Classical Numerical Analysis, Chapter 04

Abner J. Salgado and Steven M. Wise

asalgad1@utk.edu swise1@.utk.edu
University of Tennessee

Chapter 04
Norms and Matrix Conditioning

Perturbed Linear Systems and Condition Number

Suppose, for example, that A is invertible, and $\mathbf{x} \in \mathbb{C}^n$ is the solution to the (ideal) system

$$\mathsf{A}\mathbf{x} = \mathbf{f}.$$

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

$$(\mathsf{A} + \delta\mathsf{A})\left(\mathbf{x} + \delta\mathbf{x}\right) = \mathbf{f} + \delta\mathbf{f}.$$

Clearly, $\delta\mathbf{x}$ measures the error resulting from the perturbations to our data. How large is this error vector? How large is the relative error, $\frac{\|\delta \mathsf{x}\|}{\|\mathsf{x}\|}$? How does 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

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

The Spectral Radius

Spectral Radius

Definition

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

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

Analogously, for any square matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$, the **spectral radius** of A

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

Spectral Radius and Induced Norms

Proposition

Suppose $A \in \mathfrak{L}(\mathbb{V})$, where $\mathbb{V}$ is a complex n–dimensional normed vector space. Assume that $\|\cdot\|$ is the induced norm with respect to the base vector norm. Then,

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

Proof.

Let (λ, w) be an eigen-pair of the linear operator A. We can assume that $\|w\| = 1$. Then, using consistency of induced norms,

$$|\lambda| = \|\lambda w\| = \|Aw\| \leq \|A\| \|w\| = \|A\| .$$

Therefore,

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\| .$$

$\square$

Norm of a Self–Adjoint Operator in the Euclidean Norm

Theorem

Let $\mathbb{V}$ be an n–dimensional complex inner product space and suppose $\|x\| = (x, x)^{1/2}$ is the Euclidean norm. Let $A \in \mathfrak{L}(\mathbb{V})$ be self–adjoint. Then, the induced norm satisfies

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

Proof.

Since $A : \mathbb{V} \to \mathbb{V}$ is self–adjoint there exists an orthonormal basis of eigenvectors $S = \{e_1, \ldots, e_n\}$, i.e., $(e_i, e_j) = \delta_{i,j}$, $\mathbb{V} = \textbf{span}(S)$ 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

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

Proof, Cont.

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

$$\|Ax\| \leq \max\left\{|\lambda| | \lambda \in \sigma(A)\right\} \|x\|,$$

which implies

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

gives the reverse inequality, and this conclude the proof. $\square$

Corollary

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be Hermitian, that is,* $\mathsf{A} = \mathsf{A}^{\mathsf{H}}$*. Then*

$$\|\mathsf{A}\|_2 = \rho(\mathsf{A}).$$

Theorem (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*

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

Proof.

Let $\|\cdot\|_{\mathbb{C}^n}$ be the vector norm that induces $\|\cdot\|$. Since all vector norms are equivalent, this norm is equivalent to $\|\cdot\|_2$, that is, there are constants $0 < C_1 \leq C_2$, for which

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

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

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

where $C = C_2/C_1$. Taking supremum over $\mathbf{x} \in \mathbb{C}^n_\star$ implies the upper bound. □

Spectral Radius and Norms

Theorem

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,*

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

satisfies

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

Corollary (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 resulting induced matrix norm satisfies*

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

Matrix Convergence to Zero

Definition

We say that the square matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$ is **convergent to zero** iff $\mathsf{A}^k \to \mathsf{O} \in \mathbb{C}^{n\times n}$, this is, iff

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

for any matrix norm $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, where $\mathsf{O} \in \mathbb{C}^{n\times n}$ is the $n \times n$ matrix of zeros.

Since all norms on the vector space $\mathbb{C}^{m\times n}$, whether induced or not, are equivalent, what norm appears in this last definition is irrelevant.

Theorem (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* $\mathbf{x} \in \mathbb{C}^n$,
$$\lim_{k\to\infty} \mathsf{A}^k \mathbf{x} = \mathbf{0}.$$

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

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

Proof, Cont.

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

$$\rho^k(A) \to 0.$$

This implies $\rho(A) < 1$.

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

$$\|A\|_\star \leq \rho(A) + \varepsilon,$$

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

$$\|A\|_\star \leq \rho(A) + \varepsilon < 1.$$

Then, using sub–multiplicativity,

$$\left\|A^k\right\|_\star \leq \|A\|_\star^k \leq (\rho(A) + \varepsilon)^k \to 0.$$

Proof, Cont.

