Chapter 8

Vector Norms and Matrix Norms

8.1 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 $|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}$ ($|z|$ is the *modulus* of z).

Definition 8.1. 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 $x, y, z \in E$ and $\lambda \in K$:

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

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

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

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

By (N2), setting $\lambda = -1$, we obtain

$$\|-x\| = \|(-1)x\| = |-1| \, \|x\| = \|x\| \, ;$$

that is, $\|-x\| = \|x\|$. From (N3), we have

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\| \, ,$$

which implies that

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

By exchanging x and y and using the fact that by (N2),

$$\|y - x\| = \|-(x - y)\| = \|x - y\| \, ,$$

we also have

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

Therefore,

$$|\|x\| - \|y\|| \leq \|x - y\|, \quad \text{for all } x, y \in E. \tag{$*$}$$

Observe that setting $\lambda = 0$ in (N2), we deduce that $\|0\| = 0$ without assuming (N1). Then by setting $y = 0$ in $(*)$, we obtain

$$|\|x\|| \leq \|x\| \, , \quad \text{for all } x \in E.$$

Therefore, the condition $\|x\| \geq 0$ in (N1) follows from (N2) and (N3), and (N1) can be replaced by the weaker condition

(N1') For all $x \in E$, if $\|x\| = 0$, then $x = 0$,

A function $\| \, \| \colon E \to \mathbb{R}$ satisfying Axioms (N2) and (N3) is called a *seminorm*. From the above discussion, a seminorm also has the properties

$\|x\| \geq 0$ for all $x \in E$, and $\|0\| = 0$.

However, there may be nonzero vectors $x \in E$ such that $\|x\| = 0$.

Let us give some examples of normed vector spaces.

Example 8.1.

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 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

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

See Figures 8.1 through 8.4.

There are other norms besides the ℓ^p-norms. Here are some examples.

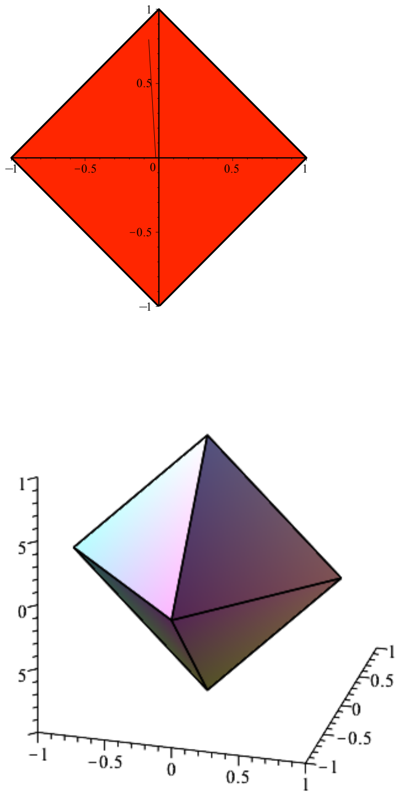

Figure 8.1: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_1 \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_1 \leq 1\}$.

1. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = |u_1| + 2|u_2|.$$
See Figure 8.5.

2. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = \left((u_1 + u_2)^2 + u_1^2\right)^{1/2}.$$
See Figure 8.6.

3. For $E = \mathbb{C}^2$,
$$\|(u_1, u_2)\| = |u_1 + iu_2| + |u_1 - iu_2|.$$

The reader should check that they satisfy all the axioms of a norm.

Some work is required to show the triangle inequality for the ℓ^p-norm.

Proposition 8.1. *If $E = \mathbb{C}^n$ or $E = \mathbb{R}^n$, for every real number $p \geq 1$, the ℓ^p-norm is indeed a norm.*

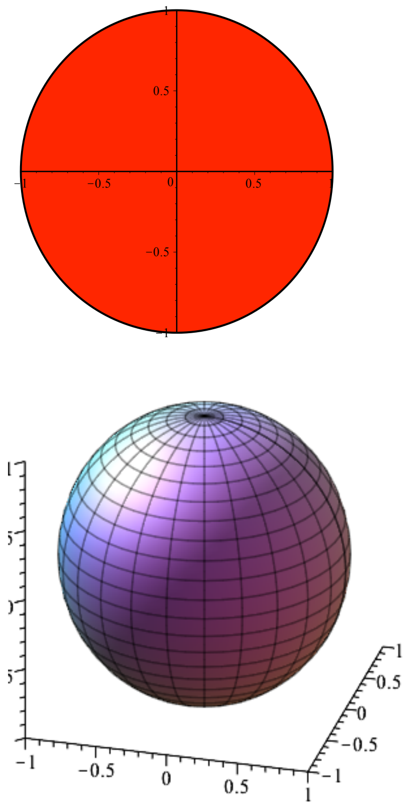

Figure 8.2: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_2 \leq 1\}$.

Proof. The cases $p = 1$ and $p = \infty$ are easy and left to the reader. If $p > 1$, then let $q > 1$ such that

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

We will make use of the following fact: for all $\alpha, \beta \in \mathbb{R}$, if $\alpha, \beta \geq 0$, then

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

To prove the above inequality, we use the fact that the exponential function $t \mapsto e^t$ satisfies the following convexity inequality:

$$e^{\theta x + (1-\theta)y} \leq \theta e^x + (1-\theta)e^y,$$

for all $x, y \in \mathbb{R}$ and all θ with $0 \leq \theta \leq 1$.

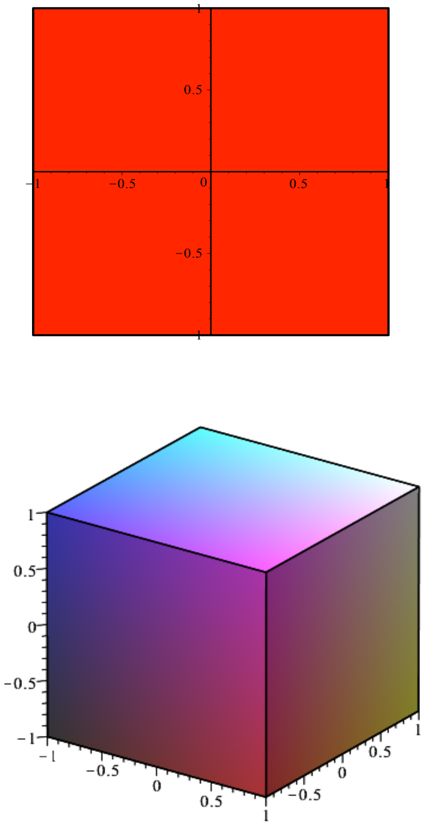

Figure 8.3: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_\infty \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_\infty \leq 1\}$.

Since the case $\alpha\beta = 0$ is trivial, let us assume that $\alpha > 0$ and $\beta > 0$. If we replace θ by $1/p$, x by $p \log \alpha$ and y by $q \log \beta$, then we get

$$e^{\frac{1}{p} p \log \alpha + \frac{1}{q} q \log \beta} \leq \frac{1}{p} e^{p \log \alpha} + \frac{1}{q} e^{q \log \beta},$$

which simplifies to

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q},$$

as claimed.

We will now prove that for any two vectors $u, v \in E$, (where E is of dimension n), we have

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

Since the above is trivial if $u = 0$ or $v = 0$, let us assume that $u \neq 0$ and $v \neq 0$. Then

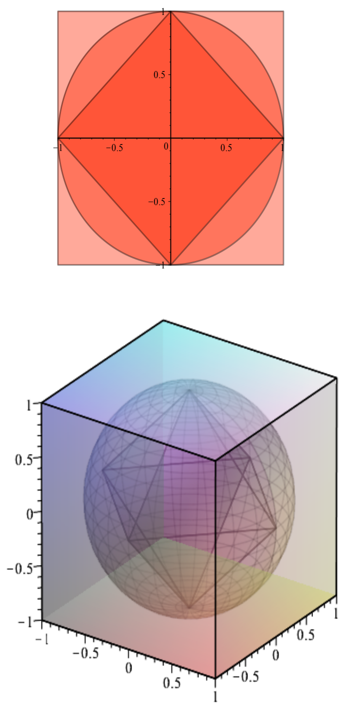

Figure 8.4: The relationships between the closed unit balls from the ℓ^1-norm, the Euclidean norm, and the sup-norm.

