MATH 5524: Matrix Theory

Working Notes

Mark Embree

Virginia Tech

Draft of 29 March 2017

1 · Basic Spectral Theory

Matrices prove so useful in applications because of the insight one gains from eigenvalues and eigenvectors. A first course in matrix theory should thus be devoted to basic *spectral theory*: the development of eigenvalues, eigenvectors, diagonalization, and allied concepts. This material – buttressed by the foundational ideas of subspaces, bases, ranges and null spaces – typically fills the entire course, and a semester ends before students have much opportunity to reap the rich harvest that mathematicians, scientists, and engineers grow from the seeds of spectral theory.

Our purpose here is to concisely recapitulate the highlights of a first course, then build up the theory in a variety of directions, each of which is illustrated by a practical application from physical science, engineering, or social science.

One can build up the spectral decomposition of a matrix through two quite distinct – though ultimately equivalent – routes, one "analytic" (for it develops fundamental quantities in terms of contour integration in the complex plane), the other "algebraic" (for it is grounded in algebraic notions of nested invariant subspaces). We shall tackle both approaches here, for each provides distinct insights in a variety of situations (as will be evidenced throughout the present course), and we firmly believe that the synthesis that comes from reconciling the two perspectives deeply enriches one's understanding.

Before embarking on this our discussion of spectral theory, we first must pause to establish notational conventions and recapitulate basic concepts from elementary linear algebra.

1.1 Notation and Preliminaries

Our basic building blocks are complex numbers (*scalars*), which we write as italicized Latin or Greek letters, e.g., $z, \zeta \in \mathbb{C}$.

From these scalar numbers we build column vectors of length n, denoted by lower-case bold-faced letters, e.g., $\mathbf{v} \in \mathbb{C}^n$. The jth entry of $\mathbf{v}$ is denoted by $v_j \in \mathbb{C}$. (Sometimes we will emphasize that a scalar or vector has real-valued entries: $\mathbf{v} \in \mathbb{R}^n$, $v_j \in \mathbb{R}$.)

A set of vectors $\mathcal{S}$ is a *subspace* provided it is closed under vector addition and scalar multiplication: (i) if $\mathbf{x}, \mathbf{y} \in \mathcal{S}$, then $\mathbf{x} + \mathbf{y} \in \mathcal{S}$ and (ii) if $\mathbf{x} \in \mathcal{S}$ and $\alpha \in \mathbb{C}$, then $\alpha\mathbf{x} \in \mathcal{S}$.

A set of vectors is *linearly independent* provided no one vector in that set can be written as a nontrivial linear combination of the other vectors in the set; equivalently, the zero vector cannot be written as a nontrivial linear combination of vectors in the set.

The *span* of a set of vectors is the set of all linear combinations:

$$\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} = \{\gamma_1\mathbf{v}_1 + \cdots + \gamma_d\mathbf{v}_d : \gamma_1, \ldots, \gamma_d \in \mathbb{C}\}.$$

The span is always a subspace.

A *basis* for a subspace $\mathcal{S}$ is a smallest collection of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} \subseteq \mathcal{S}$ whose span equals all of $\mathcal{S}$; bases are not unique, but every basis for $\mathcal{S}$ contains the same number of vectors; we call this number the *dimension* of the subspace $\mathcal{S}$, written $\dim(\mathcal{S})$. If $\mathcal{S} \subseteq \mathbb{C}^n$, then $\dim(\mathcal{S}) \le n$.

We shall often approach matrix problems *analytically*, which implies we have some way to measure *distance*. One can advance this study to a high art, but for the moment we will take a familiar approach. (A more general discussion follows in Chapter 4.) We shall impose some geometry, a system for measuring lengths and angles. To extend Euclidean geometry naturally to $\mathbb{C}^n$, we define the *inner product* between two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ to be

$$\mathbf{u}^*\mathbf{v} := \sum_{j=1}^{n} \overline{u_j} v_j,$$

where $\mathbf{u}^*$ denotes the conjugate-transpose of $\mathbf{u}$:

$$\mathbf{u}^* = [\,\overline{u_1}, \overline{u_2}, \ldots, \overline{u_n}\,] \in \mathbb{C}^{1\times n},$$

a row vector made up of the complex-conjugates of the individual entries of $\mathbf{u}$. (Occasionally it is convenient to turn a column vector into a row vector without conjugating the entries; we write this using the *transpose*: $\mathbf{u}^{\mathrm{T}} = [\,u_1, u_2, \ldots, u_n\,]$. Note that $\mathbf{u}^* = \overline{\mathbf{u}}^{\mathrm{T}}$, and if $\mathbf{u} \in \mathbb{R}^n$, then $\mathbf{u}^* = \mathbf{u}^{\mathrm{T}}$.)

Euclidean vector geometry: norms and angles between vectors

The inner product provides a notion of magnitude, or *norm*, of $\mathbf{v} \in \mathbb{C}^n$:

$$\|\mathbf{v}\| = \Big(\sum_{j=1}^{n} |v_j|^2\Big)^{1/2} = \sqrt{\mathbf{v}^*\mathbf{v}}.$$

A vector of norm 1 is called a *unit vector*.

From the definition of the norm immediately follows the *positivity* property,

- $\|\mathbf{v}\| \ge 0$, with $\|\mathbf{v}\| = 0$ only when $\mathbf{v} = \mathbf{0}$

and the *scaling* property,

- $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$ for all $\alpha \in \mathbb{C}$.

The norm also obeys the *triangle inequality*,

- $\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$ ('triangle inequality').

The proof of the triangle inequality is not immediately obvious. The standard approach relies on another critical relationship between the inner product and norm, the *Cauchy–Schwarz inequality*.

Lemma 1.1 (Cauchy–Schwarz inequality). *For any* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$,

$$|\mathbf{u}^*\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|. \tag{1.1}$$

Proof. Enough proofs of this inequality exist to fill a book [Ste04]. We shall follow a standard proof for general inner products [You88, Ch. 1].

If $\mathbf{u}^*\mathbf{v} = 0$, the result is trivial. Otherwise, express the (complex) scalar $\mathbf{u}^*\mathbf{v}$ in the polar form $\mathbf{u}^*\mathbf{v} = \mathrm{e}^{\mathrm{i}\theta}|\mathbf{u}^*\mathbf{v}|$ for some $\theta \in [0, 2\pi)$. Now, for any $t \in \mathbb{C}$, note that

$$\begin{aligned} 0 \le \|\mathbf{u} + t\mathbf{v}\|^2 &= \mathbf{u}^*\mathbf{u} + \mathbf{u}^*(t\mathbf{v}) + (t\mathbf{v})^*\mathbf{u} + (t\mathbf{v})^*(t\mathbf{v}) \\ &= \mathbf{u}^*\mathbf{u} + t\mathbf{u}^*\mathbf{v} + \overline{t\mathbf{u}^*\mathbf{v}} + (t\mathbf{v})^*(t\mathbf{v}) \\ &= \|\mathbf{u}\|^2 + 2\mathrm{Re}(t\mathbf{u}^*\mathbf{v}) + |t|^2\|\mathbf{v}\|^2, \end{aligned}$$

where we have used the fact that $z + \overline{z} = 2\mathrm{Re}(z)$ for any $z \in \mathbb{C}$. Now set t to have the form $t = \mathrm{e}^{-\mathrm{i}\theta}r$ for some $r \in \mathbb{R}$. Now recalling that $\mathbf{u}^*\mathbf{v} = \mathrm{e}^{\mathrm{i}\theta}|\mathbf{u}^*\mathbf{v}|$, we have

$$\mathrm{Re}(t\mathbf{u}^*\mathbf{v}) = \mathrm{Re}(\mathrm{e}^{-\mathrm{i}\theta}r\mathrm{e}^{\mathrm{i}\theta}|\mathbf{u}^*\mathbf{v}|) = r|\mathbf{u}^*\mathbf{v}| \in \mathbb{R},$$

and $|t|^2 = r^2$, so

$$0 \le \|\mathbf{u}\|^2 + 2r|\mathbf{u}^*\mathbf{v}| + r^2\|\mathbf{v}\|^2.$$

Notice that the right-hand side is *a quadratic equation in the variable r with real coefficients*. In particular, as $r \in \mathbb{R}$ is varied, the right-hand side traces out a parabola, opening up. Since this parabola is bounded below by 0, it must have either (a) a double root, or (b) a complex-conjugate pair of roots. In particular, this means that the *discriminant* (the "$b^2 - 4ac$" term in the quadratic formula) must be less than or equal to zero:

$$0 \ge (2|\mathbf{u}^*\mathbf{v}|)^2 - 4\|\mathbf{u}\|^2\|\mathbf{v}\|^2,$$

which can be rearranged to yield the inequality $|\mathbf{u}^*\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|$. □

With the Cauchy–Schwarz inequality at our disposal, it is simple to prove the triangle inequality. The standard proof (see [You88, Ch.]) goes as follows. Given $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$, expand $\|\mathbf{u}+\mathbf{v}\|^2$ to obtain

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\mathrm{Re}(\mathbf{u}^*\mathbf{v}) + \|\mathbf{v}\|^2.$$

Using the fact that $\mathrm{Re}(z) \le |z|$ for any $z \in \mathbb{C}$, along with the Cauchy–Schwarz inequality,

$$\begin{aligned} \|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2 + 2|\mathbf{u}^*\mathbf{v}| + \|\mathbf{v}\|^2 \\ &\le |\mathbf{u}\|^2 + 2\|\mathbf{u}\|\,\|\mathbf{v}\| + \|\mathbf{v}\|^2 \;=\; (\|\mathbf{u}\| + \|\mathbf{v}\|)^2. \end{aligned}$$

Then take square roots to get the triangle inequality: $\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$.

The Cauchy–Schwarz inequality gives us a way to establish a *geometry* for $\mathbb{C}^n$. One can show that equality holds in (1.1) if and only if $\mathbf{u}$ and $\mathbf{v}$ are collinear (one can be written as a scaling of the other). Otherwise, the quantity

$$\frac{|\mathbf{u}^*\mathbf{v}|}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \in [0,1]$$

can be regarded as a measure of the alignment of the vectors $\mathbf{u}$ and $\mathbf{v}$. More specifically, the acute *angle* between $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ is defined to be

$$\angle(\mathbf{u},\mathbf{v}) := \cos^{-1}\Big(\frac{|\mathbf{u}^*\mathbf{v}|}{\|\mathbf{u}\|\|\mathbf{v}\|}\Big) \in [0, \pi/2].$$

(If $\mathbf{u} = \mathbf{0}$ or $\mathbf{v} = \mathbf{0}$, $\angle(\mathbf{u},\mathbf{v})$ is undefined.)

A notable special case occurs when $\angle(\mathbf{u},\mathbf{v}) = \pi/2$, meaning $\mathbf{u}^*\mathbf{v} = 0$. Any $\mathbf{u}$ and $\mathbf{v}$ for which $\mathbf{u}^*\mathbf{v} = 0$ are said to be *orthogonal*, written $\mathbf{u} \perp \mathbf{v}$.

Orthogonal vectors obey a special relationship that is useful in many settings, the *Pythagorean Theorem*: If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, $\mathbf{u}^*\mathbf{v} = 0$, then

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

The proof follows immediately from expanding $\|\mathbf{u}+\mathbf{v}\|^2$.

A set comprising mutually orthogonal unit vectors is called an *orthonormal set*. Two sets $\mathcal{U}, \mathcal{V} \subseteq \mathbb{C}^n$ are *orthogonal* provided $\mathbf{u} \perp \mathbf{v}$ for all $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$. The *orthogonal complement* of a set $\mathcal{V} \subseteq \mathbb{C}^n$ is the set of vectors orthogonal to all $\mathbf{v} \in \mathcal{V}$, denoted

$$\mathcal{V}^\perp := \{\mathbf{u} \in \mathbb{C}^n : \mathbf{u}^*\mathbf{v} = 0 \text{ for all } \mathbf{v} \in \mathcal{V}\}.$$

The *sum* of two subspaces $\mathcal{U}, \mathcal{V} \subseteq \mathbb{C}^n$ is given by

$$\mathcal{U} + \mathcal{V} = \{\mathbf{u}+\mathbf{v} : \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}\}.$$

A special notation is used for this sum when the two subspaces intersect trivially, $\mathcal{U} \cap \mathcal{V} = \{\mathbf{0}\}$; we call this a *direct sum*, and write

$$\mathcal{U} \oplus \mathcal{V} = \{\mathbf{u}+\mathbf{v} : \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}\}.$$

This extra notation is justified by an important consequence: each $\mathbf{x} \in \mathcal{U} \oplus \mathcal{V}$ can be written *uniquely* as $\mathbf{x} = \mathbf{u} + \mathbf{v}$ for $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$.

Matrix notation, submatrix multiplication, the fundamental subspaces

Matrices with m rows and n columns of scalars are denoted by a bold capital letter, e.g., $\mathbf{A} \in \mathbb{C}^{m\times n}$. The (j,k) entry of $\mathbf{A} \in \mathbb{C}^{m\times n}$ is written as $a_{j,k}$; It is often more useful to access $\mathbf{A}$ by its n columns, $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Thus, we write $\mathbf{A} \in \mathbb{C}^{3\times 2}$ as

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{bmatrix} = [\,\mathbf{a}_1 \quad \mathbf{a}_2\,].$$

We have the conjugate-transpose and transpose of matrices: the (j,k) entry of $\mathbf{A}^*$ is $\overline{a_{k,j}}$, while the (j,k) entry of $\mathbf{A}^{\mathrm{T}}$ is $a_{k,j}$. For our 3×2 matrix $\mathbf{A}$,

$$\mathbf{A}^* = \begin{bmatrix} \overline{a_{1,1}} & \overline{a_{2,1}} & \overline{a_{3,1}} \\ \overline{a_{1,2}} & \overline{a_{2,2}} & \overline{a_{3,2}} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^* \\ \mathbf{a}_2^* \end{bmatrix}, \qquad \mathbf{A}^{\mathrm{T}} = \begin{bmatrix} a_{1,1} & a_{2,1} & a_{3,1} \\ a_{1,2} & a_{2,2} & a_{3,2} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \end{bmatrix}.$$

Through direct computation, one can verify that

$$(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*, \qquad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}.$$

We often will build matrices out of vectors and submatrices, and to multiply such matrices together, block-by-block, in the same manner we multiply matrices entry-by-entry. For example, if $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$ and $\mathbf{v} \in \mathbb{C}^n$, then

$$[\,\mathbf{A} \quad \mathbf{B}\,]\,\mathbf{v} = [\,\mathbf{Av} \quad \mathbf{Bv}\,] \in \mathbb{C}^{m\times 2}.$$

Another pair of examples might be helpful. For example, if

$$\mathbf{A} \in \mathbb{C}^{m\times k}, \quad \mathbf{B} \in \mathbb{C}^{m\times \ell}, \quad \mathbf{C} \in \mathbb{C}^{k\times n}, \quad \mathbf{D} \in \mathbb{C}^{\ell\times n},$$

then we have

$$[\,\mathbf{A} \quad \mathbf{B}\,]\begin{bmatrix}\mathbf{C}\\ \mathbf{D}\end{bmatrix} = [\,\mathbf{AC}+\mathbf{BD}\,] \in \mathbb{C}^{m\times n},$$

while if

$$\mathbf{A} \in \mathbb{C}^{m\times k}, \quad \mathbf{B} \in \mathbb{C}^{n\times k}, \quad \mathbf{C} \in \mathbb{C}^{k\times p}, \quad \mathbf{D} \in \mathbb{C}^{k\times q},$$

then

$$\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}[\,\mathbf{C} \quad \mathbf{D}\,] = \begin{bmatrix}\mathbf{AC} & \mathbf{AD}\\ \mathbf{BC} & \mathbf{BD}\end{bmatrix} \in \mathbb{C}^{(m+n)\times(p+q)}.$$

The *range* (or *column space*) of $\mathbf{A} \in \mathbb{C}^{m\times n}$ is denoted by $\mathcal{R}(\mathbf{A})$:

$$\mathcal{R}(\mathbf{A}) := \{\mathbf{Ax} : \mathbf{x} \in \mathbb{C}^n\} \subseteq \mathbb{C}^m.$$

The *null space* (or *kernel*) of $\mathbf{A} \in \mathbb{C}^{m\times n}$ is denoted by $\mathcal{N}(\mathbf{A})$:

$$\mathcal{N}(\mathbf{A}) := \{\mathbf{x} \in \mathbb{C}^n : \mathbf{Ax} = \mathbf{0}\} \subseteq \mathbb{C}^n.$$

The range and null space of $\mathbf{A}$ are always subspaces. The span of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{C}^m$ equals the range of the matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_n$:

$$\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \mathcal{R}(\mathbf{V}), \qquad \mathbf{V} = [\,\mathbf{v}_1, \ldots, \mathbf{v}_n\,] \in \mathbb{C}^{m\times n}.$$

The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent provided $\mathcal{N}(\mathbf{V}) = \{\mathbf{0}\}$.

With any matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ we associate ‘four fundamental subspaces’: $\mathcal{R}(\mathbf{A})$, $\mathcal{N}(\mathbf{A})$, $\mathcal{R}(\mathbf{A}^*)$, and $\mathcal{N}(\mathbf{A}^*)$. These spaces are related in a beautiful manner that STRANG calls the *Fundamental Theorem of Linear Algebra* [Str93]: For any $\mathbf{A} \in \mathbb{C}^{m\times n}$,

$$\begin{aligned}\mathcal{R}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}^*) &= \mathbb{C}^m, \qquad \mathcal{R}(\mathbf{A}) \perp \mathcal{N}(\mathbf{A}^*)\\ \mathcal{R}(\mathbf{A}^*) \oplus \mathcal{N}(\mathbf{A}) &= \mathbb{C}^n, \qquad \mathcal{R}(\mathbf{A}^*) \perp \mathcal{N}(\mathbf{A}).\end{aligned}$$

Hence, for example, each $\mathbf{v} \in \mathbb{C}^m$ can be written *uniquely* as $\mathbf{v} = \mathbf{v}_R + \mathbf{v}_N$, for orthogonal vectors $\mathbf{v}_R \in \mathcal{R}(\mathbf{A})$ and $\mathbf{v}_N \in \mathcal{N}(\mathbf{A}^*)$.

A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is *invertible* provided that given any $\mathbf{b} \in \mathbb{C}^n$, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a *unique* solution $\mathbf{x}$, which we write as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$; $\mathbf{A}$ is invertible provided its range coincides with the entire ambient space, $\mathcal{R}(\mathbf{A}) = \mathbb{C}^n$, or equivalently, its null space is trivial, $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$. A square matrix that is not invertible is said to be *singular*. Strictly rectangular matrices $\mathbf{A} \in \mathbb{C}^{m\times n}$ with $m \neq n$ are never invertible. When the inverse exists, it is unique. Thus for invertible $\mathbf{A} \in \mathbb{C}^{n\times n}$, we have $(\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})^*$, and if both $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ are invertible, then so too is their product:

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

The induced matrix norm

Given the ability to measure the lengths of vectors, we can immediately gauge the magnitude of a matrix by the maximum amount it stretches a vector. For any $\mathbf{A} \in \mathbb{C}^{m\times n}$, the *matrix norm* is defined (for now) as

$$\|\mathbf{A}\| := \max_{\substack{\mathbf{v}\in\mathbb{C}^n \\ \mathbf{v}\neq\mathbf{0}}} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\substack{\mathbf{v}\in\mathbb{C}^n \\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\|. \tag{1.2}$$

The matrix norm enjoys properties similar to the vector norm:

- $\|\mathbf{A}\| \geq 0$, and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$ (*positivity*);
- $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$ for all $\alpha \in \mathbb{C}$ (*scaling*);
- $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (*triangular inequality*).

As in the vector case, the proof of the triangle inequality merits a closer look. Start with the definition of the matrix norm and apply the vector triangle inequality:

$$\|\mathbf{A}+\mathbf{B}\| = \max_{\substack{\mathbf{v}\in\mathbb{C}^n \\ \|\mathbf{v}\|=1}} \|(\mathbf{A}+\mathbf{B})\mathbf{v}\| \leq \max_{\substack{\mathbf{v}\in\mathbb{C}^n \\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\| + \|\mathbf{B}\mathbf{v}\|.$$

Suppose this maximization is attained by some vector, call it $\mathbf{x}$:

$$\|\mathbf{A}+\mathbf{B}\| \leq \max_{\substack{\mathbf{v}\in\mathbb{C}^n \\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\| + \|\mathbf{B}\mathbf{v}\| = \|\mathbf{A}\mathbf{x}\| + \|\mathbf{B}\mathbf{x}\|. \tag{1.3}$$

Now this $\mathbf{x}$ need not be the vector that maximizes $\|\mathbf{A}\mathbf{v}\|$ over all choices for the unit vector $\mathbf{v}$, so

$$\|\mathbf{A}\mathbf{x}\| \le \max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\|, \qquad \|\mathbf{B}\mathbf{x}\| \le \max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \|\mathbf{B}\mathbf{v}\|.$$

Inserting these inequalities on the right-hand side of (1.3) gives

$$\|\mathbf{A}+\mathbf{B}\| \le \Big(\max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\|\Big) + \Big(\max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \|\mathbf{A}\mathbf{v}\|\Big) = \|\mathbf{A}\| + \|\mathbf{B}\|. \tag{1.4}$$

We would not usually be so deliberate about this proof, but students sometimes get tripped up over the move from a single maximization in (1.3) and the two maximizations in (1.4). Explicitly introducing the vector $\mathbf{x}$ makes this transition more transparent.

Matrix norms obey several other essential inequalities. Given the maximization in (1.2), it immediately follows that for *any* $\mathbf{v} \in \mathbb{C}^n$,

$$\|\mathbf{A}\mathbf{v}\| \le \|\mathbf{A}\|\|\mathbf{v}\|.$$

This result even holds when $\mathbf{v} \in \mathbb{C}^n$ is replaced by any matrix $\mathbf{B} \in \mathbb{C}^{n\times k}$,

$$\|\mathbf{A}\mathbf{B}\| \le \|\mathbf{A}\|\|\mathbf{B}\|. \tag{1.5}$$

This is the *submultiplicative* property of the matrix norm. (Prove it!)

Important matrices

Several special classes of matrices play a particularly distinguished role.

- A matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ is *diagonal* provided all off-diagonal elements are zero: $a_{j,k} = 0$ if $j \ne k$. We say $\mathbf{A} \in \mathbb{C}^{m\times n}$ is *upper triangular* provided all entries below the main diagonal are zero, $a_{j,k} = 0$ if $j > k$; *lower triangular* matrices are defined similarly.

- A square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is *Hermitian* provided $\mathbf{A}^* = \mathbf{A}$, and *symmetric* provided $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$. (For real matrices, these notions are identical, and such matrices are usually called 'symmetric' in the literature. In the complex case, Hermitian matrices are both far more common and far more convenient than symmetric matrices – though the latter do arise in certain applications, like scattering problems.) One often also encounters *skew-Hermitian* and *skew-symmetric* matrices, where $\mathbf{A}^* = -\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}} = -\mathbf{A}$.

- A square matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ is *unitary* provided that $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, which is simply a concise way of stating that the columns of $\mathbf{U}$ are orthonormal. Since $\mathbf{U}$ is square (it has n columns, each in $\mathbb{C}^n$), the columns of $\mathbf{U}$ form an *orthonormal basis* for $\mathbb{C}^n$: this fact makes unitary matrices so important. Because the inverse of a matrix is unique, $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ implies that $\mathbf{U}\mathbf{U}^* = \mathbf{I}$.

