Linear Algebra

Linear Algebra

Kanyes Thaker

Last updated: September 29, 2022

Note to the Reader

Any exploration into a sufficiently technical field should be rooted in a strong understanding of the fundamentals. Linear algebra forms the backbone of much of the scientific world. Linear equations are used to approximate models in practically every area of modern science, and extends naturally to topology, geometry, abstract algebra, machine learning, quantum physics, optimization theory, chemistry, computer graphics, biology, probability theory, and more.

Learning is a continuous process, and often times it is useful to learn concepts *pointwise* – to learn the tool needed to complete a task only when that task arises, and to stop at exactly the tool you need. But for some fundamental principles, those which can be universally applied across domains and across disciplines, it is worth the investment in spending the time to learn the breadth of the subject fully and deeply. Linear algebra is one such subject.

Contents

※ Introduction

※ Vector Spaces

A **vector**, in a geometric sense, is an object that has properties of both direction and magnitude. We often represent vectors as collections of objects, drawn from a field $\mathcal{F}$, either in **row form** or **column form**:

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Each entry in the vector is known as a **component**. With this view, the addition of two vectors $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ is the component-wise sum $\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2)$. Likewise, for a scalar value $a \in \mathcal{F}$, scalar multiplication is $a\mathbf{x} = (ax_1, ax_2)$.

A set whose elements are vectors, with the operations of addition and scalar multiplication, is known as a **vector space** $\mathcal{V}$ over the field $\mathcal{F}$ if it satisfies the following eight axioms for all $a, b \in \mathcal{F}$ and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$:

- Commutativity of vector addition: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
- Associativity of vector addition: $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$
- Zero vector: there exists a unique $\mathbf{0}$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$
- Additive inverse: for each $\mathbf{x}$, there exists a unique $\mathbf{y}$ such that $\mathbf{x} + \mathbf{y} = 0$
- Unity: There exists a unique $1 \in \mathcal{F}$ such that $1\mathbf{x} = \mathbf{x}$
- Distribution of scalar addition over a vector: $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$
- Distribution of a scalar over vector addition: $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$

The uniqueness of the zero vector and additive inverse follow from the (somewhat trivial) idea that if $\mathbf{x} + \mathbf{z} = \mathbf{y} + \mathbf{z}$ then $\mathbf{x} = \mathbf{y}$.

A **matrix** is a two-dimensional $m \times n$ array whose elements are also drawn from the field $\mathcal{F}$. We denote such a matrix as $\mathcal{M}_{m\times n}(\mathcal{F})$, and access the elements of a matrix $\mathbf{A}$ as a_{ij}, where i indicates the row index and j is the column index. We may write such an object as:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

If $m = n$, we call the matrix **square**. We may also obtain a matrix $\mathbf{B}$ in $\mathcal{M}_{n\times m}(\mathcal{F})$ by

swapping the indices of $\mathbf{A}$ (the **transpose** $\mathbf{A}^\top$) i.e. $b_{ij} = a_{ji}$. If $\mathbf{A}^\top = \mathbf{A}$, the matrix is **symmetric**. If $\mathbf{A}^\top = -\mathbf{A}$, we instead call it **skew-symmetric**. The set of values a_{ii} form the **diagonal** of the matrix; a matrix which only has non-zero entries on the diagonal is a **diagonal matrix**, and a matrix that has no nonzero entries *below* the diagonal is **upper triangular**. Any matrix where the number of non-zero entries is small (with respect to the total size of the matrix) is called **sparse**. The sum of the diagonal entries of a matrix, $\sum_{i=1}^{\min(m,n)} a_{ii}$, is the **trace** of the matrix trace($\mathbf{A}$).

2.1 Subspaces

A subset $\mathcal{W}$ of $\mathcal{V}$ is a **subspace** if $\mathcal{W}$ is also a vector space, with the same definitions of multiplication and scalar addition as $\mathcal{V}$. Most of the vector space axioms are general enough to be satisfied for any subset of $\mathcal{V}$, so there are only a few conditions that need to be checked explicitly – $\mathcal{W}$ must contain the same zero vector as $\mathcal{V}$; every vector in $\mathcal{W}$ must also have an additive inverse in $\mathcal{W}$; and $\mathcal{W}$ must be closed under addition and scalar multiplication. From these properties it can also be deduced that the intersection of subspaces of a vector space is also a subspace.

2.2 Linear Dependence

The key operations of vector addition and scalar multiplication allow us to consider vectors created via these operations. For a vector space $\mathcal{V}$ and a subset $\mathcal{S}$, a vector $\mathbf{v}$ that can be expressed as the sum of some other scaled vectors in $\mathcal{V}$ is a **linear combination** of those vectors, i.e. if $\mathbf{v} = \sum_i a_i \mathbf{v}_i$. The set of all possible linear combinations of the vectors in $\mathcal{S}$ is the **span** of that group of vectors, denoted **span**(S). Since (from above) a subspace is any subset that is closed under the vector operations and contains the zero vector, the span of any subset of $\mathcal{V}$ has to be a subspace, and (conversely) any subspace containing any subset must also contain the span of that subset. If **span**($\mathcal{S}$) = $\mathcal{V}$ then we say that $\mathcal{S}$ **generates** $\mathcal{V}$.

If a linear combination of a set of vectors is zero, one vector can be written as a linear combination of the others. Trivially, any linear combination of any vectors can be zero if their scaling coefficients a_i are all zero. If it is possible for a linear combination of a set of vectors to be zero with *not* all the $a_i = 0$, that set is **linearly dependent**. Otherwise it is **linearly independent**. Note that by this definition, removing a linearly dependent vector (a vector that is a linear combination of other vectors in the set) will not change the span of that set. So a linearly independent set $\mathcal{S}$ can become linearly *dependent* by adding any vector $\mathbf{v} \in$ **span**($\mathcal{S}$) – this also means that if a set is already linearly dependent, no superset can be linearly *in*dependent.

2.3 Basis, Dimension

We've introduced ideas of linear independence and of generating sets (usually called spanning sets). From the previous section, it appears as though there is an "upper bound" on linear independence, and after some point (when we add a vector in $\textbf{span}(\mathcal{S})$) our set won't be linearly independent anymore. Likewise, it feels like if $\textbf{span}(\mathcal{S}) = \mathcal{V}$, there must be some "smallest" size of $\mathcal{S}$ below which it is not a spanning set.

