4 Norms and Matrix Conditioning

In this chapter, we look at some qualitative aspects of the solution to (3.1), which we studied in Chapter 3. While it is usually assumed that the coefficient matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$ and the right-hand-side vector $\boldsymbol{f} \in \mathbb{C}^n$ are known, fixed, and perfectly represented in our computational device, here we are interested to see what happens to the solution to our linear system if these are somehow perturbed.

To fix ideas, suppose, for example, that A is invertible and that $\boldsymbol{x} \in \mathbb{C}^n$ is the solution to the (ideal) system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{f}. \tag{4.1}$$

Now suppose that, in some hypothetical computing device, A and $\boldsymbol{f}$ are perturbed in storage: $\mathsf{A} \to \mathsf{A} + \delta\mathsf{A}$ and $\boldsymbol{f} \to \boldsymbol{f} + \delta\boldsymbol{f}$, where $\delta\mathsf{A} \in \mathbb{C}^{n\times n}$ and $\delta\boldsymbol{f} \in \mathbb{C}^n$. Assuming that $\mathsf{A} + \delta\mathsf{A}$ is invertible, there is some $\delta\boldsymbol{x} \in \mathbb{C}^n$ such that $\boldsymbol{x} + \delta\boldsymbol{x} \in \mathbb{C}^n$ is the solution to the perturbed system

$$(\mathsf{A} + \delta\mathsf{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{f} + \delta\boldsymbol{f}. \tag{4.2}$$

Clearly, $\delta\boldsymbol{x}$ measures the error resulting from the perturbations to our data. How large is this error vector? How large is the relative error, $\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$? How do the error vector and relative error relate to the sizes of the perturbations? It turns out that the answers to our questions depend upon the so-called condition number of the matrix A, defined as

$$\kappa(\mathsf{A}) = \|\mathsf{A}\| \left\|\mathsf{A}^{-1}\right\|.$$

Of course, it is not practical to compute $\kappa(\mathsf{A})$ according to the formula above, as it involves the inverse of the coefficient matrix. But, often, the condition number can be accurately estimated. We will see that, if the condition number is large, the relative error can be quite large, even when other measures of error are actually small.

Now, since all the ideas in this chapter depend heavily upon the notions of vector norms on $\mathbb{C}^n$ and induced matrix norms on $\mathbb{C}^{n\times n}$, we urge the reader to, if necessary, review these concepts in Sections A.3 and 1.2.

4.1 The Spectral Radius

We begin with a definition.

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

Definition 4.1 (spectral radius)**.** Suppose that $A \in \mathfrak{L}(\mathbb{V})$, where $\mathbb{V}$ is a complex n-dimensional vector space. The **spectral radius** of A is

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \sigma(A)\}.$$

An analogous definition is made for any matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$.

For self-adjoint operators, the spectral radius has a very precise meaning.

Theorem 4.2 (self-adjoint operator)**.** *Let $\mathbb{V}$ be an n-dimensional complex inner product space. Suppose that $\|x\| = (x,x)^{1/2}$ is the Euclidean norm. Let $A \in \mathfrak{L}(\mathbb{V})$ be self-adjoint. Then the induced norm satisfies*

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

Proof. Since $A\colon \mathbb{V} \to \mathbb{V}$ is self-adjoint, there exists an orthonormal basis of eigenvectors $S = \{e_1, \dots, e_n\}$, i.e., $(e_i, e_j) = \delta_{i,j}$, $\mathbb{V} = \operatorname{span}(S)$, and $Ae_i = \lambda_i e_i$. Expanding $x \in \mathbb{V}$ in this basis, i.e., $x = \sum_{i=1}^n x_i e_i$ with $x_i \in \mathbb{C}$, we see that

$$Ax = \sum_{i=1}^n \lambda_i x_i e_i.$$

Since this basis is orthonormal,

$$\|x\|^2 = \sum_{i=1}^n |x_i|^2 \quad \text{and} \quad \|Ax\|^2 = \sum_{i=1}^n |\lambda_i|^2 |x_i|^2.$$

With this at hand we notice that

$$\|Ax\| \le \max\{|\lambda| \,|\lambda \in \sigma(A)\}\,\|x\|,$$

which implies that

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

Problem 4.2 gives the reverse inequality, and this concludes the proof. $\square$

For more general operators and norms, all that can be established is the following.

Theorem 4.3 (norms and spectral radius)**.** *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is any induced matrix norm. Then there is a constant $C > 0$ such that*

$$\rho(\mathsf{A}) \le \|\mathsf{A}\| \le C\sqrt{\rho(\mathsf{A}^\mathsf{H}\mathsf{A})}, \quad \forall \mathsf{A} \in \mathbb{C}^{n\times n}.$$

Proof. Let $\|\cdot\|_{\mathbb{C}^n}$ be the vector norm that induces $\|\cdot\|$. Owing to Problem A.12, this norm is equivalent to $\|\cdot\|_2$, i.e., there are constants $0 < C_1 \le C_2$ for which

$$C_1\|\boldsymbol{x}\|_{\mathbb{C}^n} \le \|\boldsymbol{x}\|_2 \le C_2\|\boldsymbol{x}\|_{\mathbb{C}^n}, \quad \forall \boldsymbol{x} \in \mathbb{C}^n.$$

This, in turn, implies that, if $\boldsymbol{x} \in \mathbb{C}^n_\star$,

$$\frac{\|\mathsf{A}\boldsymbol{x}\|_{\mathbb{C}^n}}{\|\boldsymbol{x}\|_{\mathbb{C}^n}} \le \frac{C_2}{C_1}\frac{\|\mathsf{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \le C\|\mathsf{A}\|_2 \le C\sqrt{\rho(\mathsf{A}^\mathsf{H}\mathsf{A})},$$

where we denoted $C = C_2/C_1$ and used Problem 1.29. Taking supremum over $\boldsymbol{x} \in \mathbb{C}^n_\star$ implies the upper bound. The lower bound is an exercise; see Problem 4.2. $\square$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

While the spectral radius may not necessarily be a norm it is almost one, as the following, rather technical, result shows.

