| \leq \|A\|^n \to 0$.
- $1 \Rightarrow 4$: for any induced norm, $\|A^n \underline{v}\| \leq \|A^n\|\|\underline{v}\| \to 0$.
- $4 \Rightarrow 3$: Take an eigenpair $\lambda, \underline{v}$ where $|\lambda| = \rho(A)$, then if $|\lambda| \geq 1$ we have $\|A^n \underline{v}\| = |\lambda|^n \|\underline{v}\| \nrightarrow 0$.

□

2 Iterative methods

We seek methods to solving $A\underline{x} = \underline{b}$ with $A \in \mathbb{R}^{n\times n}$ invertible and $\underline{b} \in \mathbb{R}^n$. The so-called *direct* methods produce a solution in a finite number of steps (for example, Gaussian elimination). Here we think about situations where $n \gg 1$ and so the time complexity of direct methods (which is $O(n^3)$) is prohibitive. Instead, we seek *iterative* methods of the form

$$\underline{x}^{k+1} = B\underline{x}^k + \underline{c}, \tag{1}$$

starting with some initial estimate $\underline{x}^0$. If the iterations converge to, say, $\underline{x}$ then taking $k \to \infty$ in the above equation we conclude that $\underline{x}$ is the fixed point of

$$\underline{x} = B\underline{x} + c. \tag{2}$$

The first question is this: given $A\underline{x} = \underline{b}$, how can we choose $B, \underline{c}$ such that $\underline{x}$ satifies (2)?

A general recipe is the following: write A as $A = A_1 + A_2$ where A_1 is invertible and it is "easy" to solve $A_1\underline{y} = \underline{z}$, and then put

$$\begin{aligned}(A_1 + A_2)\,\underline{x} &= \underline{b}\\ A_1\underline{x} &= -A_2\underline{x} + \underline{b}\\ \underline{x} &= \underbrace{-A_1^{-1}A_2}_{B}\,\underline{x} + \underbrace{A_1^{-1}\underline{b}}_{\underline{c}}\end{aligned}$$

2.1 Gauss-Jacobi method

Consider the splitting

$$A = L + D + U,$$

where L (resp. U) is the lower triangular (resp. upper triangular) part of A, and D is the diagonal matrix containing the diagonal elements of A:

$$L = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n-1,1} & \cdot & . & 0 & 0 \\ a_{n,1} & a_{n,2} & \dots & a_{n,n-1} & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1,n-1} & a_{1,n} \\ 0 & 0 & \cdot & \cdot & a_{2,n} \\ 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & 0 & a_{n-1,n} \\ 0 & \cdot & 0 & 0 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} a_{11} & 0 & \cdot & 0 & 0 \\ 0 & a_{22} & 0 & \cdot & 0 \\ \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & a_{n-1,n-1} & \cdot \\ 0 & \cdot & \cdot & 0 & a_{nn} \end{pmatrix}.$$

Suppose that D is invertible (i.e. $a_{ii} \neq 0$ for all $i = 1, \dots, n$). The **Gauss-Jacobi method** puts $A_1 = D$ and $A_2 = L + U$. This gives the iteration

$$\underline{x}^{k+1} = \underbrace{-D^{-1}(L+U)}_{B_J}\underline{x}^k + \underbrace{D^{-1}\underline{b}}_{c_J}$$

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[b_i - \sum_{j \neq i} a_{i,j}x_j^k\right]. \tag{3}$$

This results in $O(n^2)$ operations per iteration, which can be further parallelized (computation of each x_i^{k+1} is independent of x_j^{k+1} for $j \neq i$).

2.2 Gauss-Seidel method

A slight modification to Gauss-Jacobi iteration is to replace x_j^k in (3) with x_j^{k+1} for $j = 1, \dots, i-1$, since these were computed in the previous steps at iteration $k+1$:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1,\dots,i-1} a_{i,j}x_j^{k+1} - \sum_{j=i+1,\dots,n} a_{i,j}x_j^k\right], \quad i = 1, \dots, n. \tag{4}$$