A **basis** β for $\mathcal{V}$ is any linearly independent set of vectors that spans $\mathcal{V}$. The basis formed exclusively with vectors with exactly one nonzero element is known as the **standard basis** for that vector space (in the case of $\mathcal{F}_n$, the standard basis is $\{e_0 = (1, 0, \ldots, 0), ..., e_n = (0, 0, \ldots, 1)\}$). β is a basis for $\mathcal{V}$ only if every vector in $\mathcal{V}$ is a *unique* linear combination of the basis vectors. Interestingly, since each vector's representation with respect to a basis β is unique, we can equivalently represent any vector with respect to β by the n coefficients $a_1, ..., a_n$ applied to those basis vectors (this will come up later).

Replacement Theorem

Suppose $\mathcal{V}$ is a vector space, where $\mathcal{L} \subseteq \mathcal{V}$ is a linearly independent subset with m elements and $\mathcal{G} \subseteq \mathcal{V}$ is a spanning set for $\mathcal{V}$ with n elements. Then the following statements are true:

1. $m \leq n$; no linearly independent subset is larger than a generating subset
2. There exists a subset $\mathcal{H} \subseteq \mathcal{G}$ with at most $n - m$ elements such that $L \cup H$ is a spanning set.

The replacement theorem is a powerful tool in linear algebra – it states that any linearly independent subset can be *extended* to become a spanning set, and likewise that any spanning set can be *reduced* to a linearly independent one. This has some natural consequences – namely, it means that all sets that are linearly independent and spanning (all bases) must have the same size (else condition (1) is violated). This "size" of the bases is known as the **dimension** of the vector space, denoted $\textbf{dim}(\mathcal{V})$. Conversely, any linearly independent set of size $\textbf{dim}(\mathcal{V})$ must be a basis, as must any spanning set of size $\textbf{dim}(\mathcal{V})$.

2.3.1 Lagrange Interpolation

Let's explore a consequence of this idea that all size-$\textbf{dim}(\mathcal{V})$ linearly independent subsets are bases (we prove later that all vector spaces have a basis). Suppose we have a dataset $\{(c_1, b_1), \ldots, (c_n, b_n)\}$. A typical task would be to find some intermediate value for an unknown c_i; to do so, we could find a polynomial interpolating all these points. For each $i = 1, ..., n$, the corresponding **Lagrange basis function** is

$$\ell_i(x) = \prod_{k \neq i} \frac{x - c_k}{c_i - c_k}.$$

This form is chosen due to its special behavior at $x = c_j$, where $\ell_i(c_j) = \mathbf{1}_{i=j}$ (easy to verify). Because of this, all the ℓ_i are linearly independent. Since there are n of them, they form a basis for $P_n(\mathcal{F})$ – so *any* degree n polynomial over $\mathcal{F}$ is a linear combination of these basis functions. To interpolate our desired datapoints, we need to make it so that $f(c_i) = b_i$. We use the "indicator" property of these basis functions and set the coefficient of ℓ_i to be b_i (so when c_i is our input, b_i is our output). Our end polynomial is

$$f(x) = b_1\ell_1(x) + b_2\ell_2(x) + \ldots + b_n\ell_n(x) = \sum_{i=1}^{n} b_i\ell_i(x).$$

2.4 Infinite Spaces

Aside: this section is an optional exploration of the reasoning behind why every vector space admits a basis. It expects some basic familiarity with order theory.

It is often difficult to characterize bases for infinite spaces, for instance the vector space of real numbers over the field of rationals. For a family of sets $\mathcal{F}$ (unrelated to our use of $\mathcal{F}$ to describe fields) a member M is **maximal** if no other set in $\mathcal{F}$ contains M. A **chain** is a collection of sets that admits a total ordering, in this case, via the subset relation. A maximal chain is then a chain that admits no proper superset. The **Hausdorff maximal principle** claims that any poset admits a maximal chain (this is equivalent to the **axiom of choice**, more obviously via **Zorn's lemma**). A linearly independent subset of a vector space that is also maximal must be a basis (from the previous section). Therefore, it just remains to show that such a maximal linearly independent subset exists.

Let $\mathcal{F}$ be the family of all linearly independent subsets of $\mathcal{V}$. $\mathcal{F}$ is a poset under the subset relation, and we can partition it into chains – more specifically, from the Hausdorff maximal principle we know it admits a maximal chain C. Now suppose we have a set $\mathcal{U}$ which is the union of all elements in C. Since C is maximal, $\mathcal{U} \in C$, so no other set can contain $\mathcal{U}$ (so $\mathcal{U}$ is a maximal subset). Now we must show $\mathcal{U}$ is linearly independent (for an infinite set, this just means all finite subsets are linearly independent). Assume a finite subset $\{u_1, \ldots, u_n\}$, and state that $\sum a_i u_i = 0$. Since $\mathcal{U} = \bigcup C$, each $u_i \in A_i$ for some $A_i \in C$. Since C is a chain it is totally ordered, which means (WLOG) we can arrange the sets $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n$ – then these $\{u_n\} \in A_n$ which is linearly independent, meaning all $a_i = 0$ and $\mathcal{U}$ is linearly independent. Since $\mathcal{U}$ is a maximal linearly independent subset, $\mathcal{V}$ admits a basis.

※ Linear Transformations, Matrices

In the previous section we briefly discussed arbitrary transformations on vector spaces. In particular, we are interested in those transformations which have some "structure-preserving" property on these spaces, namely that of addition and scalar multiplication.

To that end, a **linear transformation** $T : \mathcal{V} \mapsto \mathcal{W}$ is a mapping from $\mathcal{V}$ to $\mathcal{W}$ such that, for any $c \in \mathcal{F}$ and $\mathbf{x}, \mathbf{y} \in \mathcal{V}$, $T(c\mathbf{x} + \mathbf{y}) = cT(\mathbf{x}) + T(\mathbf{y})$. Note this means $T(0) = 0$. If we have a map where $T(0) = c \in \mathcal{F}$ and $T - c$ is linear, we call T an **affine** transformation. Examples of such linear transformations include integration, differentiation, rotation, reflection, and projection.