Theorem 4.4 (spectral radius and norms). *For every matrix* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *and any* $\varepsilon > 0$, *there is a norm* $\|\cdot\|_{\mathsf{A},\varepsilon} : \mathbb{C}^n \to \mathbb{R}$ *such that the induced matrix norm*

$$\|\mathsf{M}\|_{\mathsf{A},\varepsilon} = \sup_{\boldsymbol{x}\in\mathbb{C}^n_\star} \frac{\|\mathsf{M}\boldsymbol{x}\|_{\mathsf{A},\varepsilon}}{\|\boldsymbol{x}\|_{\mathsf{A},\varepsilon}} = \sup_{\|\boldsymbol{x}\|_{\mathsf{A},\varepsilon}=1} \|\mathsf{M}\boldsymbol{x}\|_{\mathsf{A},\varepsilon}, \quad \forall \mathsf{M} \in \mathbb{C}^{n\times n},$$

satisfies

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

Proof. Appealing to the Schur factorization, Lemma 1.46, there is a unitary matrix $\mathsf{P} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathsf{B} \in \mathbb{C}^{n\times n}$ such that

$$\mathsf{A} = \mathsf{P}^{\mathsf{H}}\mathsf{B}\mathsf{P}.$$

The diagonal elements of B are the eigenvalues of A. Let us write

$$\mathsf{B} = \Lambda + \mathsf{U},$$

where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ with

$$\sigma(\mathsf{A}) = \{\lambda_1, \dots, \lambda_n\} = \sigma(\mathsf{B}),$$

and $\mathsf{U} = [u_{i,j}] \in \mathbb{C}^{n\times n}$ is strictly upper triangular. Let $\delta > 0$ be arbitrary. Define

$$\mathsf{D} = \operatorname{diag}\left(1, \delta^{-1}, \dots, \delta^{1-n}\right).$$

Next, define

$$\mathsf{C} = \mathsf{D}\mathsf{B}\mathsf{D}^{-1} = \Lambda + \mathsf{E},$$

where

$$\mathsf{E} = \mathsf{D}\mathsf{U}\mathsf{D}^{-1}.$$

Now observe that E, like U, must be strictly upper triangular, and the elements of $\mathsf{E} = [e_{i,j}]$ must satisfy

$$e_{i,j} = \begin{cases} 0, & j \le i, \\ \delta^{j-i} u_{i,j}, & j > i. \end{cases}$$

Consequently, the elements of E can be made arbitrarily small in modulus, depending on our choice of δ.

Now notice that

$$\mathsf{A} = \mathsf{P}^{-1}\mathsf{D}^{-1}\mathsf{C}\mathsf{D}\mathsf{P},$$

and, since $\mathsf{D}\mathsf{P}$ is nonsingular, the following defines a norm and an induced matrix norm: for any $\boldsymbol{x} \in \mathbb{C}^n$,

$$\|\boldsymbol{x}\|_\star = \|\mathsf{D}\mathsf{P}\boldsymbol{x}\|_2,$$

and, for any $\mathsf{M} \in \mathbb{C}^{n\times n}$,

$$\|\mathsf{M}\|_\star = \sup_{\|\boldsymbol{x}\|_\star=1} \|\mathsf{M}\boldsymbol{x}\|_\star.$$

Observe that

$$\|\mathsf{A}\boldsymbol{y}\|_\star = \|\mathsf{DPA}\boldsymbol{y}\|_2 = \|\mathsf{CDP}\boldsymbol{y}\|_2 .$$

Define $\boldsymbol{z} = \mathsf{DP}\boldsymbol{y}$. Then

$$\|\mathsf{A}\boldsymbol{y}\|_\star = \|\mathsf{C}\boldsymbol{z}\|_2 = \sqrt{\boldsymbol{z}^{\mathsf{H}}\mathsf{C}^{\mathsf{H}}\mathsf{C}\boldsymbol{z}}.$$

But

$$\mathsf{C}^{\mathsf{H}}\mathsf{C} = \left(\Lambda^{\mathsf{H}} + \mathsf{E}^{\mathsf{H}}\right)(\Lambda + \mathsf{E}) = \Lambda^{\mathsf{H}}\Lambda + \mathsf{M}(\delta),$$

where

$$\mathsf{M}(\delta) = \mathsf{E}^{\mathsf{H}}\Lambda + \Lambda^{\mathsf{H}}\mathsf{E} + \mathsf{E}^{\mathsf{H}}\mathsf{E}.$$

As an exercise, the reader should prove that, for a given matrix A, there is a constant $K_1 > 0$ such that

$$\|\mathsf{M}(\delta)\|_2 \le K_1\delta$$

for all $0 < \delta \le 1$. Thus, using the definition of the spectral radius, the Cauchy–Schwarz inequality, and induced norm consistency,

$$\begin{aligned}
\boldsymbol{z}^{\mathsf{H}}\mathsf{C}^{\mathsf{H}}\mathsf{C}\boldsymbol{z} &= \boldsymbol{z}^{\mathsf{H}}\Lambda^{\mathsf{H}}\Lambda\boldsymbol{z} + \boldsymbol{z}^{\mathsf{H}}\mathsf{M}(\delta)\boldsymbol{z}\\
&\le \max_{k=1,\dots,n} |\lambda_k|^2\boldsymbol{z}^{\mathsf{H}}\boldsymbol{z} + \|\boldsymbol{z}\|_2\,\|\mathsf{M}(\delta)\boldsymbol{z}\|_2\\
&\le \left(\rho(\mathsf{A})^2 + \|\mathsf{M}(\delta)\|_2\right)\|\boldsymbol{z}\|_2^2\\
&\le \left(\rho(\mathsf{A})^2 + K_1\delta\right)\|\boldsymbol{z}\|_2^2 .
\end{aligned}$$

To finish up, note that

$$\|\boldsymbol{y}\|_\star = \left\|(\mathsf{DP})^{-1}\boldsymbol{z}\right\|_\star = \|\boldsymbol{z}\|_2 .$$

Hence, $\|\boldsymbol{y}\|_\star = 1$ if and only if $\|\boldsymbol{z}\|_2 = 1$ and

$$\{\|\mathsf{A}\boldsymbol{y}\|_\star \mid \|\boldsymbol{y}\|_\star = 1\} = \{\|\mathsf{C}\boldsymbol{z}\|_\star \mid \|\boldsymbol{z}\|_2 = 1\} .$$