In matrix form, this corresponds to the splitting $A_1 = L + D$ and $A_2 = U$. To see this equivalence, consider

$$\underline{x}^{k+1} = \underbrace{-(L+D)^{-1}U}_{B_{GS}}\underline{x}^k + \underbrace{(L+D)^{-1}\underline{b}}_{\underline{c}_{GS}}$$

$$(L+D)\underline{x}^{k+1} = \underline{b} - U\underline{x}^k = \underline{y}^k.$$

Since U is upper triangular, we have

$$\underline{y}_i^k = b_i - \sum_{j=i+1}^{n} a_{i,j} x_j^k,$$

while the explicit solution of $(L+D)\,\underline{x}^{k+1} = \underline{y}^k$ by forward substitution gives precisely

$$x_i^{k+1} = \frac{1}{a_{ii}} \left[y_i^k - \sum_{j=1}^{i-1} a_{i,j} x_j^{k+1} \right]$$

which is the same as (4).

In contrast with Gauss-Jacobi method, the computations cannot be parallelized; however we may end up with faster convergence since we are using the values x_j^{k+1} "as soon as possible".

2.3 Convergence of iterative methods

Let us consider the general iteration (1). Define the error at iteration k as $\underline{e}^k = \underline{x}^k - \underline{x}$. Then, recalling that $\underline{x} = B\underline{x} + \underline{c}$ we obtain

$$\begin{aligned} \underline{e}^{k+1} &= \underline{x}^{k+1} - \underline{x} \\ &= B\underline{x}^k + \underline{c} - \underline{x} \\ &= B\left(\underline{x}^k - \underline{x}\right) \\ \underline{e}^{k+1} &= B\underline{e}^k = \cdots = B^{k+1}\underline{e}^0 \end{aligned}$$

Recalling theorem 2 we immediately obtain the following result.

Theorem 3. *The iteration (1) converges for every $\underline{x}^0$ if and only if $\rho(B) < 1$.*

Remark. If all eigenvalues λ of B satisfy $|\lambda| > 1$ then the iteration does not converge for any $\underline{x}^0$. In contrast, if for some λ we have $|\lambda| < 1$ then $\underline{x}^0 = v_\lambda$ will converge (disregarding roundoff errors of course).

For Gauss-Jacobi (resp. Gauss-Seidel), a necessary and sufficient condition for convergence for any initial estimate would be $\rho(B_J) < 1$ (resp. $\rho(B_{GS}) < 1$). This might be difficult to check directly. Fortunately, there is a simpler sufficient condition.

Definition 2. The matrix B is called *(row-wise) diagonally dominant* if for every $i = 1, \ldots, n$

$$|a_{ii}| > \sum_{j \neq i} |a_{i,j}|\,.$$

Theorem 4. *Suppose that A is diagonally dominant. Then Gauss-Jacobi (resp. Gauss-Seidel) converges for any initial estimate $\underline{x}^0$.*

Proof. Let us prove the claim for Gauss-Jacobi:

$$\begin{aligned}\|B_J\|_\infty &= \|D^{-1}(L+U)\|_\infty \\ &= \max_{i=1,\dots,n} \frac{\sum_{j\neq i}|a_{i,j}|}{|a_{i,i}|} \\ &< 1.\end{aligned}$$

□

2.4 Error estimate

We would like a quick method to estimate the error at iteration k. We have already seen that $\underline{e}^k = B^k \underline{e}^0$ and therefore

$$\|\underline{e}^k\| \leq \|B\|^k \|\underline{e}^0\| \tag{5}$$

but of course we don't know $\underline{e}^0$. Luckily we have the following result.

Claim 7. For any induced norm we have

$$\|\underline{e}^k\| \leq \frac{\|B\|^k}{1-\|B\|}\|\underline{x}^1 - \underline{x}^0\|. \tag{6}$$

Proof. Recall $\underline{e}^{k+1} = B\underline{e}^k$, then

$$\begin{aligned}\|\underline{e}^k\| &= \|\underline{x}^k - \underline{x}\| \\ &\leq \|\underbrace{\underline{x} - \underline{x}^{k+1}}_{=\underline{e}^{k+1}}\| + \|\underline{x}^{k+1} - \underline{x}^k\| \\ &\leq \|B\|\|\underline{e}^k\| + \|\underline{x}^{k+1} - \underline{x}^k\| \\ \|\underline{e}^k\|(1-\|B\|) &\leq \|\underline{x}^{k+1} - \underline{x}^k\|.\end{aligned}$$