- Often we encounter $k < n$ orthonormal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{C}^n$. Defining $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in \mathbb{C}^{n\times k}$, we see that $\mathbf{U}^*\mathbf{U} = \mathbf{I} \in \mathbb{C}^{k\times k}$. Since $\mathbf{U}$ is not square, it has no inverse; in fact, we must have $\mathbf{U}\mathbf{U}^* \neq \mathbf{I} \in \mathbb{C}^{n\times n}$. Though they arise quite often, there is no settled name for matrices with $k < n$ orthonormal columns; we like the term *subunitary* best.

Premultiplication of a vector by a unitary or subunitary matrix leaves the 2-norm unchanged, for if $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, then

$$\|\mathbf{U}\mathbf{x}\|^2 = (\mathbf{U}\mathbf{x})^*(\mathbf{U}\mathbf{x}) = \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|^2.$$

The intuition behind the algebra is that $\mathbf{U}\mathbf{x}$ is just a representation of $\mathbf{x}$ in a different orthonormal basis, and that change of basis should not affect the magnitude of $\mathbf{x}$. Moreover, if $\mathbf{U} \in \mathbb{C}^{k\times m}$ is unitary or subunitary, and $\mathbf{V} \in \mathbb{C}^{n\times n}$ is unitary, then for any $\mathbf{A} \in \mathbb{C}^{m\times n}$,

$$\|\mathbf{U}\mathbf{A}\mathbf{V}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{U}\mathbf{A}\mathbf{V}\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{V}\mathbf{x}\| = \max_{\substack{\|\mathbf{y}\|=1\\ \mathbf{y}=\mathbf{V}\mathbf{x}}} \|\mathbf{A}\mathbf{y}\| = \|\mathbf{A}\|.$$

These properties, $\|\mathbf{U}\mathbf{x}\| = \|\mathbf{x}\|$ and $\|\mathbf{U}\mathbf{A}\mathbf{V}\| = \|\mathbf{A}\|$, are collectively known as the *unitary invariance* of the norm.

- A square matrix $\mathbf{P} \in \mathbb{C}^{n\times n}$ is a *projector* provided $\mathbf{P}^2 = \mathbf{P}$ (note that powers mean repeated multiplication: $\mathbf{P}^2 := \mathbf{P}\mathbf{P}$). Projectors play a fundamental role in the spectral theory to follow. We say that $\mathbf{P}$ *projects onto* $\mathcal{R}(\mathbf{P})$ and *along* $\mathcal{N}(\mathbf{P})$. Notice that if $\mathbf{u} \in \mathcal{R}(\mathbf{P})$, then $\mathbf{P}\mathbf{u} = \mathbf{u}$; it follows that $\|\mathbf{P}\| \geq 1$. When the projector is Hermitian, $\mathbf{P} = \mathbf{P}^*$, the Fundamental Theorem of Linear Algebra implies that $\mathcal{R}(\mathbf{P}) \perp \mathcal{N}(\mathbf{P}^*) = \mathcal{N}(\mathbf{P})$, and so $\mathbf{P}$ is said to be an *orthogonal projector*: the space it *projects onto* is orthogonal to the space it *projects along.* In this case, $\|\mathbf{P}\| = 1$. The orthogonal projector provides an ideal framework for characterizing best approximations from a subspace: for any $\mathbf{x} \in \mathbb{C}^{n\times n}$ and orthogonal projector $\mathbf{P}$,

$$\min_{\mathbf{u}\in\mathcal{R}(\mathbf{P})} \|\mathbf{x} - \mathbf{u}\| = \|\mathbf{x} - \mathbf{P}\mathbf{x}\|.$$

A projector that is not orthogonal (i.e., $\mathbf{P} \neq \mathbf{P}^*$) is called *oblique*.

Various projectors will arise throughout this course. The simplest examples project onto a single dimension.

- If $\mathbf{u} \in \mathbb{C}^n$ is nonzero, then
$$\mathbf{P} := \frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}} = \Big(\frac{\mathbf{u}}{\|\mathbf{u}\|}\Big)\Big(\frac{\mathbf{u}}{\|\mathbf{u}\|}\Big)^*$$
is the *orthogonal* projector with $\mathcal{R}(\mathbf{P}) = \operatorname{span}\{\mathbf{u}\}$.
- If $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ are *not orthogonal* ($\mathbf{v}^*\mathbf{u} \neq 0$), then
$$\mathbf{P} := \frac{\mathbf{u}\mathbf{v}^*}{\mathbf{v}^*\mathbf{u}}$$
is the (generally *oblique*) projector *onto* $\mathcal{R}(\mathbf{P}) = \operatorname{span}\{\mathbf{u}\}$ and *along* $\mathcal{N}(\mathbf{P}) = \operatorname{span}\{\mathbf{v}\}^\perp$.

Both of these projectors onto one-dimensional subspaces have natural generalizations to higher dimensions. (In fact, all projectors can be written in one of these two following ways.)

- If $\mathbf{U} \in \mathbb{C}^{n\times k}$ is a subunitary matrix ($\mathbf{U}^*\mathbf{U} = \mathbf{I}$), then
$$\mathbf{P} := \mathbf{U}\mathbf{U}^*$$
is the *orthogonal* projector onto $\mathcal{R}(\mathbf{P}) = \mathcal{R}(\mathbf{U})$.
- If $\mathbf{U}, \mathbf{V} \in \mathbb{C}^{n\times k}$ with $\mathbf{V}^*\mathbf{U} = \mathbf{I}$ (we say the columns of $\mathbf{U}$ and $\mathbf{V}$ are *biorthogonal*), then
$$\mathbf{P} := \mathbf{U}\mathbf{V}^*$$
is the (generally *oblique*) projector onto $\mathcal{R}(\mathbf{P}) = \mathcal{R}(\mathbf{U})$ and along $\mathcal{N}(\mathbf{P}) = \mathcal{N}(\mathbf{V}^*) = \mathcal{R}(\mathbf{V})^\perp$.

Basic metric space properties

In this concise introduction we shall not go into lengthy detail about metric space properties of $\mathbb{C}^n$ and $\mathbb{C}^{n\times n}$ endowed with the norms described earlier in this chapter. However, a few properties are important to mention and used in proofs later.

Definition 1.2. *The sequence $\{\mathbf{x}_k\}_{k=1}^{\infty} \subset \mathbb{C}^n$ converges to $\mathbf{x} \in \mathbb{C}^n$ provided that, for any $\varepsilon > 0$, there exists some positive integer N_ε such that for all $k > N_\varepsilon$, $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon$. We write $\mathbf{x}_k \to \mathbf{x}$ as $k \to \infty$.*

To show that $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is convergent directly via this definition, one must identify the target $\mathbf{x}$ to which the sequence converges. Often it is much easier to simply show that all the vectors in the sequence are getting closer together, in the formal precise sense.

Definition 1.3. *The vectors $\{\mathbf{x}_k\}_{k=1}^{\infty} \subset \mathbb{C}^n$ form a Cauchy sequence provided that, for any $\varepsilon > 0$, there exists a positive integer N_ε for which $\|\mathbf{x}_j - \mathbf{x}_k\| < \varepsilon$ for all $j, k > N_\varepsilon$.*

Suppose you can show that $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Must there exist some $\mathbf{x} \in \mathbb{C}^n$ to which in converges? In finite dimensions (like $\mathbb{C}^n$) the answer is alway *yes.* (In infinite dimensions the question is more subtle, as you learn in a functional analysis class.) Normed vector spaces in which all Cauchy sequences converge are called *complete*.

Fact 1.4. *Suppose $\{\mathbf{x}_k\}_{k=1}^{\infty} \subset \mathbb{C}^n$ is a Cauchy sequence. Then there exists some $\mathbf{x} \in \mathbb{C}^n$ such that $\mathbf{x}_k \to \mathbf{x}$ as $k \to \infty$.*

These definitions can be naturally extended to sequences $\{\mathbf{A}_k\}_{k=1}^{\infty}$ of matrices in $\mathbb{C}^{m\times n}$, using the matrix norm defined in (1.2).

Definition 1.5. *The sequence $\{\mathbf{A}_k\}_{k=1}^{\infty} \subset \mathbb{C}^{m\times n}$ converges to $\mathbf{A} \in \mathbb{C}^{m\times n}$ provided that, for any $\varepsilon > 0$, there exists some positive integer N_ε such that for all $k > N_\varepsilon$, $\|\mathbf{A}_k - \mathbf{A}\| < \varepsilon$. We write $\mathbf{A}_k \to \mathbf{A}$ as $k \to \infty$.*

Definition 1.6. *The matrices $\{\mathbf{A}_k\}_{k=1}^{\infty} \subset \mathbb{C}^{m\times n}$ form a Cauchy sequence provided that, for any $\varepsilon > 0$, there exists a positive integer N_ε such that $\|\mathbf{A}_j - \mathbf{A}_k\| < \varepsilon$ for all $j, k > N_\varepsilon$.*

Fact 1.7. *Suppose $\{\mathbf{A}_k\}_{k=1}^{\infty} \subset \mathbb{C}^{m\times n}$ is a Cauchy sequence. Then there exists some $\mathbf{A} \in \mathbb{C}^{m\times n}$ such that $\mathbf{A}_k \to \mathbf{A}$ as $k \to \infty$.*

1.2 Eigenvalues and Eigenvectors

In the early 1730s Daniel Bernoulli was curious about vibrations. In the decades since the publication of Newton's *Principia* in 1686, natural philosophers across Europe fundamentally advanced their understanding of basic mechanical systems. Bernoulli was studying the motion of a compound pendulum – a massless string suspended with a few massive beads. Given the initial location of those beads, could you predict how the pendulum would swing? To keep matters simple he considered only small motions, in which case the beads only move, to first approximation, in the horizontal direction. Though Bernoulli addressed a variety of configurations (including the limit of infinitely many masses) [Ber33, Ber34], for our purposes it suffices to consider three equal masses, m, separated by equal lengths of thread, ℓ; see Fig. 1.1. If we denote by x_1, x_2, x_3 the displacement of the three masses and by g the force of gravity, the system oscillates according to the second order differential equation

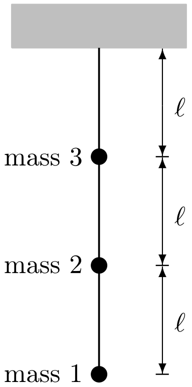

Figure 1.1. A pendulum with three equally-spaced masses.

$$\begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} = \frac{g}{m\ell} \begin{bmatrix} -1 & 1 & 0 \\ 1 & -3 & 2 \\ 0 & 2 & -5 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix},$$

which we abbreviate

$$\mathbf{x}''(t) = -\mathbf{A}\mathbf{x}(t), \tag{1.6}$$

though this convenient matrix notation did not come along until Arthur Cayley introduced it some 120 years later.

Given these equations of motion, Bernoulli asked: what nontrivial displacements $x_1(0)$, $x_2(0)$, and $x_3(0)$ produce motion where both masses pass through the vertical at the same time, as in the rest configuration? In other words, which values of $\mathbf{x}(0) = \mathbf{u} \neq \mathbf{0}$ produce, at some $t > 0$, the solution $\mathbf{x}(t) = \mathbf{0}$?

Bernoulli approached this question via basic mechanical principles (see [Tru60, 160ff]); we shall summarize the result in our modern matrix setting. For a system with n masses, there exist nonzero vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and corresponding scalars $\lambda_1, \ldots, \lambda_n$ for which $\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j$. When the initial displacement corresponds to one of these directions, $\mathbf{x}(0) = \mathbf{u}_j$, the differential equation (1.6) reduces to

$$\mathbf{x}''(t) = -\lambda_j \mathbf{x}(t),$$

which (provided the system begins at rest, $\mathbf{x}'(0) = \mathbf{0}$) has the solution

$$\mathbf{x}(t) = \cos(\sqrt{\lambda_j} t)\mathbf{u}_j.$$

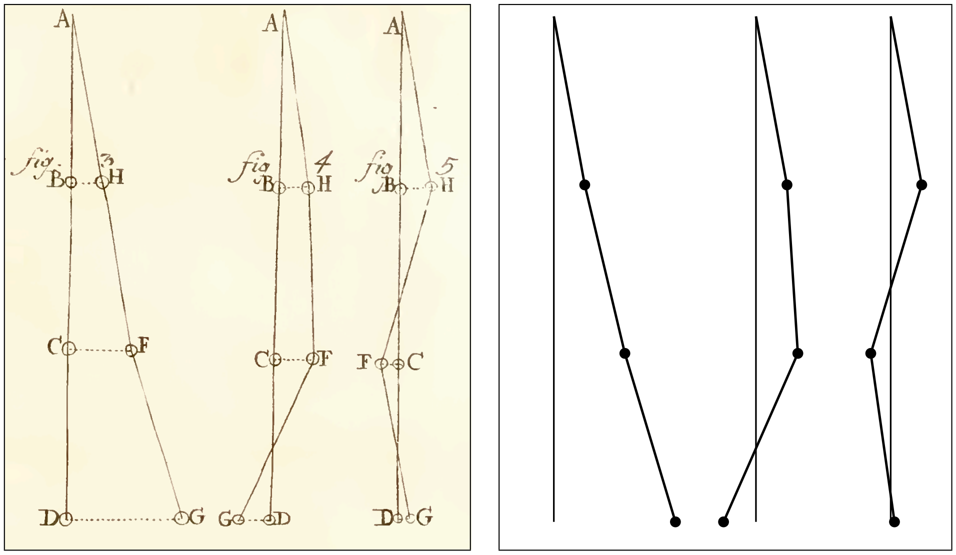

Figure 1.2. The three eigenvectors for a uniform 3-mass system: Bernoulli's 1733 illustration [Ber33], from a scan of the journal at `archive.org` (left); plotted in MATLAB (right).

Hence whenever $\sqrt{\lambda_j}t$ is a half-integer multiple of π, i.e.,

$$t = \frac{(2k+1)\pi}{\sqrt{\lambda_j}}, \qquad k = 0, 1, 2, \ldots,$$

we have $\mathbf{x}(t) = \mathbf{0}$: the masses will simultaneously have zero horizontal displacement, as Bernoulli desired. He computed these special directions for cases, and also studied the limit of infinitely many beads. His illustration of the three distinguished vectors for three equally-separated uniform masses is compared to a modern plot in Figure 1.2.

While Bernoulli's specific question may seem arcane, the ideas surrounding its resolution have had, over the course of three centuries, a fundamental influence in subjects as diverse as quantum mechanics, population ecology, and economics. As Truesdell observes [Tru60, p. 154ff], Bernoulli did not yet realize that his special solutions $\mathbf{u}_1, \ldots, \mathbf{u}_n$ were the key to understanding *all* vibrations, for if

$$\mathbf{x}(0) = \sum_{j=1}^{n} \gamma_j \mathbf{u}_j,$$

then the true solution is a superposition of the special solutions oscillating at their various frequencies:

$$\mathbf{x}(t) = \sum_{j=1}^{n} \gamma_j \cos(\sqrt{\lambda_j} t)\mathbf{u}_j. \tag{1.7}$$

With these distinguished directions $\mathbf{u}$ – for which multiplication by the matrix $\mathbf{A}$ has the same effect as multiplication by the scalar λ – we thus unlock the mysteries of linear dynamical systems. Probably you have encountered these quantities already: we call $\mathbf{u}$ an *eigenvector* associated with the *eigenvalue* λ. (Use of the latter name, a half-translation of the German *eigenwert*, was not settled until recent decades. Older English-language books favor alternatives like *latent root*, *characteristic value*, or *proper value*.)

Subtracting the left side of the equation $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ from the right yields

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0},$$

which implies $\mathbf{u} \in \mathcal{N}(\mathbf{A} - \lambda\mathbf{I})$. In the analysis to follow, this proves to be a particularly useful way to think about eigenvalues and eigenvectors.

Eigenvalues, Eigenvectors, Spectrum, Resolvent

A point $\lambda \in \mathbb{C}$ is an *eigenvalue* of $\mathbf{A} \in \mathbb{C}^{n\times n}$ if that $\lambda\mathbf{I} - \mathbf{A}$ is not invertible. Any nonzero $\mathbf{u} \in \mathcal{N}(\lambda\mathbf{I} - \mathbf{A})$ is called an *eigenvector* of $\mathbf{A}$ associated with λ, and

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}.$$

The *spectrum* of $\mathbf{A}$ is the set of all eigenvalues of $\mathbf{A}$:

$$\sigma(\mathbf{A}) := \{\lambda \in \mathbb{C} : \lambda\mathbf{I} - \mathbf{A} \text{ is not invertible}\}.$$

For any complex number $z \notin \sigma(\mathbf{A})$, the *resolvent* is the matrix

$$\mathbf{R}(z) = (z\mathbf{I} - \mathbf{A})^{-1}, \tag{1.8}$$

which can be viewed as a function, $\mathbf{R}(z) : \mathbb{C} \setminus \sigma(\mathbf{A}) \to \mathbb{C}^{n\times n}$.

1.3 Properties of the Resolvent

The resolvent is a central object in spectral theory – it reveals the existence of eigenvalues, indicates where eigenvalues can fall, and shows how sensitive these eigenvalues are to perturbations. To form the resolvent, you could follow the usual Gauss–Jordan elimination procedure that you learned for computing the inverse of a standard matrix. Now the matrix you are inverting involves the variable z, but the same procedure applies.

Transform the augmented matrix $[z\mathbf{I}-\mathbf{A} \mid \mathbf{I}]$ *into the form* $[\mathbf{I} \mid (z\mathbf{I}-\mathbf{A})^{-1}]$

using *elementary row operations*: row swaps, scaling rows (by a term the possibly includes z), and adding rows. These elementary row operations might involve multiplying or dividing by terms involving z, but they do not involve any other nonlinear functions (e.g., no $\sin(z)$ or $\sqrt{z}$ or e^z). Thus as the inverse of the matrix $z\mathbf{I}-\mathbf{A}$, the resolvent will have as its entries rational functions of z.

It is hard to appreciate the resolvent without looking at a few examples:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{R}(z) = \begin{bmatrix} \dfrac{1}{z} & 0 \\ 0 & \dfrac{1}{z-1} \end{bmatrix};$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{R}(z) = \begin{bmatrix} \dfrac{1}{z} & \dfrac{1}{z(z-1)} \\ 0 & \dfrac{1}{z-1} \end{bmatrix};$$

$$\mathbf{A} = \begin{bmatrix} 2 & -2 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{R}(z) = \begin{bmatrix} \dfrac{1+z}{z(z-1)} & \dfrac{-2}{z(z-1)} \\ \dfrac{1}{z(z-1)} & \dfrac{z-2}{z(z-1)} \end{bmatrix};$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{R}(z) = \begin{bmatrix} \dfrac{1}{z} & \dfrac{1}{z(z-1)} & \dfrac{1}{z^2} \\ 0 & \dfrac{1}{(z-1)} & 0 \\ 0 & 0 & \dfrac{1}{z} \end{bmatrix}.$$

The resolvent will fail to exist at any $z \in \mathbb{C}$ where one (or more) of its rational entries has a *pole* (division by zero) – and these points are the eigenvalues of $\mathbf{A}$. Indeed, we could have defined $\sigma(\mathbf{A})$ to be 'the set of all $z \in \mathbb{C}$ where $\mathbf{R}(z)$ does not exist'. Can you identify these points in the examples above?

In all four examples, the $\mathbf{A}$ matrices have the same eigenvalues, $\sigma(\mathbf{A}) = \{0, 1\}$. In each case, notice that $\mathbf{R}(z)$ exists for all $z \in \mathbb{C}$ except $z = 0$ and

$z = 1$, where at least one entry in the matrix $\mathbf{R}(z)$ encounters division by zero. As z approaches an eigenvalue, since at least one entry blows up, we must have $\|\mathbf{R}(z)\| \to \infty$, but the rate at which this limit is approached can vary with the matrix (e.g., $1/z$ versus $1/z^2$). Being comprised of rational functions, the resolvent has a complex derivative at any point $z \in \mathbb{C}$ that is not in the spectrum of $\mathbf{A}$, and hence on any open set in the complex plane not containing an eigenvalue, the resolvent is an *analytic function*.

How large can the eigenvalues be? Is there a threshold on $|z|$ beyond which $\mathbf{R}(z)$ must exist? A classic approach to this question provides our first example of a matrix-valued series.

Neumann Series

Theorem 1.8. *For any matrix* $\mathbf{E} \in \mathbb{C}^{n\times n}$ *with* $\|\mathbf{E}\| < 1$*, the matrix* $\mathbf{I} - \mathbf{E}$ *is invertible with*

$$(\mathbf{I} - \mathbf{E})^{-1} = \lim_{k\to\infty} \mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \mathbf{E}^3 + \cdots + \mathbf{E}^k$$

and

$$\|(\mathbf{I} - \mathbf{E})^{-1}\| \le \frac{1}{1 - \|\mathbf{E}\|}.$$

Proof. The essence of the proof is to establish a matrix generalization of the scalar Taylor series, $(1-\varepsilon)^{-1} = 1 + \varepsilon + \varepsilon^2 + \varepsilon^3 + \cdots$, which converges provided $|\varepsilon| < 1$. Formally, we wish to write

$$(\mathbf{I} - \mathbf{E})^{-1} = \mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \mathbf{E}^3 + \cdots, \tag{1.9}$$

provided $\|\mathbf{E}\| < 1$. One can see this simply by multiplying

$$\begin{aligned}(\mathbf{I} - \mathbf{E})&(\mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \mathbf{E}^3 + \cdots) \\ &= (\mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \mathbf{E}^3 + \cdots) - (\mathbf{E} + \mathbf{E}^2 + \mathbf{E}^3 + \mathbf{E}^4 + \cdots) \\ &= \mathbf{I}.\end{aligned}$$

To make this intuition rigorous, define for $k = 1, 2, \ldots$

$$\mathbf{X}_k := \mathbf{I} + \mathbf{E} + \cdots + \mathbf{E}^{k-1}.$$

We will show that $\{\mathbf{X}_k\}$ is a Cauchy sequence (Definition 1.6), and hence converges to some matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ (Fact 1.7). To show $\{\mathbf{X}_k\}$ is Cauchy, assume without loss of generality that $k \ge j$, so that

$$\mathbf{X}_k - \mathbf{X}_j = \mathbf{E}^j + \mathbf{E}^{j+1} + \cdots + \mathbf{E}^{k-1} = \mathbf{E}^j(\mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \cdots + \mathbf{E}^{k-j-1}).$$

Thus, using the submultiplicativity of the matrix norm (1.5),

$$\begin{aligned}\|\mathbf{X}_k - \mathbf{X}_j\| &\le \|\mathbf{E}\|^j(\|\mathbf{I}\| + \|\mathbf{E}\| + \|\mathbf{E}\|^2 + \cdots \|\mathbf{E}\|^{k-j-1})\\ &\le \|\mathbf{E}\|^j(\|\mathbf{I}\| + \|\mathbf{E}\| + \|\mathbf{E}\|^2 + |\mathbf{E}\|^3 + \cdots)\\ &\le \|\mathbf{E}\|^j\,\frac{1}{1-\|\mathbf{E}\|}, \qquad (1.10)\end{aligned}$$

summing the geometric series and assuming $\|\mathbf{E}\| < 1$. Thus given any $\varepsilon > 0$, we can pick N_ε in Definition 1.6 such that

$$\varepsilon < \|\mathbf{E}\|^{N_\varepsilon}\,\frac{1}{1-\|\mathbf{E}\|},$$

and then for any $j, k > N_\varepsilon$, by (1.10) we conclude that $\|\mathbf{X}_k - \mathbf{X}_j\| \le \varepsilon$ and hence $\{\mathbf{X}_k\}$ is a Cauchy sequence. Fact 1.7 ensures that there exists some $\mathbf{X} \in \mathbb{C}^{n\times n}$ such that $\mathbf{X}_k \to \mathbf{X}$.