Consequently,

$$\begin{aligned}
\|\mathsf{A}\|_\star &= \sup_{\|\boldsymbol{y}\|_\star=1} \|\mathsf{A}\boldsymbol{y}\|_\star\\
&= \sup_{\|\boldsymbol{z}\|_2=1} \|\mathsf{C}\boldsymbol{z}\|_2\\
&\le \sup_{\|\boldsymbol{z}\|_2=1} \sqrt{\rho(\mathsf{A})^2 + K_1\delta}\,\|\boldsymbol{z}\|_2\\
&= \sqrt{\rho(\mathsf{A})^2 + K_1\delta} \le \rho(\mathsf{A}) + K_2\delta,
\end{aligned}$$

for some $K_2 > 0$, for all $0 < \delta \le 1$. The result follows on choosing

$$\delta \le \min\left(1, \varepsilon/K_2\right) . \qquad \square$$

As a consequence of this result we can provide an extension of Theorem 4.2 for a broader class of matrices.

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

Corollary 4.5 (equality)**.** *Suppose that* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is diagonalizable. Then there exists a norm* $\|\cdot\|_\star : \mathbb{C}^n \to \mathbb{R}$ *such that the induced matrix norm satisfies*

$$\|\mathsf{A}\|_\star = \rho(\mathsf{A}).$$

Proof. See Problem 4.3. $\square$

We provide now a notion of convergence for matrices and, with the aid of the spectral radius, provide necessary and sufficient conditions for convergence.

Definition 4.6 (convergence)**.** We say that the square matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$ is **convergent to zero** if and only if $\mathsf{A}^k \to \mathsf{O} \in \mathbb{C}^{n\times n}$, i.e., if and only if

$$\lim_{k\to\infty} \|\mathsf{A}^k\| \to 0$$

for any matrix norm $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$.

Remark 4.7 (norm equivalence)**.** We recall that, owing to Theorem A.29, all norms on $\mathbb{C}^{m\times n}$, whether induced or not, are equivalent. For this reason, the norm in this last definition does not matter.

Theorem 4.8 (convergence criteria)**.** *Let* $\mathsf{A} \in \mathbb{C}^{n\times n}$*. The following are equivalent.*

1. A *is convergent to zero.*
2. $\rho(\mathsf{A}) < 1$*.*
3. *For all* $\boldsymbol{x} \in \mathbb{C}^n$*,*

$$\lim_{k\to\infty} \mathsf{A}^k \boldsymbol{x} = \mathbf{0}.$$

Proof. ($1 \implies 2$) We recall two facts. First, if $\lambda \in \sigma(\mathsf{A})$, then $\lambda^k \in \sigma(\mathsf{A}^k)$. This follows from the Schur factorization: if $\mathsf{A} = \mathsf{U}\mathsf{T}\mathsf{U}^\mathsf{H}$, where T is upper triangular and U is unitary, then

$$\mathsf{A}^k = \mathsf{U}\mathsf{T}^k\mathsf{U}^\mathsf{H}.$$

Second, $\rho(\mathsf{A}) \le \|\mathsf{A}\|$, for any induced matrix norm. Therefore,

$$0 \le \rho^k(\mathsf{A}) = \rho(\mathsf{A}^k) \le \left\|\mathsf{A}^k\right\|.$$

Thus, if $\left\|\mathsf{A}^k\right\| \to 0$, it follows that

$$\rho^k(\mathsf{A}) \to 0.$$

This implies that $\rho(\mathsf{A}) < 1$.

($2 \implies 1$) By Theorem 4.4, there is an induced matrix norm $\|\cdot\|_\star$ such that

$$\|\mathsf{A}\|_\star \le \rho(\mathsf{A}) + \varepsilon$$

for any $\varepsilon > 0$. Recall that the choice of $\|\cdot\|_\star$ depends upon A and $\varepsilon > 0$. Since, by assumption, $\rho(\mathsf{A}) < 1$, there is an $\varepsilon > 0$ such that $\rho(\mathsf{A}) + \varepsilon < 1$, and, therefore, an induced norm $\|\cdot\|_\star$ such that

$$\|\mathsf{A}\|_\star \le \rho(\mathsf{A}) + \varepsilon < 1.$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

Then, using sub-multiplicativity,

$$\left\|\mathsf{A}^k\right\|_\star \le \|\mathsf{A}\|_\star^k \le (\rho(\mathsf{A})+\varepsilon)^k \to 0.$$

Consequently,

$$\lim_{k\to\infty} \left\|\mathsf{A}^k\right\|_\star = 0.$$

($1 \implies 3$) Suppose that $\lim_{k\to\infty} \left\|\mathsf{A}^k\right\|_\infty = 0$. Let $\boldsymbol{x} \in \mathbb{C}^n$ be arbitrary. Then

$$\left\|\mathsf{A}^k \boldsymbol{x}\right\|_\infty \le \left\|\mathsf{A}^k\right\|_\infty \|\boldsymbol{x}\|_\infty \to 0$$

since $\left\|\mathsf{A}^k\right\|_\infty \to 0$. Hence, $\left\|\mathsf{A}^k \boldsymbol{x}\right\|_\infty \to 0$. This implies that

$$\lim_{k\to\infty} \mathsf{A}^k \boldsymbol{x} = \mathbf{0}.$$

($3 \implies 1$) Suppose that, for any $\boldsymbol{x} \in \mathbb{C}^n$,

$$\lim_{k\to\infty} \mathsf{A}^k \boldsymbol{x} = \mathbf{0}.$$

Then it follows that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$,

$$\boldsymbol{y}^{\mathsf{H}} \mathsf{A}^k \boldsymbol{x} \to 0.$$

Now suppose that $\boldsymbol{y} = \boldsymbol{e}_i$ and $\boldsymbol{x} = \boldsymbol{e}_j$, then, since

$$\boldsymbol{y}^{\mathsf{H}} \mathsf{A}^k \boldsymbol{x} = \boldsymbol{e}_i^{\mathsf{H}} \mathsf{A}^k \boldsymbol{e}_j = \left[\mathsf{A}^k\right]_{i,j},$$

it follows that

$$\lim_{k\to\infty} \left[\mathsf{A}^k\right]_{i,j} = 0.$$

This implies that

$$\lim_{k\to\infty} \left\|\mathsf{A}^k\right\|_{\max} = 0.$$

Hence, A is convergent to zero. □

As an easy consequence we obtain a sufficient criterion for convergence to zero.