There are some special linear transformations worth mentioning – the **identity** transformation $I_{\mathcal{V}} : \mathcal{V} \mapsto \mathcal{V}$ takes every element to itself ($I_{\mathcal{V}}(\mathbf{x}) = \mathbf{x}$) and the **zero** transformation $T_0 : \mathcal{V} \mapsto \{0\}$ maps every element to the zero vector ($T_0(\mathbf{x}) = 0$). If we consider $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$, with $\mathbf{x} \in \mathcal{V} = \mathbf{x}_1 \in \mathcal{W}_1 + \mathbf{x}_2 \in \mathcal{W}_2$, then the **projection** of $\mathbf{x}$ onto $\mathcal{W}_1$ is $proj_{\mathcal{W}_1}(\mathbf{x}) = \mathbf{x}_1$.

3.1 Rank, Nullity

There are two sets which help us understand the intrinsic properties of T more closely. The **null space** or **kernel** of T, denoted $\mathcal{N}(T)$ or $\ker(T)$, is the set of vectors which T sends to zero, $\{\mathbf{x} : T(\mathbf{x}) = 0\}$. The **range** or **image** of T, denoted $\mathcal{R}(T)$ or $\text{im}(T)$, is the set of vectors in $\mathcal{W}$ that result from applying T to vectors in $\mathcal{V}$, $\{\mathbf{w} \in \mathcal{W} : (\exists \mathbf{v} \in \mathcal{V} : T(\mathbf{v}) = \mathbf{w})\}$. By linearity, both the null space and range are subspaces of $\mathcal{V}$ and $\mathcal{W}$, respectively. $T(\beta)$ is a generating set for $\text{im}(T)$, but note that even though β is a basis for $\mathcal{V}$, $T(\beta)$ need not be a basis for $\mathcal{W}$ (take $T(x, y) = (x, y, 0)$). We additionally name the dimension of the kernel the **nullity** of T, and the dimension of the image the **rank**.

Rank-Nullity Theorem

$$\dim(\ker(T)) + \dim(\text{im}(T)) = \textbf{null}(T) + \textbf{rank}(T) = \dim(\mathcal{V}).$$

A transformation T is injective if and only if the nullspace is trivial (only contains the zero vector). If $T : \mathcal{V} \to \mathcal{W}$ and $\dim(\mathcal{V}) = \dim(\mathcal{W})$ then T is bijective if $\dim(\text{im}(T)) = \dim(\mathcal{V})$ (this is a natural consequence of the above theorem and definitions).

A linear transformation is completely characterized by its behavior when applied to a basis. Given a basis $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ for $\mathcal{V}$, and a set of vectors $\{\mathbf{w}_1, ..., \mathbf{w}_n\}$ in $\mathcal{W}$, there exists exactly one linear transformation T that sends $\mathbf{v}_i$ to $\mathbf{w}_i$. This means that two linear transformations T and U can only be equivalent if, for every input vector $\mathbf{v}_i$ the output the same $\mathbf{w}_i$.

3.2 Matrices

An **ordered basis** for a vector space $\mathcal{V}$ is a sequencing of basis vectors (so that two bases with the same elements but different order are not equal). The standard ordered basis for $\mathcal{F}^n$ is $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$. For a given ordered basis with basis vectors $\mathbf{v}_i$, there is a unique representation of the vectors $\mathbf{x} \in \mathcal{V}$ as $\mathbf{x} = \sum_i a_i \mathbf{v}_i$. We can group all of these coefficients

a_is to form a new size-n vector in $\mathcal{F}^n$. This **coordinate vector**, denoted $[\mathbf{x}]_\beta$, is a linear transform from $\mathcal{V} \to \mathcal{F}^n$. It uniquely characterizes the vector $\mathbf{x}$ with respect to the basis β.

We can extend this idea to propose that for a linear transform $T : \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}$ has basis $\beta = \{\mathbf{v}_i\}_{i=1}^n$ and $\mathcal{W}$ has basis $\gamma = \{\mathbf{w}_i\}_{i=1}^m$, there is an equivalent coordinate representation. In fact, each $T(\mathbf{v}_j) = \sum_i a_{ij}(\mathbf{w}_i)$. These coefficients a_{ij} correspond to a matrix $\mathbf{A}$, equivalently denoted $[T]_{\beta\to\gamma}$. The jth column of $\mathbf{A}$ is $[T(\mathbf{v}_j)]_\gamma$, i.e. the jth column of $\mathbf{A}$ consists of the coefficients of $a_{1j}\gamma_1 + ... + a_{mj}\gamma_m = T(\mathbf{v}_j)$. The zero transform has matrix $[T_0]_{\beta\to\gamma}$ consisting entirely of zeros, and the identity has matrix $[I_\mathcal{V}]_{\beta\to\gamma}$ where $a_{ij} = \delta_{i,j}$ (the Kronecker delta).

Matrices preserve linearity, and are hence linear themselves, i.e. $[aT+U]_{\beta\to\gamma} = a[T]_{\beta\to\gamma} + [U]_{\beta\to\gamma}$. Endowed with these two operations of addition and scalar multiplication, the set of all transformations from $\mathcal{V} \to \mathcal{W}$ is itself a vector space, denoted $\mathcal{L}(\mathcal{V}, \mathcal{W})$.

A natural next question involves the chaining of linear transformations, i.e. composing linear transformations with each other, and the relationship between these compositions and the matrix representation. For $T : \mathcal{V} \to \mathcal{W}$ and $U : \mathcal{W} \to \mathcal{Z}$, the composition $U \circ T$ is written as $UT : \mathcal{V} \to \mathcal{Z}$ and is linear. We can think of this as performing the transformation T on $\mathcal{V}$, and then performing U on $\text{im}(\mathcal{V})$. If $T : \mathcal{V} \to \mathcal{V}$, we can compose T with itself. $T^2 = TT$, $T^3 = T^2T$, $T^k = T^{k-1}T$.

We can think about the matrix representations of these transformations in a similar way. If $\mathbf{A} = [U]_{\beta\to\gamma}$ and $\mathbf{B} = [T]_{\alpha\to\beta}$, then $\mathbf{AB} = [UT]_{\alpha\to\gamma}$. If α has dimension n, β has dimension p, and γ has dimension m, then $\mathbf{A} \in \mathcal{F}^{m\times p}$, $\mathbf{B} \in \mathcal{F}^{p\times n}$, and $\mathbf{AB} \in \mathcal{F}^{m\times n}$.