It remains to show that this $\mathbf{X}$ is actually the inverse of $\mathbf{I} - \mathbf{E}$. This means that we need to show that $\mathbf{X}$ does what an inverse is supposed to do, i.e., that $\mathbf{X}(\mathbf{I} - \mathbf{E}) = \mathbf{I}$. We compute:

$$\begin{aligned}\|\mathbf{X}(\mathbf{I} - \mathbf{E}) - \mathbf{I})\| &= \lim_{k\to\infty} \|\mathbf{X}_k(\mathbf{I} - \mathbf{E}) - \mathbf{I}\|\\ &= \lim_{k\to\infty} \left\| \sum_{j=0}^{k} \mathbf{E}^j - \sum_{j=0}^{k} \mathbf{E}^{j+1} - \mathbf{I} \right\|\\ &= \lim_{k\to\infty} \|\mathbf{E}^{k+1}\| = \|\mathbf{0}\| = 0.\end{aligned}$$

Thus, $\mathbf{X} = (\mathbf{I} - \mathbf{E})^{-1}$. It is natural to write

$$\mathbf{X} = (\mathbf{I} - \mathbf{E})^{-1} = \mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \cdots.$$

Again using submultiplicativity of the matrix norm, we can bound the norm of the inverse:

$$\begin{aligned}\|(\mathbf{I} - \mathbf{E})^{-1}\| &= \lim_{k\to\infty} \|\mathbf{I} + \mathbf{E} + \mathbf{E}^2 + \cdots + \mathbf{E}^k\|\\ &\le \|\mathbf{I}\| + \|\mathbf{E}\| + \|\mathbf{E}\|^2 + \cdots = \frac{1}{1-\|\mathbf{E}\|}. \quad \blacksquare\end{aligned}$$

Neumann Series for Resolvent

Theorem 1.9. *For any* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and* $|z| > \|\mathbf{A}\|$*, the resolvent* $\mathbf{R}(z)$ *exists and*

$$\|\mathbf{R}(z)\| \le \frac{1}{|z| - \|\mathbf{A}\|}.$$

Proof. The statement about the resolvent $\mathbf{R}(z) := (z\mathbf{I} - \mathbf{A})^{-1}$ follows directly from Theorem 1.8. If $|z| > \|\mathbf{A}\|$, then $\|\mathbf{A}/z\| < 1$, so

$$(z\mathbf{I} - \mathbf{A})^{-1} = z^{-1}(\mathbf{I} - \mathbf{A}/z)^{-1} = z^{-1}(\mathbf{I} + \mathbf{A}/z + \mathbf{A}^2/z^2 + \cdots)$$

and

$$\|(z\mathbf{I} - \mathbf{A})^{-1}\| \le \frac{1}{|z|}\,\frac{1}{1 - \|\mathbf{A}/z\|} = \frac{1}{|z| - \|\mathbf{A}\|}. \quad \blacksquare$$

Theorems 1.8 and 1.9 prove very handy in a variety of circumstances. For example, it provides a mechanism for inverting small perturbations of arbitrary invertible matrices; see Problem 1.1. Of more immediate interest to us now, notice that since $\|\mathbf{R}(z)\| < 1/(|z| - \|\mathbf{A}\|)$ when $|z| > \|\mathbf{A}\|$, there can be no eigenvalues that exceed $\|\mathbf{A}\|$ in magnitude:

$$\sigma(\mathbf{A}) \subseteq \{z \in \mathbb{C} : |z| \le \|\mathbf{A}\|\}.$$

In fact, we could have reached this conclusion more directly, since the eigenvalue-eigenvector equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ gives

$$|\lambda|\|\mathbf{v}\| = \|\mathbf{A}\mathbf{v}\| \le \|\mathbf{A}\|\|\mathbf{v}\|.$$

At this point, we know that a matrix cannot have arbitrarily large eigenvalues. But a more fundamental question remains: does a matrix need to have any eigenvalues at all? Is it possible that the spectrum is empty? As a consequence of Theorem 1.9, we see this cannot be the case.

Theorem 1.10. *Every matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *has at least one eigenvalue.*

Proof. Liouville's Theorem (from complex analysis) states that any entire function (a function that is analytic throughout the entire complex plane) that is bounded throughout $\mathbb{C}$ must be constant. If $\mathbf{A}$ has no eigenvalues, then $\mathbf{R}(z)$ is entire and $\|\mathbf{R}(z)\|$ is bounded throughout the disk $\{z \in \mathbb{C} : |z| \le \|\mathbf{A}\|\}$. By Theorem 1.9, $\|\mathbf{R}(z)\|$ is also bounded outside this disk. Hence by Liouville's Theorem, $\mathbf{R}(z)$ must be constant. However, Theorem 1.9 implies

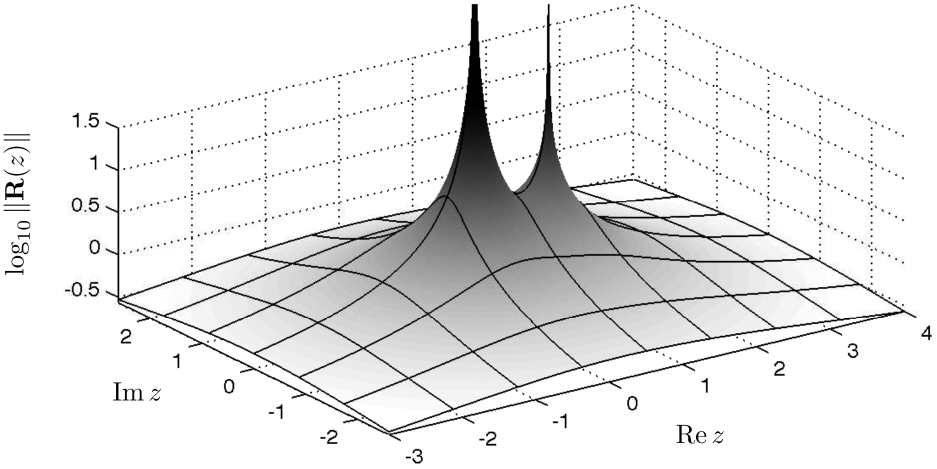

Figure 1.3. Norm of the resolvent for the 3×3 matrix on page 15.

that $\|\mathbf{R}(z)\| \to 0$ as $|z| \to \infty$. Thus, we must have $\mathbf{R}(z) = \mathbf{0}$, but this contradicts the identity $\mathbf{R}(z)(z\mathbf{I} - \mathbf{A}) = \mathbf{I}$. Hence $\mathbf{A}$ has an eigenvalue. ■

(One might favor an apparently less sophisticated approach. All entries in $\mathbf{R}(z)$ are rational functions, meaning that the (j, k) entry of $\mathbf{R}(z)$ can be written as $p_{j,k}(z)/q_{j,k}(z)$ for some polynomials $p_{j,k}$ and $q_{j,k}$. Suppose $\mathbf{A}$ does not have any eigenvalues. This means that $q_{j,k}(z) \neq 0$ for all $j, k \in \{1, \ldots, n\}$. Since each of these $q_{j,k}$ polynomials has no roots, it must be a (nonzero) constant. Without loss of generality, assume $q_{j,k}(z) = 1$ for all z. Thus $r_{j,k}(z) = p_{j,k}(z)$ is a polynomial for all j and k. However, any non-constant polynomial must satisfy $|p(z)| \to \infty$ as $|z| \to \infty$

One appeal of the proofs of Theorems 1.9 and 1.10 (e.g., [Kat80, p. 37], [You88, Ch. 7]) is that they generalize readily to bounded linear operators on infinite dimensional spaces; see, e.g., [RS80, p. 191].

We hope the explicit resolvents of those small matrices shown at the beginning of this section helped build your intuition for the important concepts that followed. It might also help to supplement this analytical perspective with a picture of the resolvent norm in the complex plane. Figure 1.3 plots $\|\mathbf{R}(z)\|$ as a function of $z \in \mathbb{C}$: eigenvalues correspond to the spikes in the plot, which go off to infinity – these are the poles of entries in the resolvent. Notice that the peak around the eigenvalue $\lambda = 0$ is a bit broader than the one around $\lambda = 1$, reflecting the quadratic growth in $1/z^2$ term, compared to linear growth in $1/(z - 1)$.

1.4 Reduction to Schur Triangular Form

We saw in Theorem 1.10 that every matrix must have at least one eigenvalue. Now we leverage this basic result to factor any square $\mathbf{A}$ into a distinguished form, the combination of a unitary matrix and an upper triangular matrix.

Schur Triangularization

Theorem 1.11. *For any matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ there exists a unitary matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathbf{T} \in \mathbb{C}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*. \tag{1.11}$$

The eigenvalues of $\mathbf{A}$ are the diagonal elements of $\mathbf{T}$:

$$\sigma(\mathbf{A}) = \{t_{1,1}, \ldots, t_{n,n}\}.$$

Hence $\mathbf{A}$ has exactly n eigenvalues (counting multiplicity).

Proof. We will prove this result by mathematical induction. If $\mathbf{A}$ is a scalar, $\mathbf{A} \in \mathbb{C}^{1\times 1}$, the result is trivial: take $\mathbf{U} = 1$ and $\mathbf{T} = \mathbf{A}$. Now make the inductive assumption that the result holds for $(n-1)\times(n-1)$ matrices. By Theorem 1.10, any matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ must have at least one eigenvalue $\lambda \in \mathbb{C}$, and consequently a corresponding unit eigenvector $\mathbf{u}$, so that

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \qquad \|\mathbf{u}\| = 1.$$

Now construct a matrix $\mathbf{Q} \in \mathbb{C}^{n\times(n-1)}$ whose columns form an orthonormal basis for the space $\text{span}\{\mathbf{u}\}^\perp$, so that $[\,\mathbf{u} \quad \mathbf{Q}\,] \in \mathbb{C}^{n\times n}$ is a unitary matrix. Then (following the template for block-matrix multiplication described on page 6), we have

$$[\,\mathbf{u} \quad \mathbf{Q}\,]^*\,\mathbf{A}\,[\,\mathbf{u} \quad \mathbf{Q}\,] = \begin{bmatrix} \mathbf{u}^*\mathbf{A}\mathbf{u} & \mathbf{u}^*\mathbf{A}\mathbf{Q} \\ \mathbf{Q}^*\mathbf{A}\mathbf{u} & \mathbf{Q}^*\mathbf{A}\mathbf{Q} \end{bmatrix}. \tag{1.12}$$

The $(1,1)$ entry is simply

$$\mathbf{u}^*\mathbf{A}\mathbf{u} = \mathbf{u}^*(\lambda\mathbf{u}) = \lambda\|\mathbf{u}\|^2 = \lambda,$$

while the orthogonality of $\mathbf{u}$ and $\mathcal{R}(\mathbf{Q})$ gives the $(2,1)$ entry

$$\mathbf{Q}^*\mathbf{A}\mathbf{u} = \lambda\mathbf{Q}^*\mathbf{u} = \mathbf{0} \in \mathbb{C}^{(n-1)\times 1}.$$

We can say nothing specific about the $(1,2)$ and $(2,2)$ entries, but give them the abbreviated names

$$\widehat{\mathbf{t}}^* := \mathbf{u}^*\mathbf{A}\mathbf{Q} \in \mathbb{C}^{1\times(n-1)}, \qquad \widehat{\mathbf{A}} := \mathbf{Q}^*\mathbf{A}\mathbf{Q} \in \mathbb{C}^{(n-1)\times(n-1)}.$$

Now rearrange (1.12) into the form

$$\mathbf{A} = [\,\mathbf{u}\quad \mathbf{Q}\,]\begin{bmatrix}\lambda & \widehat{\mathbf{t}}^* \\ \mathbf{0} & \widehat{\mathbf{A}}\end{bmatrix}[\,\mathbf{u}\quad \mathbf{Q}\,]^*, \tag{1.13}$$

which is a *partial* SCHUR decomposition of $\mathbf{A}$: the central matrix on the right has zeros below the diagonal in the first column, but we must extend this. By the inductive hypothesis we can compute a SCHUR factorization for the $(n-1)\times(n-1)$ matrix $\widehat{\mathbf{A}}$:

$$\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{T}}\widehat{\mathbf{U}}^*.$$

Substitute this formula into (1.13) to obtain

$$\begin{aligned}
\mathbf{A} &= [\,\mathbf{u}\quad \mathbf{Q}\,]\begin{bmatrix}\lambda & \widehat{\mathbf{t}}^* \\ \mathbf{0} & \widehat{\mathbf{U}}\widehat{\mathbf{T}}\widehat{\mathbf{U}}^*\end{bmatrix}\begin{bmatrix}\mathbf{u}^* \\ \mathbf{Q}^*\end{bmatrix} \\
&= [\,\mathbf{u}\quad \mathbf{Q}\,]\begin{bmatrix}1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{U}}\end{bmatrix}\begin{bmatrix}\lambda & \widehat{\mathbf{t}}^*\widehat{\mathbf{U}} \\ \mathbf{0} & \widehat{\mathbf{T}}\end{bmatrix}\begin{bmatrix}1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{U}}^*\end{bmatrix}\begin{bmatrix}\mathbf{u}^* \\ \mathbf{Q}^*\end{bmatrix} \\
&= [\,\mathbf{u}\quad \mathbf{Q}\widehat{\mathbf{U}}\,]\begin{bmatrix}\lambda & \widehat{\mathbf{t}}^*\widehat{\mathbf{U}} \\ \mathbf{0} & \widehat{\mathbf{T}}\end{bmatrix}[\,\mathbf{u}\quad \mathbf{Q}\widehat{\mathbf{U}}\,]^*. \qquad (1.14)
\end{aligned}$$

The central matrix in this last arrangement is upper triangular,

$$\mathbf{T} := \begin{bmatrix}\lambda & \widehat{\mathbf{t}}^*\widehat{\mathbf{U}} \\ \mathbf{0} & \widehat{\mathbf{T}}\end{bmatrix}.$$

We label the matrices on either side of $\mathbf{T}$ in (1.14) as $\mathbf{U}$ and $\mathbf{U}^*$, where

$$\mathbf{U} := [\,\mathbf{u}\quad \mathbf{Q}\widehat{\mathbf{U}}\,],$$

so that $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$. Our last task is to confirm that $\mathbf{U}$ is unitary: to see this, note that the orthogonality of $\mathbf{u}$ and $\mathcal{R}(\mathbf{Q})$, together with the fact that $\widehat{\mathbf{U}}$ is unitary, implies

$$\mathbf{U}^*\mathbf{U} = \begin{bmatrix}\mathbf{u}^*\mathbf{u} & \mathbf{u}^*\mathbf{Q}\widehat{\mathbf{U}} \\ \widehat{\mathbf{U}}^*\mathbf{Q}^*\mathbf{u} & \widehat{\mathbf{U}}^*\mathbf{Q}^*\mathbf{Q}\widehat{\mathbf{U}}\end{bmatrix} = \begin{bmatrix}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix}.$$

Given the SCHUR factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$, we wish to show that $\mathbf{A}$ and $\mathbf{T}$ have the same eigenvalues. If $(\lambda, \mathbf{v})$ is a eigenpair of $\mathbf{A}$, so $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, then $\mathbf{U}\mathbf{T}\mathbf{U}^*\mathbf{v} = \lambda\mathbf{v}$. Premultiply that last equation by $\mathbf{U}^*$ to obtain

$$\mathbf{T}(\mathbf{U}^*\mathbf{v}) = \lambda(\mathbf{U}^*\mathbf{v}).$$

In other words, λ is an eigenvalue of $\mathbf{T}$ with eigenvector $\mathbf{U}^*\mathbf{v}$. (Note that $\|\mathbf{U}^*\mathbf{v}\| = \|\mathbf{v}\|$ by unitary invariance of the norm, so $\mathbf{U}^*\mathbf{v} \neq \mathbf{0}$.) What then are the eigenvalues of $\mathbf{T}$? As an upper triangular matrix, $z\mathbf{I} - \mathbf{T}$ will be singular (or equivalently, have linearly dependent columns) if and only if one of its diagonal elements is zero, i.e., $z = t_{j,j}$ for some $1 \leq j \leq n$. It follows that

$$\sigma(\mathbf{A}) = \sigma(\mathbf{T}) = \{t_{1,1}, t_{2,2}, \ldots, t_{n,n}\}. \qquad \blacksquare$$

The SCHUR factorization is far from unique: for example, multiplying any column of $\mathbf{U}$ by -1 will yield a different Schur form. The following corollary describes a more significant (and useful) way in which Schur factorizations can differ. More significantly, as we will address later, the eigenvalues can appear *in any desired order* on the diagonal of $\mathbf{T}$. (Note that the order of the eigenvalues will affect both $\mathbf{U}$ and the upper triangular entries of $\mathbf{T}$.) Next we extract a simple but important corollary from the Schur form.

Corollary 1.12. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ has at most n distinct eigenvalues.*

Proof. Compute a SCHUR factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$. All values on the diagonal of $\mathbf{T}$ are eigenvalues of $\mathbf{A}$. (This is evident from the proof of Theorem 1.11, of from the fact that $(\mathbf{T} - t_{j,j}\mathbf{I})\mathbf{x} = \mathbf{b}$ is not uniquely solvable for all $\mathbf{b} \in \mathbb{C}^n$, due to the failure of back substitution.) The SCHUR factor $\mathbf{T}$ has at most n distinct diagonal entries. Might $\mathbf{A}$ have additional eigenvalues? Suppose $\lambda \in \mathbb{C}$ does not equal any of the diagonal entries of $\mathbf{T}$. Then

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{U}\mathbf{T}\mathbf{U}^* - \lambda\mathbf{I} = \mathbf{U}(\mathbf{T} - \lambda\mathbf{I})\mathbf{U}^*,$$

and so $\mathbf{A} - \lambda\mathbf{I}$ is invertible if and only if $\mathbf{T} - \lambda\mathbf{I}$ is invertible. Since $\lambda \neq t_{j,j}$ for all $j = 1, \ldots, n$, $\mathbf{T} - \lambda\mathbf{I}$ has no zero entries on its main diagonal, and hence is invertible. (For example, you can use back substitution to solve $(\mathbf{T} - \lambda\mathbf{I})\mathbf{x} = \mathbf{b}$ uniquely for all $\mathbf{b} \in \mathbb{C}^n$.) Thus λ is not an eigenvalue of $\mathbf{A}$. Thus $\mathbf{A}$ can have at most n distinct eigenvalues. $\blacksquare$

One subtle point remains to clear up: What about multiple eigenvalues? Suppose $\mathbf{A}$ has $p < n$ distinct eigenvalues. Do these eigenvalues need to appear with the *same multiplicity* on the diagonal of $\mathbf{T}$ in all SCHUR factorizations? For example, is it possible that a 3×3 matrix $\mathbf{A}$ with eigenvalues $\lambda_1 = 1$ and λ_2 could have two SCHUR factorizations

$$\mathbf{A} = \mathbf{U}_1\mathbf{T}_1\mathbf{U}_1^* = \mathbf{U}_2\mathbf{T}_2\mathbf{U}_2^*$$

with SCHUR factors

$$\mathbf{T}_1 = \begin{bmatrix} 1 & \times & \times \\ 0 & 1 & \times \\ 0 & 0 & 2 \end{bmatrix}, \qquad \mathbf{T}_2 = \begin{bmatrix} 1 & \times & \times \\ 0 & 2 & \times \\ 0 & 0 & 2 \end{bmatrix}$$

where $\lambda_1 = 1$ and $\lambda_2 = 2$ appear with conflicting multiplicities? (It is common practice to use × to represent a generic value we do not care about.)

It turns out that such situations are *impossible*: the eigenvalues must appear with the same multiplicity on the diagonal of all SCHUR factors. We will pin that down precisely in Section 1.8.

The SCHUR factorization has widespread and important implications. We can view any matrix $\mathbf{A}$ as an upper triangular matrix, provided we cast it in the correct orthonormal basis. This means that a wide variety of phenomena can be understood for all matrices, if only we can understand them for upper triangular matrices. (Unfortunately, even this poses a formidable challenge for many situations!)

1.5 Spectral Theorem for Hermitian Matrices

The Schur factorization has special implications for Hermitian matrices, $\mathbf{A}^* = \mathbf{A}$. In this case,

$$\mathbf{U}\mathbf{T}\mathbf{U}^* = \mathbf{A} = \mathbf{A}^* = (\mathbf{U}\mathbf{T}\mathbf{U}^*)^* = \mathbf{U}\mathbf{T}^*\mathbf{U}^*.$$

Premultiply this equation by $\mathbf{U}^*$ and postmultiply by $\mathbf{U}$ to see that

$$\mathbf{T} = \mathbf{T}^*.$$

Thus, $\mathbf{T}$ is both *Hermitian* and *upper triangular*: in other words, $\mathbf{T}$ must be diagonal! Furthermore,

$$t_{j,j} = \overline{t_{j,j}};$$

in other words, the diagonal entries of $\mathbf{T}$ – the eigenvalues of $\mathbf{A}$ – must be real numbers. It is customary in this case to write $\mathbf{\Lambda}$ in place of $\mathbf{T}$.

Unitary Diagonalization of a Hermitian Matrix

Theorem 1.13. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *be Hermitian,* $\mathbf{A}^* = \mathbf{A}$*. Then there exists a unitary matrix*

$$\mathbf{U} = [\,\mathbf{u}_1, \ldots, \mathbf{u}_n\,] \in \mathbb{C}^{n\times n}$$

and diagonal matrix

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{C}^{n\times n}$$

such that

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*. \tag{1.15}$$

The orthonormal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ *are eigenvectors of* $\mathbf{A}$*, corresponding to eigenvalues* $\lambda_1, \ldots, \lambda_n$*:*

$$\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j, \qquad j = 1, \ldots, n.$$

The eigenvalues of $\mathbf{A}$ *are all real:* $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.

Often it helps to consider equation (1.15) in a slightly different form. Writing that equation out by components gives

$$\begin{aligned} \mathbf{A} &= [\,\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_n\,] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_n^* \end{bmatrix} \\ &= [\,\lambda_1\mathbf{u}_1 \quad \cdots \quad \lambda_n\mathbf{u}_n\,] \begin{bmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_n^* \end{bmatrix} = \lambda_1\mathbf{u}_1\mathbf{u}_1^* + \cdots \lambda_n\mathbf{u}_n\mathbf{u}_n^*. \end{aligned}$$

Notice that the matrices

$$\mathbf{P}_j := \mathbf{u}_j\mathbf{u}_j^* \in \mathbb{C}^{n\times n}$$

are orthogonal projectors, since $\mathbf{P}_j^* = \mathbf{P}_j$ and

$$\mathbf{P}_j^2 = \mathbf{u}_j(\mathbf{u}_j^*\mathbf{u}_j)\mathbf{u}_j^* = \mathbf{u}_j\mathbf{u}_j^* = \mathbf{P}_j.$$

Hence we write

$$\mathbf{A} = \sum_{j=1}^{n} \lambda_j \mathbf{P}_j. \tag{1.16}$$

Any Hermitian matrix is a weighted sum of orthogonal projectors! Not just any orthogonal projectors, though; these obey a special property:

$$\sum_{j=1}^{n} \mathbf{P}_j = \sum_{j=1}^{n} \mathbf{u}_j\mathbf{u}_j^* = [\,\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_n\,] \begin{bmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_n^* \end{bmatrix} = \mathbf{U}\mathbf{U}^* = \mathbf{I}.$$

Thus, we call this collection $\{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$ a *resolution of the identity*. Also note that if $j \neq k$, then the orthogonality of the eigenvectors implies

$$\mathbf{P}_j\mathbf{P}_k = \mathbf{u}_j\mathbf{u}_j^*\mathbf{u}_k\mathbf{u}_k^* = \mathbf{0}.$$

In summary, via the Schur form we have arrived at two beautifully simple ways to write a Hermitian matrix:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^* = \sum_{j=1}^{n} \lambda_j \mathbf{P}_j.$$

Which factorization is better? That varies with mathematical circumstance and personal taste.