Corollary 4.9 (convergence condition)**.** *Let* $\mathsf{M} \in \mathbb{C}^{n\times n}$. *Assume that, for some induced matrix norm* $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$,

$$\|\mathsf{M}\| < 1,$$

then M *is convergent to zero.*

Proof. Recall that, for each and every induced matrix norm $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$,

$$\rho(\mathsf{M}) \le \|\mathsf{M}\|,$$

where $\rho(\mathsf{M})$ is the spectral radius of M. □

The spectral radius of a matrix can also be estimated via powers of this matrix.

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

Proposition 4.10 (upper bound). *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm. Then, for all $A \in \mathbb{C}^{n\times n}$ and $k \in \mathbb{N}$, we have*

$$\rho(A) \le \|A^k\|^{1/k}.$$

Proof. See Problem 4.4. □

In fact, for large values of k the previous upper bound is tight.

Theorem 4.11 (Gelfand[1]). *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm. Then, for all $A \in \mathbb{C}^{n\times n}$, we have*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}.$$

Proof. Let $0 < \varepsilon < \rho(A)/2$. We define two matrices

$$A_\pm = \frac{1}{\rho(A) \pm \varepsilon} A.$$

Clearly,

$$\rho(A_\pm) = \frac{\rho(A)}{\rho(A) \pm \varepsilon},$$

which implies that

$$\rho(A_+) < 1 < \rho(A_-).$$

Therefore, A_+ is convergent to zero; consequently, there is a number $K_+ \in \mathbb{N}$ such that, for $k \ge K_+$, we have

$$\|A_+^k\| < 1 \quad \implies \quad \frac{1}{(\rho(A)+\varepsilon)^k}\|A^k\| < 1 \quad \implies \quad \|A^k\| < (\rho(A)+\varepsilon)^k.$$

On the other hand, the bound of Proposition 4.10 implies that

$$1 < \rho(A_-)^k \le \|A_-^k\| = \frac{1}{(\rho(A)-\varepsilon)^k}\|A^k\|.$$

In conclusion, for sufficiently large k, we have shown that

$$\rho(A) - \varepsilon < \|A^k\|^{1/k} < \rho(A) + \varepsilon$$

and the result follows. □

Corollary 4.12 (product of matrices). *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm. Let $\{A_i\}_{i=1}^k \subset \mathbb{C}^{n\times n}$ be a family of matrices that commute, i.e.,*

$$A_iA_j = A_jA_i, \qquad \forall i,j = 1,\dots,k.$$

Then

$$\rho\left(\prod_{i=1}^k A_i\right) \le \prod_{i=1}^k \rho(A_i).$$

Proof. See Problem 4.6. □

[1] This result is due to the Ukrainian–American mathematician Izrail Moiseevic Gelfand (1913–2009).

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

4.2 Condition Number

We can now introduce the notion of the condition number of a matrix.

Definition 4.13 (condition number)**.** Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is invertible. The **condition number** of A with respect to the matrix norm $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is

$$\kappa(\mathsf{A}) = \|\mathsf{A}\| \left\|\mathsf{A}^{-1}\right\|.$$

Before we get on to the meaning and utility of the condition number, let us present some elementary properties of this quantity.

Proposition 4.14 (properties of κ)**.** *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm and that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is invertible. Then*

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

Furthermore,

$$\frac{1}{\left\|\mathsf{A}^{-1}\right\|} \leq \|\mathsf{A} - \mathsf{B}\|$$

for any $\mathsf{B} \in \mathbb{C}^{n\times n}$ that is singular. Consequently,

$$\frac{1}{\kappa(\mathsf{A})} \leq \inf_{\det(\mathsf{B})=0} \frac{\|\mathsf{A} - \mathsf{B}\|}{\|\mathsf{A}\|}. \tag{4.3}$$

Proof. See Problem 4.10. □

Remark 4.15 (interpretation of $\kappa(\mathsf{A})$)**.** Estimate (4.3) is useful in a couple of ways. First, it says that if A is close in norm to a singular matrix B, then $\kappa(\mathsf{A})$ will be very large. Thus, nearly singular matrices are ill-conditioned. Second, this formula gives an upper bound on $\kappa(\mathsf{A})^{-1}$.

There are some nice formulas for and estimates of the condition number with respect to the induced matrix 2-norm, which is usually called the *spectral condition number* and denoted κ_2.

Proposition 4.16 (spectral condition number)**.** *Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is invertible and $\|\cdot\|_2 : \mathbb{C}^{n\times n} \to \mathbb{R}$ is the induced matrix 2-norm.*

1. *If the singular values of A are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$,*
$$\kappa_2(\mathsf{A}) = \|\mathsf{A}\|_2 \left\|\mathsf{A}^{-1}\right\|_2 = \frac{\sigma_1}{\sigma_n}.$$
2. *If the eigenvalues of $\mathsf{B} = \mathsf{A}^{\mathsf{H}}\mathsf{A}$ are $0 < \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$, then*
$$\kappa_2(\mathsf{A}) = \sqrt{\frac{\mu_n}{\mu_1}}. \tag{4.4}$$
3. *Let, for $p \in [1, \infty]$, $\kappa_p(\mathsf{A}) = \|\mathsf{A}\|_p \cdot \left\|\mathsf{A}^{-1}\right\|_p$, where $\|\cdot\|_p$ is the induced matrix norm with respect to the p-norm. We have*
$$\kappa_2(\mathsf{A}) \leq \sqrt{\kappa_1(\mathsf{A})\kappa_\infty(\mathsf{A})}.$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

4.

$$\frac{1}{\kappa_2(\mathsf{A})} = \inf_{\det(\mathsf{B})=0} \frac{\|\mathsf{A}-\mathsf{B}\|_2}{\|\mathsf{A}\|_2}.$$

5. *If* A *is Hermitian, then*

$$\kappa_2(\mathsf{A}) = \frac{\max_{\lambda\in\sigma(\mathsf{A})}|\lambda|}{\min_{\lambda\in\sigma(\mathsf{A})}|\lambda|}.$$

6. *If* A *is Hermitian positive definite with eigenvalues* $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, *then*

$$\kappa_2(\mathsf{A}) = \frac{\lambda_n}{\lambda_1}.$$

Proof. See Problem 4.12. □

The first and last items in Proposition 4.16 give an easy geometric interpretation of the spectral condition number. It is the ratio of the *maximal stretching* to the *minimal stretching* under the action of the matrix A.

