Chapter 4

Vector Norms and Matrix Norms

4.1 Normed Vector Spaces

In order to define how close two vectors or two matrices are, and in order to define the convergence of sequences of vectors or matrices, we can use the notion of a *norm*.

Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Also recall that if $z = a + ib \in \mathbb{C}$ is a complex number, with $a, b \in \mathbb{R}$, then $\overline{z} = a - ib$ and $|z| = \sqrt{a^2 + b^2}$ ($|z|$ is the *modulus* of z).

Definition 4.1. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm* on E is a function $\| \,\| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \,\|$ is called a *normed vector space*.

From (N3), we easily get

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

Example 4.1.

1. Let $E = \mathbb{R}$, and $\|x\| = |x|$, the absolute value of x.
2. Let $E = \mathbb{C}$, and $\|z\| = |z|$, the modulus of z.
3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms.

   For every $(x_1, \ldots, x_n) \in E$, we have the *1-norm* $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

   we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

   and the *sup-norm* $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

   More generally, we define the *ℓ_p-norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

There are other norms besides the ℓ_p-norms; we urge the reader to find such norms.

Some work is required to show the triangle inequality for the ℓ_p-norm.

Proposition 4.1. *If E is a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$, for every real number $p \geq 1$, the ℓ_p-norm is indeed a norm.*

The proof uses the following facts:

If $q \geq 1$ is given by

$$\frac{1}{p} + \frac{1}{q} = 1,$$

then

(1) For all $\alpha, \beta \in \mathbb{R}$, if $\alpha, \beta \geq 0$, then

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}. \tag{$*$}$$

(2) For any two vectors $u, v \in E$, we have

$$\sum_{i=1}^{n} |u_i v_i| \leq \|u\|_p \|v\|_q . \tag{$**$}$$

For $p > 1$ and $1/p + 1/q = 1$, the inequality

$$\sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |v_i|^q \right)^{1/q}$$

is known as *Hölder's inequality*.

For $p = 2$, it is the *Cauchy–Schwarz inequality*.

Actually, if we define the *Hermitian inner product* $\langle -, - \rangle$ on $\mathbb{C}^n$ by

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i \overline{v}_i,$$

where $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$, then

$$|\langle u, v \rangle| \leq \sum_{i=1}^{n} |u_i \overline{v}_i| = \sum_{i=1}^{n} |u_i v_i|,$$

so Hölder's inequality implies the inequality

$$|\langle u, v \rangle| \leq \|u\|_p \|v\|_q$$

also called *Hölder's inequality*, which, for $p = 2$ is the standard Cauchy–Schwarz inequality.

The triangle inequality for the ℓ_p-norm,

$$\left(\sum_{i=1}^{n}(|u_i+v_i|)^p\right)^{1/p} \leq \left(\sum_{i=1}^{n}|u_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n}|v_i|^q\right)^{1/q},$$

is known as *Minkowski's inequality*.

When we restrict the Hermitian inner product to real vectors, $u, v \in \mathbb{R}^n$, we get the *Euclidean inner product*

$$\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i.$$

It is very useful to observe that if we represent (as usual) $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ (in $\mathbb{R}^n$) by column vectors, then their Euclidean inner product is given by

$$\langle u, v\rangle = u^\top v = v^\top u,$$

and when $u, v \in \mathbb{C}^n$, their Hermitian inner product is given by

$$\langle u, v\rangle = v^* u = \overline{u^* v}.$$

In particular, when $u = v$, in the complex case we get

$$\|u\|_2^2 = u^* u,$$

and in the real case, this becomes

$$\|u\|_2^2 = u^\top u.$$

As convenient as these notations are, we still recommend that you do not abuse them; the notation $\langle u, v\rangle$ is more intrinsic and still "works" when our vector space is infinite dimensional.

Proposition 4.2. *The following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):*

$$\begin{aligned} \|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty, \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \\ \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\|x\|_2. \end{aligned}$$

Proposition 4.2 is actually a special case of a very important result: in a finite-dimensional vector space, any two norms are equivalent.