1.5.1 Revisiting Bernoulli's Pendulum

We pause for a moment to note that we now have all the tools needed to derive the general solution (1.7) to Bernoulli's oscillating pendulum problem. His matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is always Hermitian, so following (1.15) we write $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$, reducing the differential equation (1.6) to

$$\mathbf{x}''(t) = -\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*\mathbf{x}(t).$$

Premultiply by $\mathbf{U}^*$ to obtain

$$\mathbf{U}^*\mathbf{x}''(t) = -\boldsymbol{\Lambda}\mathbf{U}^*\mathbf{x}(t). \tag{1.17}$$

Notice that the vector $\mathbf{x}(t)$ represents the solution to the equation in the standard coordinate system

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Noting that these vectors form the columns of the identity matrix $\mathbf{I}$, we could (in a fussy mood), write

$$\mathbf{x}(t) = \mathbf{I}\mathbf{x}(t) = x_1(t)\mathbf{e}_1 + x_2(t)\mathbf{e}_2 + \cdots + x_n(t)\mathbf{e}_n.$$

In the same fashion, we could let $\mathbf{y}(t)$ denote the representation of $\mathbf{x}(t)$ in the coordinate system given by the orthonormal eigenvectors of $\mathbf{A}$:

$$\mathbf{x}(t) = \mathbf{U}\mathbf{y}(t) = y_1(t)\mathbf{u}_1 + y_2(t)\mathbf{u}_2 + \cdots + y_n(t)\mathbf{u}_n.$$

In other words,

$$\mathbf{y}(t) = \mathbf{U}^*\mathbf{x}(t),$$

so in the eigenvector coordinate system the initial conditions are

$$\mathbf{y}(0) = \mathbf{U}^*\mathbf{x}(0), \qquad \mathbf{y}'(0) = \mathbf{U}^*\mathbf{x}'(0) = \mathbf{0}.$$

Then (1.17) becomes

$$\mathbf{y}''(t) = -\mathbf{\Lambda}\mathbf{y}(t), \qquad \mathbf{y}(0) = \mathbf{U}^*\mathbf{x}(0), \quad \mathbf{y}'(0) = \mathbf{0},$$

which decouples into n independent scalar equations

$$y_j''(t) = -\lambda_j y_j(t), \qquad y_j(0) = \mathbf{u}_j^*\mathbf{x}(0), \quad y_j'(0) = 0, \quad j = 1, \ldots, n,$$

each of which has the simple solution

$$y_j(t) = \cos(\sqrt{\lambda_j}t)y_j(0).$$

Transforming back to our original coordinate system,

$$\begin{aligned}\mathbf{x}(t) = \mathbf{U}\mathbf{y}(t) &= \sum_{j=1}^{n} \cos(\sqrt{\lambda_j}t)y_j(0)\mathbf{u}_j \\ &= \sum_{j=1}^{n} \cos(\sqrt{\lambda_j}t)(\mathbf{u}_j^*\mathbf{x}(0))\mathbf{u}_j, \qquad (1.18)\end{aligned}$$

which now makes entirely explicit the equation given earlier in (1.7).

1.6 Diagonalizable Matrices

The unitary diagonalization enjoyed by all Hermitian matrices is absolutely ideal: the eigenvectors provide an alternative orthogonal coordinate system in which the matrix becomes diagonal, hence reducing matrix-vector equations to uncoupled scalar equations that can be dispatched independently.

We now aim to extend this arrangement to all square matrices. Two possible ways of generalizing $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$ seem appealing:

(i) keep $\mathbf{U}$ unitary, but relax the requirement that $\boldsymbol{\Lambda}$ be diagonal;

(ii) keep $\boldsymbol{\Lambda}$ diagonal, but relax the requirement that $\mathbf{U}$ be unitary.

Obviously we have already addressed approach (i): this simply yields the Schur triangular form. What about approach (ii)?

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ with corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. First make the quite reasonable assumption that these eigenvectors are linearly independent. For example, this assumption holds whenever the eigenvalues $\lambda_1, \ldots, \lambda_n$ are distinct (i.e., $\lambda_j \neq \lambda_k$ when $j \neq k$), which we pause to establish in the following lemma.

Lemma 1.14. *Eigenvectors associated with distinct eigenvalues are linearly independent.*

Proof. Let the nonzero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be a set of $m \leq n$ eigenvectors of $\mathbf{A} \in \mathbb{C}^{n\times n}$ associated with distinct eigenvalues $\lambda_1, \ldots, \lambda_m$. Without loss of generality, suppose we can write

$$\mathbf{v}_1 = \sum_{j=2}^{m} \gamma_j \mathbf{v}_j.$$

Premultiply each side successively by $\mathbf{A} - \lambda_k\mathbf{I}$ for $k = 2, \ldots, m$ to obtain

$$\begin{aligned}\Big(\prod_{k=2}^{m}(\mathbf{A} - \lambda_k\mathbf{I})\Big)\mathbf{v}_j &= \Big(\sum_{j=2}^{m}\gamma_j\prod_{k=2}^{m}(\mathbf{A} - \lambda_k\mathbf{I})\Big)\mathbf{v}_j \\ &= \sum_{j=2}^{m}\gamma_j\Big(\prod_{\substack{k=2\\k\neq j}}^{m}(\mathbf{A} - \lambda_k\mathbf{I})\Big)(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{v}_j = \mathbf{0},\end{aligned}$$

where the last equality follows from the fact that $\mathbf{A}\mathbf{v}_j = \lambda_j\mathbf{v}_j$. Note, however, that the left side of this equation equals

$$\Big(\prod_{k=2}^{m}(\mathbf{A}-\lambda_k\mathbf{I})\Big)\mathbf{v}_1 = \prod_{k=2}^{m}(\mathbf{A}\mathbf{v}_1-\lambda_k\mathbf{v}_1) = \prod_{k=2}^{m}(\lambda_1-\lambda_k)\mathbf{v}_1,$$

which can only be zero if $\mathbf{v}_1 = \mathbf{0}$ or $\lambda_1 = \lambda_k$ for some $k > 1$, both of which are contradictions. (Adapted from [HK71, page ??]). ■

We return to our assumption that $\mathbf{A} \in \mathbb{C}^{n\times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ with linearly independent eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$:

$$\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1, \qquad \ldots, \qquad \mathbf{A}\mathbf{v}_n = \lambda_n\mathbf{v}_n.$$

Organizing these n vector equations into one matrix equation, we find

$$[\,\mathbf{A}\mathbf{v}_1 \quad \mathbf{A}\mathbf{v}_2 \quad \cdots \quad \mathbf{A}\mathbf{v}_n\,] = [\,\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_n\mathbf{v}_n\,].$$

Recalling that postmultiplication by a diagonal matrix scales the columns of the preceding matrix, factor each side into the product of a pair of matrices:

$$\mathbf{A}\,[\,\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n\,] = [\,\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n\,]\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix}, \tag{1.19}$$

We summarize equation (1.19) as

$$\mathbf{A}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}.$$

The assumption that the eigenvectors are linearly independent enables us to invert $\mathbf{V}$, hence

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}. \tag{1.20}$$

This equation gives an analogue of Theorem 1.13. The class of matrices that admit n linearly independent eigenvectors – and hence the factorization $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$ – are known as *diagonalizable* matrices. We now seek a variant of the projector-based representation (1.16). Toward this end, write $\mathbf{V}^{-1}$ out by *rows*,

$$\mathbf{V}^{-1} = \begin{bmatrix}\widehat{\mathbf{v}}_1^* \\ \widehat{\mathbf{v}}_2^* \\ \vdots \\ \widehat{\mathbf{v}}_n^*\end{bmatrix} \in \mathbb{C}^{n\times n},$$

so that (1.20) takes the form

$$\mathbf{A} = [\,\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n\,] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{v}}_1^* \\ \widehat{\mathbf{v}}_2^* \\ \vdots \\ \widehat{\mathbf{v}}_n^* \end{bmatrix} = \sum_{j=1}^{n} \lambda_j \mathbf{v}_j \widehat{\mathbf{v}}_j^*. \quad (1.21)$$

The rows of $\mathbf{V}^{-1}$ are called *left eigenvectors*, since $\widehat{\mathbf{v}}_j^*\mathbf{A} = \lambda_j \widehat{\mathbf{v}}_j^*$. (In this context the normal eigenvectors $\mathbf{v}_j$ that give $\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j$ are *right eigenvectors*.) Also note that while in general the set of right eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are not orthogonal, the left and right eigenvectors are *biorthogonal*:

$$\widehat{\mathbf{v}}_j^*\mathbf{v}_k = (\mathbf{V}^{-1}\mathbf{V})_{j,k} = \begin{cases} 1, & j = k; \\ 0, & j \neq k. \end{cases}$$

This motivates our definition

$$\mathbf{P}_j := \mathbf{v}_j \widehat{\mathbf{v}}_j^*. \quad (1.22)$$

Three facts follow directly from $\mathbf{V}^{-1}\mathbf{V} = \mathbf{V}\mathbf{V}^{-1} = \mathbf{I}$:

$$\mathbf{P}_j^2 = \mathbf{P}_j, \qquad \mathbf{P}_j\mathbf{P}_k = \mathbf{0} \text{ if } j \neq k, \qquad \sum_{j=1}^{n} \mathbf{P}_j = \mathbf{I}.$$

As in the Hermitian case, we have a set of projectors that give a resolution of the identity. These observations are cataloged in the following Theorem.

Diagonalization

Theorem 1.15. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ and associated linearly independent eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ can be written as*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \quad (1.23)$$

for

$$\mathbf{V} = [\,\mathbf{v}_1, \ldots, \mathbf{v}_n\,] \in \mathbb{C}^{n\times n}$$

and diagonal matrix

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{C}^{n\times n}.$$

Denoting the jth row of $\mathbf{V}^{-1}$ as $\widehat{\mathbf{v}}_j^*$, the matrices $\mathbf{P}_j := \mathbf{v}_j \widehat{\mathbf{v}}_j^*$ are projectors, $\mathbf{P}_j\mathbf{P}_k = \mathbf{0}$ if $j \neq k$, and

$$\mathbf{A} = \sum_{j=1}^{n} \lambda_j \mathbf{P}_j. \quad (1.24)$$

While equation (1.24) looks identical to the expression (1.16) for Hermitian matrices, notice an important distinction: the projectors $\mathbf{P}_j$ here will not, in general, be orthogonal projectors, since there is no guarantee that $\mathbf{P}_j$ is Hermitian. Furthermore, the eigenvalues λ_j of diagonalizable matrices need not be real, even when $\mathbf{A}$ has real entries.

On the surface, diagonalizable matrices provide many of the same advantages in applications as Hermitian matrices. For example, if $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ is a diagonalization, then the differential equation

$$\mathbf{x}'(t) = -\mathbf{A}\mathbf{x}(t), \qquad \mathbf{x}'(0) = \mathbf{0}$$

has the solution

$$\mathbf{x}(t) = \sum_{j=1}^{n} \cos(\sqrt{\lambda_j}\,t)(\widehat{\mathbf{v}}_j^*\mathbf{x}(0))\mathbf{v}_j,$$

which generalizes the Hermitian formula (1.18). The n linearly independent eigenvectors provide a coordinate system for $\mathbb{C}^n$ in which the matrix $\mathbf{A}$ has a diagonal representation. Unlike the Hermitian case, this new coordinate system will not generally be orthogonal, and these new coordinates can distort physical space in surprising ways that have important implications in applications – as we shall see in Chapter **??**.

1.7 Illustration: Damped Mechanical System

We now have a complete spectral theory for all matrices having n linearly independent eigenvectors. Do there exist matrices that lack this property? Indeed there do – we shall encounter such a case in a natural physical setting.

An extensible spring vibrates according to the differential equation

$$x''(t) = -x(t) - 2ax'(t), \tag{1.25}$$

where the term $-2ax'(t)$ corresponds to viscous damping (e.g., a dashpot) that effectively removes energy from the system when the constant $a > 0$. To write this second-order equation as a system of first-order equations, we introduce the variable $y(t) := x'(t)$, so that

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2a \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \tag{1.26}$$

which we write as $\mathbf{x}'(t) = \mathbf{A}(a)\mathbf{x}(t)$. We wish to tune the constant a to achieve the *fastest* energy decay in this system. As we shall later see, this can be accomplished (for a specific definition of ‘fastest’) by minimizing the

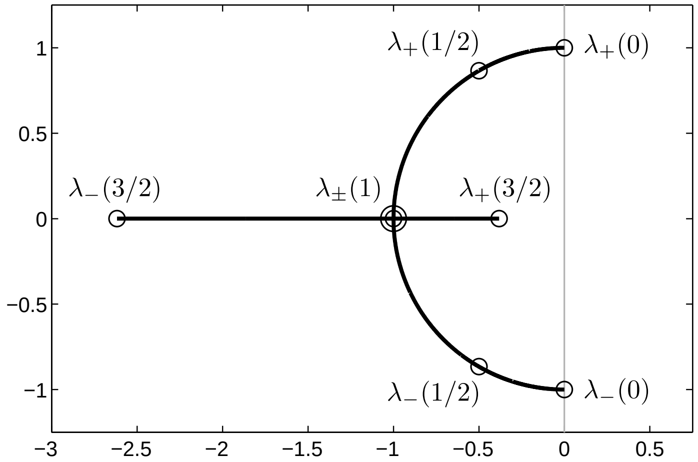

Figure 1.4. Eigenvalues $\lambda_\pm(a)$ of the damped spring for $a \in [0, 3/2]$; for reference, four values a are marked with circles, and the gray line shows the imaginary axis.

real part of the rightmost eigenvalue of $\mathbf{A}(a)$. For any fixed a, the eigenvalues are given by

$$\lambda_\pm = -a \pm \sqrt{a^2 - 1}$$

with associated eigenvectors

$$\mathbf{v}_\pm = \begin{bmatrix} 1 \\ \lambda_\pm \end{bmatrix}.$$

For $a \in [0, 1)$, these eigenvalues form a complex conjugate pair; for $a \in (1, \infty)$, $\lambda_\pm$ form a pair of real eigenvalues. In both cases, the eigenvalues are distinct, so its eigenvectors are linearly independent and $\mathbf{A}(a)$ is diagonalizable.

At the point of transition between the complex and real eigenvalues, $a = 1$, something interesting happens: the eigenvalues match, $\lambda_+ = \lambda_-$, as do the eigenvectors, $\mathbf{v}_+ = \mathbf{v}_-$: the eigenvectors are linearly dependent! Moreover, *this is the value of a that gives the most rapid energy decay*, in the sense that the decay rate

$$\alpha(a) := \max_{\lambda \in \sigma(\mathbf{A}(a))} \operatorname{Re} \lambda = \begin{cases} -a, & a \in [0, 1]; \\ -a + \sqrt{a^2 - 1}, & a \in [1, \infty) \end{cases}$$

is minimized when $a = 1$. This is evident from the plotted eigenvalues in Figure 1.4, and solutions in Figure 1.5.

The need to fully understand physical situations such as this motivates our development of a spectral theory for matrices that are not diagonalizable.

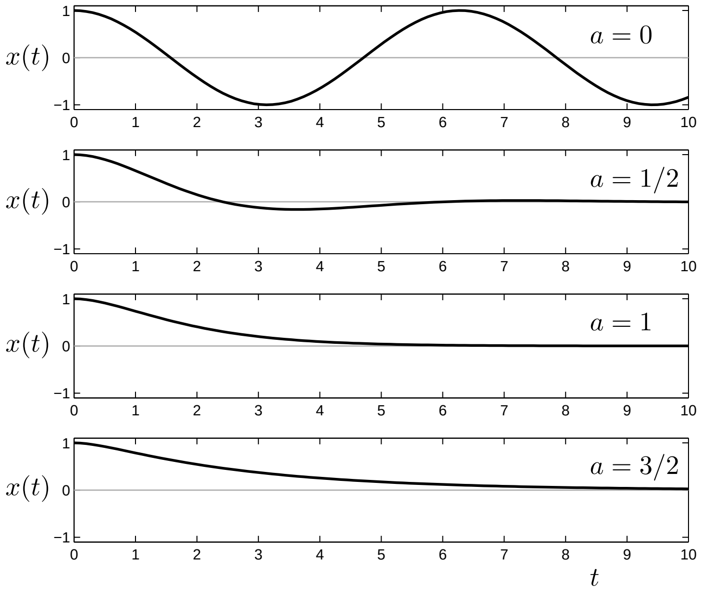

Figure 1.5. Solution $x(t)$ of the damped spring equation with $x(0) = 1$ and $x'(0) = 0$, for various values of a. The critical value $a = 1$ causes the most rapid decay of the solution, yet for this parameter the matrix $\mathbf{A}(a)$ is not diagonalizable.

1.8 Nondiagonalizable Matrices: The Jordan Form

Matrices $\mathbf{A} \in \mathbb{C}^{n\times n}$ that are not diagonalizable are rare (in a sense that can be made quite precise), but do arise in applications. Our goal in this section is to derive something as close as possible to diagonal form for such monsters, the famous Jordan canonical form. While the process we shall describe has a constructive flavor, we must emphasize this point: the "algorithm" to follow is not a stable computational procedure that applied mathematicians use in the heat of battle. In fact, the Jordan form is a fragile object that is rarely ever computed. So why work through the details? Comprehension of the steps that follow gives important insight into how matrices behave in a variety of settings, so while you may never need to build the Jordan form of a matrix, your understanding of its general properties will pay rich dividends.

We take two approaches here. In this section, we follow a constructive technique due to Fletcher & Sorensen [FS83], giving a (relatively) clean proof of the existence of the Jordan form. The next section gives a less rigorous but more concrete approach, illustrated with several examples.

1.8.1 Sylvester equation preliminaries

The first approach begins with a warm-up exercise, a quick introduction to Sylvester equations. Suppose that $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{B} \in \mathbb{C}^{m\times m}$, and $\mathbf{C} \in \mathbb{C}^{m\times n}$. We will soon have cause to think about solving equations of the form

$$\boxed{\mathbf{A}}\ \boxed{\mathbf{X}} - \boxed{\mathbf{X}}\ \boxed{\mathbf{B}} = \boxed{\mathbf{C}} \tag{1.27}$$

for the unknown $\mathbf{X} \in \mathbb{C}^{m\times n}$. Under what conditions on the square matrices $\mathbf{A}$ and $\mathbf{B}$ will this *Sylvester equation* have a unique solution? We will prove the following theorem constructively.

Unique solvability of Sylvester equations

Theorem 1.16. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{B} \in \mathbb{C}^{m\times m}$, *and* $\mathbf{C} \in \mathbb{C}^{m\times n}$. *There exists a unique matrix* $\mathbf{X} \in \mathbb{C}^{m\times n}$ *that solves the Sylvester equation* $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ *if and only if* $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) = \emptyset$, *i.e.,* $\mathbf{A}$ *and* $\mathbf{B}$ *have no eigenvalues in common.*

Proof. Following Sorensen [ES, Chap. 4], we shall prove this theorem constructively, using an algorithm for solving Sylvester equations called the Bartels–Stewart method [BS72]. Start by computing Schur factorizations $\mathbf{A} = \mathbf{UTU}^*$ and $\mathbf{B} = \mathbf{QRQ}^*$, so the Sylvester equation becomes

$$\mathbf{UTU}^*\mathbf{X} - \mathbf{XQRQ}^* = \mathbf{C}. \tag{1.28}$$

Premultiply both sides by $\mathbf{U}^*$ and postmultiply by $\mathbf{Q}$. Since $\mathbf{U}$ and $\mathbf{Q}$ are unitary matrices, we transform equation (1.28) to

$$\mathbf{TU}^*\mathbf{XQ} - \mathbf{U}^*\mathbf{XQR} = \mathbf{U}^*\mathbf{CQ}.$$

Define $\mathbf{Z} := \mathbf{U}^*\mathbf{XQ}$ and $\mathbf{D} := \mathbf{U}^*\mathbf{CQ}$. Now we can solve $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ uniquely if and only if we can solve

$$\mathbf{TZ} - \mathbf{ZR} = \mathbf{D} \tag{1.29}$$

uniquely, where $\mathbf{T}$ and $\mathbf{R}$ are upper-triangular matrices. We will solve for each column of $\mathbf{Z}$ successively, which will require us to handle $\mathbf{D}$ column-by-column. Thus we partition $\mathbf{Z}$ and $\mathbf{D}$ into columns

$$\mathbf{Z} = [\,\mathbf{z}_1 \quad \mathbf{z}_2 \quad \cdots \quad \mathbf{z}_m\,], \qquad \mathbf{D} = [\,\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_m\,],$$

where $\mathbf{z}_j, \mathbf{d}_j \in \mathbb{C}^n$ for $j = 1, \ldots, m$. As usual, let $\mathbf{e}_k$ denote the kth column of the identity matrix. Notice that

$$\mathbf{Z}\mathbf{e}_k = \mathbf{z}_k, \qquad \mathbf{R}\mathbf{e}_k = \begin{bmatrix} r_{1,k} \\ \vdots \\ r_{k,k} \\ \mathbf{0} \end{bmatrix} = \sum_{j=1}^{k} r_{j,k}\mathbf{e}_k, \qquad \mathbf{D}\mathbf{e}_k = \mathbf{d}_k,$$

noting the upper-triangular structure of $\mathbf{R}$. To solve for $\mathbf{z}_1$, the first column of $\mathbf{Z}$, premultiply (1.29) by $\mathbf{e}_1$ to get

$$\mathbf{T}\mathbf{Z}\mathbf{e}_1 - \mathbf{Z}\mathbf{R}\mathbf{e}_1 = \mathbf{D}\mathbf{e}_1,$$

which simplifies to

$$\mathbf{T}\mathbf{z}_1 - \mathbf{Z}\begin{bmatrix} r_{1,1} \\ \mathbf{0} \end{bmatrix} = \mathbf{d}_1.$$

The structure of that first column of $\mathbf{R}$ allows for a further simplification,

$$\mathbf{T}\mathbf{z}_1 - r_{1,1}\mathbf{z}_1 = \mathbf{d}_1.$$

Factor the right-hand side to give

$$(\mathbf{T} - r_{1,1}\mathbf{I})\mathbf{z}_1 = \mathbf{d}_1. \tag{1.30}$$

Notice that $\mathbf{T} - r_{1,1}\mathbf{I}$ is an upper triangular matrix, and so we can solve (1.30) uniquely for $\mathbf{z}_1$ if and only if none of the diagonal entries of $\mathbf{T} - r_{1,1}\mathbf{I}$ is zero, or, equivalently, if $r_{1,1}$ (an eigenvalue of $\mathbf{B}$) does not equal a diagonal entry of $\mathbf{T}$ (the eigenvalues of $\mathbf{A}$). Thus, we can compute a unique $\mathbf{z}_1$ if and only if $r_{1,1} \notin \sigma(\mathbf{A})$.