Consequently,

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

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

$$\left\| A^k \mathbf{x} \right\|_\infty \leq \left\| A^k \right\|_\infty \left\| \mathbf{x} \right\|_\infty \to 0,$$

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

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

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

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

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

$$\mathbf{y}^{\mathrm{H}} A^k \mathbf{x} \to 0.$$

Proof, Cont.

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

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

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. $\square$

Sufficient Condition for Matrix Convergence to Zero

Corollary (convergence condition)

Let $\mathsf{M} \in \mathbb{C}^{n\times n}$*, and 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 $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$,

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

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

Gelfand Relation

Proposition (upper bound)

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

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

Theorem (Gelfand relation)

Suppose that $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ *is an induced matrix norm. Then, for all* $\mathsf{A} \in \mathbb{C}^{n\times n}$*, we have*

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

Condition Number

Condition Number

Definition

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

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

Elementary Properties of Condition Number

Proposition (properties of κ)

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

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

Furthermore,

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

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

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

Proof.

Since the norm is of induced type, it satisfies the sub-multiplicative property. Thus

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

Proof, Cont.

But, the induced norm of the identity matrix is always 1. To see this, observe that

$$\|\mathsf{I}_n\| = \sup_{\mathbf{x}\in\mathbb{C}^n_\star} \frac{\|\mathsf{I}_n\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{x}\in\mathbb{C}^n_\star} \frac{\|\mathbf{x}\|}{\|\mathbf{x}\|} = 1.$$

To get the next inequality, since B is singular, there is a non-zero vector $\mathbf{w} \in \mathbb{C}^n$ such that $\mathsf{B}\mathbf{w} = \mathbf{0}$. Thus,

$$\mathbf{w} = \mathsf{A}^{-1}\mathsf{A}\mathbf{w} = \mathsf{A}^{-1}(\mathsf{A}-\mathsf{B})\mathbf{w}.$$

Using sub-multiplicativity and consistency,

$$\|\mathbf{w}\| \le \left\|\mathsf{A}^{-1}\right\| \|\mathsf{A}-\mathsf{B}\| \|\mathbf{w}\|.$$

Since $\|\mathbf{w}\| \neq 0$, by cancellation, we get

$$\frac{1}{\|\mathsf{A}^{-1}\|} \le \|\mathsf{A}-\mathsf{B}\|.$$

Proof, Cont.

Next, observe that, for any singular matrix B,

$$\frac{1}{\kappa(A)} \leq \frac{\|A - B\|}{\|A\|}.$$

Note that the left hand side is a lower bound for quotients on the left. The infimum is the greatest lower bound. Therefore,

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

□

Spectral Condition Number

Proposition

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* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$,
$$\kappa_2(\mathsf{A}) = \|\mathsf{A}\|_2 \|\mathsf{A}^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}.$$
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*
$$\kappa_2(\mathsf{A}) = \sqrt{\frac{\mu_n}{\mu_1}}. \tag{2}$$
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,*
$$\kappa_2(\mathsf{A}) \leq \sqrt{\kappa_1(\mathsf{A})\kappa_\infty(\mathsf{A})}.$$

Spectral Condition Number (Cont.)

Proposition (Cont.)

4. $$\frac{1}{\kappa_2(\mathsf{A})} = \inf_{\substack{\mathsf{B} \in \mathbb{C}^{n \times n} \\ \det(\mathsf{B})=0}} \frac{\|\mathsf{A}-\mathsf{B}\|_2}{\|\mathsf{A}\|_2}.$$

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 HPD with eigenvalues* $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ *then*
$$\kappa_2(\mathsf{A}) = \frac{\lambda_n}{\lambda_1}.$$

Proof.

Let us prove (1) and (4). To prove (1), we need only show that $\left\|\mathsf{A}^{-1}\right\|_2 = \sigma_n^{-1}$. Suppose that $\mathsf{A} = \mathsf{U}\boldsymbol{\Sigma}\mathsf{V}^{\mathsf{H}}$ is an SVD for A.

Proof, Cont.

Then $A^{-1} = V\Sigma^{-1}U^H$. It follows by unitary invariance that

$$\left\|A^{-1}\right\|_2 = \left\|V\Sigma^{-1}U^H\right\|_2 = \left\|\Sigma^{-1}\right\|_2 = \sigma_n^{-1},$$

noting that σ_n^{-1} is the largest diagonal element and, therefore, the largest diagonal element.