Definition 4.2. Given any (real or complex) vector space E, two norms $\| \|_a$ and $\| \|_b$ are *equivalent* iff there exists some positive reals $C_1, C_2 > 0$, such that

$$\|u\|_a \leq C_1 \|u\|_b \quad \text{and} \quad \|u\|_b \leq C_2 \|u\|_a \,, \text{ for all } u \in E.$$

Given any norm $\| \|$ on a vector space of dimension n, for any basis $(e_1, \ldots, e_n)$ of E, observe that for any vector $x = x_1e_1 + \cdots + x_ne_n$, we have

$$\|x\| = \|x_1e_1 + \cdots + x_ne_n\| \leq C \|x\|_1 \,,$$

with $C = \max_{1 \leq i \leq n} \|e_i\|$ and

$$\|x\|_1 = \|x_1e_1 + \cdots + x_ne_n\| = |x_1| + \cdots + |x_n|.$$

The above implies that

$$| \, \|u\| - \|v\| \, | \leq \|u - v\| \leq C \|u - v\|_1 \,,$$

which means that the map $u \mapsto \|u\|$ is *continuous* with respect to the norm $\| \|_1$.

Let S_1^{n-1} be the unit ball with respect to the norm $\| \ \|_1$, namely

$$S_1^{n-1} = \{x \in E \mid \|x\|_1 = 1\}.$$

Now, S_1^{n-1} is a closed and bounded subset of a finite-dimensional vector space, so by Heine–Borel (or equivalently, by Bolzano–Weiertrass), S_1^{n-1} is compact.

On the other hand, it is a well known result of analysis that any continuous real-valued function on a nonempty compact set has a minimum and a maximum, and that they are achieved.

Using these facts, we can prove the following important theorem:

Theorem 4.3. *If E is any real or complex vector space of finite dimension, then any two norms on E are equivalent.*

Next, we will consider norms on matrices.

4.2 Matrix Norms

For simplicity of exposition, we will consider the vector spaces $\mathrm{M}_n(\mathbb{R})$ and $\mathrm{M}_n(\mathbb{C})$ of square $n \times n$ matrices.

Most results also hold for the spaces $\mathrm{M}_{m,n}(\mathbb{R})$ and $\mathrm{M}_{m,n}(\mathbb{C})$ of rectangular $m \times n$ matrices.

Since $n \times n$ matrices can be multiplied, the idea behind matrix norms is that they should behave "well" with respect to matrix multiplication.

Definition 4.3. A *matrix norm* $\| \|$ on the space of square $n \times n$ matrices in $\mathrm{M}_n(K)$, with $K = \mathbb{R}$ or $K = \mathbb{C}$, is a norm on the vector space $\mathrm{M}_n(K)$ with the additional property that

$$\|AB\| \leq \|A\| \, \|B\| \, ,$$

for all $A, B \in \mathrm{M}_n(K)$.

Since $I^2 = I$, from $\|I\| = \left\|I^2\right\| \leq \|I\|^2$, we get $\|I\| \geq 1$, for every matrix norm.

Before giving examples of matrix norms, we need to review some basic definitions about matrices.

Given any matrix $A = (a_{ij}) \in \mathrm{M}_{m,n}(\mathbb{C})$, the *conjugate* $\overline{A}$ of A is the matrix such that

$$\overline{A}_{ij} = \overline{a}_{ij}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

The *transpose* of A is the $n \times m$ matrix $A^{\top}$ such that

$$A^{\top}_{ij} = a_{ji}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

The *adjoint* of A is the $n \times m$ matrix A^* such that

$$A^* = \overline{(A^{\top})} = (\overline{A})^{\top}.$$

When A is a real matrix, $A^* = A^{\top}$.