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj},\ 1 \le i \le m,\ 1 \le j \le n.$$

We refer to this as **matrix multiplication**, and it is not commutative in general. It is associative (by definition) and also distributive, meaning $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$ and $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. A matrix which, when multiplied by itself k times is zero (i.e. $\mathbf{M}$ s.t. $\mathbf{M}^k = \mathbf{0}$) is referred to as being **nilpotent**.

We can think of each column of $\mathbf{AB}$ as the result when we apply the linear transform represented by $\mathbf{A}$ to each of the column vectors in $\mathbf{B}$. For column $\mathbf{u}_j \in \mathbf{AB}$, $\mathbf{u}_j = \mathbf{Av}_j$ for $\mathbf{v}_j \in \mathbf{B}$, where $\mathbf{v}_j = \mathbf{Be}_j$ where $\mathbf{e}_j \in \alpha$ is a standard basis vector for $\mathcal{V}$. In a linear transformation sense, we write this as $[T(\mathbf{u})]_{\beta\to\gamma} = [T]_{\beta\to\gamma}[\mathbf{u}]_\gamma$.

3.2.1 Invertibility, Isomorphism

For a linear transformation $T : \mathcal{V} \mapsto \mathcal{W}$, a function $U : \mathcal{W} \mapsto \mathcal{V}$ is an **inverse** of T if $TU = I_\mathcal{W}$ and $UT = I_\mathcal{V}$. We write such a transformation U as T^{-1}, and note that T^{-1} must also be invertible. This means that for any $\mathbf{v} \in \mathcal{V}$, $(T^{-1}T)(\mathbf{v}) = \mathbf{v}$ – and for any

$\mathbf{w} \in \mathcal{W}$, $(TT^{-1})(\mathbf{w}) = \mathbf{w}$. T must be bijective, else it would be impossible to uniquely recover the input vector. From the rank-nullity theorem, this has the implication that $\dim(\mathcal{V}) = \dim(\mathcal{W})$. Note $(TU)^{-1} = U^{-1}T^{-1}$. From a matrix point of view, the $n \times n$ matrix $\mathbf{A}$ is invertible if there exists some $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$.

Orthogonal, Unitary Matrices

If a matrix's inverse is also its transpose, i.e. $\mathbf{AA}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$, we say that the matrix is **orthogonal**.

If we take $\mathbf{A}$ to be a complex-valued matrix and let $\mathbf{A}^*$ be the **conjugate transpose** of $\mathbf{A}$ (the transpose with all elements replaced with their complex conjugates), and $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}_n$, we refer to $\mathbf{A}$ as being **unitary** (unitary real-valued matrices are orthogonal).

As a side note, the matrix $\mathbf{A}^*$ is sometimes written as $\mathbf{A}^\dagger$ (not to be confused with $\mathbf{A}^+$ for pseudoinverse), or $\mathbf{A}^\mathbf{H}$, and is the **Hermitian adjoint** for a finite-dimensional matrix, i.e. the matrix $\mathbf{B}$ such that $\langle \mathbf{Ax}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{By} \rangle$. In $\mathbb{R}^n$, the transpose is the adjoint matrix.

Invertibility helps formalize the notion that certain vector spaces "resemble" other vector spaces, with the exception of how the vectors are visually represented on the page. Adding two matrices in $\mathcal{M}_{2\times2}(\mathcal{F})$ looks very similar to adding two polynomials in $P_4(\mathcal{F})$; in fact, every transformation on $P_4(\mathcal{F})$ could be just as well applied to $\mathcal{M}_{2\times2}(\mathcal{F})$. This means there is some "structure-preserving" transformation between the two – we call this an **isomorphism**.

Isomorphism

Two vector spaces $\mathcal{V}, \mathcal{W}$ are **isomorphic** if there exists an invertible linear transformation $T : \mathcal{V} \mapsto \mathcal{W}$. Such a T is an **isomorphism** from $\mathcal{V}$ onto $\mathcal{W}$.

From the previous discussion, it is clear that two vector spaces are isomorphic if and only if they are the same dimension. The concept of an isomorphism allows us to draw an explicit relationship between the space of transformations from $\mathcal{V}$ to $\mathcal{W}$, and the set of matrices corresponding to those transformations. In fact, the transformation $\Phi_{\beta\to\gamma} : \mathcal{L}(\mathcal{V}, \mathcal{W}) \mapsto \mathcal{M}_{m\times n}(\mathcal{F})$ is an isomorphism, indicating that these two vector spaces carry the same structure. In particular, the map $\phi_\beta : \mathcal{V} \mapsto \mathcal{F}^n$ representing the coordinate representation $[\mathbf{v}]_\beta$ is an isomorphism. This relationship has the advantage of allowing us to describe *any* transformation between two arbitrary vector spaces through an "equivalent" matrix multiplication from $\mathcal{F}^n \to \mathcal{F}^m$. In particular, the **left matrix multiplication** $L_\mathbf{A} : \mathcal{F}^m \to \mathcal{F}^n$, $L_\mathbf{A}(\mathbf{x}) = \mathbf{Ax}$ allows us to capture any linear

transformation T as a matrix-vector multiplication:

$$\mathbf{v} \in \mathcal{V} \xrightarrow{\phi_\beta} [\mathbf{v}]_\beta \in \mathcal{F}^m \xrightarrow{L_\mathbf{A}} [\mathbf{v}]_\gamma \in \mathcal{F}^n \xrightarrow{\phi_\gamma^{-1}} \mathbf{w} \in \mathcal{W}.$$

3.2.2 Change of Coordinates

Given the idea that there exists an isomorphism for any $\mathcal{V}$ onto $\mathcal{F}^n$, we might wonder how to represent a coordinate vector in terms of different bases, i.e. how to relate $[\mathbf{x}]_\beta$ and $[\mathbf{x}]_{\beta'}$. In particular, there exists an invertible matrix $\mathbf{Q} = [I_\mathcal{V}]_{\beta' \to \beta}$ for which $[\mathbf{x}]_\beta = \mathbf{Q}[\mathbf{x}]_{\beta'}$. We call such a $\mathbf{Q}$ the **change-of-coordinate** matrix from β' to β. In particular, the jth column of $\mathbf{Q}$ consists of the coefficients $\{a_{ij}\}$ such that $\beta_i' = \sum a_{ij}\beta_j$.