For (4), we already know from a previous result that

$$\frac{1}{\kappa_2(A)} \leq \inf_{\substack{B\in\mathbb{C}^{n\times n}\\ \det(B)=0}} \frac{\|A-B\|_2}{\|A\|_2}.$$

Suppose that $B = U\tilde{\Sigma}V^H$, where

$$\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n-1}, 0).$$

Thus, B is singular, with $\mathrm{rank}(B) = n - 1$. Then,

$$\frac{\|A-B\|_2}{\|A\|_2} = \frac{\sigma_n}{\sigma_1} = \frac{1}{\kappa_2(A)}.$$

Condition Number and Determinant

Example

We have at least two "practical" measures for how close a matrix is being singular. The first is $|\det(\mathsf{A})| \ll 1$. The second is $\kappa(\mathsf{A}) \gg 1$. But, unfortunately, these two measures can be wildly different, and it is difficult to know which measure to trust.

Let $\mathsf{A} \in \mathbb{R}^{n\times n}$ have the SVD

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

Then,

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

which is very small, for, say $n = 100$, but

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

which is not considered to be very large.

Residual Vector Versus Error Vector

Definition (error and residual)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is invertible and* $\mathbf{f} \in \mathbb{C}^n$. *Let* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. *The* **residual vector** *with respect to* $\mathbf{x}' \in \mathbb{C}^n$ *is defined as*

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

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

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

Consequently,

$$\mathsf{A}\mathbf{e} = \mathbf{r},$$

which is sometimes called the **error equation**.

Theorem (relative error estimate)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be invertible,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. 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*

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

Proof.

Let us prove the upper bound and leave the lower as an exercise. Since $\mathbf{e} = \mathsf{A}^{-1}\mathbf{r}$, using consistency of the induced norm

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

Likewise,

$$\|\mathbf{f}\| = \|\mathsf{A}\mathbf{x}\| \leq \|\mathsf{A}\| \|\mathbf{x}\| \implies \frac{1}{\|\mathbf{x}\|} \leq \|\mathsf{A}\| \frac{1}{\|\mathbf{f}\|}.$$

Combining the inequalities, we obtain the claimed upper bound. □

Practical Implications for Residual Calculations

Example

Suppose that

$$A = \begin{bmatrix} 1.0000 & 2.0000 \\ 1.0001 & 2.0000 \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} 3.0000 \\ 3.0001 \end{bmatrix}.$$

This matrix is almost singular. The condition number with respect to the infinity norm is $\kappa_\infty = 60002$, pretty large. The true solution to $A\mathbf{x} = \mathbf{f}$ is, of course, $\mathbf{x} = [1.0000, 1.0000]^\intercal$. Suppose you estimate the solution to be $\mathbf{x}' = [0.0000, 1.5000]^\intercal$. The error and residual are, respectively,

$$\mathbf{e} = \begin{bmatrix} 1.0000 \\ -0.5000 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0.0000 \\ 0.0001 \end{bmatrix}.$$

This is a big discrepancy. It results from the large condition number for A.

Practical Implications for Residual Calculations (Cont.)

Example (Cont.)

In this case, $\|\mathbf{e}\|_\infty = 1.0$ and $\|\mathbf{r}\|_\infty = 1.0 \times 10^{-4}$. Thus

$$\frac{\|\mathbf{e}\|_\infty}{\|\mathbf{x}\|_\infty} = 1 \qquad \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{f}\|_\infty} \approx 3.3332 \times 10^{-5}.$$

The last theorem guarantees that

$$1 = \frac{\|\mathbf{e}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \kappa_\infty(\mathsf{A}) \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{f}\|_\infty} \approx (60002) \times \left(3.3332 \times 10^{-5}\right) \approx 2.0000.$$

Moral: the fact that the residual is small in norm does not imply that the error will be small in norm.

Perturbations and Matrix Conditioning

Theorem (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,*

$$\|(\mathsf{I}_n - \mathsf{M})^{-1}\| \leq \frac{1}{1 - \|\mathsf{M}\|},$$

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.*

Proof.

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

$$\|(\mathsf{I}_n - \mathsf{M})\mathbf{x}\| \geq \left| \|\mathbf{x}\| - \|\mathsf{M}\mathbf{x}\| \right| \geq (1 - \|\mathsf{M}\|)\|\mathbf{x}\|.$$

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

Proof, Cont.

To obtain the norm estimate notice that

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

where we have used the reverse triangle inequality and sub–multiplicativity. The upper bound of the quantity $\|(\mathsf{I}_n - \mathsf{M})^{-1}\|$ 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 $N \to \infty$.

Proof, Cont.