Inequality $(*)$ with $\alpha = |u_i| / \left\lVert u \right\rVert_p$ and $\beta = |v_i| / \left\lVert v \right\rVert_q$ yields

$$\frac{|u_i v_i|}{\left\lVert u \right\rVert_p \left\lVert v \right\rVert_q} \leq \frac{|u_i|^p}{p \left\lVert u \right\rVert_p^p} + \frac{|v_i|^q}{q \left\lVert u \right\rVert_q^q},$$

for $i = 1, \ldots, n$, and by summing up these inequalities, we get

$$\sum_{i=1}^{n} |u_i v_i| \leq \left\lVert u \right\rVert_p \left\lVert v \right\rVert_q,$$

as claimed. To finish the proof, we simply have to prove that property (N3) holds, since (N1) and (N2) are clear. For $i = 1, \ldots, n$, we can write

$$(|u_i| + |v_i|)^p = |u_i|(|u_i| + |v_i|)^{p-1} + |v_i|(|u_i| + |v_i|)^{p-1},$$

so that by summing up these equations we get

$$\sum_{i=1}^{n} (|u_i| + |v_i|)^p = \sum_{i=1}^{n} |u_i|(|u_i| + |v_i|)^{p-1} + \sum_{i=1}^{n} |v_i|(|u_i| + |v_i|)^{p-1},$$

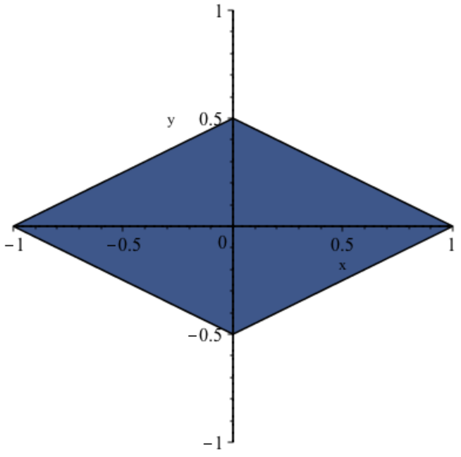

Figure 8.5: The unit closed unit ball $\{(u_1, u_2) \in \mathbb{R}^2 \mid \|(u_1, u_2)\| \leq 1\}$, where $\|(u_1, u_2)\| = |u_1| + 2|u_2|$.

and using Inequality $(**)$, with $V \in E$ where $V_i = (|u_i| + |v_i|)^{p-1}$, we get

$$\begin{aligned}\sum_{i=1}^n (|u_i| + |v_i|)^p &\leq \|u\|_p \|V\|_q + \|v\|_p \|V\|_q \\ &= (\|u\|_p + \|v\|_p)\left(\sum_{i=1}^n (|u_i| + |v_i|)^{(p-1)q}\right)^{1/q}.\end{aligned}$$

However, $1/p + 1/q = 1$ implies $pq = p + q$, that is, $(p-1)q = p$, so we have

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

which yields

$$\left(\sum_{i=1}^n (|u_i| + |v_i|)^p\right)^{1-1/q} = \left(\sum_{i=1}^n (|u_i| + |v_i|)^p\right)^{1/p} \leq \|u\|_p + \|v\|_p.$$

Since $|u_i + v_i| \leq |u_i| + |v_i|$, the above implies the triangle inequality $\|u + v\|_p \leq \|u\|_p + \|v\|_p$, as claimed. □

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*.

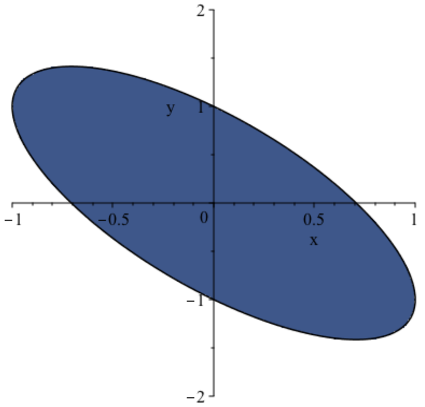

Figure 8.6: The unit closed unit ball $\{(u_1, u_2) \in \mathbb{R}^2 \mid \|(u_1, u_2)\| \leq 1\}$, where $\|(u_1, u_2)\| = \big((u_1 + u_2)^2 + u_1^2\big)^{1/2}$.

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 following inequalities.

Corollary 8.2. *(Hölder's inequalities) For any real numbers p, q, such that $p, q \geq 1$ and*

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

(with $q = +\infty$ if $p = 1$ and $p = +\infty$ if $q = 1$), we have the inequalities

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

and

$$|\langle u, v \rangle| \leq \|u\|_p \|v\|_q, \qquad u, v \in \mathbb{C}^n.$$

For $p = 2$, this 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.

Remark: If $0 < p < 1$, then $x \mapsto \|x\|_p$ is not a norm because the triangle inequality *fails*. For example, consider $x = (2, 0)$ and $y = (0, 2)$. Then $x + y = (2, 2)$, and we have $\|x\|_p = (2^p + 0^p)^{1/p} = 2$, $\|y\|_p = (0^p + 2^p)^{1/p} = 2$, and $\|x + y\|_p = (2^p + 2^p)^{1/p} = 2^{(p+1)/p}$. Thus

$$\|x + y\|_p = 2^{(p+1)/p}, \quad \|x\|_p + \|y\|_p = 4 = 2^2.$$

Since $0 < p < 1$, we have $2p < p + 1$, that is, $(p + 1)/p > 2$, so $2^{(p+1)/p} > 2^2 = 4$, and the triangle inequality $\|x + y\|_p \leq \|x\|_p + \|y\|_p$ fails.

Observe that

$$\|(1/2)x\|_p = (1/2)\|x\|_p = \|(1/2)y\|_p = (1/2)\|y\|_p = 1, \qquad \|(1/2)(x + y)\|_p = 2^{1/p},$$

and since $p < 1$, we have $2^{1/p} > 2$, so

$$\|(1/2)(x + y)\|_p = 2^{1/p} > 2 = (1/2)\|x\|_p + (1/2)\|y\|_p,$$

and the map $x \mapsto \|x\|_p$ is not convex.

For $p = 0$, for any $x \in \mathbb{R}^n$, we have

$$\|x\|_0 = |\{i \in \{1, \ldots, n\} \mid x_i \neq 0\}|,$$

the number of nonzero components of x. The map $x \mapsto \|x\|_0$ is not a norm this time because Axiom (N2) fails. For example,

$$\|(1,0)\|_0 = \|(10,0)\|_0 = 1 \neq 10 = 10\,\|(1,0)\|_0\,.$$

The map $x \mapsto \|x\|_0$ is also not convex. For example,

$$\|(1/2)(2,2)\|_0 = \|(1,1)\|_0 = 2,$$

and

$$\|(2,0)\|_0 = \|(0,2)\|_0 = 1,$$

but

$$\|(1/2)(2,2)\|_0 = 2 > 1 = (1/2)\,\|(2,0)\|_0 + (1/2)\,\|(0,2)\|_0\,.$$

Nevertheless, the “zero-norm” $x \mapsto \|x\|_0$ is used in machine learning as a regularizing term which encourages sparsity, namely increases the number of zero components of the vector x.

The following proposition is easy to show.

Proposition 8.3. *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 8.3 is actually a special case of a very important result: *in a finite-dimensional vector space, any two norms are equivalent.*

Definition 8.2. 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

$$\begin{aligned} \|x\| = \|x_1e_1 + \cdots + x_ne_n\| &\leq |x_1|\,\|e_1\| + \cdots + |x_n|\,\|e_n\| \\ &\leq C(|x_1| + \cdots + |x_n|) = C\,\|x\|_1\,, \end{aligned}$$

with $C = \max_{1 \leq i \leq n} \|e_i\|$ and with the norm $\|x\|_1$ defined as

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

The above implies that

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

and this implies the following corollary.

Corollary 8.4. *For any norm $u \mapsto \|u\|$ on a finite-dimensional (complex or real) vector space E, the map $u \mapsto \|u\|$ is continuous with respect to the norm $\| \|_1$.*

Let S_1^{n-1} be the unit sphere 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 8.5. *If E is any real or complex vector space of finite dimension, then any two norms on E are equivalent.*

Proof. It is enough to prove that any norm $\| \|$ is equivalent to the 1-norm. We already proved that the function $x \mapsto \|x\|$ is continuous with respect to the norm $\| \|_1$, and we observed that the unit sphere S_1^{n-1} is compact. Now we just recalled that because the function $f \colon x \mapsto \|x\|$ is continuous and because S_1^{n-1} is compact, the function f has a minimum m and a maximum M, and because $\|x\|$ is never zero on S_1^{n-1}, we must have $m > 0$. Consequently, we just proved that if $\|x\|_1 = 1$, then

$$0 < m \leq \|x\| \leq M,$$

so for any $x \in E$ with $x \neq 0$, we get

$$m \leq \|x/ \|x\|_1\| \leq M,$$

which implies

$$m \|x\|_1 \leq \|x\| \leq M \|x\|_1 .$$

Since the above inequality holds trivially if $x = 0$, we just proved that $\| \|$ and $\| \|_1$ are equivalent, as claimed. □

Remark: Let P be a $n \times n$ symmetric positive definite matrix. It is immediately verified that the map $x \mapsto \|x\|_P$ given by

$$\|x\|_P = (x^\top P x)^{1/2}$$

is a norm on $\mathbb{R}^n$ called a *quadratic norm*. Using some convex analysis (the Löwner–John ellipsoid), it can be shown that *any* norm $\| \|$ on $\mathbb{R}^n$ can be approximated by a quadratic norm in the sense that there is a quadratic norm $\| \|_P$ such that

$$\|x\|_P \leq \|x\| \leq \sqrt{n} \|x\|_P \qquad \text{for all } x \in \mathbb{R}^n;$$

see Boyd and Vandenberghe [11], Section 8.4.1.

Next we will consider norms on matrices.

8.2 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 8.3. 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 called *submultiplicativity* that

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

for all $A, B \in \mathrm{M}_n(K)$. A norm on matrices satisfying the above property is often called a *submultiplicative* matrix norm.

Since $I^2 = I$, from $\|I\| = \|I^2\| \leq \|I\|^2$, 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

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

We also review eigenvalues and eigenvectors. We content ourselves with definition involving matrices. A more general treatment will be given later on (see Chapter 14).

Definition 8.4. 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 8.4 *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. The fact that eigenvectors are nonzero is implicitly used in all the arguments involving them, so it seems safer (but perhaps not as elegant) to stipulate that eigenvectors should be nonzero.

If A is a square real matrix $A \in \mathrm{M}_n(\mathbb{R})$, then we restrict Definition 8.4 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 6.11, $\lambda I - A$ 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 8.5. 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.