We can extend this idea to linear transformations that stay within $\mathcal{V}$ – these are known as **linear operators**, $T : \mathcal{V} \mapsto \mathcal{V}$. To represent the relationship between $[T]_{\beta'}$ and $[T]_\beta$, we can consider the following sequence of operations. Given an input $[\mathbf{x}]_{\beta'}$ represented in β', we can use $\mathbf{Q}$ to transform it into $[\mathbf{x}]_\beta$. Then we can apply our operator $[T]_\beta$; finally, we convert the output back into a β'-representation with $\mathbf{Q}^{-1}$.

Similarity and Conjugacy

We represent the change of basis for a linear operator $T : \mathcal{V} \to \mathcal{W}$ from β to β', assuming the existence of an *invertible* change of coordinate matrix $\mathbf{Q} = [I_\mathcal{V}]_{\beta' \to \beta}$, as

$$[T]_{\beta'} = [I_\mathcal{V}]_{\beta \to \beta'}[T]_\beta[I_\mathcal{V}]_{\beta' \to \beta} = \mathbf{Q}^{-1}[T]_\beta\mathbf{Q}.$$

If there exists an $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$, we say that $\mathbf{A}$ and $\mathbf{B}$ are **similar** – more specifically, if we consider the **general linear group** $GL_n(\mathcal{F})$ (the group of $n \times n$ invertible matrices equipped with matrix multiplication), we call $\mathbf{B}$ a **conjugate** of $\mathbf{A}$, and the transform $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is known as **conjugacy**.

3.3 Dual Spaces

Consider now the set of transformations, here denoted as f, $f : \mathcal{V} \mapsto \mathcal{F}$. A transformation that sends $\mathcal{V}$ to $\mathcal{F}$ is known as a **linear functional** on $\mathcal{V}$. The vector space of linear functionals, $\mathcal{L}(\mathcal{V}, \mathcal{F})$, is the **dual space** of $\mathcal{V}$, denoted $\mathcal{V}^*$. $\mathcal{L}(\mathcal{V}, \mathcal{F})$ has dimension $\dim(\mathcal{V}) \times \dim(\mathcal{F}) = \dim(\mathcal{V})$, meaning $\mathcal{V}$ and $\mathcal{V}^*$ are isomorphic.

Suppose $\mathcal{V}$ has basis $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$. Then for any vector $\mathbf{x} \in \mathcal{V}$, we know $\mathbf{x} = \sum_i a_i\mathbf{v}_i$. The linear functional $f_i(\mathbf{x}) = a_i$ is the **ith coordinate function with respect to** β. For the basis vectors themselves, this choice of function has the unique property that $f_i(\mathbf{v}_j) = \delta_{ij}$. $\beta^* = \{f_1, ..., f_n\}$ is thus linearly independent and has dimension n, so it is an ordered basis for $\mathcal{V}^*$, known as the **dual basis**.

Now, consider a linear transformation $T : \mathcal{V} \to \mathcal{W}$. $\mathcal{W}$ itself also has a dual space, $\mathcal{W}^*$, corresponding to the set of all functionals with domain $\mathcal{W}$. For $g \in \mathcal{W}^*$, we can

represent the transformation of a vector $\mathbf{v} \in \mathcal{V} \to \mathcal{W} \to \mathcal{F}$ by the composition gT. A classic question in category theory is to try and find, for a composition $f(g(x))$, a function f^* with the same domain as f such that $f(g(x)) = f^*(x)$ (a **pullback**) – how we can get the same result as gT without needing to directly invoke T. In this case, we are interested in the transformation $U : \mathcal{W}^* \mapsto \mathcal{V}^*$ such that $(U(g))(\mathbf{v}) = (gT)(\mathbf{v})$. We call such a map U the **transpose** of T, denoted $T^\top$. In a more abstract sense, the transpose is a map between dual spaces induced by the transformation T – it is the **algebraic adjoint** of T, meaning it fulfills the property that $\langle T^\top(f), \mathbf{x}\rangle = \langle f, T(\mathbf{x})\rangle$, where the inner product is the natural composition $\langle z, h\rangle = z(h)$ (see section 7).

Finally, we can show that there is also an isomorphism between $\mathcal{V}^{**}$, the **double dual**, and $\mathcal{V}$. We define $\hat{\mathbf{x}}(f) = f(\mathbf{x})$ for all $f \in \mathcal{V}^*$, making $\hat{\mathbf{x}} : \mathcal{V}^* \mapsto \mathcal{F}$ a functional on $\mathcal{V}^*$. The function $(\mathbf{x}) = \hat{\mathbf{x}}$ is an isomorphism from $\mathcal{V}^{**}$ onto $\mathcal{V}$ – the dual of the dual space is "equivalent" to the original space.

※ Systems of Linear Equations

One of the most powerful (and common) applications of linear algebra is to find solutions to systems of linear equations – that is, find solutions to sets of equations of the form $\sum_i a_i x_i - b = 0$. To do this, we use a series of **elementary operations** which we can apply to rows (or columns) of matrices in order to manipulate them into forms which yield the solutions to these equations. The elementary operations are:

1. Swapping rows/columns of the matrix
2. Multiplying rows/columns in-place by a scalar
3. Adding a scalar multiple of a row/column to another row/column.

When we apply a single one of these operations to the identity matrix $\mathbf{I}$, we dub the result an **elementary matrix**. In fact, applying a row-wise elementary operation to a matrix $\mathbf{A}$ is equivalent to left-multiplying it with the corresponding elementary matrix (likewise, right-multiplying for column-wise operations). The inverse of the operation done to create an elementary matrix is another elementary matrix, meaning that all elementary matrices are invertible.

4.1 Rank, Inverse

The rank of a matrix $\mathbf{A}$ is the same as the rank of $L_\mathbf{A} : \mathcal{F}^m \to F^n$ – $\mathbf{A} \in \mathcal{M}_{n\times n}(\mathcal{F})$ is invertible precisely if $\mathbf{rank}(\mathbf{A}) = n$ (we call $\mathbf{A}$ **full-rank**). In general, the rank of a linear transformation is the rank of its corresponding matrix. Rank is preserved through multiplication with invertible matrices (composing $L_\mathbf{A}$ with an invertible transformation will not change the dimension of its image). Since elementary operations are invertible, they are therefore also rank-preserving – so a matrix transformed through this method

has the same rank as the original.