To compute $\mathbf{z}_2$, we follow a similar strategy. Multiply (1.29) against $\mathbf{e}_2$ to obtain $\mathbf{T}\mathbf{Z}\mathbf{e}_2 - \mathbf{Z}\mathbf{R}\mathbf{e}_2 = \mathbf{D}\mathbf{e}_2$, i.e.,

$$\mathbf{T}\mathbf{z}_2 - \mathbf{Z}(r_{1,2}\mathbf{e}_1 + r_{2,2}\mathbf{e}_2) = \mathbf{d}_2,$$

which further simplifies to

$$\mathbf{T}\mathbf{z}_2 - r_{1,2}\mathbf{z}_1 - r_{2,2}\mathbf{z}_2 = \mathbf{d}_2.$$

Since we have computed $\mathbf{z}_1$ already, move the $r_{1,2}\mathbf{z}_1$ term to the right-hand side to yield

$$(\mathbf{T} - r_{2,2}\mathbf{I})\mathbf{z}_2 = \mathbf{d}_2 + r_{1,2}\mathbf{z}_1.$$

This equation has a unique solution $\mathbf{z}_2$ if and only if the diagonal of $\mathbf{T}-r_{2,2}\mathbf{I}$ is nonzero, i.e., if $r_{2,2} \notin \sigma(\mathbf{A})$.

Continue this same procedure for the subsequent columns of $\mathbf{Z}$. At the kth step, we have

$$(\mathbf{T} - r_{k,k}\mathbf{I})\mathbf{z}_k = \mathbf{d}_k + \sum_{j=1}^{k-1} r_{j,k}\mathbf{z}_j,$$

which has a unique solution $\mathbf{z}_k$ if and only if $r_{k,k} \notin \sigma(\mathbf{A})$.

Continuing this process for $k = 2, \dots m$ gives $\mathbf{z}_1, \dots, \mathbf{z}_m$. Each step of the process produces a unique solution if and only if $r_{k,k} \notin \sigma(\mathbf{A})$ for $k = 1, \dots, m$. Since $\{r_{1,1}, \dots, r_{m,m}\} = \sigma(\mathbf{B})$, this means that a unique solution $\mathbf{Z}$ to (1.29) exists if and only if $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) = \emptyset$.

From the solution $\mathbf{Z}$, form $\mathbf{X} = \mathbf{U}\mathbf{Z}\mathbf{Q}^*$ to solve $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$. ■

Sylvester equations (and especially the special case of Lyapunov *equations*, $\mathbf{AX} + \mathbf{XA}^* = \mathbf{C}$) arise in many contexts in control theory. They will also be essential to the next step of our development.

1.8.2 Block diagonalization

We seek to transform any square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ into something as close as possible to diagonal form. The first (and, in a sense, most important) step is to transform $\mathbf{A}$ into *block diagonal form*. Suppose $\mathbf{A}$ has $p \le n$ distinct eigenvalues. We seek some invertible matrix $\mathbf{V} \in \mathbb{C}^{n\times n}$ such that

$$\mathbf{A} = \mathbf{V}\begin{bmatrix} \mathbf{T}_1 & & \\ & \ddots & \\ & & \mathbf{T}_p \end{bmatrix}\mathbf{V}^{-1}, \tag{1.31}$$

where the off-diagonal blocks are all zero, and each diagonal block $\mathbf{T}_k$ has a single eigenvalue $\sigma(\mathbf{T}_k) = \{\lambda_k\}$. (Since there are p blocks and p distinct eigenvalues, this implies that $\sigma(\mathbf{T}_j)\cap\sigma(\mathbf{T}_k) = \emptyset$ when $j \neq k$.) Our approach follows Fletcher & Sorensen [FS83].

Block Diagonalization

Theorem 1.17. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ have p distinct eiegnvalues $\lambda_1, \dots, \lambda_p$. There exists an invertible matrix $\mathbf{V} \in \mathbb{C}^{n\times n}$ such that (1.31) holds, where each block $\mathbf{T}_k$ has the single eigenvalue λ_k.*

Proof. Start with a SCHUR factorization, $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$. The proof will be inductive on the number of eigenvalues, p. If $p = 1$, the SCHUR factorization is trivially of the form (1.31). Inductively assume a factorization of the form (1.31) exists for matrices with $p - 1$ eigenvalues.

We can construct the SCHUR factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ so that eigenvalue λ_1 occurs in the leading a_1 entries on the main diagonal of $\mathbf{T}$, and in none of the subsequent diagonal entries. Thus, the SCHUR form has the form

$$\mathbf{A} = \mathbf{U}\begin{bmatrix} \mathbf{T}_1 & \mathbf{S} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}\mathbf{U}^*,$$

where $\mathbf{T}_1 \in \mathbb{C}^{a_1 \times a_1}$ with $\sigma(\mathbf{T}_1) = \{\lambda_1\}$ and $\sigma(\mathbf{T}_1) \cap \sigma(\widehat{\mathbf{T}}_1) = \emptyset$. (In what follows we will not emphasize the dimensions of the various block matrices that appear; presume they are all set to the proper dimensions that make matrix multiplication sensible.) We can think of block-diagonalizing as applying a series of transformations to $\mathbf{A}$ (a matrix on the left, its inverse on the right) to progressively move $\mathbf{A}$ to block-diagonal form. To start, the SCHUR factorization gives

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{C} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}.$$

Now we seek to apply some new transformation ($\mathbf{S}$ and $\mathbf{S}^{-1}$) to give

$$\mathbf{S}^{-1}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{S} = \mathbf{S}^{-1}\begin{bmatrix} \mathbf{T}_1 & \mathbf{C} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}\mathbf{S} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}. \tag{1.32}$$

To leave those $\mathbf{T}_1$ and $\widehat{\mathbf{T}}_1$ matrices on the main diagonal fixed by the transformation, we will look for $\mathbf{S}$ of the form

$$\mathbf{S} := \begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

for some $\mathbf{X}$ that remains to be determined. Notice that $\mathbf{S}$ has an easy inverse:

$$\begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

To figure out how to choose $\mathbf{X}$, compute

$$\begin{aligned} \mathbf{S}^{-1}\begin{bmatrix} \mathbf{T}_1 & \mathbf{C} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}\mathbf{S} &= \begin{bmatrix} \mathbf{I} & \mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{T}_1 & \mathbf{C} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}\begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{T}_1 & \mathbf{C} + \mathbf{X}\widehat{\mathbf{T}}_1 - \mathbf{T}_1\mathbf{X} \\ \mathbf{0} & \widehat{\mathbf{T}}_1 \end{bmatrix}. \end{aligned}$$

To conform with the goal (1.32), we need $\mathbf{X}$ to give $\mathbf{C}+\mathbf{X}\widehat{\mathbf{T}}_1-\mathbf{T}_1\mathbf{X}=\mathbf{0}$:

$$\mathbf{T}_1\mathbf{X}-\mathbf{X}\widehat{\mathbf{T}}_1=\mathbf{C}.$$

Conveniently enough, this is a SYLVESTER equation, as we have just studied! Theorem 1.16 ensures we can find a unique $\mathbf{X}$ that solves this equation provided $\sigma(\mathbf{T}_1)\cap\sigma(\widehat{\mathbf{T}}_1)=\emptyset$. The ordering of the eigenvalues in the SCHUR factor ensures that this is the case. Hence, we can find invertible $\mathbf{S}\in\mathbb{C}^{n\times n}$ so that

$$\mathbf{S}^{-1}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{S}=\begin{bmatrix}\mathbf{T}_1 & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{T}}_1\end{bmatrix}.$$

Notice that $\widehat{\mathbf{T}}_1$ has $p-1$ distinct eigenvalues. By the inductive assumption, we can find some invertible $\widehat{\mathbf{V}}\in\mathbb{C}^{(n-a_1)\times(n-a_1)}$ such that

$$\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{T}}_1\widehat{\mathbf{V}}=\begin{bmatrix}\mathbf{T}_2 & & \\ & \ddots & \\ & & \mathbf{T}_p\end{bmatrix},$$

with $\sigma(\mathbf{T}_j)=\{\lambda_j\}$ for $j=2,\ldots,p$. Thus,

$$\begin{aligned}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}^{-1}\end{bmatrix}\mathbf{S}^{-1}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{S}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}\end{bmatrix} &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{T}_1 & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{T}}_1\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{T}_1 & & & \\ & \mathbf{T}_2 & & \\ & & \ddots & \\ & & & \mathbf{T}_p\end{bmatrix}.\end{aligned}$$

Define

$$\mathbf{V}:=\mathbf{U}\mathbf{S}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}\end{bmatrix}\qquad\text{with}\qquad\mathbf{V}^{-1}=\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{V}}^{-1}\end{bmatrix}\mathbf{S}^{-1}\mathbf{U}^*,$$

so that

$$\mathbf{A}=\mathbf{V}\begin{bmatrix}\mathbf{T}_1 & & & \\ & \mathbf{T}_2 & & \\ & & \ddots & \\ & & & \mathbf{T}_p\end{bmatrix}\mathbf{V}^{-1},$$

thus establishing the theorem. ∎

1.8.3 The spectral representation of a matrix

Recall that in equation (1.16) we wrote a Hermitian $\mathbf{A} \in \mathbb{C}^{n\times n}$ in the form

$$\mathbf{A} = \sum_{j=1}^{n} \lambda_j \mathbf{P}_j,$$

where $\mathbf{P}_j$ was an orthogonal projector, and in (1.24) we obtained the same formula for a general *diagonalizable* $\mathbf{A}$, though in that case the projectors $\mathbf{P}_j$ need not be orthogonal. In this subsection we will use the block diagonalization to develop a similar formula that holds for *all matrices.*

Start with the block diagonalization (1.31), where the diagonal block $\mathbf{T}_j$ has dimension $a_j \times a_j$. The key step is to partition the matrix $\mathbf{V} \in \mathbb{C}^{n\times n}$ into *columns* of commensurate size,

$$\mathbf{V} = \left.\begin{array}{|c|c|c|c|} \hline & & & \\ \mathbf{V}_1 & \mathbf{V}_2 & \cdots & \mathbf{V}_p \\ & & & \\ \hline \end{array}\right\} n$$

$$\qquad\quad \underbrace{\phantom{\mathbf{V}_1}}_{a_1}\ \underbrace{\phantom{\mathbf{V}_2}}_{a_2}\ \ \underbrace{\phantom{\mathbf{V}_p}}_{a_p}$$

and to partition the *rows* of $\mathbf{V}^{-1}$ similarly:

$$\widehat{\mathbf{V}}^* := \mathbf{V}^{-1} = \underbrace{\begin{array}{|c|} \hline \widehat{\mathbf{V}}_1^* \\ \hline \widehat{\mathbf{V}}_2^* \\ \hline \cdots \\ \hline \widehat{\mathbf{V}}_p^* \\ \hline \end{array}}_{n} \begin{array}{l} \}\, a_1 \\ \}\, a_2 \\ \\ \}\, a_p \end{array}$$

The simple identity $\widehat{\mathbf{V}}^*\mathbf{V} = \mathbf{V}^{-1}\mathbf{V} = \mathbf{I}$ has an important implication:

$$\widehat{\mathbf{V}}^*\mathbf{V} = \begin{bmatrix} \widehat{\mathbf{V}}_1^*\mathbf{V}_1 & \cdots & \widehat{\mathbf{V}}_1^*\mathbf{V}_p \\ \vdots & \ddots & \vdots \\ \widehat{\mathbf{V}}_p^*\mathbf{V}_1 & \cdots & \widehat{\mathbf{V}}_p^*\mathbf{V}_p \end{bmatrix}$$

implies that

$$\widehat{\mathbf{V}}_j^*\mathbf{V}_k = \left\{ \begin{array}{ll} \mathbf{0} \in \mathbb{C}^{a_j \times a_k}, & j \neq k; \\ \mathbf{I} \in \mathbb{C}^{a_j \times a_j}, & j = k. \end{array} \right. \tag{1.33}$$

Flip the identity around in the opposite order, $\mathbf{V}\widehat{\mathbf{V}}^* = \mathbf{V}\mathbf{V}^{-1} = \mathbf{I}$, and we learn another interesting fact:

$$\mathbf{I} = \mathbf{V}\widehat{\mathbf{V}}^* = \mathbf{V}_1\widehat{\mathbf{V}}_1^* + \cdots + \mathbf{V}_p\widehat{\mathbf{V}}_p^*. \tag{1.34}$$

This statement means *the identity matrix can be expressed as the sum of the* $n\times n$ *matrices* $\mathbf{V}_1\widehat{\mathbf{V}}_1^*, \ldots, \mathbf{V}_p\widehat{\mathbf{V}}_p^*$. We say that these $\mathbf{V}_j\widehat{\mathbf{V}}_j^*$ form *a resolution of the identity.*

We shall now use the partitioning of $\mathbf{V}$ and $\widehat{\mathbf{V}}^*$ into columns and rows to express the block diagonalization (1.31) in a new manner. We write

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{V}_1 & \cdots & \mathbf{V}_p \end{array}\right] \begin{bmatrix} \mathbf{T}_1 & & \\ & \ddots & \\ & & \mathbf{T}_p \end{bmatrix} \left[\begin{array}{c} \widehat{\mathbf{V}}_1^* \\ \hline \vdots \\ \hline \widehat{\mathbf{V}}_p^* \end{array}\right]. \tag{1.35}$$

Multiply the first two matrices together to get

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{V}_1\mathbf{T}_1 & \cdots & \mathbf{V}_p\mathbf{T}_p \end{array}\right] \left[\begin{array}{c} \widehat{\mathbf{V}}_1^* \\ \hline \vdots \\ \hline \widehat{\mathbf{V}}_p^* \end{array}\right],$$

and then take the product of these last two matrices to arrive at

$$\mathbf{A} = \mathbf{V}_1\mathbf{T}_1\widehat{\mathbf{V}}_1^* + \cdots + \mathbf{V}_p\mathbf{T}_p\widehat{\mathbf{V}}_p^*. \tag{1.36}$$

Let us take a close look at one of the terms in this sum,

$$\mathbf{V}_j\mathbf{T}_j\widehat{\mathbf{V}}_j^* \in \mathbb{C}^{n\times n}. \tag{1.37}$$

Since $\mathbf{T}_j$ is upper triangular, with eigenvalue λ_j on the main diagonal, we can write

$$\mathbf{T}_j = \lambda_j\mathbf{I} + \mathbf{R}_j,$$

where $\mathbf{R}_j \in \mathbb{C}^{a_j\times a_j}$ is *strictly upper triangular*, meaning that it has zero on and below the main diagonal. Thus we can write (1.37) as

$$\begin{aligned} \mathbf{V}_j\mathbf{T}_j\widehat{\mathbf{V}}_j^* &= \mathbf{V}_j(\lambda_j\mathbf{I} + \mathbf{R}_j)\widehat{\mathbf{V}}_j^* \\ &= \lambda_j\mathbf{V}_j\widehat{\mathbf{V}}_j^* + \mathbf{V}_j\mathbf{R}_j\widehat{\mathbf{V}}_j^*. \end{aligned}$$

Substituting such expressions for each j into the sum (1.36) gives

$$\mathbf{A} = \sum_{j=1}^{p}\Big(\lambda_j\mathbf{V}_j\widehat{\mathbf{V}}_j^* + \mathbf{V}_j\mathbf{R}\widehat{\mathbf{V}}_j^*\Big). \tag{1.38}$$

Compare this sum to the expression (1.16) for Hermitian $\mathbf{A}$ and (1.24) for diagonalizable $\mathbf{A}$. Inspired by the similarity, we establish the following notation.

Spectral Projectors and Nilpotents

Definition 1.18. *Using the notation of this subsection, the spectral projector* $\mathbf{P}_j$ *and the spectral nilpotent associated with the eigenvalue* λ_j *are defined by*

$$\mathbf{P}_j := \mathbf{V}_j\widehat{\mathbf{V}}_j^*, \qquad \mathbf{D}_j := \mathbf{V}_j\mathbf{R}_j\widehat{\mathbf{V}}_j^*.$$

The names of $\mathbf{P}_j$ and $\mathbf{D}_j$ were not selected by accident! Notice that

$$\mathbf{P}_j^2 = \mathbf{V}_j\widehat{\mathbf{V}}_j^*\mathbf{V}_j\widehat{\mathbf{V}}_j^* = \mathbf{V}_j(\widehat{\mathbf{V}}_j^*\mathbf{V}_j)\widehat{\mathbf{V}}_j^* = \mathbf{V}_j(\mathbf{I})\widehat{\mathbf{V}}_j^* = \mathbf{P}_j,$$

so $\mathbf{P}_j$ is a projector. (Recall that (1.33) implies that $\widehat{\mathbf{V}}_j^*\mathbf{V}_j = \mathbf{I}$.) Similarly, since $\widehat{\mathbf{V}}_j^*\mathbf{V}_k = \mathbf{0}$ when $j \neq k$, we also have the important fact that

$$\mathbf{P}_j\mathbf{P}_k = \mathbf{0}, \qquad j \neq k.$$

We call $\mathbf{D}_j$ the *spectral nilpotent* because it is a *nilpotent matrix*, meaning that $\mathbf{D}_j^m = 0$ for some positive integer m. This fact follows easily from the strictly upper triangular structure of $\mathbf{R}_j$:

$$\begin{aligned}\mathbf{D}_j^{a_j} &= (\mathbf{V}_j\mathbf{R}_j\widehat{\mathbf{V}}_j^*)^{a_j}\\ &= \mathbf{V}_j\mathbf{R}_j^{a_j}\widehat{\mathbf{V}}_j^* = \mathbf{V}_j\mathbf{0}\widehat{\mathbf{V}}_j^* = \mathbf{0}.\end{aligned}$$

We formally collect the sum (1.38) and the related facts.

Spectral Representation of a Square Matrix

Theorem 1.19. *Suppose the matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *has* p *distinct eigenvalues* $\lambda_1, \ldots, \lambda_p$. *There exist projectors* $\mathbf{P}_1, \ldots, \mathbf{P}_p$ *and nilpotent matrices* $\mathbf{D}_1, \ldots, \mathbf{D}_p$ *such that*

$$\mathbf{A} = \sum_{j=1}^{p} \lambda_j\mathbf{P}_j + \sum_{j=1}^{p} \mathbf{D}_j.$$

1.8.4 The Jordan form

For many purposes, the spectral representation in Theorem 1.19 gives entirely adequate insight into the spectral structure of a matrix. However, it is possible to push a bit harder, to impose some extra structure upon the upper triangular matrices $\mathbf{T}_j$. This extra structure will demand some work, but the effort will give you excellent practice working with block-matrix multiplication and similarity transformations. We start with some definitions.

Algebraic and Geometric Multiplicity

Definition 1.20. *Let λ_j be an eigenvalue of $\mathbf{A} \in \mathbb{C}^{n\times n}$.*

The algebraic multiplicity *a_j of λ_j is the number of times λ_j appears on the diagonal of $\mathbf{T}$ in the* Schur *factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$.*

The geometric multiplicity *g_j of λ_j is the number of linearly independent eigenvectors of $\mathbf{A}$ corresponding to λ_j, i.e., $g_j = \dim(\mathcal{N}(\mathbf{A} - \lambda_j\mathbf{I}))$.*

To obtain the Jordan form, we simply need to *independently* transform each of the $\mathbf{T}_j \in \mathbb{C}^{a_j\times a_j}$ matrices into something close to diagonal form. For each of these upper triangular matrices we shall build an invertible $\mathbf{W}_j \in \mathbb{C}^{a_j\times a_j}$ such that

$$\mathbf{W}_j\mathbf{T}_j\mathbf{W}_j^{-1} = \mathbf{J}_j, \tag{1.39}$$

where $\mathbf{J}_j \in \mathbb{C}^{a_j\times a_j}$ is a block diagonal matrix with g_j blocks:

$$\mathbf{J}_j = \begin{bmatrix} \mathbf{J}_{j,1} & & & \\ & \mathbf{J}_{j,2} & & \\ & & \ddots & \\ & & & \mathbf{J}_{j,g_j} \end{bmatrix}. \tag{1.40}$$

Each of these sub-blocks is nearly diagonal; in fact, they are *bidiagonal*, with nonzero entries only on the main diagonal and the first superdiagonal:

$$\mathbf{J}_{j,k} = \begin{bmatrix} \lambda_j & 1 & & \\ & \lambda_j & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}. \tag{1.41}$$

A matrix of this form is called a *Jordan block*.

Our ultimate goal is to find an invertible matrix $\mathbf{X}$ such that $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ is a block diagonal matrix in which all the submatrices on the diagonal are Jordan blocks. This arrangement is the *Jordan canonical form* of $\mathbf{A}$.

From (1.39) we have $\mathbf{T}_j = \mathbf{W}_j\mathbf{J}_j\mathbf{W}_j^{-1}$. Substitute this form of $\mathbf{T}_J$ into the block-diagonalization (1.31) to obtain Substitute the transformations (1.39) of the diagonal blocks $\mathbf{T}_j$ to obtain

$$\begin{aligned}
\mathbf{A} &= \mathbf{V}\begin{bmatrix} \mathbf{T}_1 & & \\ & \ddots & \\ & & \mathbf{T}_p \end{bmatrix}\mathbf{V}^{-1} = \mathbf{V}\begin{bmatrix} \mathbf{W}_1\mathbf{J}_1\mathbf{W}_1^{-1} & & \\ & \ddots & \\ & & \mathbf{W}_p\mathbf{J}_p\mathbf{W}_p^{-1} \end{bmatrix}\mathbf{V}^{-1} \\
&= \mathbf{V}\begin{bmatrix} \mathbf{W}_1 & & \\ & \ddots & \\ & & \mathbf{W}_p \end{bmatrix}\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\begin{bmatrix} \mathbf{W}_1^{-1} & & \\ & \ddots & \\ & & \mathbf{W}_p^{-1} \end{bmatrix}\mathbf{V}^{-1} \\
&= \mathbf{X}\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\mathbf{X}^{-1},
\end{aligned}$$

where

$$\mathbf{X} = \mathbf{V}\begin{bmatrix} \mathbf{W}_1 & & \\ & \ddots & \\ & & \mathbf{W}_p \end{bmatrix}.$$

This factorization

$$\mathbf{A} = \mathbf{X}\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\mathbf{X}^{-1} \tag{1.42}$$

is the *Jordan canonical form* of $\mathbf{A}$. Given what we already know, we "simply" need to discover these $\mathbf{W}_j$ transformations to complete our derivation of this pivotal factorization.

The derivation of these $\mathbf{W}_j$ transformations turns out to be rather intricate. Before plunging in, pause for a moment to think about a special case (not just a special case; in fact, the most common case).

What if there are as many (linearly independent) eigenvectors as there are eigenvalues, i.e., $a_j = g_j$? Then

$$a_j = g_j = \dim(\mathcal{N}(\mathbf{A} - \lambda_j\mathbf{I})) = \dim(\mathcal{N}(\mathbf{T} - \lambda_j\mathbf{I})).$$

Since $\lambda_j \notin \sigma(\mathbf{T}_k)$ for $k \neq j$, the submatrix $\mathbf{T}_k - \lambda_j\mathbf{I}$ is invertible for $k \neq j$. Thus, null space of $\mathbf{T} - \lambda_j\mathbf{I}$ all comes from the $\mathbf{T}_j - \lambda_j\mathbf{I}$ block:

$$a_j = \dim(\mathcal{N}(\mathbf{T} - \lambda_j\mathbf{I})) = \dim(\mathcal{N}(\mathbf{T}_j - \lambda_j\mathbf{I})),$$

so $\mathbf{T}_j - \lambda_j \mathbf{I}$ is a $a_j \times a_j$ matrix whose null space has dimension a_j. This means that $(\mathbf{T}_j - \lambda_j \mathbf{I})\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{C}^{a_j}$, which is only possible if $\mathbf{T}_j - \lambda_j \mathbf{I} = \mathbf{0}$: hence $\mathbf{T}_j = \lambda_j \mathbf{I}$, and the upper triangular block is already diagonal, in the form (1.40). The transformation $\mathbf{W}_j$ is trivial in this common, important case: $\mathbf{W}_j = \mathbf{I}$. Thus, the proof below is only necessary for those monsters where there is a deficit of linearly independent eigenvectors, $g_j < a_j$.