Since the eigenvalue $\lambda_1, \ldots, \lambda_n$ of A are the zeros of the polynomial

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

we deduce (see Section 14.1 for details) that

$$\begin{aligned} \mathrm{tr}(A) &= \lambda_1 + \cdots + \lambda_n \\ \det(A) &= \lambda_1 \cdots \lambda_n. \end{aligned}$$

Proposition 8.6. *For any matrix norm $\| \, \|$ on $\mathrm{M}_n(\mathbb{C})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, we have*

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

Proof. Let λ be some eigenvalue of A for which $|\lambda|$ is maximum, that is, such that $|\lambda| = \rho(A)$. If u ($\neq 0$) is any eigenvector associated with λ and if U is the $n \times n$ matrix whose columns are all u, then $Au = \lambda u$ implies

$$AU = \lambda U,$$

and since

$$|\lambda| \, \|U\| = \|\lambda U\| = \|AU\| \leq \|A\| \, \|U\|$$

and $U \neq 0$, we have $\|U\| \neq 0$, and get

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

as claimed. □

Proposition 8.6 also holds for any real matrix norm $\| \, \|$ on $\mathrm{M}_n(\mathbb{R})$ but the proof is more subtle and requires the notion of induced norm. We prove it after giving Definition 8.7.

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

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

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. See Chapter 16 for the proof of these results.

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 $A = (a_{ij})$,

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

Definition 8.6. 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 8.7. *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* $A \in \mathrm{M}_n(\mathbb{C})$.

Proof. (1) The only property that requires a proof is the fact $\|AB\|_F \leq \|A\|_F \|B\|_F$. This follows from the Cauchy–Schwarz inequality:

$$\begin{aligned}\|AB\|_F^2 &= \sum_{i,j=1}^n \left|\sum_{k=1}^n a_{ik}b_{kj}\right|^2 \\ &\leq \sum_{i,j=1}^n \left(\sum_{h=1}^n |a_{ih}|^2\right)\left(\sum_{k=1}^n |b_{kj}|^2\right) \\ &= \left(\sum_{i,h=1}^n |a_{ih}|^2\right)\left(\sum_{k,j=1}^n |b_{kj}|^2\right) = \|A\|_F^2 \|B\|_F^2\,.\end{aligned}$$

(2) We have

$$\|A\|_F^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(VV^*A^*A) = \mathrm{tr}(V^*A^*AV) = \|AV\|_F^2\,,$$

and

$$\|A\|_F^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(A^*U^*UA) = \|UA\|_F^2\,.$$

The identity

$$\|A\|_F = \|UAV\|_F$$

follows from the previous two.

(3) It is well known that the trace of a matrix is equal to the sum of its eigenvalues. Furthermore, A^*A is symmetric positive semidefinite (which means that its eigenvalues are nonnegative), so $\rho(A^*A)$ is the largest eigenvalue of A^*A and

$$\rho(A^*A) \leq \mathrm{tr}(A^*A) \leq n\rho(A^*A),$$

which yields (3) by taking square roots. □

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!

8.3 Subordinate Norms

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 8.8. *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 $C_A \geq 0$, 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).

Proof. For every basis $(e_1, \ldots, e_n)$ of $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every vector $u = u_1 e_1 + \cdots + u_n e_n$, we have

$$\begin{aligned} \|Au\| &= \|u_1 A(e_1) + \cdots + u_n A(e_n)\| \\ &\leq |u_1| \, \|A(e_1)\| + \cdots + |u_n| \, \|A(e_n)\| \\ &\leq C_1(|u_1| + \cdots + |u_n|) = C_1 \|u\|_1, \end{aligned}$$

where $C_1 = \max_{1 \leq i \leq n} \|A(e_i)\|$. By Theorem 8.5, the norms $\| \ \|$ and $\| \ \|_1$ are equivalent, so there is some constant $C_2 > 0$ so that $\|u\|_1 \leq C_2 \|u\|$ for all u, which implies that

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

where $C_A = C_1 C_2$. □

Proposition 8.8 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 8.8 implies that for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$),

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

Since $\|\lambda u\| = |\lambda| \, \|u\|$, for every nonzero vector x, we have

$$\frac{\|Ax\|}{\|x\|} = \frac{\|x\| \, \|A(x/\|x\|)\|}{\|x\|} = \|A(x/\|x\|)\|,$$

which implies 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

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

The above considerations justify the following definition.

Definition 8.7. If $\|\ \|$ is any norm on $\mathbb{C}^n$, we define the function $\|\ \|_{\mathrm{op}}$ on $\mathrm{M}_n(\mathbb{C})$ by

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

The function $A \mapsto \|A\|_{\mathrm{op}}$ is called the *subordinate matrix norm* or *operator norm* induced by the norm $\|\ \|$.

Another notation for the operator norm of a matrix A (in particular, used by Horn and Johnson [36]), is $|\!|\!|A|\!|\!|$.

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

$$\|Ax\| \leq \|A\|_{\mathrm{op}} \|x\|, \quad \text{for all } x \in \mathbb{C}^n.$$

A norm $\|\ \|_{\mathrm{op}}$ on $\mathrm{M}_n(\mathbb{C})$ satisfying the above property is said to be *subordinate* to the vector norm $\|\ \|$ on $\mathbb{C}^n$. As a consequence of the above inequality, we have

$$\|ABx\| \leq \|A\|_{\mathrm{op}} \|Bx\| \leq \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \|x\|,$$

for all $x \in \mathbb{C}^n$, which implies that

$$\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \quad \text{for all } A, B \in \mathrm{M}_n(\mathbb{C}),$$

showing that $A \mapsto \|A\|_{\mathrm{op}}$ is a matrix norm (it is submultiplicative).

Observe that the operator norm is also defined by

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

Since the function $x \mapsto \|Ax\|$ is continuous (because $|\,\|Ay\| - \|Ax\|\,| \leq \|Ay - Ax\| \leq C_A \|x - y\|$) and the unit sphere $S^{n-1} = \{x \in \mathbb{C}^n \mid \|x\| = 1\}$ is compact, there is some $x \in \mathbb{C}^n$ such that $\|x\| = 1$ and

$$\|Ax\| = \|A\|_{\mathrm{op}}.$$

Equivalently, there is some $x \in \mathbb{C}^n$ such that $x \neq 0$ and

$$\|Ax\| = \|A\|_{\mathrm{op}} \|x\|.$$

The definition of an operator norm also implies that

$$\|I\|_{\mathrm{op}} = 1.$$

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

If $\| \;\|$ is a vector norm on $\mathbb{C}^n$, the operator norm $\| \;\|_{\mathrm{op}}$ that it induces applies to matrices in $\mathrm{M}_n(\mathbb{C})$. If we are careful to denote vectors and matrices so that no confusion arises, for example, by using lower case letters for vectors and upper case letters for matrices, it should be clear that $\|A\|_{\mathrm{op}}$ is the operator norm of the matrix A and that $\|x\|$ is the vector norm of x. Consequently, following common practice to alleviate notation, we will drop the subscript "op" and simply write $\|A\|$ instead of $\|A\|_{\mathrm{op}}$.

The notion of subordinate norm can be slightly generalized.

Definition 8.8. If $K = \mathbb{R}$ or $K = \mathbb{C}$, for any norm $\| \;\|$ on $\mathrm{M}_{m,n}(K)$, and for any two norms $\| \;\|_a$ on K^n and $\| \;\|_b$ on K^m, we say that the norm $\| \;\|$ is *subordinate* to the norms $\| \;\|_a$ and $\| \;\|_b$ if

$$\|Ax\|_b \leq \|A\| \, \|x\|_a \quad \text{for all } A \in \mathrm{M}_{m,n}(K) \text{ and all } x \in K^n.$$

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 $\| \;\|_\infty$.

We now prove Proposition 8.6 for real matrix norms.

Proposition 8.9. *For any matrix norm $\| \;\|$ on $\mathrm{M}_n(\mathbb{R})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have*

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

Proof. We follow the proof in Denis Serre's book [57]. If A is a real matrix, the problem is that the eigenvectors associated with the eigenvalue of maximum modulus may be complex. We use a trick based on the fact that for every matrix A (real or complex),

$$\rho(A^k) = (\rho(A))^k,$$

which is left as an exercise (use Proposition 14.7 which shows that if $(\lambda_1, \ldots, \lambda_n)$ are the (not necessarily distinct) eigenvalues of A, then $(\lambda_1^k, \ldots, \lambda_n^k)$ are the eigenvalues of A^k, for $k \geq 1$).

Pick any complex matrix norm $\| \, \|_c$ on $\mathbb{C}^n$ (for example, the Frobenius norm, or any subordinate matrix norm induced by a norm on $\mathbb{C}^n$). The restriction of $\| \, \|_c$ to real matrices is a real norm that we also denote by $\| \, \|_c$. Now by Theorem 8.5, since $\mathrm{M}_n(\mathbb{R})$ has finite dimension n^2, there is some constant $C > 0$ so that

$$\|B\|_c \leq C \|B\|, \quad \text{for all} \quad B \in \mathrm{M}_n(\mathbb{R}).$$

Furthermore, for every $k \geq 1$ and for every real $n \times n$ matrix A, by Proposition 8.6, $\rho(A^k) \leq \left\|A^k\right\|_c$, and because $\| \, \|$ is a matrix norm, $\left\|A^k\right\| \leq \|A\|^k$, so we have

$$(\rho(A))^k = \rho(A^k) \leq \left\|A^k\right\|_c \leq C \left\|A^k\right\| \leq C \|A\|^k,$$

for all $k \geq 1$. It follows that

$$\rho(A) \leq C^{1/k} \|A\|, \quad \text{for all} \quad k \geq 1.$$

However because $C > 0$, we have $\lim_{k \mapsto \infty} C^{1/k} = 1$ (we have $\lim_{k \mapsto \infty} \frac{1}{k} \log(C) = 0$). Therefore, we conclude that

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

as desired. □

We now determine explicitly what are the subordinate matrix norms associated with the vector norms $\| \, \|_1$, $\| \, \|_2$, and $\| \, \|_\infty$.