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 $N \to \infty$,

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

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

The Set of Invertible Matrices is an Open Set!

Corollary

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* $\mathsf{T} \in \mathbb{C}^{n\times n}$ *satifies*

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

then T *is invertible.*

Proof.

Notice that

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

and, therefore, T will be invertible provided $\mathsf{I}_n - (\mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T})$ is invertible. Define $\mathsf{M} = \mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T}$. Then,

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

Using the last theorem, T is invertible, since $\mathsf{I}_n - (\mathsf{I}_n - \mathsf{R}^{-1}\mathsf{T})$ is invertible. $\square$

Theorem (relative error estimate, case $\delta\mathbf{f} = \mathbf{0}$)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be invertible,* $\mathbf{f} \in \mathbb{C}^n$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. 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* $\mathbf{x} + \delta\mathbf{x} \in \mathbb{C}^n$ *solves the perturbed problem*

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

Then $\delta\mathbf{x}$ *is uniquely determined and*

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

Proof.

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 a previous theorem, we conclude that $\mathsf{A} + \delta\mathsf{A}$ is invertible. Therefore, $\delta\mathbf{x}$ exists and is unique.

Proof, Cont.

In addition, we have

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

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

$$\|\mathsf{M}\| \leq \|\mathsf{A}^{-1}\|\|\delta\mathsf{A}\| < 1$$

implies

$$\|(\mathsf{I}_n - \mathsf{M})^{-1}\| \leq \frac{1}{1 - \|\mathsf{M}\|} \leq \frac{1}{1 - \|\mathsf{A}^{-1}\|\|\delta\mathsf{A}\|}.$$

Now

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

Proof, Cont.

Consequently,

$$\begin{aligned}
\|\delta \mathbf{x}\| &\leq \left\|(\mathsf{I}_n - \mathsf{M})^{-1}\right\| \; \|\mathsf{M}\| \; \|\mathbf{x}\| \\
&\leq \frac{\|\mathsf{A}^{-1}\|\|\delta \mathsf{A}\|}{1 - \|\mathsf{A}^{-1}\|\|\delta \mathsf{A}\|} \|\mathbf{x}\| \\
&= \frac{\kappa(\mathsf{A})}{1 - \kappa(\mathsf{A}) \frac{\|\delta \mathsf{A}\|}{\|\mathsf{A}\|}} \frac{\|\delta \mathsf{A}\|}{\|\mathsf{A}\|} \|\mathbf{x}\|.
\end{aligned}$$

The result follows after dividing by $\|\mathbf{x}\|$. □

Theorem (relative error estimate, general case)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be invertible,* $\mathbf{f} \in \mathbb{C}^n$, *and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. *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* $\left\|\mathsf{A}^{-1}\right\| \|\delta\mathsf{A}\| < 1$, $\delta\mathbf{f} \in \mathbb{C}^n$ *is given, and* $\mathbf{x} + \delta\mathbf{x} \in \mathbb{C}^n$ *satisfies the perturbed problem*

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

Then $\delta\mathbf{x}$ *is uniquely determined and*

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

Proof.

Let $\mathsf{M} = -\mathsf{A}^{-1}\delta\mathsf{A}$. Then, $\mathsf{I}_n - \mathsf{M}$ is invertible. Furthermore, $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$ and $\mathbf{x} + \delta\mathbf{x} = (\mathsf{I}_n - \mathsf{M})^{-1}\mathsf{A}^{-1}(\mathbf{f} + \delta\mathbf{f})$.

Proof, Cont.

Therefore,

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

This shows that

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

Notice also that

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

and

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

Proof, Cont.

The previous three inequalities, when combined, yield

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

as we intended to show. □

Computations with Ill Conditioned Matrices: $n = 10$

```
>>  n = 10;
    A = rand(n);
    x = ones(n,1);
    f = A*x;
    [xx, err] = Solve(A,f);
    norm(x-xx)

ans = 2.4651e-14

>>  A = hilb(n);
    x = ones(n,1);
    f = A*x;
    [xx, err] = Solve(A,f);
    norm(x-xx)

ans = 0.0015
```

Computations with Ill Conditioned Matrices: $n = 100$

```
>>  n = 100;
    A = rand(n);
    x = ones(n,1);
    f = A*x;
    [xx, err] = Solve(A,f);
    norm(x-xx)

ans = 5.4012e-11

>>  A = hilb(n);
    x = ones(n,1);
    f = A*x;
    [xx, err] = Solve(A,f);
    norm(x-xx)

ans = 1.6254e+04
```