1.8.5 Proof of the Jordan form

We continue to follow the proof of Fletcher & Sorensen [FS83].

When a Jordan block of the form $\mathbf{J}_{j,k}$ in (1.41) has a zero eigenvalue, notice that it can be written in a concise block form as

$$\mathbf{E} = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \mathbf{0} \end{bmatrix} \in \mathbb{C}^{d \times d}, \tag{1.43}$$

where the identity matrix is of size dimension $d-1 \times d-1$. For example, when $d = 4$,

$$\mathbf{E} = \left[\begin{array}{c|ccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 \end{array}\right] = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \mathbf{0} \end{bmatrix}.$$

It will be helpful to notice that

$$\mathbf{E}^*\mathbf{E} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \mathbf{I} - \mathbf{E}^*\mathbf{E} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{1.44}$$

In this subsection we will focus on transforming a single triangular block $\mathbf{T}_j$ with eigenvalue λ_j into the Jordan form (1.40). *Since we focus on the one block, we will drop the subscript j.* We shall thus deal with

$$\mathbf{T} = \lambda \mathbf{I} + \mathbf{R},$$

where $\mathbf{R}$ is *strictly upper triangular* (it has zero on the main diagonal).

Theorem 1.21. *For any strictly upper triangular matrix* $\mathbf{R} \in \mathbb{C}^{a\times a}$ *there exists an invertible* $\mathbf{W} \in \mathbb{C}^{a\times a}$ *such that*

$$\mathbf{R} = \mathbf{W}\mathbf{J}\mathbf{W}^{-1},$$

where $\mathbf{N} = \mathrm{diag}(\mathbf{E}_1, \ldots, \mathbf{E}_g)$ *with the blocks*

$$\mathbf{E}_k = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \mathbf{0} \end{bmatrix} \in \mathbb{C}^{d_k\times d_k}$$

arranged by decreasing dimension, $d_1 \ge d_2 \ge \cdots \ge d_m$*, that sum to* a*:* $d_1 + \cdots + d_m = a$*.*

(Here the notation $\mathbf{N}$ is meant to denote *nilpotent*, since $\mathbf{N}^{d_1} = \mathbf{N}^a = \mathbf{0}$.)

Proof. We prove this theorem by induction on the dimension, a. If $a = 1$, the strictly upper triangular $\mathbf{R} \in \mathbb{C}^{1\times 1}$ is simply $\mathbf{R} = 0$, so take $\mathbf{W} = 1$ and $\mathbf{N} = 0$.

Now suppose the result holds for all strictly upper triangular matrices of dimension $a - 1$ or less. Partition $\mathbf{R} \in \mathbb{C}^{a\times a}$ as

$$\mathbf{R} = \begin{bmatrix} 0 & \mathbf{r}^* \\ \mathbf{0} & \widehat{\mathbf{R}} \end{bmatrix}.$$

We will execute a series of similarity transformations that take $\mathbf{R}$ to the desired form.

• **First similarity transformation**. Using the inductive assumption, write

$$\widehat{\mathbf{R}} = \mathbf{W}_1\mathbf{N}_1\mathbf{W}_1^{-1}, \qquad \mathbf{N}_1 = \mathrm{diag}(\mathbf{E}_1, \ldots, \mathbf{E}_q),$$

with the $\mathbf{E}_j$ blocks of decreasing dimension,$d_1 \ge d_2 \ge \cdots \ge d_q$. Define

$$\mathbf{S}_1 := \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_1 \end{bmatrix}, \quad \text{with} \quad \mathbf{S}_1^{-1} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_1^{-1} \end{bmatrix},$$

so that

$$\mathbf{R}_1 := \mathbf{S}_1^{-1}\mathbf{R}\mathbf{S}_1 = \begin{bmatrix} 0 & \mathbf{r}^*\mathbf{W}_1^{-1} \\ 0 & \mathbf{N}_1 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{s}^* & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix}, \tag{1.45}$$

where we have introduced the notation

$$[\,\mathbf{s}^* \quad \mathbf{t}^*\,] := \mathbf{r}^*\mathbf{W}_1^{-1}, \qquad \mathbf{N}_1 =: \mathrm{diag}(\mathbf{E}_1, \mathbf{N}_2).$$

• **Second similarity transformation**. We seek to hammer the $\mathbf{s}^*$ and $\mathbf{t}^*$ vectors into the correct form. First, we will transform $\mathbf{s}^*$ so that it only has a nonzero in its first entry. Define

$$\mathbf{S}_2 := \begin{bmatrix} 1 & \mathbf{s}^*\mathbf{E}_1^* & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{with} \quad \mathbf{S}_2^{-1} = \begin{bmatrix} 1 & -\mathbf{s}^*\mathbf{E}_1^* & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}$$

so that

$$\mathbf{R}_2 := \mathbf{S}_2^{-1}\mathbf{R}_1\mathbf{S}_2 = \begin{bmatrix} 0 & \mathbf{s}^* - \mathbf{s}^*\mathbf{E}_1^*\mathbf{E}_1 & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix} = \begin{bmatrix} 0 & \rho\mathbf{e}_1^* & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix},$$

where the $(1,2)$ entry is $\mathbf{s}^* - \mathbf{s}^*\mathbf{E}_1^*\mathbf{E}_1 = \mathbf{s}^*(\mathbf{I} - \mathbf{E}_1^*\mathbf{E}_1) = \mathbf{s}^*\mathbf{e}_1\mathbf{e}_1^*$ (see (1.44)), and we have defined $\rho := \mathbf{s}^*\mathbf{e}_1$.

The proof now bifurcates into two cases, depending on the value of ρ.

• **$\boldsymbol{\rho = 0}$, third similarity transformation**. When $\rho = 0$, $\mathbf{R}_2$ has the form

$$\mathbf{R}_2 = \begin{bmatrix} 0 & \mathbf{0} & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

It will take a few transformations to make this fact evident, but this structure implies that $\mathbf{E}_1$ is a JORDAN block of $\mathbf{R}$. First swap rows and columns to place $\mathbf{E}_1$ in the $(1,1)$ block. Define

$$\mathbf{S}_3 := \begin{bmatrix} 0 & \mathbf{I} & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{with} \quad \mathbf{S}_3^{-1} = \begin{bmatrix} 0 & 1 & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Premultiplication by $\mathbf{S}_3$ swaps the first two (block) rows; postmultiplication by $\mathbf{S}_3^{-1}$ swaps the first two (block columns. Thus

$$\mathbf{R}_3 := \mathbf{S}_3^{-1}\mathbf{R}_2\mathbf{S}_3 = \left[\begin{array}{c|cc} \mathbf{E}_1 & \mathbf{0} & \mathbf{0} \\ \cline{2-3} \mathbf{0} & 0 & \mathbf{t}^* \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{array}\right].$$

• **$\boldsymbol{\rho = 0}$, fourth similarity transformation**. Notice that 2×2 block matrix at the bottom-right of $\mathbf{R}_3$: it has dimension no larger than $(a-1) \times (a-1)$ (since $\mathbf{E}_1$ has dimension *at least* 1×1). Thus we can apply the inductive assumption again to factor

$$\begin{bmatrix} 0 & \mathbf{t}^* \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix} = \widehat{\mathbf{W}}\mathbf{N}_3\widehat{\mathbf{W}}^{-1},$$

where

$$\mathbf{N}_3 = \mathrm{diag}(\widehat{\mathbf{E}}_2, \ldots, \widehat{\mathbf{E}}_g),$$

which implicitly defines the integer g. Thus

$$\mathbf{R}_3 = \begin{bmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{W}}\mathbf{N}_3\widehat{\mathbf{W}}^{-1} \end{bmatrix}.$$

Define the transformation

$$\mathbf{S}_4 := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{W}} \end{bmatrix}, \quad \text{with} \quad \mathbf{S}_4^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{W}}^{-1} \end{bmatrix},$$

so

$$\mathbf{R}_4 := \mathbf{S}_4^{-1}\mathbf{R}_3\mathbf{S}_4 = \begin{bmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_3 \end{bmatrix}.$$

Now this series of transformations have taken $\mathbf{R}$ into a block diagonal matrix with the required diagonal blocks. One small detail remains: We want to arrange the blocks in decreasing size. It might be possible that $\mathbf{E}_1$ is smaller than one of the $\widehat{\mathbf{E}}_j$ matrices that fall on the diagonal of $\mathbf{N}_3$.

- **$\boldsymbol{\rho = 0}$, fifth similarity transformation**. Introduce a matrix $\mathbf{S}_5$ that is a permutation of the rows and columns of the identity matrix (akin to $\mathbf{S}_3$ above) so that

$$\mathbf{R}_5 = \mathbf{S}_5^{-1}\mathbf{R}_4\mathbf{S}_5 = \mathrm{diag}(\mathbf{E}_1, \ldots, \mathbf{E}_g) =: \mathbf{N},$$

where we have relabelled the matrices $\mathbf{E}_1, \widehat{\mathbf{E}}_2, \ldots, \widehat{\mathbf{E}}_g$ to have new names $\mathbf{E}_1, \ldots, \mathbf{E}_g$ with $\mathbf{E}_j \in \mathbb{C}^{d_j \times d_j}$ and $d_1 \geq d_2 \geq \cdots \geq d_g$, as required.

- **$\boldsymbol{\rho = 0}$, wrap-up**. We can now wrap-up the $\rho = 0$ case. We have built invertible matrices $\mathbf{S}_1, \ldots, \mathbf{S}_5$ so that

$$\begin{aligned} \mathbf{N} = \mathrm{diag}(\mathbf{E}_1, \ldots, \mathbf{E}_g) &= \mathbf{S}_5^{-1}\mathbf{R}_4\mathbf{S}_5 \\ &= \mathbf{S}_5^{-1}\mathbf{S}_4^{-1}\mathbf{R}_3\mathbf{S}_4\mathbf{S}_5 \\ &= \mathbf{S}_5^{-1}\mathbf{S}_4^{-1}\mathbf{S}_3^{-1}\mathbf{R}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5 \\ &= \mathbf{S}_5^{-1}\mathbf{S}_4^{-1}\mathbf{S}_3^{-1}\mathbf{S}_2^{-1}\mathbf{R}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5 \\ &= \mathbf{S}_5^{-1}\mathbf{S}_4^{-1}\mathbf{S}_3^{-1}\mathbf{S}_2^{-1}\mathbf{S}_1^{-1}\mathbf{R}\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5 \\ &= (\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5)^{-1}\mathbf{R}(\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5) = \mathbf{W}^{-1}\mathbf{R}\mathbf{W}, \end{aligned}$$

where $\mathbf{W} := \mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5$. The proof is complete for the $\rho = 0$ case.

• **$\boldsymbol{\rho \neq 0}$, third similarity transformation**. Now we attend to the remaining case, $\rho \neq 0$. Recall that the first two similarity transformations left us with

$$\mathbf{R}_2 = \mathbf{S}_2^{-1}\mathbf{R}_1\mathbf{S}_2 = \begin{bmatrix} 0 & \rho\mathbf{e}_1^* & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

The $\rho \neq 0$ case corresponds to the situation where the $\mathbf{E}_1$ block will be enlarged by one dimension. The specific value of $\rho \neq 0$ will not matter, so scale it out via the transformation

$$\mathbf{S}_3 := \begin{bmatrix} \rho & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ 0 & \mathbf{0} & \rho \end{bmatrix}, \quad \text{with} \quad \mathbf{S}_3^{-1} = \begin{bmatrix} 1/\rho & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ 0 & \mathbf{0} & 1/\rho \end{bmatrix}.$$

Premultiplication of $\mathbf{R}_2$ by $\mathbf{S}_3^{-1}$ scales the first and third *rows* of $\mathbf{R}_2$ by $1/\rho$; postmultiplication by $\mathbf{S}_3$ scales the first and third *columns* by ρ. As a result,

$$\mathbf{R}_3 := \mathbf{S}_3^{-1}\mathbf{R}_2\mathbf{S}_3 = \begin{bmatrix} 0 & \mathbf{e}_1^* & \mathbf{t}^* \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

Notice that the upper-left 2×2 block forms an enlarged Jordan block of the form (1.43). If $\mathbf{E}_1 \in \mathbb{C}^{d_1 \times d_1}$, then by the structure of $\mathbf{E}_1$ (see (1.43)),

$$\begin{bmatrix} 0 & \mathbf{e}_1^* \\ \mathbf{0} & \mathbf{E}_1 \end{bmatrix} = \left[\begin{array}{c|cc} 0 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \hline \mathbf{0} & 0 & \mathbf{0} \end{array}\right] =: \mathbf{E} \in \mathbb{C}^{(d_1+1)\times(d_1+1)}. \tag{1.46}$$

Thus, $\mathbf{E}_1$ has been enlarged by one dimension to give $\mathbf{E}$.

• **$\boldsymbol{\rho \neq 0}$, remaining similarity transformation**. With the definition for $\mathbf{E}$ in (1.46) we can write

$$\mathbf{R}_3 = \begin{bmatrix} \mathbf{E} & \mathbf{e}_1\mathbf{t}^* \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

The rest of the proof executes successive transformations that progressively clear $\mathbf{e}_1\mathbf{t}^*$ out of the $(1,2)$ block of $\mathbf{R}_3$. First we need to think carefully about dimensions. Recall that $\mathbf{N}_2$ emerged in (1.45) through invocation of the inductive assumption, with $\mathbf{N}_2 = \text{diag}(\mathbf{E}_2, \ldots, \mathbf{E}_q)$. That inductive assumption gave blocks $\mathbf{E}_j$ of dimension $d_j \times d_j$ with $d_1 \geq d_2 \geq \cdots \geq d_q$. Thus

$$\mathbf{N}_2^{d_1} = \text{diag}(\mathbf{E}_2^{d_1}, \ldots, \mathbf{E}_q^{d_1}) = \mathbf{0}, \tag{1.47}$$

due to the dimension of the $\mathbf{E}_j$ blocks and their nilpotent structure. Introduce the new transformation

$$\widetilde{\mathbf{S}}_1 := \begin{bmatrix} 1 & \mathbf{e}_2\mathbf{t}^* \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \widetilde{\mathbf{S}}_1^{-1} := \begin{bmatrix} 1 & -\mathbf{e}_2\mathbf{t}^* \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

and form

$$\widetilde{\mathbf{R}}_1 := \widetilde{\mathbf{S}}_1^{-1}\mathbf{R}_3\widetilde{\mathbf{S}}_1 = \begin{bmatrix} \mathbf{E} & \mathbf{e}_2\mathbf{t}^*\mathbf{N}_2 + \mathbf{e}_1\mathbf{t}^* - \mathbf{E}\mathbf{e}_2\mathbf{t}^* \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

Notice from the structure of $\mathbf{E}$ in (1.46) that

$$\mathbf{E}\mathbf{e}_1 = \mathbf{e}_2$$

and hence we can simplify $\widetilde{\mathbf{R}}_1$ to

$$\widetilde{\mathbf{R}}_1 = \begin{bmatrix} \mathbf{E} & \mathbf{e}_2\mathbf{t}^*\mathbf{N}_2 \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

Notice the alteration we have made to the $(1,2)$ block. The next step makes the pattern clear: proceed with

$$\widetilde{\mathbf{S}}_2 := \begin{bmatrix} 1 & \mathbf{e}_3\mathbf{t}^*\mathbf{N}_2 \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \widetilde{\mathbf{S}}_2^{-1} := \begin{bmatrix} 1 & -\mathbf{e}_3\mathbf{t}^*\mathbf{N}_2 \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

so, noting that $\mathbf{E}\mathbf{e}_3 = \mathbf{e}_2$, we arrive at

$$\widetilde{\mathbf{R}}_2 := \widetilde{\mathbf{S}}_2^{-1}\widetilde{\mathbf{R}}_1\widetilde{\mathbf{S}}_2 = \begin{bmatrix} \mathbf{E} & \mathbf{e}_3\mathbf{t}^*\mathbf{N}_2^2 \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

Notice that higher power of $\mathbf{N}_2$ in the new $(1,2)$ block. Now proceed apace: for $k \geq 3$, define

$$\widetilde{\mathbf{S}}_k := \begin{bmatrix} 1 & \mathbf{e}_{k+1}\mathbf{t}^*\mathbf{N}_2^{k-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \widetilde{\mathbf{S}}_k^{-1} := \begin{bmatrix} 1 & -\mathbf{e}_{k+1}\mathbf{t}^*\mathbf{N}_2^{k-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

and with $\mathbf{E}\mathbf{e}_{k+1} = \mathbf{e}_k$, we arrive at

$$\widetilde{\mathbf{R}}_k := \widetilde{\mathbf{S}}_k^{-1}\widetilde{\mathbf{R}}_{k-1}\widetilde{\mathbf{S}}_k = \begin{bmatrix} \mathbf{E} & \mathbf{e}_{k+1}\mathbf{t}^*\mathbf{N}_2^k \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix}.$$

Repeat these steps for $k = 3, \ldots, d_1$. The last step delivers, by invoking $\mathbf{N}_2^{d_1} = \mathbf{0}$ in (1.47),

$$\widetilde{\mathbf{R}}_{d_1} = \begin{bmatrix} \mathbf{E} & \mathbf{e}_{d_1+1}\mathbf{t}^*\mathbf{N}_2^{d_1} \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 \end{bmatrix} = \operatorname{diag}(\mathbf{E}, \mathbf{E}_2, \ldots, \mathbf{E}_q). \tag{1.48}$$

Check that $\mathbf{e}_{d_1+1}$ in the $(1,2)$ block makes sense: The first block row of $\widetilde{\mathbf{R}}_{d_1}$ has d_1+1 rows, so the vector $\mathbf{e}_{d_1+1}$ is the last column of the $(d_1+1)\times(d_1+1)$ identity matrix: we have not exceeded the allowable dimensions.

Finally, we have reduced $\mathbf{R}$ to a block diagonal form. Renaming $\mathbf{E}_1 := \mathbf{E}$ and setting $g = q$, we have

$$\begin{aligned}
\mathbf{N} = \mathrm{diag}(\mathbf{E}_1,\ldots,\mathbf{E}_g) &= \widetilde{\mathbf{R}}_{d_1} \\
&= \widetilde{\mathbf{S}}_{d_1}^{-1}\widetilde{\mathbf{R}}_{d_1-1}\widetilde{\mathbf{S}}_{d-1} \\
&= \widetilde{\mathbf{S}}_{d_1}^{-1}\cdots\widetilde{\mathbf{S}}_1^{-1}\mathbf{R}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1} \\
&= \widetilde{\mathbf{S}}_{d_1}^{-1}\cdots\widetilde{\mathbf{S}}_1^{-1}\mathbf{S}_3^{-1}\mathbf{R}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1} \\
&= \widetilde{\mathbf{S}}_{d_1}^{-1}\cdots\widetilde{\mathbf{S}}_1^{-1}\mathbf{S}_3^{-1}\mathbf{S}_2^{-1}\mathbf{R}_1\mathbf{S}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1} \\
&= \widetilde{\mathbf{S}}_{d_1}^{-1}\cdots\widetilde{\mathbf{S}}_1^{-1}\mathbf{S}_3^{-1}\mathbf{S}_2^{-1}\mathbf{S}_1^{-1}\mathbf{R}\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1} \\
&= \Big(\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1}\Big)^{-1}\mathbf{R}\Big(\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1}\Big) \\
&= \mathbf{W}^{-1}\mathbf{R}\mathbf{W},
\end{aligned}$$

where $\mathbf{W} := \mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\widetilde{\mathbf{S}}_1\cdots\widetilde{\mathbf{S}}_{d-1}$. This completes the proof for the $\rho \neq 0$ case.

Having handled the $\rho = 0$ case (where $\mathbf{E}_1$ is a stand-alone Jordan block) and the $\rho \neq 0$ case (where $\mathbf{E}_1$ is expanded by one dimension), we have completed the proof for all strictly upper triangular $\mathbf{R} \in \mathbb{C}^{a\times a}$. ■

The number of Jordan blocks, g, in the factor $\mathbf{N} = \mathbf{W}^{-1}\mathbf{R}\mathbf{W}$ in Theorem 1.21 reveals the number of (linearly independent) eigenvectors of $\mathbf{T} = \lambda\mathbf{I} + \mathbf{R}$ associated with the eigenvalue λ. To see this, we should assess the null space of $\mathbf{T} - \lambda\mathbf{I}$. Since $\mathbf{W}$ is invertible,

$$\dim(\mathcal{N}(\mathbf{T} - \lambda\mathbf{I})) = \dim(\mathcal{N}(\mathbf{R})) = \dim(\mathcal{N}(\mathbf{W}\mathbf{N}\mathbf{W}^{-1})) = \dim(\mathcal{N}(\mathbf{N})).$$

Recall from basic linear algebra that $\mathbf{N}$ is in reduced row-echelon form, and the first column of every Jordan block of $\mathbf{N}$ is a "free column"; the other columns of $\mathbf{N}$ are "pivot columns". From this one intuits that $\dim(\mathcal{N}(\mathbf{N})) = g$ and $\dim(\mathcal{R}(\mathbf{N})) = a-g$. Notice that $\mathbf{T}\mathbf{v} = \lambda\mathbf{v}$ if and only if $\mathbf{v} \in \mathcal{N}(\mathbf{R})$, which is true if and only if $\mathbf{W}^{-1}\mathbf{v} \in \mathcal{N}(\mathbf{N})$. Hence $\mathbf{T}$ has g linearly independent eigenvectors associated with λ. Thus, g is the *geometric multiplicity* of λ, as given in Definition 1.20.

Assembling the Jordan from from Theorem 1.21

Return to the full matrix $\mathbf{A}$ with its p diagonal blocks, $\mathbf{T}_j = \lambda_j\mathbf{I}+\mathbf{R}_j$. Apply Theorem 1.21 to $\mathbf{R}_j$ to get $\mathbf{R}_j = \mathbf{W}_j\mathbf{N}_j\mathbf{W}_j^{-1}$. Then

$$\mathbf{T}_j = \lambda_j\mathbf{I} + \mathbf{W}_j\mathbf{N}_j\mathbf{W}_j^{-1} = \mathbf{W}_j(\lambda_j\mathbf{I} + \mathbf{N}_j)\mathbf{W}_j^{-1}. \tag{1.49}$$

Define $\mathbf{J}_j := \lambda_j\mathbf{I} + \widehat{\mathbf{J}}_j$, and we have

$$\mathbf{T}_j = \mathbf{W}_j\mathbf{J}\mathbf{W}_j^{-1},$$

thus finishing the construction of the Jordan form

$$\mathbf{A} = \mathbf{X}\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\mathbf{X}^{-1}$$

with $\mathbf{J} := \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_p)$.