A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *Hermitian* if

$$A^* = A.$$

If A is a real matrix ($A \in \mathrm{M}_n(\mathbb{R})$), we say that A is *symmetric* if

$$A^{\top} = A.$$

A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *normal* if

$$AA^* = A^*A,$$

and if A is a real matrix, it is *normal* if

$$AA^\top = A^\top A.$$

A matrix $U \in \mathrm{M}_n(\mathbb{C})$ is *unitary* if

$$UU^* = U^*U = I.$$

A real matrix $Q \in \mathrm{M}_n(\mathbb{R})$ is *orthogonal* if

$$QQ^\top = Q^\top Q = I.$$

Given any matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, the *trace* $\mathrm{tr}(A)$ of A is the sum of its diagonal elements

$$\mathrm{tr}(A) = a_{11} + \cdots + a_{nn}.$$

It is easy to show that the trace is a linear map, so that

$$\mathrm{tr}(\lambda A) = \lambda \mathrm{tr}(A)$$

and

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$

Moreover, if A is an $m \times n$ matrix and B is an $n \times m$ matrix, it is not hard to show that

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

We also review eigenvalues and eigenvectors. We content ourselves with definition involving matrices. A more general treatment will be given later on (see Chapter 9).

Definition 4.4. Given any square matrix $A \in \mathrm{M}_n(\mathbb{C})$, a complex number $\lambda \in \mathbb{C}$ is an *eigenvalue* of A if there is some *nonzero* vector $u \in \mathbb{C}^n$, such that

$$Au = \lambda u.$$

If λ is an eigenvalue of A, then the *nonzero* vectors $u \in \mathbb{C}^n$ such that $Au = \lambda u$ are called *eigenvectors of A associated with λ*; together with the zero vector, these eigenvectors form a subspace of $\mathbb{C}^n$ denoted by $E_\lambda(A)$, and called the *eigenspace associated with λ*.

Remark: Note that Definition 4.4 *requires an eigenvector to be nonzero*.

A somewhat unfortunate consequence of this requirement is that the set of eigenvectors is *not* a subspace, since the zero vector is missing!

On the positive side, whenever eigenvectors are involved, there is no need to say that they are nonzero.

If A is a square real matrix $A \in \mathrm{M}_n(\mathbb{R})$, then we restrict Definition 4.4 to real eigenvalues $\lambda \in \mathbb{R}$ and real eigenvectors.

However, it should be noted that although every complex matrix always has at least some complex eigenvalue, a real matrix may not have any real eigenvalues. For example, the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has the complex eigenvalues i and $-i$, but no real eigenvalues.

Thus, typically, even for real matrices, we consider complex eigenvalues.

Observe that $\lambda \in \mathbb{C}$ is an eigenvalue of A
iff $Au = \lambda u$ for some nonzero vector $u \in \mathbb{C}^n$
iff $(\lambda I - A)u = 0$
iff the matrix $\lambda I - A$ defines a linear map which has a nonzero kernel, that is,
iff $\lambda I - A$ not invertible.

However, from Proposition 3.10, $\lambda I - A$ is not invertible iff

$$\det(\lambda I - A) = 0.$$

Now, $\det(\lambda I - A)$ is a polynomial of degree n in the indeterminate λ, in fact, of the form

$$\lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A).$$

Thus, we see that the eigenvalues of A are the zeros (also called *roots*) of the above polynomial.

Since every complex polynomial of degree n has exactly n roots, counted with their multiplicity, we have the following definition:

Definition 4.5. Given any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, the polynomial

$$\det(\lambda I - A) = \lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A)$$

is called the *characteristic polynomial* of A. The n (not necessarily distinct) roots $\lambda_1, \ldots, \lambda_n$ of the characteristic polynomial are all the *eigenvalues* of A and constitute the *spectrum* of A.

We let

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

be the largest modulus of the eigenvalues of A, called the *spectral radius* of A.

Proposition 4.4. *For any matrix norm $\| \ \|$ on $\mathrm{M}_n(\mathbb{C})$ and for any square $n \times n$ matrix A, we have*

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

Remark: Proposition 4.4 still holds for real matrices $A \in \mathrm{M}_n(\mathbb{R})$, but a different proof is needed since in the above proof the eigenvector u may be complex.