Example 4.1 It is well known that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is singular if and only if $\det(\mathsf{A}) = 0$. Thus, it may be thought that, similar to item 4 of Proposition 4.16, the quantity $|\det(\mathsf{A})|$ may also be used to quantify how close to singular a matrix can be. The following example shows that this is not necessarily the case.

Let $\mathsf{A} \in \mathbb{R}^{n\times n}$ have the singular value decomposition

$$\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^{\mathsf{H}}, \qquad \sigma_j = \frac{1}{j}, \quad j = 1, \dots, n.$$

Then

$$|\det(\mathsf{A})| = \prod_{j=1}^{n} \frac{1}{j} = \frac{1}{n!},$$

but, owing to the first item in Proposition 4.16,

$$\kappa_2(\mathsf{A}) = \frac{\sigma_1}{\sigma_n} = \frac{1}{1/n} = n.$$

Definition 4.17 (error and residual)**.** Given a matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$ and a vector $\boldsymbol{f} \in \mathbb{C}^n$ with A nonsingular, let $\boldsymbol{x} \in \mathbb{C}^n$ solve (3.1). The **residual vector** with respect to $\boldsymbol{x}' \in \mathbb{C}^n$ is defined as

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{x}') = \boldsymbol{f} - \mathsf{A}\boldsymbol{x}' = \mathsf{A}(\boldsymbol{x} - \boldsymbol{x}').$$

The **error vector** with respect to $\boldsymbol{x}'$ is defined as

$$\boldsymbol{e} = \boldsymbol{e}(\boldsymbol{x}') = \boldsymbol{x} - \boldsymbol{x}'.$$

Consequently,

$$\mathsf{A}\boldsymbol{e} = \boldsymbol{r}.$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

It often happens that we have obtained an approximate solution $\boldsymbol{x}' \in \mathbb{C}^n$. We would like to have some measure of the error, but a direct measurement of the error would require the exact solution $\boldsymbol{x}$. The next best thing is the residual, which is an indirect measurement of the error, as the last definition suggests. The next theorem tells us how useful the residual is in determining the relative size of the error.

Theorem 4.18 (relative error estimate). *Let* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be invertible,* $\boldsymbol{f} \in \mathbb{C}^n_\star$*, and* $\boldsymbol{x}$ *solves* (3.1)*. Assume that* $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ *is the induced matrix norm with respect to the vector norm* $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$*. Then*

$$\frac{1}{\kappa(\mathsf{A})}\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{f}\|} \le \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \le \kappa(\mathsf{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{f}\|}.$$

Proof. Since $\boldsymbol{e} = \mathsf{A}^{-1}\boldsymbol{r}$, using consistency of the induced norm

$$\|\boldsymbol{e}\| = \left\|\mathsf{A}^{-1}\boldsymbol{r}\right\| \le \left\|\mathsf{A}^{-1}\right\| \|\boldsymbol{r}\|.$$

Likewise,

$$\|\boldsymbol{f}\| = \|\mathsf{A}\boldsymbol{x}\| \le \|\mathsf{A}\| \|\boldsymbol{x}\|,$$

which implies that

$$\frac{1}{\|\boldsymbol{x}\|} \le \|\mathsf{A}\| \frac{1}{\|\boldsymbol{f}\|}.$$

Combining the first and third inequalities, we obtain the claimed upper bound,

$$\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \le \|\mathsf{A}\| \left\|\mathsf{A}^{-1}\right\| \frac{\|\boldsymbol{r}\|}{\|\boldsymbol{f}\|} = \kappa(\mathsf{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{f}\|}.$$

The rest of the proof is left to the reader as an exercise; see Problem 4.13. □

4.3 Perturbations and Matrix Conditioning

Let us now return to the motivating problem with which we began the chapter, i.e., trying to estimate how much the solution to (3.1) changes under perturbations to the data A and $\boldsymbol{f}$. It is not difficult to imagine a scenario where the data are perturbed. Perturbations may come from measurement errors, and so they are not exactly known. Or, perhaps, perturbations may be introduced when numbers are stored in finite precision in the computer. Let $\delta\mathsf{A} \in \mathbb{C}^{n\times n}$ and $\delta\boldsymbol{f} \in \mathbb{C}^n$ be known (or estimable) perturbations of the data. The problem that is actually solved then is

$$(\mathsf{A} + \delta\mathsf{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{f} + \delta\boldsymbol{f}.$$

We then wish to provide an estimate for how large is the relative error $\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$. Formally, the perturbation $\delta\boldsymbol{x} \in \mathbb{C}^n$ is

$$\delta\boldsymbol{x} = (\mathsf{A} + \delta\mathsf{A})^{-1}(\boldsymbol{f} + \delta\boldsymbol{f}) - \boldsymbol{x},$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

provided that $\mathsf{A} + \delta\mathsf{A}$ is invertible. Observe that, since A is invertible, we have

$$(\mathsf{A} + \delta\mathsf{A})^{-1} = (\mathsf{A}(\mathsf{I}_n + \mathsf{A}^{-1}\delta\mathsf{A}))^{-1} = (\mathsf{I}_n + \mathsf{A}^{-1}\delta\mathsf{A})^{-1}\mathsf{A}^{-1}.$$

Therefore, we have reduced the question of the invertibility of $\mathsf{A} + \delta\mathsf{A}$ to a more general question: Given $\mathsf{M} \in \mathbb{C}^{n\times n}$, when is $\mathsf{I}_n \pm \mathsf{M}$ invertible?