From the previous definitions, we may conclude that the rank of a matrix is the same as the number of its linearly independent rows/columns. This results from the fact that $\dim(\text{im}(\mathbf{L_A}))$ is the same as $\dim(\mathbf{span}(\mathbf{L_A}(\beta)))$ for a basis β of $\mathcal{F}^n$; but for a basis vector $\mathbf{e}_i$, $\mathbf{L_A}(\mathbf{e}_i)$ is the jth column of $\mathbf{A}$ (by definition). So the span of $L_\mathbf{A}(\beta)$ is then the span of the columns of $\mathbf{A}$. A symmetric argument over $\mathbf{A}^\top$ reveals the same about rows.

The challenge then becomes how to identify how many linearly independent rows/-columns there are. It so happens that any matrix can be transformed, using a series of elementary row operations $\mathbf{B}$ and elementary column operators $\mathbf{C}$, into a matrix of form

$$\mathbf{D} = \mathbf{BAC} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r,n-r} \\ \mathbf{0}_{m-r,r} & \mathbf{0}_{m-r,n-r} \end{pmatrix}.$$

In other words, we can transform $\mathbf{A}$ into a matrix whose top left corner is the identity matrix of dimension $\mathbf{rank}(\mathbf{A})$, and whose other entries are all zero. Note a common point of confusion – even though $\mathbf{B}$ and $\mathbf{C}$ are invertible, this is *not* equivalent to saying that $\mathbf{A}$ is *similar* to $\mathbf{D}$! This representation makes it clear that the rank of a matrix is dependent only the number of linearly independent rows/columns, and further implies that $\mathbf{rank}(\mathbf{AB}) \leq \min(\mathbf{rank}(\mathbf{A}), \mathbf{rank}(\mathbf{B}))$. We can use all this information to use elementary row operations to find the inverse of a matrix. Given two matrices $\mathbf{A}$ and $\mathbf{B}$, we define the **augmented matrix** $\mathbf{C} = (\mathbf{A}|\mathbf{B})$ as the column-wise concatenation of $\mathbf{A}$ and $\mathbf{B}$. If we augment $\mathbf{A}$ with the identity matrix, and perform row operations to transform $\mathbf{A}$ into $\mathbf{I}$, we will retrieve the inverse of $\mathbf{A}$:

$$\mathbf{A}^{-1}(\mathbf{A}|\mathbf{I}_n) = (\mathbf{A}^{-1}\mathbf{A}|\mathbf{A}^{-1}\mathbf{I}) = (\mathbf{I}|\mathbf{A}^{-1}).$$

4.2 Solving Systems of Linear Equations

Consider the system of equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

We call such a system a **system of linear equations over** $\mathcal{F}$. From our knowledge of the definition of matrix multiplication, we may equivalently write this system as the equation $\mathbf{Ax} = \mathbf{b}$, where:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ & & \ddots & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Our solution is a vector $\mathbf{s} \in \mathcal{F}^n$ such that $\mathbf{As} = \mathbf{b}$; we denote the set of all solutions as $\mathcal{S}$. If $\mathcal{S}$ is empty, we call the system **inconsistent**.

If $\mathbf{b} = \mathbf{0}$, we call the system **homogenous**. By definition, the solution set $\mathcal{S}_0$ for a homogenous system of linear equations is nothing more than $\ker(L_\mathbf{A})$. In fact, we can generate the solution to a non-homogenous system of equations $\mathbf{Ax} = \mathbf{b}$ by finding a single solution $\mathbf{s}$ – we can then generate all other solutions by adding $\{\mathbf{s}\}$ to each solution of the corresponding homogenous system $\mathbf{Ax} = \mathbf{0}$ (i.e. $\mathcal{S}_\mathbf{b} = \{\mathbf{s} + \mathbf{s}_0 | \mathbf{s}_0 \in \mathcal{S}_0\}$). If $\mathbf{A}$ is full-rank, there is only one solution to the system (on account of the fact that $\mathcal{S}_0$ is empty), and that solution is $\mathbf{A}^{-1}\mathbf{b}$.

The obvious question: *how can we use this information to find a solution*? Our algorithm hinges on a crucial observation – multiplying a linear equation by an invertible matrix preserves the solution set. This is because if $\mathbf{CAx} = \mathbf{Cb}$ admits solution $\mathbf{s}$, then left-multiplying by $\mathbf{C}^{-1}$ shows that $\mathbf{Ax} = \mathbf{b}$ also admits solution $\mathbf{s}$ (the converse is true by a symmetric argument). Since our elementary matrices from before are invertible, we maintain our solution vector through any number of multiplications with elementary matrices. If the augmented matrix $(\mathbf{A}|\mathbf{b})$ can be transformed into $(\mathbf{A}'|\mathbf{b}')$ via elementary operations, then the solutions to $\mathbf{Ax} = \mathbf{b}$ are the same as the solutions to $\mathbf{A}'\mathbf{x} + \mathbf{b}'$.

Reduced-row echelon form is a target form for our augmented matrix that makes solving the resulting system extremely convenient.

1. All non-zero rows are above all zero rows,
2. The first nonzero entry in each row is the *only* nonzero entry in that column
3. The first nonzero entry in each row is to the right of the nonzero entry before it, and its value is 1.

The following augmented matrix is in reduced row echelon form.

$$\left(\begin{array}{cccc|c} 1 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)$$

The corresponding system of linear equations admits no unique solution. It consists of the equations $x_1 + 2x_3 = 1$, $x_2 = 2$, and $x_4 = 3$. There are no unique values for x_1 and x_3 that we can determine, so there are infinitely many possible solutions. The reduced-row echelon form is unique for each matrix (there exists a bijection between a matrix and its reduced-row echelon form, defined by the specific set of elementary operations performed).

The most efficient elementary method to turn a matrix into its reduced-row echelon form takes two steps – the **forward pass** uses elementary operations to transform the augmented matrix into an upper triangular matrix, where we satisfy condition (3). The

backward pass uses the "1"s on the diagonal to satisfy conditions (1) and (2) above the diagonal. This method is known as **Gaussian elimination** – it is guaranteed to put *any* matrix into reduced-row echelon form.

In this form we get some information – if any row has more than one nonzero entry, or if either a row or column has *only* zeros, the system has infinitely many solutions. If the row has only zeros *except* for the entry in the very last column, the system has no solutions. If the left hand side has exactly one entry in each row and column (the left side of the augmented matrix is the identity matrix), the system admits exactly one solution, which is easy to find by looking at the corresponding simplified system of equations.