On the other hand,

$$\underline{x}^{k+1} - \underline{x}^k = \underline{e}^{k+1} - \underline{e}^k = B\left(\underline{e}^k - \underline{e}^{k-1}\right) = B\left(\underline{x}^k - \underline{x}^{k-1}\right) = \cdots = B^k\left(\underline{x}^1 - \underline{x}^0\right)$$

and so

$$\begin{aligned}\|\underline{e}^k\| &\leq \frac{\|\underline{x}^{k+1} - \underline{x}^k\|}{1-\|B\|} \\ &\leq \frac{\|B\|^k}{1-\|B\|}\|\underline{x}^1 - \underline{x}^0\|,\end{aligned}$$

finishing the proof. □

Example 1. Consider solving the following system by Gauss-Jacobi method:

$$\begin{pmatrix} 2 & -1 & 0 \\ 1 & 4 & 2 \\ -2 & 1 & 5 \end{pmatrix} \underline{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad , \underline{x}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

how many iterations are needed to ensure $\left\|\underline{x} - \underline{x}^k\right\|_\infty \leq 10^{-4}$?

Solution. Using (6), we compute:

$$B = \begin{pmatrix} 0 & 1/2 & 0 \\ -1/4 & 0 & -2/4 \\ 2/5 & -1/5 & 0 \end{pmatrix} \Rightarrow \|B\|_\infty = \frac{3}{4}$$

$$\underline{c} = D^{-1}\underline{b} = \begin{pmatrix} 1/2 \\ 1/4 \\ 1/5 \end{pmatrix}$$

$$\underline{x}^{(1)} = \begin{pmatrix} 0 & 1/2 & 0 \\ -1/4 & 0 & -2/4 \\ 2/5 & -1/5 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1/2 \\ 1/4 \\ 1/5 \end{pmatrix} = \begin{pmatrix} 1 \\ -1/2 \\ 2/5 \end{pmatrix}$$

and therefore

$$\left\|\underline{x}^{(1)} - \underline{x}^{(0)}\right\|_\infty = \left\| \begin{pmatrix} 0 \\ -3/2 \\ -3/5 \end{pmatrix} \right\|_\infty = 3/2$$

$$\left\|\underline{e}^{(k)}\right\|_\infty \leq \frac{\|B\|_\infty^k}{1 - \|B\|_\infty} \left\|\underline{x}^{(1)} - \underline{x}^{(0)}\right\|_\infty = \frac{(3/4)^k}{(1/4)} \cdot \frac{3}{2} = 6(3/4)^k$$

$$6(3/4)^k \leq 10^{-4}$$

$$\lceil k \rceil \geq 39.$$

Example 2. Consider the linear system with a parameter p.

$$\begin{pmatrix} 1/3 & p \\ 1/6 & 1/4 \end{pmatrix} \underline{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

1. Give a sufficient condition on p to ensure convergence of Gauss-Seidel method.
2. Write down the Gauss-Seidel iteration for $p = 1/4$.
3. For $p = 1/4$, find a bound for the error after k iterations of Gauss-Seidel in the ∞-norm.

Solution. We can use theorem 4. To ensure diagonal dominance, we must have $|p| < 1/3$. For $p = 1/4$, we compute

$$\begin{aligned} B_{GS} &= -(L+D)^{-1} U \\ &= -\begin{pmatrix} 1/3 & 0 \\ 1/6 & 1/4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1/4 \\ 0 & 0 \end{pmatrix} = -\begin{pmatrix} 3 & 0 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1/4 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{3}{4} \\ 0 & \frac{1}{2} \end{pmatrix} \\ \underline{c}_{GS} &= (L+D)^{-1} \underline{b} \\ &= -\begin{pmatrix} 3 & 0 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \end{pmatrix}. \end{aligned}$$

According to (5) we have

$$\begin{aligned} \|B_{GS}\|_\infty &= 3/4 \\ \|\underline{e}^k\| &\leq \left(\frac{3}{4}\right)^k \|\underline{e}^0\|_\infty. \end{aligned}$$

3 Condition numbers

Question: what is the influence of small changes in the data on the accuracy of the computed solution in $Ax = b$?