Theorem 4.19 (Neumann series)**.** *Suppose that* $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ *is an induced matrix norm with respect to the vector norm* $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$*. Let* $\mathsf{M} \in \mathbb{C}^{n\times n}$ *with* $\|\mathsf{M}\| < 1$*. Then* $\mathsf{I}_n - \mathsf{M}$ *is invertible,*

$$\left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \le \frac{1}{1 - \|\mathsf{M}\|}$$

and

$$(\mathsf{I}_n - \mathsf{M})^{-1} = \sum_{k=0}^{\infty} \mathsf{M}^k.$$

The series $\sum_{k=0}^{\infty} \mathsf{M}^k$ *is known as the* Neumann series.[2]

Proof. Using the reverse triangle inequality and consistency, since $\|\mathsf{M}\| < 1$, for any $\boldsymbol{x} \in \mathbb{C}^n$,

$$\|(\mathsf{I}_n - \mathsf{M})\boldsymbol{x}\| \ge \big|\|\boldsymbol{x}\| - \|\mathsf{M}\boldsymbol{x}\|\big| \ge (1 - \|\mathsf{M}\|)\|\boldsymbol{x}\|.$$

This inequality implies that if $(\mathsf{I}_n - \mathsf{M})\boldsymbol{x} = \boldsymbol{0}$, then $\boldsymbol{x} = \boldsymbol{0}$. Therefore, $\mathsf{I}_n - \mathsf{M}$ is invertible.

To obtain the norm estimate, notice that

$$\begin{aligned} 1 &= \|\mathsf{I}_n\| \\ &= \left\|(\mathsf{I}_n - \mathsf{M})(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \\ &= \left\|(\mathsf{I}_n - \mathsf{M})^{-1} - \mathsf{M}(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \\ &\ge \left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\| - \|\mathsf{M}\| \left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\|, \end{aligned}$$

where we have used the reverse triangle inequality and sub-multiplicativity. The upper bound of the quantity $\left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\|$ now follows.

Finally, for $N \in \mathbb{N}$, define

$$\mathsf{R}_N = \sum_{k=0}^{N} \mathsf{M}^k.$$

Let us show that $\mathsf{R}_N(\mathsf{I}_n - \mathsf{M}) \to \mathsf{I}_n$ as $N \to \infty$. Indeed,

$$\mathsf{R}_N(\mathsf{I}_n - \mathsf{M}) = \sum_{k=0}^{N} \mathsf{M}^k(\mathsf{I}_n - \mathsf{M}) = \sum_{k=0}^{N} \mathsf{M}^k - \sum_{k=0}^{N} \mathsf{M}^{k+1} = \mathsf{I}_n - \mathsf{M}^{N+1},$$

which shows that, as $N \to \infty$,

$$\|\mathsf{R}_N(\mathsf{I}_n - \mathsf{M}) - \mathsf{I}_n\| = \left\|\mathsf{M}^{N+1}\right\| \le \|\mathsf{M}\|^{N+1} \to 0,$$

using the sub-multiplicativity of the induced norm and the fact that $\|\mathsf{M}\| < 1$. $\square$

[2] Named in honor of the German mathematician Carl Gottfried Neumann (1832–1925).

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

A consequence of Theorem 4.19 is that the set of invertible matrices is *open*. In this context, this means that any matrix that is sufficiently close to an invertible one will also be invertible.

Corollary 4.20 (inverse of a perturbation). *Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$. If $\mathsf{R} \in \mathbb{C}^{n\times n}$ is invertible and $\mathsf{T} \in \mathbb{C}^{n\times n}$ satisfies*

$$\left\|\mathsf{R}^{-1}\right\| \|\mathsf{R} - \mathsf{T}\| < 1,$$

then T is invertible.

Proof. Notice that

$$\mathsf{T} = \mathsf{R}(\mathsf{I}_n - (\mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T}));$$

therefore, T will be invertible provided that $\mathsf{I}_n - (\mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T})$ is invertible. Define $\mathsf{M} = \mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T}$ to conclude that, according to Theorem 4.19, we need $\|\mathsf{M}\| < 1$. Observe that

$$\|\mathsf{M}\| = \left\|\mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T}\right\| = \left\|\mathsf{R}^{-1}(\mathsf{R} - \mathsf{T})\right\| \le \left\|\mathsf{R}^{-1}\right\| \|\mathsf{R} - \mathsf{T}\| < 1,$$

and so T is invertible. □

With these results at hand we can give an estimate for the relative size of the error in the problem we were originally interested in. Let us first begin by assuming that $\delta\boldsymbol{f} = \mathbf{0}$.

Theorem 4.21 (relative error estimate, case $\delta\boldsymbol{f} = \mathbf{0}$). *Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be invertible, $\boldsymbol{f} \in \mathbb{C}^n$, and $\boldsymbol{x} \in \mathbb{C}^n$ solves (3.1). Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is the induced matrix norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$. Assume that $\delta\mathsf{A} \in \mathbb{C}^{n\times n}$ satisfies $\|\mathsf{A}^{-1}\delta\mathsf{A}\| < 1$ and that $\boldsymbol{x} + \delta\boldsymbol{x} \in \mathbb{C}^n$ solves the perturbed problem*

$$(\mathsf{A} + \delta\mathsf{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{f}.$$

Then $\delta\boldsymbol{x}$ is uniquely determined and

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \frac{\kappa(\mathsf{A})}{1 - \kappa(\mathsf{A})\frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}} \frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}.$$