※ Determinants

The **determinant** is a scalar-valued function of the entries in a square matrix that helps explain some of its properties. Determinants are controversial functions, as most texts do not provide a motivation for their existence or importance. Sheldon Axler's *Linear Algebra Done Right* emphasizes how most results in elementary linear algebra can be deduced without determinants, and instead be determined from the (very related) general idea of the minimal polynomial (see the last section of these notes).[1] In practice, the determinant is still widely used, and the its rejection is on more philosophical versus practical grounds.

In general, the determinant is *not* a linear function of $\mathcal{M}_{n\times n}(\mathcal{F})$, although it *is* linear with respect to each *row* of $\mathbf{A} \in \mathcal{M}_{n\times n}(\mathcal{F})$ (this is known as n**-linearity**). For a matrix $\mathbf{A} \in \mathcal{M}_{2\times 2}(\mathcal{F})$, we define the determinant as:

$$\det(\mathbf{A}) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \quad \mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Determinant (Cofactor)

For a general matrix $\mathbf{A} \in \mathcal{M}_{n\times n}(\mathbf{F})$, the determinant of $\mathbf{A}$ is:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det(\tilde{\mathbf{A}}_{ij}).$$

Here, $\tilde{\mathbf{A}}_{ij}$ represents the matrix $\mathbf{A}$ with row i and column j removed. From this definition, it becomes clear that any matrix with a row (or column) of all zeros has a determinant of zero (we call such a function **alternating**).

The scalar $(-1)^{i+j} \det(\tilde{\mathbf{A}}_{ij})$ is called a **cofactor**. The matrix formed by all

[1] Note that Axler's approaches are only true in general over $\mathbb{C}$ and $\mathbb{R}$, and not arbitrary fields; however this is usually sufficient.

cofactors is known as the **cofactor matrix C**; its transpose $\text{adj}(\mathbf{A}) = \mathbf{C}^\top$ is the **classical adjoint** or **adjugate matrix** to **A**, with the property that $\mathbf{A}\,\text{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{I}_n$. Note that this is *not* equivalent to the "adjoint operator" (the **Hermitian adjoint**) discussed in section 7.

The determinant is impacted by performing elementary row operations on the original matrix. In fact, the only n-linear, alternating function which carries these properties under elementary row operations is the determinant.

1. If **B** is obtained by swapping two rows of **A**, then $\det(\mathbf{B}) = -\det(\mathbf{A})$.
2. If **B** is obtained by multiplying a row of **A** by a scalar k, then $\det(\mathbf{B}) = k\det(\mathbf{A})$.
3. If **B** is obtained by adding a scalar multiple of a row of **A** to another row of **A**, then $\det(\mathbf{B}) = \det(\mathbf{A})$.
4. The determinant of an upper-triangular matrix is the product of its diagonal entries.
5. If two rows or columns are linearly dependent (one of them can be row/column-reduced to all zeros) then the determinant is zero.
6. $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.
7. $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$; a matrix is invertible if and only if $\det(\mathbf{A}) \neq 0$.
8. $\det(\mathbf{A}^\top) = \det(\mathbf{A})$.
9. If **A** and **B** are conjugate, (there exists **Q** such that $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{AQ}$), then $\det(\mathbf{B}) = \det(\mathbf{A})$.

※ Diagonalization

Linear transformations can look like a lot of things – an integral, a derivative, a rotation, a projection, a reflection, a stretch, a shear. These operations might look different (geometrically) depending on the basis the transformation is represented in. A natural question, then, is whether there exists a basis β for $\mathcal{V}$ such that $[T]_\beta$ is *only* a "stretching" transformation (the simplest geometric transformation); equivalently, such that $[T]_\beta$ is a diagonal matrix. We may also say that we are curious whether **A** is similar to a diagonal matrix **D**.

6.1 Eigenbasis

A linear operator T for which there exists an ordered basis β such that $[T]_\beta$ is diagonal is called **diagonalizable**. If $\mathbf{D} = [T]_\beta$ is a diagonal matrix, then applying **D** to a basis vector **v** is then $T(v_j) = \sum_{i=1}^{n} d_{ij}\mathbf{v}_i = d_{jj}\mathbf{v}_j = \lambda_j\mathbf{v}_j$ where $\lambda_j = d_{jj}$. Such a vector **v** where

$T(\mathbf{v}) = \lambda\mathbf{v}$ for a scalar λ is called an **eigenvector**, and the corresponding λ is called an **eigenvalue** (equivalently, $\mathbf{Av} = \lambda\mathbf{v}$). There can be multiple eigenvectors for each eigenvalue (they form a subspace). The basis of each subspace (**eigenspaces**, $\mathcal{E}_\lambda$) is an **eigenbasis** of $\mathcal{E}_\lambda$.

The process of finding such a basis is known as **diagonalization**. Since $\mathbf{Av} = \lambda\mathbf{v}$, rearranging terms shows us that $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v} = 0$; this of course implies that the matrix $(\mathbf{A} - \lambda\mathbf{I}_n)$ has a nontrivial kernel, meaning it is not invertible and $\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0$. This form, $f(t) = \det(\mathbf{A} - t\mathbf{I}_n)$, is the **characteristic polynomial** of $\mathbf{A}$, and its roots are the eigenvalues of $\mathbf{A}$. As an aside, note that although the entries of $(\mathbf{A} - \lambda\mathbf{I}_n)$ are not in $\mathcal{F}$, they *are* in $\mathcal{F}(t)$, the field of quotients of polynomials in t with coefficients in $\mathcal{F}$. Since the determinant is preserved under similarity transformations, the eigenvalues of all similar matrices to $\mathbf{A}$ are the same (but the corresponding eigenvectors are different).

The characteristic polynomial has degree n, and, by the Fundamental Theorem of Algebra, it has n roots (although all the n possible roots need not be unique, nor do they need to be in $\mathcal{F}$). For example, consider the transformation $\mathbf{T}$ that performs a rotation of an angle θ about the zero vector. Clearly there is *no* vector we can find where T can be reduced to a scaling operation – for *any* input vector, the , so such a matrix is not diagonalizable. When a matrix *is* diagonalizable, we can also write $\mathbf{AQ} = \mathbf{DQ}$ where $\mathbf{Q}$ is the matrix of eigenvectors and $\mathbf{D}$ is the matrix of eigenvalues, where $\mathbf{D}_{ii}$ is the eigenvalue corresponding to the ith column of $\mathbf{Q}$. Then $\mathbf{A} = \mathbf{QDQ}^{-1}$ (and $\mathbf{D} = \mathbf{Q}^{-1}\mathbf{AQ}$). This is known as the **eigendecomposition** of $\mathbf{A}$.