We use Theorem 4.3 and a trick based on the fact that

$$\rho(A^k) = (\rho(A))^k \quad \text{for all} \quad k \geq 1.$$

Now, it turns out that if A is a real $n \times n$ *symmetric* matrix, then the eigenvalues of A are all *real* and there is some orthogonal matrix Q such that

$$A = Q^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_n) Q,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A.

Similarly, if A is a complex $n \times n$ *Hermitian* matrix, then the eigenvalues of A are all *real* and there is some unitary matrix U such that

$$A = U^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) U,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A.

We now return to matrix norms. We begin with the so-called *Frobenius norm*, which is just the norm $\| \, \|_2$ on $\mathbb{C}^{n^2}$, where the $n \times n$ matrix A is viewed as the vector obtained by concatenating together the rows (or the columns) of A.

The reader should check that for any $n \times n$ complex matrix $A = (a_{ij})$,

$$\left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(A^* A)} = \sqrt{\mathrm{tr}(A A^*)}.$$

Definition 4.6. The *Frobenius norm* $\| \ \|_F$ is defined so that for every square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$,

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(AA^*)} = \sqrt{\mathrm{tr}(A^*A)}.$$

The following proposition show that the Frobenius norm is a matrix norm satisfying other nice properties.

Proposition 4.5. *The Frobenius norm* $\| \ \|_F$ *on* $\mathrm{M}_n(\mathbb{C})$ *satisfies the following properties:*

(1) It is a matrix norm; that is, $\|AB\|_F \leq \|A\|_F \|B\|_F$*, for all* $A, B \in \mathrm{M}_n(\mathbb{C})$*.*

(2) It is unitarily invariant, which means that for all unitary matrices U, V*, we have*

$$\|A\|_F = \|UA\|_F = \|AV\|_F = \|UAV\|_F.$$

(3) $\sqrt{\rho(A^*A)} \leq \|A\|_F \leq \sqrt{n}\sqrt{\rho(A^*A)}$*, for all* $A \in \mathrm{M}_n(\mathbb{C})$*.*

Remark: The Frobenius norm is also known as the *Hilbert-Schmidt norm* or the *Schur norm*. So many famous names associated with such a simple thing!

We now give another method for obtaining matrix norms using subordinate norms.

First, we need a proposition that shows that in a finite-dimensional space, the linear map induced by a matrix is bounded, and thus continuous.

Proposition 4.6. *For every norm $\| \;\|$ on $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$), there is a real constant $C_A > 0$, such that*

$$\|Au\| \leq C_A \|u\|,$$

for every vector $u \in \mathbb{C}^n$ (or $u \in \mathbb{R}^n$ if A is real).

Proposition 4.6 says that every linear map on a finite-dimensional space is *bounded*.

This implies that every linear map on a finite-dimensional space is continuous.

Actually, it is not hard to show that a linear map on a normed vector space E is bounded iff it is continuous, regardless of the dimension of E.

Proposition 4.6 implies that for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$),

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} \leq C_A.$$

Now, since $\|\lambda u\| = |\lambda| \, \|u\|$, it is easy to show that

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\| = 1}} \|Ax\|.$$

Similarly

$$\sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\| = 1}} \|Ax\|.$$

Definition 4.7. If $\| \,\|$ is any norm on $\mathbb{C}^n$, we define the function $\| \,\|$ on $\mathrm{M}_n(\mathbb{C})$ by

$$\|A\| = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|=1}} \|Ax\| .$$

The function $A \mapsto \|A\|$ is called the *subordinate matrix norm* or *operator norm* induced by the norm $\| \,\|$.

It is easy to check that the function $A \mapsto \|A\|$ is indeed a norm, and by definition, it satisfies the property

$$\|Ax\| \leq \|A\| \, \|x\| ,$$

for all $x \in \mathbb{C}^n$.

This implies that

$$\|AB\| \leq \|A\| \, \|B\|$$

for all $A, B \in \mathrm{M}_n(\mathbb{C})$, showing that $A \mapsto \|A\|$ is a matrix norm.

Observe that the subordinate matrix norm is also defined by

$$\|A\| = \inf\{\lambda \in \mathbb{R} \mid \|Ax\| \leq \lambda \|x\|, \text{ for all } x \in \mathbb{C}^n\}.$$

The definition also implies that

$$\|I\| = 1.$$

The above show that the Frobenius norm is not a subordinate matrix norm (why?).