Proof. Let us begin by repeating a previous computation. Since A is invertible, we can write $\mathsf{A} + \delta\mathsf{A} = \mathsf{A}(\mathsf{I}_n + \mathsf{A}^{-1}\delta\mathsf{A})$. Define $\mathsf{M} = -\mathsf{A}^{-1}\delta\mathsf{A}$, which satisfies $\|\mathsf{M}\| < 1$. Invoking Theorem 4.19 we conclude that $\mathsf{A} + \delta\mathsf{A}$ is invertible. Therefore, $\delta\boldsymbol{x}$ exists and is unique. In addition, we have

$$(\mathsf{A} + \delta\mathsf{A})^{-1} = (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{A}^{-1}$$

and $\left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \le \frac{1}{1-\|\mathsf{M}\|}$. Moreover, the obvious estimate

$$\|\mathsf{M}\| \le \left\|\mathsf{A}^{-1}\right\| \|\delta\mathsf{A}\|$$

implies that

$$\frac{1}{1 - \|\mathsf{M}\|} \le \frac{1}{1 - \left\|\mathsf{A}^{-1}\right\|\|\delta\mathsf{A}\|}.$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

Now

$$\begin{aligned}\delta\boldsymbol{x} &= (\mathsf{A}+\delta\mathsf{A})^{-1}\boldsymbol{f} - \mathsf{A}^{-1}\boldsymbol{f}\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{A}^{-1}\boldsymbol{f} - \mathsf{A}^{-1}\boldsymbol{f}\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}(\mathsf{A}^{-1}\boldsymbol{f} - (\mathsf{I}_n - \mathsf{M})\mathsf{A}^{-1}\boldsymbol{f})\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{M}\mathsf{A}^{-1}\boldsymbol{f}\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{M}\boldsymbol{x}.\end{aligned}$$

Consequently,

$$\|\delta\boldsymbol{x}\| \le \left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \ \|\mathsf{M}\| \ \|\boldsymbol{x}\| \le \frac{\left\|\mathsf{A}^{-1}\right\| \|\delta\mathsf{A}\|}{1 - \|\mathsf{A}^{-1}\| \, \|\delta\mathsf{A}\|}\|\boldsymbol{x}\| = \frac{\kappa(\mathsf{A})}{1-\kappa(\mathsf{A})\frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}} \frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}\|\boldsymbol{x}\|.$$

The result follows. □

To conclude our discussion, let us see what happens when we perturb both A and $\boldsymbol{f}$.

Theorem 4.22 (relative error estimate, general case)**.** *Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be invertible, $\boldsymbol{f} \in \mathbb{C}^n$, and $\boldsymbol{x} \in \mathbb{C}^n$ solves* (3.1)*. Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is an induced matrix norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$. Assume that $\delta\mathsf{A} \in \mathbb{C}^{n\times n}$ satisfies $\left\|\mathsf{A}^{-1}\delta\mathsf{A}\right\| < 1$, $\delta\boldsymbol{f} \in \mathbb{C}^n$ is given, and $\boldsymbol{x}+\delta\boldsymbol{x} \in \mathbb{C}^n$ satisfies the perturbed problem*

$$(\mathsf{A}+\delta\mathsf{A})(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{f}+\delta\boldsymbol{f}.$$

Then $\delta\boldsymbol{x}$ is uniquely determined and

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \frac{\kappa(\mathsf{A})}{1-\kappa(\mathsf{A})\frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}}\left(\frac{\|\delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|} + \frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}\right).$$

Proof. Let $\mathsf{M} = -\mathsf{A}^{-1}\delta\mathsf{A}$. We then have that $\boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{f}$ and $\boldsymbol{x}+\delta\boldsymbol{x} = (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{A}^{-1}(\boldsymbol{f}+\delta\boldsymbol{f})$. Therefore,

$$\begin{aligned}\delta\boldsymbol{x} &= (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{A}^{-1}(\boldsymbol{f}+\delta\boldsymbol{f}) - \mathsf{A}^{-1}\boldsymbol{f}\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}\left(\mathsf{A}^{-1}\boldsymbol{f} + \mathsf{A}^{-1}\delta\boldsymbol{f} - (\mathsf{I}_n - \mathsf{M})\mathsf{A}^{-1}\boldsymbol{f}\right)\\ &= (\mathsf{I}_n - \mathsf{M})^{-1}(\mathsf{A}^{-1}\delta\boldsymbol{f} + \mathsf{M}\mathsf{A}^{-1}\boldsymbol{f}).\end{aligned}$$

This shows that

$$\|\delta\boldsymbol{x}\| \le \frac{1}{1-\kappa(\mathsf{A})\frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}}\left(\left\|\mathsf{A}^{-1}\delta\boldsymbol{f}\right\| + \left\|\mathsf{M}\mathsf{A}^{-1}\boldsymbol{f}\right\|\right).$$

Notice also that

$$\left\|\mathsf{M}\mathsf{A}^{-1}\boldsymbol{f}\right\| = \|\mathsf{M}\boldsymbol{x}\| \le \|\mathsf{M}\|\|\boldsymbol{x}\| \le \left\|\mathsf{A}^{-1}\right\| \ \|\delta\mathsf{A}\| \ \|\boldsymbol{x}\| = \kappa(\mathsf{A})\frac{\|\delta\mathsf{A}\|}{\|\mathsf{A}\|}\|\boldsymbol{x}\|$$

and

$$\left\|\mathsf{A}^{-1}\delta\boldsymbol{f}\right\| \le \left\|\mathsf{A}^{-1}\right\| \|\delta\boldsymbol{f}\| \frac{\|\mathsf{A}\boldsymbol{x}\|}{\|\mathsf{A}\boldsymbol{x}\|} \le \kappa(\mathsf{A})\frac{\|\delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}\|\boldsymbol{x}\|.$$

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

The previous three inequalities, when combined, yield

$$\|\delta \boldsymbol{x}\| \leq \frac{\kappa(\mathsf{A})}{1-\kappa(\mathsf{A})\frac{\|\delta \mathsf{A}\|}{\|\mathsf{A}\|}}\left(\frac{\|\delta \boldsymbol{f}\|}{\|\boldsymbol{f}\|}+\frac{\|\delta \mathsf{A}\|}{\|\mathsf{A}\|}\right)\|\boldsymbol{x}\|,$$

as we intended to show. □

Problems

4.1 Does the spectral radius, introduced in Definition 4.1, define a norm?

4.2 Suppose that $\mathbb{V}$ is a finite-dimensional complex normed vector space. Show that if $\|\cdot\|$ is any induced operator norm, then $\rho(A) \leq \|A\|$ for all $A \in \mathcal{L}(\mathbb{V})$.