The Jordan form is just a special kind of block diagonalization of $\mathbf{A}$, so should not affect the spectral projectors $\mathbf{P}_j$ and spectral nilpotents discussed in Theorem 1.8.3. Indeed, substituting in the factorization (1.49) for $\mathbf{T}_j$ into the block-diagonalization (1.35) gives

$$\begin{aligned}
\mathbf{A} &= \left[\begin{array}{c|c|c} \mathbf{V}_1 & \cdots & \mathbf{V}_p \end{array}\right]\begin{bmatrix} \mathbf{T}_1 & & \\ & \ddots & \\ & & \mathbf{T}_p \end{bmatrix}\left[\begin{array}{c} \widehat{\mathbf{V}}_1^* \\ \hline \vdots \\ \hline \widehat{\mathbf{V}}_p^* \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{V}_1 & \cdots & \mathbf{V}_p \end{array}\right]\begin{bmatrix} \mathbf{W}_1\mathbf{J}_1\mathbf{W}_1^{-1} & & \\ & \ddots & \\ & & \mathbf{W}_p\mathbf{J}_p\mathbf{W}_p^{-1} \end{bmatrix}\left[\begin{array}{c} \widehat{\mathbf{V}}_1^* \\ \hline \vdots \\ \hline \widehat{\mathbf{V}}_p^* \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{V}_1\mathbf{W}_1 & \cdots & \mathbf{V}_p\mathbf{W}_p \end{array}\right]\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\left[\begin{array}{c} \mathbf{W}_1^{-1}\widehat{\mathbf{V}}_1^* \\ \hline \vdots \\ \hline \mathbf{W}_p^{-1}\widehat{\mathbf{V}}_p^* \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{array}\right]\begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}\left[\begin{array}{c} \widehat{\mathbf{X}}_1^* \\ \hline \vdots \\ \hline \widehat{\mathbf{X}}_p^* \end{array}\right] = \mathbf{X}\mathbf{J}\mathbf{X}^{-1},
\end{aligned}$$

where the matrices $\mathbf{X}_j := \mathbf{V}_j\mathbf{W}_j$ form the columns of $\mathbf{X}$, and $\widehat{\mathbf{X}}_j^* := \mathbf{W}_j^{-1}\widehat{\mathbf{V}}_p^*$ for the rows of $\mathbf{X}$. This form form for $\mathbf{A}$ can be multiplied out to give

$$\mathbf{A} = \sum_{j=1}^{p} \mathbf{X}_j\mathbf{J}_j\widehat{\mathbf{X}}_j^*.$$

Notice that

$$\mathbf{X}_j\widehat{\mathbf{X}}_j^* = \mathbf{V}_j\mathbf{W}_j\mathbf{W}_j^{-1}\widehat{\mathbf{V}}_j^* = \mathbf{V}_j\widehat{\mathbf{V}}_j^* = \mathbf{P}_j.$$

Similarly, writing $\mathbf{J}_j = \lambda_j\mathbf{I} + \mathbf{N}_j$, we have

$$\mathbf{X}_j\mathbf{N}_j\widehat{\mathbf{X}}_j^* = \mathbf{V}_j\mathbf{W}_j\mathbf{N}_j\mathbf{W}_j^{-1}\widehat{\mathbf{V}}_j^* = \mathbf{V}_j\mathbf{R}_j\widehat{\mathbf{V}}_j^* = \mathbf{D}_j.$$

so we can form the spectral projectors and nilpotents just as well using the ingredients from the block diagonalization or the JORDAN form.

Observe that $\mathcal{R}(\mathbf{X}_k)$ is an *invariant subspace*, meaning that if $\mathbf{x} \in \mathcal{R}(\mathbf{X}_k)$, then $\mathbf{A}\mathbf{x} \subset \mathcal{R}(\mathbf{X}_k)$. To see this, write such an $\mathbf{x}$ as a linear combination of the columns of $\mathbf{X}_k$, giving $\mathbf{x} = \mathbf{X}_k\mathbf{c}$. Then

$$\mathbf{A}\mathbf{x} = \sum_{j=1}^{p} \mathbf{X}_j\mathbf{J}_j\widehat{\mathbf{X}}_j^*\mathbf{X}_k\mathbf{c} = \mathbf{X}_k\mathbf{J}_k\mathbf{c} \in \mathcal{R}(\mathbf{X}_k).$$

Thus $\mathbf{A}\mathcal{R}(\mathbf{X}_k) \subseteq \mathcal{R}(\mathbf{X}_k)$. Observe that $\mathbf{P}_k$ is a projector *onto* this invariant subspace $\mathcal{R}(\mathbf{X}_k)$ and *along* $\mathcal{N}(\widehat{\mathbf{X}}_k^*)$.

Index, Defective, Derogatory

The *index*, i_j, of the eigenvalue λ_j equals the length of the longest JORDAN chain (equivalently, the dimension of the largest Jordan block) associated with λ_j. Note that $1 \le i_j \le a_j$.

An eigenvalue λ_j is called *defective* if $i_j > 1$ (i.e., it has a JORDAN block of dimension larger than one, or, equivalently, lacking a_j linearly independent eigenvectors, $a_j > g_j$).

An eigenvalue λ_j is called *derogatory* if $g_j > 1$ (i.e., it has more than one linearly independent eigenvector).

A matrix with at least one *defective* (or *derogatory*) eigenvalue is sometimes casually called a *defective* (or *dergatory*) *matrix*.

In the notation of Theorem 1.21, the index i_j corresponds to the largest Jordan block $\mathbf{E}_1 \in \mathbb{C}^{d_1\times d_1}$.

Jordan Canonical Form

Theorem 1.22. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ have distinct eigenvalues $\lambda_1, \ldots, \lambda_p$ with algebraic multiplicities $a_1, \ldots, a_p$, geometric multiplicities $g_1, \ldots, g_p$, and indices $i_1, \ldots, i_p$. There exist $\mathbf{J} \in \mathbb{C}^{n\times n}$ and invertible $\mathbf{X} \in \mathbb{C}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1},$$

which can be partitioned in the form

$$\mathbf{X} = [\,\mathbf{X}_1 \quad \cdots \quad \mathbf{X}_p\,], \quad \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix}, \quad \mathbf{X}^{-1} = \begin{bmatrix} \widehat{\mathbf{X}}_1^* \\ \vdots \\ \widehat{\mathbf{X}}_p^* \end{bmatrix}$$

with $\mathbf{X}_j \in \mathbb{C}^{n\times a_j}$, $\mathbf{J}_j \in \mathbb{C}^{a_j\times a_j}$. Each matrix $\mathbf{J}_j$ is of the form

$$\mathbf{J}_j = \begin{bmatrix} \mathbf{J}_{j,1} & & \\ & \ddots & \\ & & \mathbf{J}_{j,g_j} \end{bmatrix} = \lambda_j\mathbf{I} + \mathbf{N}_j,$$

where $\mathbf{N}_j \in \mathbb{C}^{a_j\times a_j}$ is nilpotent of degree i_j: $\mathbf{N}_j^{i_j} = \mathbf{0}$. The submatrices $\mathbf{J}_{j,k}$ are Jordan blocks of the form

$$\mathbf{J}_{j,k} = \begin{bmatrix} \lambda_j & 1 & & \\ & \lambda_j & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}.$$

The spectral projectors and spectral nilpotents of $\mathbf{A}$ are defined as

$$\mathbf{P}_j := \mathbf{X}_j\widehat{\mathbf{X}}_j^* \in \mathbb{C}^{n\times n}, \qquad \mathbf{D}_j := \mathbf{X}_j\mathbf{N}_j\widehat{\mathbf{X}}_j^* \in \mathbb{C}^{n\times n}.$$

For $j \neq k$ they satisfy

$$\mathbf{P}_j^2 = \mathbf{P}_j, \quad \mathbf{P}_j\mathbf{P}_k = \mathbf{0}, \qquad \sum_{j=1}^{p} \mathbf{P}_j = \mathbf{I},$$
$$\mathbf{D}_j^{i_j} = \mathbf{0}, \quad \mathbf{P}_j\mathbf{D}_j = \mathbf{D}_j\mathbf{P}_j = \mathbf{D}_j, \quad \mathbf{P}_j\mathbf{D}_k = \mathbf{D}_k\mathbf{P}_j = \mathbf{0},$$

and give the spectral representation of $\mathbf{A}$:

$$\mathbf{A} = \sum_{j=1}^{p} (\lambda_j\mathbf{P}_j + \mathbf{D}_j).$$

Some pitfalls of Jordan form computation

One could imagine using the proof of Theorem 1.21 to *numerically compute* the Jordan form, but, as Fletcher & Sorensen [FS83] note, (citing the seminal work of Golub & Wilkinson [GW76]), computation of the Jordan form is fraught with challenges. First, you would need to compute the block-triangulation of $\mathbf{A}$. This requires that you identify the distinct eigenvalues of $\mathbf{A}$. When you numerically compute the eigenvalues of a matrix, you actually compute the eigenvalues of $\mathbf{A}+\mathbf{E}$ for some matrix $\mathbf{E}$ that is very small in norm: this makes it impossible to know if two eigenvalues really are the same, or are just distinct eigenvalues that are close together. Even if you have a block triangularlization, then to execute the proof of Theorem 1.21, one needs to determine if $\rho = 0$. What is we merely have $\rho \approx 10^{-10}$? Should that be regarded as zero? Decisions like this affect the size of the Jordan blocks, and thus the ultimate Jordan form.

One can always construct an arbitrarily small perturbation $\mathbf{E}$ that will turn a nondiagonalizable matrix $\mathbf{A}$ into a diagonalizable matrix $\mathbf{A}+\mathbf{E}$. This suggests the tempting idea that, 'Nondiagonalizable matrices never arise in practical physical situations.' However, we have seen in Section 1.7 that reality is more nuanced. An enriched understanding requires the tools of perturbation theory, to be developed in Chapter **??**.

Cayley–Hamilton Theorem

The Jordan form provides one convenient approach to defining and analyzing functions of matrices, $f(\mathbf{A})$, as we shall examine in detail in Chapter **??**. For the moment, consider two special polynomials that bear a close connection to the spectral characterization of Theorem 1.22.

Characteristic Polynomial and Minimal Polynomial

Suppose the matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ has the Jordan structure detailed in Theorem 1.22. The *characteristic polynomial* of $\mathbf{A}$ is given by

$$p_{\mathbf{A}}(z) = \prod_{j=1}^{p} (z-\lambda_j)^{a_j},$$

a degree-n polynomial; the *minimal polynomial* of $\mathbf{A}$ is

$$m_{\mathbf{A}}(z) = \prod_{j=1}^{p} (z - \lambda_j)^{d_j}.$$

Since the algebraic multiplicity can never exceed the index of an eigenvalue (the size of the largest Jordan block), we see that $m_{\mathbf{A}}$ is a polynomial of degree no greater than n. If $\mathbf{A}$ is not derogatory, then $p_{\mathbf{A}} = m_{\mathbf{A}}$; if $\mathbf{A}$ is derogatory, then the degree of $m_{\mathbf{A}}$ is strictly less than the degree of $p_{\mathbf{A}}$, and $m_{\mathbf{A}}$ divides $p_{\mathbf{A}}$. Since

$$(\mathbf{J}_j - \lambda_j \mathbf{I})^{d_j} = \mathbf{N}_j^{d_j} = \mathbf{0},$$

notice that

$$p_{\mathbf{A}}(\mathbf{J}_j) = m_{\mathbf{A}}(\mathbf{J}_j) = \mathbf{0}, \qquad j = 1, \ldots, p,$$

and so

$$p_{\mathbf{A}}(\mathbf{A}) = m_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}.$$

Thus, both the characteristic polynomial and minimal polynomials *annihilate* $\mathbf{A}$. In fact, $m_{\mathbf{A}}$ is the lowest degree polynomial that annihilates $\mathbf{A}$. The first fact, proved in the 2×2 and 3×3 case by Cayley [Cay58], is known as the *Cayley–Hamilton Theorem*

Cayley–Hamilton Theorem

Theorem 1.23. *The characteristic polynomial $p_{\mathbf{A}}$ of $\mathbf{A} \in \mathbb{C}^{n\times n}$ annihilates $\mathbf{A}$:*

$$p_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}.$$

1.9 A Concrete Approach to the Jordan Form

The preceding section gave a detailed proof of the Jordan form, but that formal approach disguises some nuances of the form that comes from examining some concrete examples. In this section we will build the Jordan form up gradually for a few small matrices.

1.9.1 Take 1: one eigenvalue, one block

Suppose $\mathbf{A}$ is a matrix with one eigenvalue λ, but only one eigenvector, i.e., $\dim(\mathcal{N}(\mathbf{A}-\lambda\mathbf{I}))=1$, as in the illustrative example

$$\mathbf{A}=\begin{bmatrix}1&1&1\\0&1&1\\0&0&1\end{bmatrix}. \tag{1.50}$$

Notice that $\mathbf{A}-\lambda\mathbf{I}$ has zeros on the main diagonal, and this structure has interesting implications for higher powers of $\mathbf{A}-\lambda\mathbf{I}$:

$$\mathbf{A}-\lambda\mathbf{I}=\begin{bmatrix}0&1&1\\0&0&1\\0&0&0\end{bmatrix},\quad (\mathbf{A}-\lambda\mathbf{I})^2=\begin{bmatrix}0&0&2\\0&0&0\\0&0&0\end{bmatrix},\quad (\mathbf{A}-\lambda\mathbf{I})^3=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}.$$

Zeros cascade up successive diagonals with every higher power, and thus the null space of $(\mathbf{A}-\lambda\mathbf{I})^k$ grows in dimension with k. In particular, we must have that $(\mathbf{A}-\lambda\mathbf{I})^n=\mathbf{0}$, so

$$\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^n)=\mathbb{C}^n.$$

For the moment *assume* that $(\mathbf{A}-\lambda\mathbf{I})^{n-1}\neq\mathbf{0}$, and hence we can always find some vector

$$\mathbf{x}_n\notin\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{n-1}),\qquad \mathbf{x}_n\in\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{n}).$$

For example (1.50), we could take

$$\mathbf{x}_3=\begin{bmatrix}0\\0\\1\end{bmatrix}.$$

(This assumption equates to the fact that $\mathbf{A}$ has only one linearly independent eigenvector. Can you explain why?) Now *define*

$$\mathbf{x}_{n-1}:=(\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_n.$$

Notice that $(\mathbf{A}-\lambda\mathbf{I})^{n-2}\mathbf{x}_{n-1}=(\mathbf{A}-\lambda\mathbf{I})^{n-1}\mathbf{x}_n\neq\mathbf{0}$, yet $(\mathbf{A}-\lambda\mathbf{I})^{n-1}\mathbf{x}_{n-1}=(\mathbf{A}-\lambda\mathbf{I})^{n}\mathbf{x}_n=\mathbf{0}$, and so

$$\mathbf{x}_{n-1}\notin\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{n-2}),\qquad \mathbf{x}_{n-1}\in\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{n-1}).$$

Repeat this procedure, defining

$$\begin{aligned}\mathbf{x}_{n-2} &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{n-1},\\ &\vdots\\ \mathbf{x}_1 &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_2.\end{aligned}$$

For each $k = 1, \ldots, n$, we have

$$\mathbf{x}_k \not\in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{k-1}), \qquad \mathbf{x}_k \in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{k}).$$

The $k = 1$ case is particularly interesting: $\mathbf{x}_1 \not\in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^0) = \mathcal{N}(\mathbf{I}) = \{\mathbf{0}\}$, so $\mathbf{x}_1 \neq \mathbf{0}$, and $\mathbf{x}_1 \in \mathcal{N}(\mathbf{A}-\lambda\mathbf{I})$, so $\mathbf{A}\mathbf{x}_1 = \lambda\mathbf{x}_1$, so $\mathbf{x}_1$ is an eigenvector of $\mathbf{A}$.

To be concrete, return to (1.50): we compute

$$\mathbf{x}_2 := (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_3 = \begin{bmatrix}1\\1\\0\end{bmatrix}, \qquad \mathbf{x}_1 := (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_2 = \begin{bmatrix}1\\0\\0\end{bmatrix}.$$

Notice that we now have three equations relating $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$:

$$\begin{aligned}(\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_1 &= \mathbf{0},\\ (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_2 &= \mathbf{x}_1,\\ (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_3 &= \mathbf{x}_2.\end{aligned}$$

Confirm that we can rearrange these into the matrix form

$$[\,\mathbf{A}\mathbf{x}_1 \quad \mathbf{A}\mathbf{x}_2 \quad \mathbf{A}\mathbf{x}_3\,] = [\,\mathbf{0} \quad \mathbf{x}_1 \quad \mathbf{x}_2\,] + [\,\lambda\mathbf{x}_1 \quad \lambda\mathbf{x}_2 \quad \lambda\mathbf{x}_3\,],$$

which we factor as

$$\mathbf{A}\,[\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3\,] = [\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3\,]\begin{bmatrix}\lambda & 1 & \\ & \lambda & 1\\ & & \lambda\end{bmatrix}, \tag{1.51}$$

and abbreviate as

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{J}.$$

By construction, the columns of $\mathbf{X}$ are linearly independent, so

$$\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1}. \tag{1.52}$$

This is known as the *Jordan canonical form* of $\mathbf{A}$: it is not a diagonalization, since $\mathbf{J}$ has nonzeros on the superdiagonal – but this is as close as we get.

For more general $\mathbf{A}$ with one eigenvalue and one eigenvector, equation (1.51) generalizes as expected, giving

$$\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1}$$

with

$$\mathbf{X} = [\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n\,], \qquad \mathbf{J} = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{C}^{n\times n},$$

with unspecified entries equal to zero.

Notice that $\mathbf{x}_1$ is the only eigenvector in $\mathbf{X}$. The kth column of $\mathbf{X}$ is called a *generalized eigenvector of grade k*, and the collection $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a *Jordan chain*. The matrix $\mathbf{J}$ is called a *Jordan block*.

With this basic structure in hand, we are prepared to add the next layer of complexity.

1.9.2 Take 2: one eigenvalue, multiple blocks

Again presume that $\mathbf{A}$ is upper triangular with a single eigenvalue λ, and define $d_1 \in \{1, \ldots, n\}$ to be the largest integer such that

$$\dim(\mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_1-1}) \neq \dim(\mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_1}).$$

The assumptions of the last section required $d = n$. Now we relax that assumption, and we will have to work a bit harder to get a full set of n generalized eigenvectors. Consider the following enlargement of our last example:

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

for which

$$\mathbf{A} - \lambda\mathbf{I} \;=\; \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad (\mathbf{A} - \lambda\mathbf{I})^2 = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$(\mathbf{A} - \lambda\mathbf{I})^3 \;=\; \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad (\mathbf{A} - \lambda\mathbf{I})^4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

In this case, $d_1 = 3 < 4 = n$. Now define $\mathbf{x}_{1,d_1} \in \mathbb{C}^n$ so that

$$\mathbf{x}_{1,d_1} \notin \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_1-1}), \quad \mathbf{x}_{1,d_1} \in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_1}).$$

In the 4×4 example, take

$$\mathbf{x}_{1,3} := \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \notin \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^2), \quad \mathbf{x}_{1,3} \in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^3).$$

Now build out the Jordan chain just as in Section 1.9.1:

$$\begin{aligned} \mathbf{x}_{1,d_1-1} &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,d_1}, \\ &\vdots \\ \mathbf{x}_{1,1} &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,2}. \end{aligned}$$

For the example,

$$\mathbf{x}_{1,2} := (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,3} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_{1,1} := (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,2} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

This chain $\{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,d_1}\}$ has length d_1; if $d_1 < n$, we cannot build from this chain alone an invertible matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ such that $\mathbf{A} = \mathbf{XJX}^{-1}$, as in (1.52). We must derive $n - d_1$ more generalized eigenvectors associated with λ: this is the most subtle aspect of the Jordan form. (As a payback for this difficulty, matrices with this structure are called *derogatory*.)

We defined $d_1 \in \{1, \ldots, n\}$ to be the largest integer such that

$$\dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_1})) - \dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_1-1})) \ge 1.$$

If $d_1 < n$, define $d_2 \in \{1, \ldots, d_1\}$ to be the largest integer such that

$$\dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_2})) - \dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_2-1})) \ge 2,$$

and pick $\mathbf{x}_{2,d_2}$ such that

$$\mathbf{x}_{2,d_2} \notin \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_2-1}), \quad \mathbf{x}_{2,d_2} \in \mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^{d_2})$$

and

$$\mathbf{x}_{2,d_2} \notin \operatorname{span}\{\mathbf{x}_{1,d_2}, \ldots, \mathbf{x}_{1,d_1}\}.$$

The last condition ensures that $\mathbf{x}_{2,d_2}$ is linearly independent of the first Jordan chain. Now build out the rest of the second Jordan chain.

$$\begin{aligned} \mathbf{x}_{2,d_2-1} &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{2,d_2}, \\ &\vdots \\ \mathbf{x}_{2,1} &:= (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{2,2}. \end{aligned}$$

In the 4×4 example, we have

$$\dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^0)) = 0, \quad \dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^1)) = 2, \quad \dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^2)) = 3,$$

$$\dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^3)) = 4, \quad \dim(\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^4)) = 4,$$

and thus $d_2 = 1$. Since

$$\mathcal{N}(\mathbf{A}-\lambda\mathbf{I}) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\},$$

and $\mathbf{x}_{1,1} = [1,0,0,0]^{\mathrm{T}}$, we select

$$\mathbf{x}_{2,1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

In this case, $d_1 + d_2 = 3 + 1 = 4 = n$, so we have a full set of generalized eigenvectors, associated with the equations

$$\begin{aligned} (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,1} &= \mathbf{0}, \\ (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,2} &= \mathbf{x}_{1,1}, \\ (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{1,3} &= \mathbf{x}_{1,2}, \\ (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_{2,1} &= \mathbf{0}. \end{aligned}$$

We arrange these in matrix form,

$$\mathbf{A}\,[\,\mathbf{x}_{1,1}\;\;\mathbf{x}_{1,2}\;\;\mathbf{x}_{1,3}\;\;\mathbf{x}_{2,1}\,] = [\,\mathbf{x}_{1,1}\;\;\mathbf{x}_{1,2}\;\;\mathbf{x}_{1,3}\;\;\mathbf{x}_{2,1}\,] \left[\begin{array}{ccc|c} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \lambda & \\ \hline & & & \lambda \end{array}\right]. \qquad (1.53)$$

The last matrix has horizontal and vertical lines drawn to help you see the block-diagonal structure, with the jth diagonal block of size $d_j \times d_j$ corresponding to the jth Jordan chain.

In the example, we were able to stop at this stage because $d_1 + d_2 = n$. If $d_1 + d_2 < n$, we find another Jordan chain by repeating the above procedure: Define $d_3 \in \{1, \ldots, d_2\}$ to be the largest integer such that

$$\dim(\mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_3})) - \dim(\mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_3-1})) \geq 3,$$

and pick $\mathbf{x}_{3,d_3}$ such that

$$\mathbf{x}_{3,d_3} \notin \mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_3-1}), \quad \mathbf{x}_{3,d_3} \in \mathcal{N}((\mathbf{A} - \lambda\mathbf{I})^{d_3})$$

and

$$\mathbf{x}_{3,d_3} \notin \operatorname{span}\{\mathbf{x}_{1,d_3}, \ldots, \mathbf{x}_{1,d_1}, \mathbf{x}_{2,d_3}, \ldots, \mathbf{x}_{2,d_2}\}.$$

Then complete the Jordan chain

$$\begin{aligned} \mathbf{x}_{3,d_3-1} &:= (\mathbf{A} - \lambda\mathbf{I})\mathbf{x}_{3,d_3}, \\ &\vdots \\ \mathbf{x}_{3,1} &:= (\mathbf{A} - \lambda\mathbf{I})\mathbf{x}_{3,2}. \end{aligned}$$

Repeat this pattern until $d_1 + \cdots + d_k = n$, which will lead to a factorization of the form (1.53), but now with k individual Jordan blocks on the main diagonal of the last matrix. (We have not explicitly justified that all must end this well, with the matrix $\mathbf{X} = [\,\mathbf{x}_{1,1} \quad \cdots \quad \mathbf{x}_{k,d_k}\,] \in \mathbb{C}^{n \times n}$ invertible; such formalities can be set with a little reflection on the preceding steps.)