Proposition 8.10. *For every square matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\begin{aligned}
\|A\|_1 &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_1 = 1}} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}| \\
\|A\|_\infty &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_\infty = 1}} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \\
\|A\|_2 &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_2 = 1}} \|Ax\|_2 = \sqrt{\rho(A^* A)} = \sqrt{\rho(AA^*)}.
\end{aligned}$$

Note that $\|A\|_1$ is the maximum of the ℓ^1-norms of the columns of A and $\|A\|_\infty$ is the maximum of the ℓ^1-norms of the rows of A. 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 $\|A\|_2 = \rho(A)$.

Proof. For every vector u, we have

$$\|Au\|_1 = \sum_i \left| \sum_j a_{ij} u_j \right| \leq \sum_j |u_j| \sum_i |a_{ij}| \leq \left(\max_j \sum_i |a_{ij}| \right) \|u\|_1 ,$$

which implies that

$$\|A\|_1 \leq \max_j \sum_{i=1}^{n} |a_{ij}|.$$

It remains to show that equality can be achieved. For this let j_0 be some index such that

$$\max_j \sum_i |a_{ij}| = \sum_i |a_{ij_0}|,$$

and let $u_i = 0$ for all $i \neq j_0$ and $u_{j_0} = 1$.

In a similar way, we have

$$\|Au\|_\infty = \max_i \left| \sum_j a_{ij} u_j \right| \leq \left(\max_i \sum_j |a_{ij}| \right) \|u\|_\infty ,$$

which implies that

$$\|A\|_\infty \leq \max_i \sum_{j=1}^{n} |a_{ij}|.$$

To achieve equality, let i_0 be some index such that

$$\max_i \sum_j |a_{ij}| = \sum_j |a_{i_0 j}|.$$

The reader should check that the vector given by

$$u_j = \begin{cases} \frac{\overline{a}_{i_0 j}}{|a_{i_0 j}|} & \text{if } a_{i_0 j} \neq 0 \\ 1 & \text{if } a_{i_0 j} = 0 \end{cases}$$

works.

We have

$$\|A\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} \|Ax\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} x^* A^* A x.$$

Since the matrix A^*A is symmetric, it has real eigenvalues and it can be diagonalized with respect to a unitary matrix. These facts can be used to prove that the function $x \mapsto x^* A^* A x$ has a maximum on the sphere $x^* x = 1$ equal to the largest eigenvalue of A^*A, namely, $\rho(A^*A)$. We postpone the proof until we discuss optimizing quadratic functions. Therefore,

$$\|A\|_2 = \sqrt{\rho(A^*A)}.$$

Let use now prove that $\rho(A^*A) = \rho(AA^*)$. First assume that $\rho(A^*A) > 0$. In this case, there is some eigenvector $u\,(\neq 0)$ such that

$$A^*Au = \rho(A^*A)u,$$

and since $\rho(A^*A) > 0$, we must have $Au \neq 0$. Since $Au \neq 0$,

$$AA^*(Au) = A(A^*Au) = \rho(A^*A)Au$$

which means that $\rho(A^*A)$ is an eigenvalue of AA^*, and thus

$$\rho(A^*A) \leq \rho(AA^*).$$

Because $(A^*)^* = A$, by replacing A by A^*, we get

$$\rho(AA^*) \leq \rho(A^*A),$$

and so $\rho(A^*A) = \rho(AA^*)$.

If $\rho(A^*A) = 0$, then we must have $\rho(AA^*) = 0$, since otherwise by the previous reasoning we would have $\rho(A^*A) = \rho(AA^*) > 0$. Hence, in all case

$$\|A\|_2^2 = \rho(A^*A) = \rho(AA^*) = \|A^*\|_2^2.$$

For any unitary matrices U and V, it is an easy exercise to prove that V^*A^*AV and A^*A have the same eigenvalues, so

$$\|A\|_2^2 = \rho(A^*A) = \rho(V^*A^*AV) = \|AV\|_2^2,$$

and also

$$\|A\|_2^2 = \rho(A^*A) = \rho(A^*U^*UA) = \|UA\|_2^2.$$

Finally, if A is a normal matrix ($AA^* = A^*A$), it can be shown that there is some unitary matrix U so that

$$A = UDU^*,$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix consisting of the eigenvalues of A, and thus

$$A^*A = (UDU^*)^*UDU^* = UD^*U^*UDU^* = UD^*DU^*.$$

However, $D^*D = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$, which proves that

$$\rho(A^*A) = \rho(D^*D) = \max_i |\lambda_i|^2 = (\rho(A))^2,$$

so that $\|A\|_2 = \rho(A)$. □

Definition 8.9. For $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, the norm $\|A\|_2 =$ is often called the *spectral norm*.

Observe that Property (3) of Proposition 8.7 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 spectral norm.

The reader will check that the above proof still holds if the matrix A is real (change unitary to orthogonal), 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* $m \times n$ matrices, with the same formulae. Similarly, the Frobenius norm given by

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

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

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

Remark: It can be shown that for any two real numbers $p, q \geq 1$ such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$, we have

$$\begin{aligned}\|A^*\|_q = \|A\|_p &= \sup\{\Re(y^*Ax) \mid \|x\|_p = 1, \|y\|_q = 1\}\\ &= \sup\{|\langle Ax, y\rangle| \mid \|x\|_p = 1, \|y\|_q = 1\},\end{aligned}$$

where $\|A^*\|_q$ and $\|A\|_p$ are the operator norms.

Remark: Let $(E, \|\ \|)$ and $(F, \|\ \|)$ be two normed vector spaces (for simplicity of notation, we use the same symbol $\|\ \|$ for the norms on E and F; this should not cause any confusion). Recall that a function $f\colon E \to F$ is *continuous* if for every $a \in E$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for all $x \in E$,

$$\text{if} \quad \|x - a\| \leq \eta \quad \text{then} \quad \|f(x) - f(a)\| \leq \epsilon.$$

It is not hard to show that a *linear map* $f\colon E \to F$ is continuous iff there is some constant $C \geq 0$ such that

$$\|f(x)\| \leq C\,\|x\| \text{ for all } x \in E.$$

If so, we say that f is *bounded* (or a *linear bounded operator*). We let $\mathcal{L}(E; F)$ denote the set of all continuous (equivalently, bounded) linear maps from E to F. Then we can define the *operator norm* (or *subordinate norm*) $\|\ \|$ on $\mathcal{L}(E; F)$ as follows: for every $f \in \mathcal{L}(E; F)$,

$$\|f\| = \sup_{\substack{x \in E\\ x \neq 0}} \frac{\|f(x)\|}{\|x\|} = \sup_{\substack{x \in E\\ \|x\| = 1}} \|f(x)\|\,,$$

or equivalently by

$$\|f\| = \inf\{\lambda \in \mathbb{R} \mid \|f(x)\| \leq \lambda \|x\|, \text{ for all } x \in E\}.$$

It is not hard to show that the map $f \mapsto \|f\|$ is a norm on $\mathcal{L}(E; F)$ satisfying the property

$$\|f(x)\| \leq \|f\| \|x\|$$

for all $x \in E$, and that if $f \in \mathcal{L}(E; F)$ and $g \in \mathcal{L}(F; G)$, then

$$\|g \circ f\| \leq \|g\| \|f\|.$$

Operator norms play an important role in functional analysis, especially when the spaces E and F are *complete*.

8.4 Inequalities Involving Subordinate Norms

In this section we discuss two technical inequalities which will be needed for certain proofs in the last three sections of this chapter. First we prove a proposition which will be needed when we deal with the condition number of a matrix.

Proposition 8.11. *Let $\| \|$ be any matrix norm, and let $B \in \mathrm{M}_n(\mathbb{C})$ 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).

Proof. (1) Observe that $(I + B)u = 0$ implies $Bu = -u$, so

$$\|u\| = \|Bu\|.$$

Recall that

$$\|Bu\| \leq \|B\| \|u\|$$

for every subordinate norm. Since $\|B\| < 1$, if $u \neq 0$, then

$$\|Bu\| < \|u\|,$$

which contradicts $\|u\| = \|Bu\|$. Therefore, we must have $u = 0$, which proves that $I + B$ is injective, and thus bijective, i.e., invertible. Then we have

$$(I + B)^{-1} + B(I + B)^{-1} = (I + B)(I + B)^{-1} = I,$$

so we get

$$(I+B)^{-1} = I - B(I+B)^{-1},$$

which yields

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

and finally,

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

(2) If $I + B$ is singular, then -1 is an eigenvalue of B, and by Proposition 8.6, we get $\rho(B) \leq \|B\|$, which implies $1 \leq \rho(B) \leq \|B\|$. □

The second inequality is a result is that is needed to deal with the convergence of sequences of powers of matrices.