4.3 Prove Corollary 4.5.

4.4 Prove Proposition 4.10.

4.5 Prove, using induction on $n \in \mathbb{N}$, that if $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ commute, then they are simultaneously triangularizable, i.e., there is a nonsingular $\mathsf{P} \in \mathbb{C}^{n\times n}$ for which $\mathsf{P}^{-1}\mathsf{A}\mathsf{P}$ and $\mathsf{P}^{-1}\mathsf{B}\mathsf{P}$ are upper triangular.

Hint: Show that if $(\lambda, \boldsymbol{x})$ is an eigenpair of A, then so is $(\lambda, \mathsf{B}\boldsymbol{x})$.

4.6 Prove Corollary 4.12.

Hint: See the previous problem.

4.7 Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ and $\mu \in \mathbb{C}$ be such that $|\mu| > \rho(\mathsf{A})$. Show that the series

$$\sum_{k=0}^{\infty} \frac{1}{\mu^k}\mathsf{A}^k$$

converges to $(\mathsf{I}_n - \mu^{-1}\mathsf{A})^{-1}$.

4.8 Let $\mathsf{A} \in \mathbb{C}^{n\times n}$. Define

$$\mathsf{S}_k = \mathsf{I}_n + \mathsf{A} + \cdots + \mathsf{A}^k.$$

a) Prove that the sequence $\{\mathsf{S}_k\}_{k=0}^{\infty}$ converges if and only if A is convergent to zero.

b) Prove that if A is convergent to zero, then $\mathsf{I} - \mathsf{A}$ is nonsingular and

$$\lim_{k\to\infty} \mathsf{S}_k = (\mathsf{I} - \mathsf{A})^{-1} \ .$$

4.9 Show that if $\|\mathsf{A}\| < 1$ for some induced matrix norm, then $\mathsf{I} - \mathsf{A}$ is nonsingular and

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

4.10 Prove Proposition 4.14.

4.11 Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is the induced norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$. Show that if λ is an eigenvalue of $\mathsf{A}^{\mathsf{H}}\mathsf{A}$, where $\mathsf{A} \in \mathbb{C}^{n\times n}$, then

$$0 \leq \lambda \leq \left\|\mathsf{A}^{\mathsf{H}}\right\| \|\mathsf{A}\| .$$

4.12 Prove Proposition 4.16.

4.13 Complete the proof of Theorem 4.18.

4.14 Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be nonsingular. Show that the condition numbers $\kappa_\infty(\mathsf{A})$ and $\kappa_1(\mathsf{A})$ will not change after permutation of rows or columns.

https://doi.org/10.1017/9781108942607.007 Published online by Cambridge University Press

4.15 Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is a matrix norm and κ is the condition number defined with respect to it. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $0 \neq \alpha \in \mathbb{C}$. Show that $\kappa(\alpha A) = \kappa(A)$.
4.16 Show that if $Q \in \mathbb{C}^n$ is unitary, then $\kappa_2(Q) = 1$.
4.17 Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is the induced norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ and κ is the condition number defined with respect to this norm. Let $A \in \mathbb{C}^n$ be invertible. Show that

$$\kappa(A) \geq \frac{\max_{\lambda\in\sigma(A)} |\lambda|}{\min_{\lambda\in\sigma(A)} |\lambda|}.$$

4.18 Let $A = R^H R$ with $R \in \mathbb{C}^{n\times n}$ nonsingular. Give an expression for $\kappa_2(A)$ in terms of $\kappa_2(R)$.
4.19 Let $A \in \mathbb{C}^{n\times n}$ be invertible, $\boldsymbol{f} \in \mathbb{C}^n_\star$, and $\boldsymbol{x} \in \mathbb{C}^n$ solves (3.1). Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is the induced norm with respect to the vector norm $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$. Let the perturbations $\delta\boldsymbol{x}, \delta\boldsymbol{f} \in \mathbb{C}^n$ satisfy $A\delta\boldsymbol{x} = \delta\boldsymbol{f}$, so that $A(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{f} + \delta\boldsymbol{f}$.

a) Prove the error (or perturbation) estimate

$$\frac{1}{\kappa(A)}\frac{\|\delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|} \leq \frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \kappa(A)\frac{\|\delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}.$$

b) Show that, for any invertible matrix A, the upper bound for $\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$ above can be attained for suitable choices of $\boldsymbol{f}$ and $\delta\boldsymbol{f}$.

4.20 Show that, for every nonsingular $A \in \mathbb{C}^{n\times n}$, we have

$$\frac{1}{n} \leq \frac{\kappa_\infty(A)}{\kappa_2(A)} \leq n.$$

4.21 Let

$$A = \begin{bmatrix} 1.0000 & 2.0000 \\ 1.0001 & 2.0000 \end{bmatrix}.$$

a) Calculate $\kappa_1(A)$ and $\kappa_\infty(A)$.
b) Use (4.3) to obtain upper bounds on $\kappa_1(A)^{-1}$ and $\kappa_\infty(A)^{-1}$.
c) Suppose that you wish to solve $A\boldsymbol{x} = \boldsymbol{f}$, where $\boldsymbol{f} = \begin{bmatrix} 3.0000 \\ 3.0001 \end{bmatrix}$. Instead of $\boldsymbol{x}$ you obtain the approximation $\boldsymbol{x}' = \boldsymbol{x} + \delta\mathbf{x} = \begin{bmatrix} 0.0000 \\ 1.5000 \end{bmatrix}$. For this approximation you discover $\boldsymbol{f}' = \boldsymbol{f} + \delta\boldsymbol{f} = \begin{bmatrix} 3.0000 \\ 3.0000 \end{bmatrix}$, where $A\boldsymbol{x}' = \boldsymbol{f}'$. Calculate $\|\delta\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_1$ exactly. (You will need the exact solution, of course.) Then use the general estimate

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \kappa(A)\frac{\|\delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}$$

to obtain an upper bound for $\|\delta\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_1$. How good is $\|\delta\boldsymbol{f}\|_1/\|\boldsymbol{f}\|_1$ as indicator of the size of $\|\delta\boldsymbol{x}\|_1 / \|\boldsymbol{x}\|_1$?