Remark: For any norm $\| \ \|$ on $\mathbb{C}^n$, we can define the function $\| \ \|_{\mathbb{R}}$ on $\mathrm{M}_n(\mathbb{R})$ by

$$\|A\|_{\mathbb{R}} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\| = 1}} \|Ax\| .$$

The function $A \mapsto \|A\|_{\mathbb{R}}$ is a matrix norm on $\mathrm{M}_n(\mathbb{R})$, and

$$\|A\|_{\mathbb{R}} \leq \|A\| ,$$

for all real matrices $A \in \mathrm{M}_n(\mathbb{R})$.

However, it is possible to construct vector norms $\| \ \|$ on $\mathbb{C}^n$ and *real* matrices A such that

$$\|A\|_{\mathbb{R}} < \|A\| .$$

In order to avoid this kind of difficulties, we define subordinate matrix norms over $\mathrm{M}_n(\mathbb{C})$.

Luckily, it turns out that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms, $\| \ \|_1$, $\| \ \|_2$, and $\| \ \|_\infty$.

Proposition 4.7. *For every square matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\|A\|_1 = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_1=1}} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

$$\|A\|_\infty = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_\infty=1}} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

$$\|A\|_2 = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_2=1}} \|Ax\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}.$$

Furthermore, $\|A^*\|_2 = \|A\|_2$, the norm $\| \ \|_2$ is unitarily invariant, which means that

$$\|A\|_2 = \|UAV\|_2$$

for all unitary matrices U, V, and if A is a normal matrix, then $\|A\|_2 = \rho(A)$.

The norm $\|A\|_2$ is often called the *spectral norm*.

Observe that property (3) of proposition 4.5 says that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\, \|A\|_2 \,,$$

which shows that the Frobenius norm is an upper bound on the spectral norm. The Frobenius norm is much easier to compute than the spectal norm.

The reader will check that the above proof still holds if the matrix A is real, confirming the fact that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms $\| \;\|_1$, $\| \;\|_2$, and $\| \;\|_\infty$.

It is also easy to verify that the proof goes through for *rectangular* matrices, with the same formulae.

Similarly, the Frobenius norm is also a norm on rectangular matrices. For these norms, whenever AB makes sense, we have

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

The following proposition will be needed when we deal with the condition number of a matrix.

Proposition 4.8. *Let $\| \ \|$ be any matrix norm and let B be a matrix such that $\|B\| < 1$.*

(1) If $\| \ \|$ is a subordinate matrix norm, then the matrix $I + B$ is invertible and

$$\left\|(I + B)^{-1}\right\| \leq \frac{1}{1 - \|B\|}.$$

(2) If a matrix of the form $I + B$ is singular, then $\|B\| \geq 1$ for every matrix norm (not necessarily subordinate).

The following result is needed to deal with the convergence of sequences of powers of matrices.

Proposition 4.9. *For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \, \|$ such that*

$$\|A\| \leq \rho(A) + \epsilon.$$

The proof uses Theorem 9.4, which says that there exists some invertible matrix U and some upper triangular matrix T such that

$$A = UTU^{-1}.$$

Note that equality is generally not possible; consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

for which $\rho(A) = 0 < \|A\|$, since $A \neq 0$.

4.3 Condition Numbers of Matrices

Unfortunately, there exist linear systems $Ax = b$ whose solutions are *not stable under small perturbations of either b or A*.

For example, consider the system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

The reader should check that it has the solution $x = (1, 1, 1, 1)$. If we perturb slightly the right-hand side, obtaining the new system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$

the new solutions turns out to be $x = (9.2, -12.6, 4.5, -1.1)$.

In other words, a relative error of the order $1/200$ in the data (here, b) produces a relative error of the order $10/1$ in the solution, which represents an amplification of the relative error of the order 2000.

Now, let us perturb the matrix slightly, obtaining the new system

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.98 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

This time, the solution is $x = (-81, 137, -34, 22)$.