Proposition 8.12. *For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \ \|$ such that*

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

Proof. By Theorem 14.5, there exists some invertible matrix U and some upper triangular matrix T such that

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

and say that

$$T = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. For every $\delta \neq 0$, define the diagonal matrix

$$D_\delta = \mathrm{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1}),$$

and consider the matrix

$$(UD_\delta)^{-1} A (UD_\delta) = D_\delta^{-1} T D_\delta = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1} t_{1n} \\ 0 & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2} t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \delta t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Now define the function $\| \ \| \colon \mathrm{M}_n(\mathbb{C}) \to \mathbb{R}$ by

$$\|B\| = \left\|(UD_\delta)^{-1} B (UD_\delta)\right\|_\infty,$$

for every $B \in \mathrm{M}_n(\mathbb{C})$. Then it is easy to verify that the above function is the matrix norm subordinate to the vector norm

$$v \mapsto \left\| (UD_\delta)^{-1} v \right\|_\infty.$$

Furthermore, for every $\epsilon > 0$, we can pick δ so that

$$\sum_{j=i+1}^{n} |\delta^{j-i} t_{ij}| \leq \epsilon, \quad 1 \leq i \leq n-1,$$

and by definition of the norm $\| \ \|_\infty$, we get

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

which shows that the norm that we have constructed satisfies the required properties. □

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 $A \neq 0$.

8.5 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 as $b + \Delta b$, where

$$\Delta b = \begin{pmatrix} 0.1 \\ -0.1 \\ 0.1 \\ -0.1 \end{pmatrix},$$

we obtain 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 solution turns out to be $x + \Delta x = (9.2, -12.6, 4.5, -1.1)$, where

$$\Delta x = (9.2, -12.6, 4.5, -1.1) - (1, 1, 1, 1) = (8.2, -13.6, 3.5, -2.1).$$

Then a relative error of the data in terms of the one-norm,

$$\frac{\|\Delta b\|_1}{\|b\|_1} = \frac{0.4}{119} = \frac{4}{1190} \approx \frac{1}{300},$$

produces a relative error in the input

$$\frac{\|\Delta x\|_1}{\|x\|_1} = \frac{27.4}{4} \approx 7.$$

So a relative order of the order $1/300$ in the data produces a relative error of the order $7/1$ in the solution, which represents an amplification of the relative error of the order 2100.

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 $x + \Delta x = (-81, 137, -34, 22)$. 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:

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

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

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

The second equation yields $Ax + A\Delta x + \Delta A(x + \Delta x) = b$, and by subtracting the first equation we get

$$\Delta x = -A^{-1}\Delta A(x + \Delta x).$$

It follows that

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

which can be rewritten as

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

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* $\|A\| \left\|A^{-1}\right\|$. In fact, this factor turns out to be optimal and this suggests the following definition:

Definition 8.10. For any subordinate matrix norm $\| \, \|$, for any invertible matrix A, the number

$$\operatorname{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

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

The condition number $\operatorname{cond}(A)$ 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 8.13. *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.

Proof. We already proved the inequality. Now, because $\| \ \|$ is a subordinate matrix norm, there exist some vectors $x \neq 0$ and $\Delta b \neq 0$ for which

$$\left\|A^{-1}\Delta b\right\| = \left\|A^{-1}\right\| \|\Delta b\| \quad \text{and} \quad \|Ax\| = \|A\| \, \|x\| \, .$$

□

Proposition 8.14. *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

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

Proof. The first inequality has already been proven. To show that equality can be achieved, let w be any vector such that $w \neq 0$ and

$$\left\|A^{-1}w\right\| = \left\|A^{-1}\right\| \|w\| \, ,$$

and let $\beta \neq 0$ be any real number. Now the vectors

$$\begin{aligned} \Delta x &= -\beta A^{-1}w \\ x + \Delta x &= w \\ b &= (A + \beta I)w \end{aligned}$$

and the matrix

$$\Delta A = \beta I$$

sastisfy the equations

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b \\ \|\Delta x\| = |\beta| \left\|A^{-1}w\right\| &= \|\Delta A\| \left\|A^{-1}\right\| \|x + \Delta x\|. \end{aligned}$$

Finally we can pick β so that $-\beta$ is not equal to any of the eigenvalues of A, so that $A + \Delta A = A + \beta I$ is invertible and b is is nonzero.

If $\|\Delta A\| < 1/\left\|A^{-1}\right\|$, then

$$\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| < 1,$$

so by Proposition 8.11, the matrix $I + A^{-1}\Delta A$ is invertible and

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

Recall that we proved earlier that

$$\Delta x = -A^{-1}\Delta A(x + \Delta x),$$

and by adding x to both sides and moving the right-hand side to the left-hand side yields

$$(I + A^{-1}\Delta A)(x + \Delta x) = x,$$

and thus

$$x + \Delta x = (I + A^{-1}\Delta A)^{-1}x,$$

which yields

$$\begin{aligned} \Delta x &= ((I + A^{-1}\Delta A)^{-1} - I)x = (I + A^{-1}\Delta A)^{-1}(I - (I + A^{-1}\Delta A))x \\ &= -(I + A^{-1}\Delta A)^{-1}A^{-1}(\Delta A)x. \end{aligned}$$

From this and

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

we get

$$\|\Delta x\| \leq \frac{\|A^{-1}\| \|\Delta A\|}{1 - \|A^{-1}\| \|\Delta A\|} \|x\|,$$

which can be written as

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

which is the kind of inequality that we were seeking. □

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 $\|\Delta A\| < 1/\|A^{-1}\|$ (and $b \neq 0$), then

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

see Demmel [16], Section 2.2 and Horn and Johnson [36], Section 5.8.

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 (see Section 20.2) 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

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

Definition 8.11. Given a complex $n \times n$ matrix A, a triple (U, V, Σ) such that $A = V\Sigma U^\top$, where U and V are $n \times n$ unitary matrices and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix of real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, is called a *singular decomposition* (for short *SVD*) of A. If A is a real matrix, then U and V are orthogonal matrices The nonnegative numbers $\sigma_1, \ldots, \sigma_n$ are called the *singular values* of A.

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^*.

Since σ_1^2 is the largest eigenvalue of A^*A (and AA^*), note that $\sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \sigma_1$.

Corollary 8.15. *The spectral norm $\|A\|_2$ of a matrix A is equal to the largest singular value of A. Equivalently, the spectral norm $\|A\|_2$ of a matrix A is equal to the ℓ^∞-norm of its vector of singular values,*

$$\|A\|_2 = \max_{1 \leq i \leq n} \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_\infty.$$

Since the Frobenius norm of a matrix A is defined by $\|A\|_F = \sqrt{\mathrm{tr}(A^*A)}$ and since

$$\mathrm{tr}(A^*A) = \sigma_1^2 + \cdots + \sigma_n^2$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of A^*A, we see that

$$\|A\|_F = (\sigma_1^2 + \cdots + \sigma_n^2)^{1/2} = \|(\sigma_1, \ldots, \sigma_n)\|_2.$$

Corollary 8.16. *The Frobenius norm of a matrix is given by the ℓ^2-norm of its vector of singular values;* $\|A\|_F = \|(\sigma_1, \dots, \sigma_n)\|_2$.

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

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

Proposition 8.17. *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 $\lambda_1, \dots, \lambda_n$ 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.

Proof. The properties in (1) are immediate consequences of the properties of subordinate matrix norms. In particular, $AA^{-1} = I$ implies

$$1 = \|I\| \leq \|A\| \left\|A^{-1}\right\| = \mathrm{cond}(A).$$

(2) We showed earlier that $\|A\|_2^2 = \rho(A^*A)$, which is the square of the modulus of the largest eigenvalue of A^*A. Since we just saw that the eigenvalues of A^*A are $\sigma_1^2 \geq \cdots \geq \sigma_n^2$, where $\sigma_1, \dots, \sigma_n$ are the singular values of A, we have

$$\|A\|_2 = \sigma_1.$$

Now if A is invertible, then $\sigma_1 \geq \cdots \geq \sigma_n > 0$, and it is easy to show that the eigenvalues of $(A^*A)^{-1}$ are $\sigma_n^{-2} \geq \cdots \geq \sigma_1^{-2}$, which shows that

$$\left\|A^{-1}\right\|_2 = \sigma_n^{-1},$$

and thus

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

(3) This follows from the fact that $\|A\|_2 = \rho(A)$ for a normal matrix.

(4) If A is a unitary matrix, then $A^*A = AA^* = I$, so $\rho(A^*A) = 1$, and $\|A\|_2 = \sqrt{\rho(A^*A)} = 1$. We also have $\|A^{-1}\|_2 = \|A^*\|_2 = \sqrt{\rho(AA^*)} = 1$, and thus $\mathrm{cond}(A) = 1$.

(5) This follows immediately from the unitary invariance of the spectral norm. □

Proposition 8.17 (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 $\mathrm{cond}_2(A) = \|A\|_2 \, \|A^{-1}\|_2$, 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. For more details, see Horn and Johnson [36] (Section 5.8 and Section 7.4).

It should 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

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

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

A classical example of matrix with a very large condition number is the *Hilbert matrix* $H^{(n)}$, the $n \times n$ matrix with

$$H_{ij}^{(n)} = \left(\frac{1}{i+j-1}\right).$$

For example, when $n = 5$,

$$H^{(5)} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix}.$$

It can be shown that

$$\mathrm{cond}_2(H^{(5)}) \approx 4.77 \times 10^5.$$

Hilbert introduced these matrices in 1894 while studying a problem in approximation theory. The Hilbert matrix $H^{(n)}$ is symmetric positive definite. A closed-form formula can be given for its determinant (it is a special form of the so-called *Cauchy determinant*); see Problem 8.15. The inverse of $H^{(n)}$ can also be computed explicitly; see Problem 8.15. It can be shown that

$$\mathrm{cond}_2(H^{(n)}) = O((1+\sqrt{2})^{4n}/\sqrt{n}).$$

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.