We have now finished with the case where $\mathbf{A}$ has a single eigenvalue. Two extreme cases merit special mention: $d_1 = n$ (one Jordan block of size $n \times n$, so that $\mathbf{A}$ is *defective* but not *derogatory*) and $d_1 = d_2 = \cdots = d_n = 1$ (n Jordan blocks of size 1×1, so that $\mathbf{A}$ is *derogatory* but not *defective*). (Did you notice that a diagonalizable matrix with one eigenvalue must be a multiple of the identity matrix?)

We shall not dwell here in the case of multiple eigenvalues; the structure goes through just as for single eigenvalues, with each distinct eigenvalue now contributing one of the diagonal JORDAN factors $\mathbf{J}_j$, a column-block $\mathbf{X}_j$ of $\mathbf{X}$, and a row-block $\widehat{\mathbf{X}}_j^*$ of $\mathbf{X}^{-1}$, as discussed in Theorem 1.22.

1.10 Analytic Approach to Spectral Theory

The approach to the Jordan form in the last section was highly algebraic: Jordan chains were constructed by repeatedly applying $\mathbf{A} - \lambda\mathbf{I}$ to eigenvectors of the highest grade, and these were assembled to form a basis for the invariant subspace associated with λ. Once the hard work of constructing these Jordan chains is complete, results about the corresponding spectral projectors and nilpotents can be proved directly from the fact that $\mathbf{V}^{-1}\mathbf{V} = \mathbf{I}$.

In this section we briefly mention an entirely different approach, one based much more on analysis rather than algebra, and for that reason one more readily suitable to infinite dimensional matrices. This approach gives ready formulas for the spectral projectors and nilpotents, but more work would be required to determine the properties of these matrices.

To begin, recall that resolvent $\mathbf{R}(z) := (z\mathbf{I} - \mathbf{A})^{-1}$ defined in (1.8) on page 14 for all $z \in \mathbb{C}$ that are not eigenvalues of $\mathbf{A}$. Recall from Section 1.3 that the resolvent is a rational function of the parameter z. Suppose that $\mathbf{A}$ has p distinct eigenvalues, $\lambda_1, \ldots, \lambda_p$, and for each of these eigenvalues, let Γ_j denote a small circle in the complex plane centered at λ_j with radius sufficiently small that no other eigenvalue is on or in the interior Γ_j. Then we can then *define* the *spectral projector* and *spectral nilpotent* for λ_j:

$$\mathbf{P}_j := \frac{1}{2\pi\mathrm{i}}\int_{\Gamma_j} \mathbf{R}(z)\,\mathrm{d}z, \qquad \mathbf{D}_j := \frac{1}{2\pi\mathrm{i}}\int_{\Gamma_j} (z-\lambda_j)\mathbf{R}(z)\,\mathrm{d}z. \tag{1.54}$$

In these definitions the integrals are taken entrywise. For example, the matrix and resolvent

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \qquad \mathbf{R}(z) = \begin{bmatrix} \dfrac{1}{z-1} & 0 & 0 \\ 0 & \dfrac{1}{z-2} & 0 \\ \dfrac{1}{(z-1)^2} & 0 & \dfrac{1}{z-1} \end{bmatrix}$$

with eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$ has spectral projectors

$$\mathbf{P}_1 = \frac{1}{2\pi\mathrm{i}} \begin{bmatrix} \displaystyle\int_{\Gamma_1} \frac{1}{z-1}\,\mathrm{d}z & 0 & 0 \\ 0 & \displaystyle\int_{\Gamma_1} \frac{1}{z-2}\,\mathrm{d}z & 0 \\ \displaystyle\int_{\Gamma_1} \frac{1}{(z-1)^2}\,\mathrm{d}z & 0 & \displaystyle\int_{\Gamma_1} \frac{1}{z-1}\,\mathrm{d}z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{P}_2 = \frac{1}{2\pi\mathrm{i}}\left[\begin{array}{ccc} \displaystyle\int_{\Gamma_2}\frac{1}{z-1}\,\mathrm{d}z & 0 & 0 \\ 0 & \displaystyle\int_{\Gamma_2}\frac{1}{z-2}\,\mathrm{d}z & 0 \\ \displaystyle\int_{\Gamma_2}\frac{1}{(z-1)^2}\,\mathrm{d}z & 0 & \displaystyle\int_{\Gamma_2}\frac{1}{z-1}\,\mathrm{d}z \end{array}\right] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D}_1 = \frac{1}{2\pi\mathrm{i}}\left[\begin{array}{ccc} \displaystyle\int_{\Gamma_1} 1\,\mathrm{d}z & 0 & 0 \\ 0 & \displaystyle\int_{\Gamma_1}\frac{z-1}{z-2}\,\mathrm{d}z & 0 \\ \displaystyle\int_{\Gamma_1}\frac{1}{z-1}\,\mathrm{d}z & 0 & \displaystyle\int_{\Gamma_1} 1\,\mathrm{d}z \end{array}\right] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

and $\mathbf{D}_2 = \mathbf{0}$.

One can verify that all the properties of spectral projectors and nilpotents outlined in Theorem 1.22 hold with this alternative definition. We will not exhaustively prove the properties of these resolvent integrals (for such details, consult Cox's notes for CAAM 335, or Section I.5 of the excellent monograph by Tosio Kato [Kat80]), but we will give one representative proof to give you a taste of how these arguments proceed.

First Resolvent Identity

Lemma 1.24. *If w and z are not eigenvalues of $\mathbf{A} \in \mathbb{C}^{n\times n}$, then*

$$(z-w)\mathbf{R}(z)\mathbf{R}(w) = \mathbf{R}(w) - \mathbf{R}(z).$$

Proof. Given the identity $(z-w)\mathbf{I} = (z\mathbf{I} - \mathbf{A}) - (w\mathbf{I} - \mathbf{A})$, multiply both sides on the left by $\mathbf{R}(z)$ and on the right by $\mathbf{R}(w)$ to obtain the result. ■

Theorem 1.25. *The matrix $\mathbf{P}_j$ defined in* (1.54) *is a projector.*

Proof. We shall show that $\mathbf{P}_j^2 = \mathbf{I}$. Let Γ_j and $\widehat{\Gamma}_j$ denote two circles in the complex plane that enclose λ_j and not other eigenvalues of $\mathbf{A}$; moreover, suppose that $\widehat{\Gamma}_j$ is strictly contained in the interior of Γ_j. It follows that

$$\mathbf{P}_j\mathbf{P}_j = \Big(\frac{1}{2\pi\mathrm{i}}\int_{\Gamma_j}\mathbf{R}(z)\,\mathrm{d}z\Big)\Big(\frac{1}{2\pi\mathrm{i}}\int_{\widehat{\Gamma}_j}\mathbf{R}(w)\,\mathrm{d}w\Big)$$

$$= \left(\frac{1}{2\pi\mathrm{i}}\right)^2 \int_{\Gamma_j} \int_{\widehat{\Gamma}_j} \mathbf{R}(z)\mathbf{R}(w)\,\mathrm{d}w\,\mathrm{d}z$$

$$= \left(\frac{1}{2\pi\mathrm{i}}\right)^2 \int_{\Gamma_j} \int_{\widehat{\Gamma}_j} \frac{\mathbf{R}(w)-\mathbf{R}(z)}{z-w}\,\mathrm{d}w\,\mathrm{d}z,$$

where the last step is a consequence of the First Resolvent Identity. Now split the integrand into two components, and swap the order of integration in the first integral to obtain

$$\mathbf{P}_j\mathbf{P}_j = \left(\frac{1}{2\pi\mathrm{i}}\right)^2 \int_{\widehat{\Gamma}_j} \mathbf{R}(w) \int_{\Gamma_j} \frac{1}{z-w}\,\mathrm{d}z\,\mathrm{d}w - \left(\frac{1}{2\pi\mathrm{i}}\right)^2 \int_{\Gamma_j} \mathbf{R}(z) \int_{\widehat{\Gamma}_j} \frac{1}{z-w}\,\mathrm{d}w\,\mathrm{d}z.$$

It remains to resolve the two double integrals. These calculations require careful consideration of the contour arrangements and the respective variables of integration, $z \in \Gamma_j$ and $w \in \widehat{\Gamma}_j$. To compute

$$\int_{\widehat{\Gamma}_j} \frac{1}{z-w}\,\mathrm{d}w,$$

notice that the integrand has a pole at $w = z$, but since $w \in \widehat{\Gamma}_j$, this pole occurs *outside the contour of integration*, and hence the integral is zero. On the other hand, the integrand in

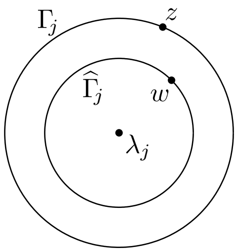

Figure 1.6. Contours use in proof that $\mathbf{P}_j^2 = \mathbf{P}_j$.

$$\int_{\Gamma_j} \frac{1}{z-w}\,\mathrm{d}z$$

has a pole at $z = w$, which occurs *inside the contour of integration* since $z \in \Gamma_j$. It follows that

$$\mathbf{P}_j\mathbf{P}_j = \left(\frac{1}{2\pi\mathrm{i}}\right)^2 \int_{\widehat{\Gamma}_j} \mathbf{R}(w)(2\pi\mathrm{i})\,\mathrm{d}w = \mathbf{P}_j. \quad \blacksquare$$

Variations on this proof can be used to derive a host of similar relationships, which we summarize below.

Theorem 1.26. *Let $\mathbf{P}_j$ and $\mathbf{D}_j$ denote the spectral projectors and nilpotents defined in* (1.54) *for eigenvalue λ_j, where $\lambda_1, \ldots, \lambda_p$ denote the distinct eigenvalues of $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $d_1, \ldots, d_p$ their indices. Then*

- $\mathbf{P}_j^2 = \mathbf{P}_j$;
- $\mathbf{D}_j^{d_j} = \mathbf{0}$ *(in particular, $\mathbf{D}_j = \mathbf{0}$ if $d_j = 1$);*
- $\mathbf{P}_j\mathbf{D}_j = \mathbf{D}_j\mathbf{P}_j = \mathbf{D}_j$;
- $\mathbf{P}_j\mathbf{P}_k = \mathbf{0}$ *when $j \neq k$;*
- $\mathbf{D}_j\mathbf{P}_k = \mathbf{P}_k\mathbf{D}_j = \mathbf{D}_j\mathbf{D}_k = \mathbf{0}$ *when $j \neq k$.*

It is an interesting exercise to relate the spectral projectors and nilpotents to the elements of the Jordan form developed in the last section. Ultimately, one arrives at a beautifully concise expansion for a generic matrix.

Spectral Representation of a Matrix

Theorem 1.27. *Any matrix $\mathbf{A}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_p$ and corresponding spectral projectors $\mathbf{P}_1, \ldots, \mathbf{P}_p$ and spectral nilpotents $\mathbf{D}_1, \ldots, \mathbf{D}_p$ can be written in the form*

$$\mathbf{A} = \sum_{j=1}^{p} \lambda_j \mathbf{P}_j + \mathbf{D}_j.$$

1.11 Coda: Simultaneous Diagonalization

We conclude this chapter with a basic result we will have cause to apply later, which serves as a nice way to tie together several concepts we have encountered.

Simultaneous Diagonalization

Theorem 1.28. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ be diagonalizable matrices. Then $\mathbf{AB} = \mathbf{BA}$ if and only if there exists some invertible $\mathbf{V} \in \mathbb{C}^{n\times n}$ such that $\mathbf{VAV}^{-1} = \mathbf{\Lambda}$ and $\mathbf{VBV}^{-1} = \mathbf{\Gamma}$ are both diagonal.*

Proof. First suppose that $\mathbf{A}$ and $\mathbf{B}$ are both simultaneously diagonalizable, i.e., there exists an invertible $\mathbf{V} \in \mathbb{C}^{n\times n}$ such that $\mathbf{V}\mathbf{A}\mathbf{V}^{-1} = \boldsymbol{\Lambda}$ and $\mathbf{V}\mathbf{B}\mathbf{V}^{-1} = \boldsymbol{\Gamma}$ are both diagonal. Then using the multiplication of *diagonal* matrices is always commutative,

$$\mathbf{A}\mathbf{B} = \mathbf{V}^{-1}\boldsymbol{\Lambda}\mathbf{V}\mathbf{V}^{-1}\boldsymbol{\Gamma}\mathbf{V} = \mathbf{V}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\mathbf{V} = \mathbf{V}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\mathbf{V} = \mathbf{V}^{-1}\boldsymbol{\Gamma}\mathbf{V}\mathbf{V}^{-1}\boldsymbol{\Lambda}\mathbf{V} = \mathbf{B}\mathbf{A}.$$

Now suppose that $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable matrices that commute: $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$. Diagonalize $\mathbf{A}$ to obtain

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \begin{bmatrix} \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix},$$

where $\lambda_1 \notin \sigma(\mathbf{D})$, i.e., we group all copies of the eigenvalue λ_1 in the upper-left corner. Now partition $\mathbf{V}^{-1}\mathbf{B}\mathbf{V}$ accordingly:

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{V} = \begin{bmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix}.$$

Since $\mathbf{A}$ and $\mathbf{B}$ commute, so too do $\mathbf{V}^{-1}\mathbf{A}\mathbf{V}$ and $\mathbf{V}^{-1}\mathbf{B}\mathbf{V}$:

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V}\mathbf{V}^{-1}\mathbf{B}\mathbf{V} = \begin{bmatrix} \lambda_1\mathbf{W} & \lambda_1\mathbf{X} \\ \mathbf{D}\mathbf{Y} & \mathbf{D}\mathbf{Z} \end{bmatrix}$$

$$= \mathbf{V}^{-1}\mathbf{B}\mathbf{V}\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \begin{bmatrix} \lambda_1\mathbf{W} & \mathbf{X}\mathbf{D} \\ \lambda_1\mathbf{Y} & \mathbf{Z}\mathbf{D} \end{bmatrix}.$$

Equating the (1,2) blocks gives $\lambda_1\mathbf{X} = \mathbf{X}\mathbf{D}$, i.e., $\mathbf{X}(\mathbf{D} - \lambda_1\mathbf{I}) = \mathbf{0}$. Since $\lambda_1 \notin \sigma(\mathbf{D})$, the resolvent $(\mathbf{D} - \lambda_1\mathbf{I})$ is invertible, so we conclude that $\mathbf{X} = \mathbf{0}$. Similarly, the (2,1) blocks imply $\mathbf{D}\mathbf{Y} = \lambda_1\mathbf{Y}$, and hence $\mathbf{Y} = \mathbf{0}$. Finally, the $(2, 2)$ blocks imply that $\mathbf{D}$ and $\mathbf{Z}$ commute. In summary, we arrive at the block-diagonal form

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{V} = \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix},$$

where $\mathbf{W}$ and $\mathbf{Z}$ are diagonalizable, since $\mathbf{B}$ is diagonalizable. We wish to further simplify $\mathbf{W}$ to diagonal form. Write $\mathbf{W} = \mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^{-1}$ for diagonal $\boldsymbol{\Gamma}$, and define

$$\mathbf{T} = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Then

$$\mathbf{T}^{-1}\mathbf{V}^{-1}\mathbf{B}\mathbf{V}\mathbf{T} = \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix},$$

which has a diagonal $(1,1)$ block. This transformation has no affect on the already-diagonalized (1,1) block of $\mathbf{V}^{-1}\mathbf{A}\mathbf{V}$:

$$\mathbf{T}^{-1}\mathbf{V}^{-1}\mathbf{A}\mathbf{V}\mathbf{T} = \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{-1}\mathbf{Z}\mathbf{S} \end{bmatrix}.$$

In summary, we have simultaneously block-diagonalized portions of $\mathbf{A}$ and $\mathbf{B}$. Apply the same strategy for each remaining eigenvalue of $\mathbf{A}$, i.e., to the commuting matrices $\mathbf{S}^{-1}\mathbf{D}\mathbf{S}^{-1}$ and $\mathbf{Z}$. ■

One rough way to summarize this result: diagonalizable matrices commute if and only if they have the same eigenvectors.

Problems

1. Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ is invertible, and $\mathbf{E} \in \mathbb{C}^{n\times n}$ is a 'small' perturbation. Use Theorem 1.9 to develop a condition on $\|\mathbf{E}\|$ to ensure that $\mathbf{A}+\mathbf{E}$ is invertible, and provide a bound on $\|(\mathbf{A}+\mathbf{E})^{-1}\|$.

2. Bernoulli's description of the compound pendulum with three equal masses (see Section 1.2) models an ideal situation: there is no energy loss in the system. When we add a viscous damping term, the displacement $x_j(t)$ of the jth mass is governed by the differential equation

$$\begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -3 & 2 \\ 0 & 2 & -5 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} - 2a \begin{bmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t) \end{bmatrix}$$

for damping constant $a \ge 0$. We write this equation in matrix form,

$$\mathbf{x}''(t) = -\mathbf{A}\mathbf{x}(t) - 2a\mathbf{x}'(t).$$

(Note the leading minus sign when you construct $\mathbf{A}$!) As with the damped harmonic oscillator in Section 1.7, we introduce $\mathbf{y}(t) := \mathbf{x}'(t)$ and write the second-order system in first-order form:

$$\begin{bmatrix} \mathbf{x}'(t) \\ \mathbf{y}'(t) \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{A} & -2a\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix}.$$

Denote the eigenvalues of $\mathbf{A}$ as γ_1, γ_2, and γ_3 with corresponding eigenvectors $\mathbf{u}_1$ $\mathbf{u}_2$, and $\mathbf{u}_3$.

(a) What are the eigenvalues and eigenvectors of the matrix

$$\mathbf{S}(a) = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{A} & 2a\mathbf{I} \end{bmatrix}$$

in terms of the constant $a \geq 0$ and the eigenvalues and eigenvectors of $\mathbf{A}$? (Give symbolic values in terms of γ_1, γ_2, and γ_3.)

(b) For what values of $a \geq 0$ does the matrix $\mathbf{S}(a)$ have a double eigenvalue? What can you say about the eigenvectors associated with this double eigenvalue? (Give symbolic values in terms of γ_1, γ_2, and γ_3.)

(c) Produce a plot in MATLAB (or the program of your choice) superimposing the eigenvalues of $\mathbf{S}(a)$ for $a \in [0, 3]$.

(d) What value of a minimizes the maximum real part of the eigenvalues? That is, find the $a \geq 0$ that minimizes the *spectral abscissa*

$$\alpha(\mathbf{S}(a)) := \max_{\lambda \in \mathbf{S}(a)} \operatorname{Re} \lambda.$$

3. Consider the 8×8 matrix

$$\mathbf{A} = \begin{bmatrix} 601 & 300 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3000 & 1201 & 0 & 0 & 0 & 0 & 0 & 300 \\ 475098 & 110888 & 301 & 100 & -15266 & -202 & 4418 & 27336 \\ 4594766 & 1185626 & -900 & -299 & -130972 & -3404 & 34846 & 272952 \\ -22800 & -6000 & 0 & 0 & 601 & 0 & 0 & -1800 \\ -3776916 & -968379 & 0 & 0 & 108597 & 2402 & -28800 & -222896 \\ -292663 & -71665 & 0 & 0 & 8996 & 200 & -2398 & -16892 \\ -37200 & -14400 & 0 & 0 & 300 & 0 & 0 & -2399 \end{bmatrix}.$$

Using MATLAB's `eig` command, do your best to approximate the true eigenvalues of $\mathbf{A}$ and the dimensions of the associated Jordan blocks. (Do not worry about the eigenvectors and generalized eigenvectors!) Justify your answer as best as you can. The matrices $\mathbf{A}$ and $\mathbf{A}^*$ have identical eigenvalues. Does MATLAB agree?

4. If a spectral projector $\mathbf{P}_j$ is orthogonal, show that the associated invariant subspace $\mathcal{R}(\mathbf{P}_j)$ is orthogonal to the complementary invariant subspace. Construct a matrix with three eigenvalues, where spectral projector is orthogonal, and the other two spectral projectors are not orthogonal.

5. Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{C}^{n \times n}$, and suppose that $\mathbf{AB} = \mathbf{BA}$ and $\mathbf{AC} = \mathbf{CA}$. It is tempting to claim that, via Theorem 1.28 on simultaneous diagonalization, we should have $\mathbf{BC} = \mathbf{CB}$. Construct a counterexample,

and explain why this does not contradict the characterization that 'the matrices $\mathbf{A}$ and $\mathbf{B}$ commute if and only if they have the same eigenvectors'.

References

[Ber33] Daniel Bernoulli. Theoremata de oscillationibus corporum filo flexili connexorum et catenae verticaliter suspensae. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 6:108–122, 1733. Reprinted and translated in [CD81].

[Ber34] Daniel Bernoulli. Demonstrationes theorematum suorum de oscillationibus corporum filo flexili connexorum et catenae verticaliter suspensae. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 7:162–173, 1734.

[BS72] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $AX+XB=C$. *Comm. ACM*, 15:820–826, 1972.

[BS09] Robert E. Bradley and C. Edward Sandifer. *Cauchy's Cours d'analyse: An Annotated Translation.* Springer, Dordrecht, 2009.

[Cau21] Augustin-Louis Cauchy. *Cours d'Analayse de l'École Royale Polytechnique.* Debure Frères, Paris, 1821.

[Cay58] Arthur Cayley. Memoir on the theory of matrices. *Phil. Trans. Royal Soc. London*, 148:17–37, 1858.

[CD81] John T. Cannon and Sigalia Dostrovsky. *The Evolution of Dynamics: Vibration Theory from 1687 to 1782.* Springer-Verlag, New York, 1981.

[Cra46] Harald Cramér. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, 1946.

[ES] Mark Embree and D. C. Sorensen. *An Introduction to Model Reduction for Linear and Nonlinear Differential Equations.* In preparation.

[FS83] R. Fletcher and D. C. Sorensen. An algorithmic derivation of the Jordan canonical form. *Amer. Math. Monthly*, 90:12–16, 1983.

[GW76] G. H. Golub and J. H. Wilkinson. Ill-conditioned eigensystems and the computation of the Jordan canonical form. *SIAM Review*, 18:578–619, 1976.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[HK71] Kenneth M. Hoffman and Ray Kunze. *Linear Algebra.* Prentice Hall, Englewood Cliffs, N.J., second edition, 1971. check this.

[Jol02] I. T. Jolliffe. *Principal Component Analysis.* Springer, New York, second edition, 2002.

[Kat80] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, corrected second edition, 1980.

[Lak76] Imre Lakatos. *Proofs and Refutations: The Logical of Mathematical Discovery*. Cambridge University Press, Cambridge, 1976.

[Par98] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, SIAM Classics edition, 1998.

[Ray78] Lord Rayleigh (John William Strutt). *The Theory of Sound*. Macmillan, London, 1877, 1878. 2 volumes.

[RS80] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, San Diego, revised and enlarged edition, 1980.

[SM03] Endre Süli and David Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, Cambridge, 2003.

[Ste04] J. Michael Steele. *The Cauchy–Schwarz Master Class*. Cambridge University Press, Cambridge, 2004.

[Str93] Gilbert Strang. The fundamental theorem of linear algebra. *Amer. Math. Monthly*, 100:848–855, 1993.

[Tru60] C. Truesdell. *The Rational Mechanics of Flexible or Elastic Bodies, 1638–1788*. Leonhardi Euleri Opera Omnia, Introduction to Volumes X and XI, Second Series. Orell Füssli, Zürich, 1960.

[vNW29] J. von Neumann and E. Wigner. Über das Verhalten von Eigenwerten bei Adiabatischen Prozessen. *Physikalische Zeit.*, 30:467–470, 1929.

[You88] Nicholas Young. *An Introduction to Hilbert Space*. Cambridge University Press, Cambridge, 1988.