Again, a small change in the data alters the result rather drastically.

Yet, the original system is symmetric, has determinant 1, and has integer entries.

The problem is that the matrix of the system is *badly conditioned*, a concept that we will now explain.

Given an invertible matrix A, first, assume that we perturb b to $b + \delta b$, and let us analyze the change between the two exact solutions x and $x + \delta x$ of the two systems

$$\begin{aligned} Ax &= b \\ A(x + \delta x) &= b + \delta b. \end{aligned}$$

We also assume that we have some norm $\| \ \|$ and we use the subordinate matrix norm on matrices. From

$$\begin{aligned} Ax &= b \\ Ax + A\delta x &= b + \delta b, \end{aligned}$$

we get

$$\delta x = A^{-1}\delta b,$$

and we conclude that

$$\begin{aligned} \|\delta x\| &\leq \left\|A^{-1}\right\| \|\delta b\| \\ \|b\| &\leq \|A\| \|x\| . \end{aligned}$$

Consequently, the relative error in the result $\|\delta x\| / \|x\|$ is bounded in terms of the relative error $\|\delta b\| / \|b\|$ in the data as follows:

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

Now let us assume that A is perturbed to $A + \delta A$, and let us analyze the change between the exact solutions of the two systems

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b. \end{aligned}$$

After some calculations, we get

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \left(\|A\| \, \|A^{-1}\| \right) \frac{\|\Delta A\|}{\|A\|}.$$

Observe that the above reasoning is valid even if the matrix $A + \Delta A$ is singular, as long as $x + \Delta x$ is a solution of the second system.

Furthermore, if $\|\Delta A\|$ is small enough, it is not unreasonable to expect that the ratio $\|\Delta x\| / \|x + \Delta x\|$ is close to $\|\Delta x\| / \|x\|$.

This will be made more precise later.

In summary, for each of the two perturbations, we see that the relative error in the result is bounded by the relative error in the data, *multiplied the number* $\|A\| \|A^{-1}\|$.

In fact, this factor turns out to be optimal and this suggests the following definition:

Definition 4.8. For any subordinate matrix norm $\| \|$, for any invertible matrix A, the number

$$\mathrm{cond}(A) = \|A\| \|A^{-1}\|$$

is called the *condition number* of A relative to $\| \|$.

The condition number $\mathrm{cond}(A)$ measures the sensitivity of the linear system $Ax = b$ to variations in the data b and A; a feature referred to as the *condition* of the system.

Thus, when we says that a linear system is *ill-conditioned*, we mean that the condition number of its matrix is large.

We can sharpen the preceding analysis as follows:

Proposition 4.10. *Let A be an invertible matrix and let x and $x + \delta x$ be the solutions of the linear systems*

$$\begin{aligned} Ax &= b \\ A(x + \delta x) &= b + \delta b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

holds and is the best possible. This means that for a given matrix A, there exist some vectors $b \neq 0$ and $\delta b \neq 0$ for which equality holds.

Proposition 4.11. *Let A be an invertible matrix and let x and $x + \Delta x$ be the solutions of the two systems*

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

holds and is the best possible. This means that given a matrix A, there exist a vector $b \neq 0$ and a matrix $\Delta A \neq 0$ for which equality holds. Furthermore, if $\|\Delta A\|$ is small enough (for instance, if $\|\Delta A\| < 1/\left\|A^{-1}\right\|$), we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}(1 + O(\|\Delta A\|));$$

in fact, we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}\left(\frac{1}{1 - \|A^{-1}\|\,\|\Delta A\|}\right).$$

Remark: If A and b are perturbed simultaneously, so that we get the "perturbed" system

$$(A + \Delta A)(x + \delta x) = b + \delta b,$$

it can be shown that if $\|\Delta A\| < 1/\left\|A^{-1}\right\|$ (and $b \neq 0$), then

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\| \, \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$

We now list some properties of condition numbers and figure out what $\mathrm{cond}(A)$ is in the case of the spectral norm (the matrix norm induced by $\| \, \|_2$).