Example 3. For example, consider

$$\begin{aligned} 1.01x_1 + 0.99x_2 &= 2 \\ 0.99x_1 + 1.01x_2 &= 2 \end{aligned} \tag{7}$$

whose exact solution is $x_1 = x_2 = 1$. If we change the right hand side a little bit:

$$\begin{aligned} 1.01x_1 + 0.99x_2 &= 2.02 \\ 0.99x_1 + 1.01x_2 &= 1.98 \end{aligned} \quad ,$$

the solution will become $x_1 = 2, x_2 = 0$, which is very far from the unperturbed case.

To study the problem in general, suppose that b is changed by δb. Let $\tilde{\underline{x}}$ denote the solution to the perturbed system, i.e.

$$A\tilde{\underline{x}} = \underline{b} + \underline{\delta b}$$

Using $A\underline{x} = \underline{b}$ and subtracting, we get

$$\begin{aligned} A(\tilde{\underline{x}} - \underline{x}) &= \underline{\delta b} \\ \tilde{\underline{x}} - \underline{x} &= A^{-1}\underline{\delta b} \\ \|\tilde{\underline{x}} - \underline{x}\| &\leq \|A^{-1}\| \|\underline{\delta b}\|. \end{aligned}$$

Put $\underline{\delta x} = \underline{\tilde{x}} - \underline{x}$. Since $\underline{b} = A\underline{x}$ we also have $\|\underline{b}\| \leq \|A\|\|\underline{x}\|$ and so the relative error in $\underline{x}$ can be expressed in terms of the relative error in $\underline{b}$ as follows:

$$\frac{\|\underline{\delta x}\|}{\|\underline{x}\|} \leq \underbrace{\|A\|\|A^{-1}\|}_{\text{cond}(A)} \frac{\|\underline{\delta b}\|}{\|\underline{b}\|}$$

The number $\|A\|\|A^{-1}\|$ (which depends on the norm) is called the *condition number* of A and is frequently denoted by $\kappa(A) = \text{cond}(A)$.

On the other hand, $\|\underline{\delta b}\| \leq \|A\|\|\underline{\delta x}\|$ and $\|\underline{x}\| \leq \|A^{-1}\|\|\underline{b}\|$ and so in the other direction we have

$$\frac{\|\underline{\delta x}\|}{\|\underline{x}\|} \geq \frac{1}{\kappa(A)} \frac{\|\underline{\delta b}\|}{\|\underline{b}\|},$$

giving that

$$\frac{\|\underline{\delta x}\|}{\|\underline{x}\|} \in \frac{\|\underline{\delta b}\|}{\|\underline{b}\|} \left\{ \frac{1}{\kappa(A)}, \kappa(A) \right\}.$$

If $\kappa(A) \gg 1$ the matrix is called **ill-conditioned**, otherwise it is called **well-conditioned**.

In the example above:

$$A = \begin{pmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{pmatrix}$$
$$A^{-1} = \frac{1}{0.04} \begin{pmatrix} 1.01 & -0.99 \\ -0.99 & 1.01 \end{pmatrix}$$
$$\kappa(A)_\infty = \|A\|_\infty \|A^{-1}\|_\infty = 100.$$

The condition number satisfies the following properties:

1. $\kappa(A) \geq \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|} \geq 1$. Indeed, for any induced norm we have $\|A\| \geq |\lambda_{\max}(A)|$, while if A is invertible, we also have $\|A^{-1}\| \geq |\lambda_{\max}(A^{-1})| = |\lambda_{\min}(A)|^{-1}$.

2. We have (without proof)

$$\kappa(A) = \max_{|B|=0} \frac{\|A\|}{\|A - B\|} = \frac{\|A\|}{\min_{|B|=0} \|A - B\|}$$

   Why is the denominator never zero?

Example 4. For $A = \begin{pmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{pmatrix}$, a close non-invertible matrix is $B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, therefore

$$\kappa(A)_\infty \geq \frac{\|A\|_\infty}{0.02} = 100.$$

The property 1) enables to define $\kappa(A)_\lambda = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$.

Is $\kappa(A) = 100$ big or small? If the relative error in $\underline{b}$ is 1%, the relative error in $\underline{x}$ might be as large as 100%.