8.6 An Application of Norms: Solving Inconsistent Linear Systems

The problem of solving an inconsistent linear system $Ax = b$ often arises in practice. This is a system where b does not belong to the column space of A, usually with more equations than variables. Thus, such a system has no solution. Yet we would still like to "solve" such a system, at least approximately.

Such systems often arise when trying to fit some data. For example, we may have a set of 3D data points

$$\{p_1, \ldots, p_n\},$$

and we have reason to believe that these points are nearly coplanar. We would like to find a plane that best fits our data points. Recall that the equation of a plane is

$$\alpha x + \beta y + \gamma z + \delta = 0,$$

with $(\alpha, \beta, \gamma) \neq (0, 0, 0)$. Thus, every plane is either not parallel to the x-axis ($\alpha \neq 0$) or not parallel to the y-axis ($\beta \neq 0$) or not parallel to the z-axis ($\gamma \neq 0$).

Say we have reasons to believe that the plane we are looking for is not parallel to the z-axis. If we are wrong, in the least squares solution, one of the coefficients, α, β, will be very large. If $\gamma \neq 0$, then we may assume that our plane is given by an equation of the form

$$z = ax + by + d,$$

and we would like this equation to be satisfied for all the p_i's, which leads to a system of n equations in 3 unknowns a, b, d, with $p_i = (x_i, y_i, z_i)$;

$$\begin{aligned} ax_1 + by_1 + d &= z_1 \\ \vdots \qquad &\quad \vdots \\ ax_n + by_n + d &= z_n. \end{aligned}$$

However, if n is larger than 3, such a system generally has *no solution*. Since the above system can't be solved exactly, we can try to find a solution (a, b, d) that *minimizes the least-squares error*

$$\sum_{i=1}^{n} (ax_i + by_i + d - z_i)^2.$$

This is what Legendre and Gauss figured out in the early 1800's!

In general, given a linear system

$$Ax = b,$$

we solve the *least squares problem*: minimize $\|Ax - b\|_2^2$.

Fortunately, every $n \times m$-matrix A can be written as

$$A = VDU^\top$$

where U and V are orthogonal and D is a rectangular diagonal matrix with non-negative entries (*singular value decomposition, or SVD*); see Chapter 20.

The SVD can be used to solve an inconsistent system. It is shown in Chapter 21 that there is a vector x of smallest norm minimizing $\|Ax - b\|_2$. It is given by the (Penrose) *pseudo-inverse* of A (itself given by the SVD).

It has been observed that solving in the least-squares sense may give too much weight to "outliers," that is, points clearly outside the best-fit plane. In this case, it is preferable to minimize (the ℓ^1-norm)

$$\sum_{i=1}^{n} |ax_i + by_i + d - z_i|.$$

This does not appear to be a linear problem, but we can use a trick to convert this minimization problem into a linear program (which means a problem involving linear constraints).

Note that $|x| = \max\{x, -x\}$. So by introducing new variables $e_1, \ldots, e_n$, our minimization problem is equivalent to the linear program (LP):

$$\begin{array}{rl} \text{minimize} & e_1 + \cdots + e_n \\ \text{subject to} & ax_i + by_i + d - z_i \leq e_i \\ & -(ax_i + by_i + d - z_i) \leq e_i \\ & 1 \leq i \leq n. \end{array}$$

Observe that the constraints are equivalent to

$$e_i \geq |ax_i + by_i + d - z_i|, \qquad 1 \leq i \leq n.$$

For an optimal solution, we must have equality, since otherwise we could decrease some e_i and get an even better solution. Of course, we are no longer dealing with "pure" linear algebra, since our constraints are inequalities.

We prefer not getting into linear programming right now, but the above example provides a good reason to learn more about linear programming!

8.7 Limits of Sequences and Series

If $x \in \mathbb{R}$ or $x \in \mathbb{C}$ and if $|x| < 1$, it is well known that the sums $\sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n$ converge to the limit $1/(1 - x)$ when n goes to infinity, and we write

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x}.$$

For example,

$$\sum_{k=0}^{\infty} \frac{1}{2^k} = 2.$$

Similarly, the sums

$$S_n = \sum_{k=0}^{n} \frac{x^k}{k!}$$

converge to e^x when n goes to infinity, for every x (in $\mathbb{R}$ or $\mathbb{C}$). What if we replace x by a real of complex $n \times n$ matrix A?

The partial sums $\sum_{k=0}^{n} A^k$ and $\sum_{k=0}^{n} \frac{A^k}{k!}$ still make sense, but we have to define what is the limit of a sequence of matrices. This can be done in any normed vector space.

Definition 8.12. Let $(E, \|\|)$ be a normed vector space. A *sequence* $(u_n)_{n\in\mathbb{N}}$ in E is any function $u\colon \mathbb{N} \to E$. For any $v \in E$, the sequence (u_n) *converges to* v (and v *is the limit of the sequence* (u_n)) if for every $\epsilon > 0$, there is some integer $N > 0$ such that

$$\|u_n - v\| < \epsilon \quad \text{for all } n \geq N.$$

Often we assume that a sequence is indexed by $\mathbb{N} - \{0\}$, that is, its first term is u_1 rather than u_0.

If the sequence (u_n) converges to v, then since by the triangle inequality

$$\|u_m - u_n\| \leq \|u_m - v\| + \|v - u_n\|,$$

we see that for every $\epsilon > 0$, we can find $N > 0$ such that $\|u_m - v\| < \epsilon/2$ and $\|u_n - v\| < \epsilon/2$, and so

$$\|u_m - u_n\| < \epsilon \quad \text{for all } m, n \geq N.$$

The above property is *necessary* for a convergent sequence, but *not necessarily* sufficient. For example, if $E = \mathbb{Q}$, there are sequences of rationals satisfying the above condition, but whose limit is not a rational number. For example, the sequence $\sum_{k=1}^{n} \frac{1}{k!}$ converges to e, and the sequence $\sum_{k=0}^{n} (-1)^k \frac{1}{2k+1}$ converges to $\pi/4$, but e and $\pi/4$ are not rational (in fact, they are transcendental). However, $\mathbb{R}$ is constructed from $\mathbb{Q}$ to guarantee that sequences with the above property converge, and so is $\mathbb{C}$.

Definition 8.13. Given a normed vector space $(E, \| \|)$, a sequence (u_n) is a *Cauchy sequence* if for every $\epsilon > 0$, there is some $N > 0$ such that

$$\|u_m - u_n\| < \epsilon \quad \text{for all } m, n \geq N.$$

If every Cauchy sequence converges, then we say that E is *complete*. A complete normed vector spaces is also called a *Banach space*.

A fundamental property of $\mathbb{R}$ is that *it is complete*. It follows immediately that $\mathbb{C}$ is also complete. If E is a finite-dimensional real or complex vector space, since any two norms are equivalent, we can pick the ℓ^∞ norm, and then by picking a basis in E, a sequence (u_n) of vectors in E converges iff the n sequences of coordinates (u_n^i) $(1 \leq i \leq n)$ converge, so *any finite-dimensional real or complex vector space is a Banach space*.

Let us now consider the convergence of series.

Definition 8.14. Given a normed vector space $(E, \| \|)$, a *series* is an infinite sum $\sum_{k=0}^{\infty} u_k$ of elements $u_k \in E$. We denote by S_n the partial sum of the first $n+1$ elements,

$$S_n = \sum_{k=0}^{n} u_k.$$

Definition 8.15. We say that the series $\sum_{k=0}^{\infty} u_k$ *converges* to the limit $v \in E$ if the sequence (S_n) converges to v, i.e., given any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$,

$$\|S_n - v\| < \epsilon.$$

In this case, we say that the series is *convergent*. We say that the series $\sum_{k=0}^{\infty} u_k$ *converges absolutely* if the series of norms $\sum_{k=0}^{\infty} \|u_k\|$ is convergent.

If the series $\sum_{k=0}^{\infty} u_k$ converges to v, since for all m, n with $m > n$ we have

$$\sum_{k=0}^{m} u_k - S_n = \sum_{k=0}^{m} u_k - \sum_{k=0}^{n} u_k = \sum_{k=n+1}^{m} u_k,$$

if we let m go to infinity (with n fixed), we see that the series $\sum_{k=n+1}^{\infty} u_k$ converges and that

$$v - S_n = \sum_{k=n+1}^{\infty} u_k.$$

There are series that are convergent but not absolutely convergent; for example, the series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}$$

converges to $\ln 2$, but $\sum_{k=1}^{\infty} \frac{1}{k}$ does not converge (this sum is infinite).

If E is complete, the converse is an enormously useful result.

Proposition 8.18. *Assume $(E, \| \|)$ is a complete normed vector space. If a series $\sum_{k=0}^{\infty} u_k$ is absolutely convergent, then it is convergent.*

Proof. If $\sum_{k=0}^{\infty} u_k$ is absolutely convergent, then we prove that the sequence (S_m) is a Cauchy sequence; that is, for every $\epsilon > 0$, there is some $p > 0$ such that for all $n \geq m \geq p$,

$$\|S_n - S_m\| \leq \epsilon.$$

Observe that

$$\|S_n - S_m\| = \|u_{m+1} + \cdots + u_n\| \leq \|u_{m+1}\| + \cdots + \|u_n\|,$$

and since the sequence $\sum_{k=0}^{\infty} \|u_k\|$ converges, it satisfies Cauchy's criterion. Thus, the sequence (S_m) also satisfies Cauchy's criterion, and since E is a complete vector space, the sequence (S_m) converges. □

Remark: It can be shown that if $(E, \| \|)$ is a normed vector space such that every absolutely convergent series is also convergent, then E must be complete (see Schwartz [54]).