First, we need to introduce a very important factorization of matrices, the *singular value decomposition*, for short, *SVD*.

It can be shown that given any $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, there exist two unitary matrices U and V, and a *real* diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, such that

$$A = V \Sigma U^*.$$

The nonnegative numbers $\sigma_1, \ldots, \sigma_n$ are called the *singular values* of A.

If A is a real matrix, the matrices U and V are orthogonal matrices.

The factorization $A = V\Sigma U^*$ implies that

$$A^*A = U\Sigma^2 U^* \quad \text{and} \quad AA^* = V\Sigma^2 V^*,$$

which shows that $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of *both* A^*A and AA^*, that the columns of U are corresponding eivenvectors for A^*A, and that the columns of V are corresponding eivenvectors for AA^*.

In the case of a normal matrix if $\lambda_1, \ldots, \lambda_n$ are the (complex) eigenvalues of A, then

$$\sigma_i = |\lambda_i|, \quad 1 \leq i \leq n.$$

Proposition 4.12. *For every invertible matrix $A \in \mathrm{M}_n(\mathbb{C})$, the following properties hold:*

(1)

$$\begin{aligned} \mathrm{cond}(A) &\geq 1, \\ \mathrm{cond}(A) &= \mathrm{cond}(A^{-1}) \\ \mathrm{cond}(\alpha A) &= \mathrm{cond}(A) \quad \textit{for all } \alpha \in \mathbb{C} - \{0\}. \end{aligned}$$

(2) If $\mathrm{cond}_2(A)$ denotes the condition number of A with respect to the spectral norm, then

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_n},$$

where $\sigma_1 \geq \cdots \geq \sigma_n$ are the singular values of A.

(3) If the matrix A is normal, then

$$\mathrm{cond}_2(A) = \frac{|\lambda_1|}{|\lambda_n|},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A sorted so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$.

(4) If A is a unitary or an orthogonal matrix, then

$$\mathrm{cond}_2(A) = 1.$$

(5) *The condition number* $\mathrm{cond}_2(A)$ *is invariant under unitary transformations, which means that*

$$\mathrm{cond}_2(A) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(AV),$$

for all unitary matrices U *and* V.

Proposition 4.12 (4) shows that unitary and orthogonal transformations are very well-conditioned, and part (5) shows that unitary transformations preserve the condition number.

In order to compute $\mathrm{cond}_2(A)$, we need to compute the top and bottom singular values of A, which may be hard.

The inequality

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\, \|A\|_2\,,$$

may be useful in getting an approximation of $\mathrm{cond}_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2$, if A^{-1} can be determined.

Remark: There is an interesting geometric characterization of $\mathrm{cond}_2(A)$.

If $\theta(A)$ denotes the *least angle* between the vectors Au and Av as u and v range over all pairs of orthonormal vectors, then it can be shown that

$$\mathrm{cond}_2(A) = \cot(\theta(A)/2)).$$

Thus, if A is nearly singular, then there will be some orthonormal pair u, v such that Au and Av are nearly parallel; the angle $\theta(A)$ will the be small and $\cot(\theta(A)/2))$ will be large.

It should also be noted that in general (if A is not a normal matrix) a matrix could have a very large condition number even if all its eigenvalues are identical!

For example, if we consider the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 2 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix},$$

it turns out that $\mathrm{cond}_2(A) \geq 2^{n-1}$.

Going back to our matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix},$$

which is a symmetric, positive, definite, matrix, it can be shown that its eigenvalues, which in this case are also its singular values because A is SPD, are

$$\lambda_1 \approx 30.2887 > \lambda_2 \approx 3.858 > \\ \lambda_3 \approx 0.8431 > \lambda_4 \approx 0.01015,$$

so that

$$\mathrm{cond}_2(A) = \frac{\lambda_1}{\lambda_4} \approx 2984.$$

The reader should check that for the perturbation of the right-hand side b used earlier, the relative errors $\|\delta x\| / \|x\|$ and $\|\delta x\| / \|x\|$ satisfy the inequality

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

and comes close to equality.