An important corollary of absolute convergence is that if the terms in series $\sum_{k=0}^{\infty} u_k$ are rearranged, then the resulting series is still absolutely convergent and has the *same sum*. More precisely, let σ be any permutation (bijection) of the natural numbers. The series $\sum_{k=0}^{\infty} u_{\sigma(k)}$ is called a *rearrangement* of the original series. The following result can be shown (see Schwartz [54]).

Proposition 8.19. *Assume $(E, \| \|)$ is a normed vector space. If a series $\sum_{k=0}^{\infty} u_k$ is convergent as well as absolutely convergent, then for every permutation σ of $\mathbb{N}$, the series $\sum_{k=0}^{\infty} u_{\sigma(k)}$ is convergent and absolutely convergent, and its sum is equal to the sum of the original series:*

$$\sum_{k=0}^{\infty} u_{\sigma(k)} = \sum_{k=0}^{\infty} u_k.$$

In particular, if $(E, \| \|)$ is a complete normed vector space, then Proposition 8.19 holds. We now apply Proposition 8.18 to the matrix exponential.

8.8 The Matrix Exponential

Proposition 8.20. *For any $n \times n$ real or complex matrix A, the series*

$$\sum_{k=0}^{\infty} \frac{A^k}{k!}$$

converges absolutely for any operator norm on $\mathrm{M}_n(\mathbb{C})$ (or $\mathrm{M}_n(\mathbb{R})$).

Proof. Pick any norm on $\mathbb{C}^n$ (or $\mathbb{R}^n$) and let $\|\|$ be the corresponding operator norm on $\mathrm{M}_n(\mathbb{C})$. Since $\mathrm{M}_n(\mathbb{C})$ has dimension n^2, it is complete. By Proposition 8.18, it suffices to show that the series of nonnegative reals $\sum_{k=0}^{n} \left\| \frac{A^k}{k!} \right\|$ converges. Since $\| \|$ is an operator norm, this a matrix norm, so we have

$$\sum_{k=0}^{n} \left\| \frac{A^k}{k!} \right\| \leq \sum_{k=0}^{n} \frac{\|A\|^k}{k!} \leq e^{\|A\|}.$$

Thus, the nondecreasing sequence of positive real numbers $\sum_{k=0}^{n} \left\| \frac{A^k}{k!} \right\|$ is bounded by $e^{\|A\|}$, and by a fundamental property of $\mathbb{R}$, it has a least upper bound which is its limit. □

Definition 8.16. Let E be a finite-dimensional real of complex normed vector space. For any $n \times n$ matrix A, the limit of the series

$$\sum_{k=0}^{\infty} \frac{A^k}{k!}$$

is the *exponential of A* and is denoted e^A.

A basic property of the exponential $x \mapsto e^x$ with $x \in \mathbb{C}$ is

$$e^{x+y} = e^x e^y, \quad \text{for all } x, y \in \mathbb{C}.$$

As a consequence, e^x is always invertible and $(e^x)^{-1} = e^{-x}$. For matrices, because matrix multiplication is not commutative, in general,

$$e^{A+B} = e^A e^B$$

fails! This result is salvaged as follows.

Proposition 8.21. *For any two $n \times n$ complex matrices A and B, if A and B commute, that is, $AB = BA$, then*

$$e^{A+B} = e^A e^B.$$

A proof of Proposition 8.21 can be found in Gallier [25].

Since A and $-A$ commute, as a corollary of Proposition 8.21, we see that e^A is always invertible and that

$$(e^A)^{-1} = e^{-A}.$$

It is also easy to see that

$$(e^A)^\top = e^{A^\top}.$$

In general, there is no closed-form formula for the exponential e^A of a matrix A, but for skew symmetric matrices of dimension 2 and 3, there are explicit formulae. Everyone should enjoy computing the exponential e^A where

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

If we write

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

then

$$A = \theta J$$

The key property is that

$$J^2 = -I.$$

Proposition 8.22. *If $A = \theta J$, then*

$$e^A = \cos\theta I + \sin\theta J = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Proof. We have

$$\begin{aligned} A^{4n} &= \theta^{4n} I_2, \\ A^{4n+1} &= \theta^{4n+1} J, \\ A^{4n+2} &= -\theta^{4n+2} I_2, \\ A^{4n+3} &= -\theta^{4n+3} J, \end{aligned}$$

and so

$$e^A = I_2 + \frac{\theta}{1!}J - \frac{\theta^2}{2!}I_2 - \frac{\theta^3}{3!}J + \frac{\theta^4}{4!}I_2 + \frac{\theta^5}{5!}J - \frac{\theta^6}{6!}I_2 - \frac{\theta^7}{7!}J + \cdots .$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) I_2 + \left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right) J.$$

We recognize the power series for $\cos\theta$ and $\sin\theta$, and thus

$$e^A = \cos\theta I_2 + \sin\theta J,$$

that is

$$e^A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

as claimed. □

Thus, we see that the exponential of a 2×2 skew-symmetric matrix is a rotation matrix. This property generalizes to any dimension. An explicit formula when $n = 3$ (the Rodrigues' formula) is given in Section 11.7.

Proposition 8.23. *If B is an $n \times n$ (real) skew symmetric matrix, that is, $B^\top = -B$, then $Q = e^B$ is an orthogonal matrix, that is*

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

Proof. Since $B^\top = -B$, we have

$$Q^\top = (e^B)^\top = e^{B^\top} = e^{-B}.$$

Since B and $-B$ commute, we have

$$Q^\top Q = e^{-B} e^B = e^{-B+B} = e^0 = I.$$

Similarly,

$$QQ^\top = e^B e^{-B} = e^{B-B} = e^0 = I,$$

which concludes the proof. □

It can also be shown that $\det(Q) = \det(e^B) = 1$, but this requires a better understanding of the eigenvalues of e^B (see Section 14.5). Furthermore, for every $n \times n$ rotation matrix Q (an orthogonal matrix Q such that $\det(Q) = 1$), there is a skew symmetric matrix B such that $Q = e^B$. This is a fundamental property which has applications in robotics for $n = 3$.

All familiar series have matrix analogs. For example, if $\|A\| < 1$ (where $\| \;\|$ is an operator norm), then the series $\sum_{k=0}^{\infty} A^k$ converges absolutely, and it can be shown that its limit is $(I - A)^{-1}$.

Another interesting series is the logarithm. For any $n \times n$ complex matrix A, if $\|A\| < 1$ (where $\| \;\|$ is an operator norm), then the series

$$\log(I + A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}$$

converges absolutely.

8.9 Summary

The main concepts and results of this chapter are listed below:

- *Norms* and *normed vector spaces*.
- The *triangle inequality*.
- The *Euclidean norm*; the *ℓ^p-norms*.
- *Hölder's inequality*; the *Cauchy–Schwarz inequality*; *Minkowski's inequality*.
- *Hermitian inner product* and *Euclidean inner product*.
- *Equivalent* norms.
- *All norms on a finite-dimensional vector space are equivalent* (Theorem 8.5).
- *Matrix norms*.
- *Hermitian, symmetric* and *normal* matrices. *Orthogonal* and *unitary* matrices.
- The *trace* of a matrix.
- *Eigenvalues* and *eigenvectors* of a matrix.
- The *characteristic polynomial* of a matrix.
- The *spectral radius* $\rho(A)$ of a matrix A.
- The *Frobenius norm*.

- The Frobenius norm is a *unitarily invariant* matrix norm.
- *Bounded* linear maps.
- *Subordinate matrix norms*.
- Characterization of the subordinate matrix norms for the vector norms $\| \|_1$, $\| \|_2$, and $\| \|_\infty$.
- The *spectral norm*.
- For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \|$ such that $\|A\| \leq \rho(A) + \epsilon$.
- *Condition numbers* of matrices.
- Perturbation analysis of linear systems.
- The *singular value decomposition* (SVD).
- Properties of conditions numbers. Characterization of $\mathrm{cond}_2(A)$ in terms of the largest and smallest singular values of A.
- The *Hilbert matrix*: a very badly conditioned matrix.
- Solving inconsistent linear systems by the method of *least-squares*; *linear programming*.
- Convergence of sequences of vectors in a normed vector space.
- Cauchy sequences, complex normed vector spaces, Banach spaces.
- Convergence of series. Absolute convergence.
- The matrix exponential.
- Skew symmetric matrices and orthogonal matrices.

8.10 Problems

Problem 8.1. Let A be the following matrix:

$$B = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 3/2 \end{pmatrix}.$$

Compute the operator 2-norm $\|A\|_2$ of A.

Problem 8.2. Prove Proposition 8.3, namely that the following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):

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

Problem 8.3. For any $p \geq 1$, prove that for all $x \in \mathbb{R}^n$,

$$\lim_{p \mapsto \infty} \|x\|_p = \|x\|_\infty .$$

Problem 8.4. Let A be an $n \times n$ matrix which is strictly row diagonally dominant, which means that

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

for $i = 1, \ldots, n$, and let

$$\delta = \min_i \left\{ |a_{i\,i}| - \sum_{j \neq i} |a_{i\,j}| \right\}.$$

The fact that A is strictly row diagonally dominant is equivalent to the condition $\delta > 0$.

(1) For any nonzero vector v, prove that

$$\|Av\|_\infty \geq \|v\|_\infty \delta.$$

Use the above to prove that A is invertible.

(2) Prove that

$$\left\|A^{-1}\right\|_\infty \leq \delta^{-1}.$$

Hint. Prove that

$$\sup_{v \neq 0} \frac{\|A^{-1}v\|_\infty}{\|v\|_\infty} = \sup_{w \neq 0} \frac{\|w\|_\infty}{\|Aw\|_\infty}.$$

Problem 8.5. Let A be any invertible complex $n \times n$ matrix.

(1) For any vector norm $\| \; \|$ on $\mathbb{C}^n$, prove that the function $\| \; \|_A : \mathbb{C}^n \to \mathbb{R}$ given by

$$\|x\|_A = \|Ax\| \quad \text{for all} \quad x \in \mathbb{C}^n,$$

is a vector norm.

(2) Prove that the operator norm induced by $\| \; \|_A$, also denoted by $\| \; \|_A$, is given by

$$\|B\|_A = \left\|ABA^{-1}\right\| \quad \text{for every } n \times n \text{ matrix} \quad B,$$

where $\|ABA^{-1}\|$ uses the operator norm induced by $\| \; \|$.

Problem 8.6. Give an example of a norm on $\mathbb{C}^n$ and of a *real* matrix A such that

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

where $\|-\|_{\mathbb{R}}$ and $\|-\|$ are the operator norms associated with the vector norm $\|-\|$.
Hint. This can already be done for $n = 2$.

Problem 8.7. Let $\| \ \|$ be any operator norm. Given an invertible $n \times n$ matrix A, if $c = 1/(2\|A^{-1}\|)$, then for every $n \times n$ matrix H, if $\|H\| \leq c$, then $A + H$ is invertible. Furthermore, show that if $\|H\| \leq c$, then $\|(A+H)^{-1}\| \leq 1/c$.

Problem 8.8. Let A be any $m \times n$ matrix and let $\lambda \in \mathbb{R}$ be any positive real number $\lambda > 0$.

(1) Prove that $A^\top A + \lambda I_n$ and $AA^\top + \lambda I_m$ are invertible.

(2) Prove that

$$A^\top (AA^\top + \lambda I_m)^{-1} = (A^\top A + \lambda I_n)^{-1} A^\top.$$

Remark: The expressions above correspond to the matrix for which the function

$$\Phi(x) = (Ax - b)^\top (Ax - b) + \lambda x^\top x$$

achieves a minimum. It shows up in machine learning (kernel methods).

Problem 8.9. Let Z be a $q \times p$ real matrix. Prove that if $I_p - Z^\top Z$ is positive definite, then the $(p+q) \times (p+q)$ matrix

$$S = \begin{pmatrix} I_p & Z^\top \\ Z & I_q \end{pmatrix}$$

is symmetric positive definite.

Problem 8.10. Prove that for any real or complex square matrix A, we have

$$\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty,$$

where the above norms are operator norms.
Hint. Use Proposition 8.10 (among other things, it shows that $\|A\|_1 = \left\|A^\top\right\|_\infty$).

Problem 8.11. Show that the map $A \mapsto \rho(A)$ (where $\rho(A)$ is the spectral radius of A) is neither a norm nor a matrix norm. In particular, find two 2×2 matrices A and B such that

$$\rho(A + B) > \rho(A) + \rho(B) = 0 \quad \text{and} \quad \rho(AB) > \rho(A)\rho(B) = 0.$$

Problem 8.12. Define the map $A \mapsto M(A)$ (defined on $n \times n$ real or complex $n \times n$ matrices) by

$$M(A) = \max\{|a_{ij}| \mid 1 \leq i, j \leq n\}.$$

(1) Prove that

$$M(AB) \leq nM(A)M(B)$$

for all $n \times n$ matrices A and B.

(2) Give a counter-example of the inequality

$$M(AB) \leq M(A)M(B).$$

(3) Prove that the map $A \mapsto \|A\|_M$ given by

$$\|A\|_M = nM(A) = n \max\{|a_{ij}| \mid 1 \leq i, j \leq n\}$$

is a matrix norm.

Problem 8.13. Let S be a real symmetric positive definite matrix.

(1) Use the Cholesky factorization to prove that there is some upper-triangular matrix C, unique if its diagonal elements are strictly positive, such that $S = C^\top C$.

(2) For any $x \in \mathbb{R}^n$, define

$$\|x\|_S = (x^\top S x)^{1/2}.$$

Prove that

$$\|x\|_S = \|Cx\|_2\,,$$

and that the map $x \mapsto \|x\|_S$ is a norm.

Problem 8.14. Let A be a real 2×2 matrix

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

(1) Prove that the squares of the singular values $\sigma_1 \geq \sigma_2$ of A are the roots of the quadratic equation

$$X^2 - \operatorname{tr}(A^\top A)X + |\det(A)|^2 = 0.$$

(2) If we let

$$\mu(A) = \frac{a_{1\,1}^2 + a_{1\,2}^2 + a_{2\,1}^2 + a_{2\,2}^2}{2|a_{1\,1}a_{2\,2} - a_{1\,2}a_{2\,1}|},$$

prove that

$$\operatorname{cond}_2(A) = \frac{\sigma_1}{\sigma_2} = \mu(A) + (\mu(A)^2 - 1)^{1/2}.$$

(3) Consider the subset $\mathcal{S}$ of 2×2 invertible matrices whose entries $a_{i\,j}$ are integers such that $0 \leq a_{ij} \leq 100$.

Prove that the functions $\operatorname{cond}_2(A)$ and $\mu(A)$ reach a maximum on the set $\mathcal{S}$ for the same values of A.

Check that for the matrix

$$A_m = \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix}$$

we have

$$\mu(A_m) = 19{,}603 \quad \det(A_m) = -1$$

and

$$\mathrm{cond}_2(A_m) \approx 39,206.$$

(4) Prove that for all $A \in \mathcal{S}$, if $|\det(A)| \geq 2$ then $\mu(A) \leq 10,000$. Conclude that the maximum of $\mu(A)$ on $\mathcal{S}$ is achieved for matrices such that $\det(A) = \pm 1$. Prove that finding matrices that maximize μ on $\mathcal{S}$ is equivalent to finding some integers n_1, n_2, n_3, n_4 such that

$$\begin{aligned}
&0 \leq n_4 \leq n_3 \leq n_2 \leq n_1 \leq 100\\
&n_1^2 + n_2^2 + n_3^2 + n_4^2 \geq 100^2 + 99^2 + 99^2 + 98^2 = 39,206\\
&|n_1 n_4 - n_2 n_3| = 1.
\end{aligned}$$

You may use without proof that the fact that the only solution to the above constraints is the multiset

$$\{100, 99, 99, 98\}.$$

(5) Deduce from part (4) that the matrices in $\mathcal{S}$ for which μ has a maximum value are

$$A_m = \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix} \quad \begin{pmatrix} 98 & 99 \\ 99 & 100 \end{pmatrix} \quad \begin{pmatrix} 99 & 100 \\ 98 & 99 \end{pmatrix} \quad \begin{pmatrix} 99 & 98 \\ 100 & 99 \end{pmatrix}$$

and check that μ has the same value for these matrices. Conclude that

$$\max_{A \in \mathcal{S}} \mathrm{cond}_2(A) = \mathrm{cond}_2(A_m).$$

(6) Solve the system

$$\begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 199 \\ 197 \end{pmatrix}.$$

Perturb the right-hand side b by

$$\Delta b = \begin{pmatrix} -0.0097 \\ 0.0106 \end{pmatrix}$$

and solve the new system

$$A_m y = b + \Delta b$$

where $y = (y_1, y_2)$. Check that

$$\Delta x = y - x = \begin{pmatrix} 2 \\ -2.0203 \end{pmatrix}.$$

Compute $\|x\|_2$, $\|\Delta x\|_2$, $\|b\|_2$, $\|\Delta b\|_2$, and estimate

$$c = \frac{\|\Delta x\|_2}{\|x\|_2} \left(\frac{\|\Delta b\|_2}{\|b\|_2} \right)^{-1}.$$

Check that

$$c \approx \mathrm{cond}_2(A_m) = 39,206.$$

Problem 8.15. Consider a real 2×2 matrix with zero trace of the form

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}.$$

(1) Prove that

$$A^2 = (a^2 + bc)I_2 = -\det(A)I_2.$$

If $a^2 + bc = 0$, prove that

$$e^A = I_2 + A.$$

(2) If $a^2 + bc < 0$, let $\omega > 0$ be such that $\omega^2 = -(a^2 + bc)$. Prove that

$$e^A = \cos\omega\, I_2 + \frac{\sin\omega}{\omega} A.$$

(3) If $a^2 + bc > 0$, let $\omega > 0$ be such that $\omega^2 = a^2 + bc$. Prove that

$$e^A = \cosh\omega\, I_2 + \frac{\sinh\omega}{\omega} A.$$

(3) Prove that in all cases

$$\det\left(e^A\right) = 1 \quad \text{and} \quad \mathrm{tr}(A) \geq -2.$$

(4) Prove that there exist some real 2×2 matrix B with $\det(B) = 1$ such that there is no real 2×2 matrix A with zero trace such that $e^A = B$.

Problem 8.16. Recall that the Hilbert matrix is given by

$$H_{ij}^{(n)} = \left(\frac{1}{i + j - 1}\right).$$

(1) Prove that

$$\det(H^{(n)}) = \frac{(1!2!\cdots(n-1)!)^4}{1!2!\cdots(2n-1)!},$$

thus the reciprocal of an integer.
Hint. Use Problem 6.13.

(2) Amazingly, the entries of the inverse of $H^{(n)}$ are integers. Prove that $(H^{(n)})^{-1} = (\alpha_{ij})$, with

$$\alpha_{ij} = (-1)^{i+j}(i + j - 1)\binom{n + i - 1}{n - j}\binom{n + j - 1}{n - i}\binom{i + j - 2}{i - 1}^2.$$