6.2 Diagonalizability

The question arises for whether we know a transformation is diagonalizable, and if it is, to find a suitable eigenbasis. In general, the act of calculating an eigenvector for each eigenvalue does not correspond to a basis. For example, $\mathbf{I}_2$ has two eigenvalues, both with a value of 1. Then the vectors $(0, 1)^\top$ and $(0, -1)^\top$ are both valid eigenvectors corresponding to each eigenvalue, but do not form a basis. If $\mathcal{S}_i$ is a set of possible eigenvectors for each of a distinct set of n λ_j, and every pair of $\mathcal{S}_i$ is linearly independent, then the union $\bigcup \mathcal{S}_i$ of all the vectors is also linearly independent. It is easy to show that eigenvectors for different eigenvalues must be linearly independent; this leads us to believe that a matrix with n distinct eigenvalues is diagonalizable. Alternatively, a matrix with n linearly independent eigenvectors (a basis of eigenvectors) is also diagonalizable.

We say that the characteristic polynomial $f(\lambda)$ **splits** if there exist values for $\lambda \in \mathcal{F}$ such that

$$f(\lambda) = c\prod_{i=1}^{n}(\lambda - a_i).$$

The characteristic polynomial of a diagonalizable matrix must split over $\mathcal{F}$ (it must have

n eigenvalues, though they need not be distinct). The number of times the term $(\lambda - a_i)$ appears in the factorization for $f(\lambda)$ is the **multiplicity** of λ. For instance, the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 4 \end{pmatrix}$$

has characteristic polynomial $-(\lambda - 3)^2(\lambda - 4)$, so it has two unique eigenvalues: 4, and 3 with multiplicity of two. Then any eigenbasis for T must contain as many linearly independent eigenvectors per eigenvalue λ as the multiplicity of that eigenvalue. So in our previous example, an eigenbasis for $\lambda = 3$ must contain two linearly independent eigenvectors. The eigenvectors of interest are then the nonzero vectors in $\ker(\mathbf{A} - \lambda\mathbf{I})$ for each λ; these vectors span the **eigenspace** $\mathcal{E}_\lambda$. For λ with multiplicity m, the corresponding eigenspace can have dimension ranging from 1 to m. In fact, T is diagonalizable if and only if every eigenspace has dimension equal to the multiplicity of its corresponding eigenvalue; then the eigenbasis for T is the union of the bases for all component eigenspaces. In the above example, $(\mathbf{A} - 3\mathbf{I})$ has rank 2, meaning its nullity is 1, so it is not diagonalizable.

We can also speak of this basis in terms of a direct sum of subspaces. If $\mathcal{V} = \bigoplus_{i=1}^{k} \mathcal{W}_i$, then each $\mathbf{v} \in \mathcal{V}$ can be writen as a sum of vectors $\mathbf{v}_i \in \mathcal{W}_i$, as $\mathbf{v} = \sum \mathbf{v}_i$. If γ_i is an ordered basis for $\mathcal{W}_i$, then $\bigcup \gamma_i$ is an ordered basis for $\mathcal{V}$. In this way, a linear operator T is diagonalizable if and only if $\mathcal{V}$ is the direct sum of the eigenspaces of T.

6.3 Markov Chains

One of the advantages of diagonal matrices is that matrix multiplication becomes trivially inexpensive, consisting of only n scalar multiplication operations. A natural question in most sciences is how systems behave in the long-run, i.e. the state of a system under a set of repeated, identical changes over a long period of time. Since a matrix represents a linear transformation, and the diagonalized form of a matrix is similar to the original, we may use the diagonal matrix to find information about this limit.

The limit of a sequence of matrices $\{\mathbf{A}_m\}$, $\lim_{m\to\infty} \mathbf{A}_m = \mathbf{L}$ if and only if $\lim_{m\to\infty} \mathbf{A}_{m_{ij}} = \mathbf{L}_{ij}$, i.e. the limit converges element-wise. For appropriate-dimension matrices $\mathbf{P}$ and $\mathbf{Q}$, this definition means $\lim \mathbf{P}\mathbf{A}_m = \mathbf{P}\mathbf{L}$ and $\lim \mathbf{A}_m\mathbf{Q} = \lim \mathbf{L}\mathbf{Q}$. Then, using our definition of similar matrices we must have that if $\mathbf{B} = \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}$, then $\lim \mathbf{B}_m = \mathbf{Q}\mathbf{L}\mathbf{Q}^{-1}$. Our sequence of repeated applications of a linear transformation can be represented by $\mathbf{A}_m = \mathbf{A}^m$, meaning if $\mathbf{D} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ then

$$\lim_{m\to\infty} \mathbf{A}^m = \lim_{m\to\infty} (\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1})^m = \lim_{m\to\infty} \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}\ldots = \lim_{m\to\infty} \mathbf{Q}\mathbf{D}^m\mathbf{Q}^{-1} = \mathbf{L}.$$

If $\mathbf{D}$ is the eigenvalue matrix, the limit $\mathbf{L}$ can only exist if $|\lambda_{max}| \leq 1$, i.e. $\lambda \in B(0, 1)$ the unit ball (otherwise $\lim \lambda^m = \pm\infty$). In fact, if $|\lambda_i| < 1$, it will shrink to 0 in the limit.

If $\mathbf{A}$ is a square matrix whose columns are non-negative and sum to 1, we call it a **transition matrix**, where entry (i, j) can represent the probability of an element of state i transitioning to an element of state j. The columns themselves are **probability vectors**. The property of having all columns sum to one means that the product of any two transition matrices will be another transition matrix.

Markov Chains

The general process where elements are classified as belonging to a state that can switch probabilistically over time is a **stochastic process**. If the probability of of transitioning from $i \to j$ is independent of how we arrived at state i, we say the process exhibits the **Markov property** and is a **Markov process** or **Markov chain**. A transition matrix can be written as

$$\begin{pmatrix} \mathbf{I} & \mathbf{B} \\ \mathbf{O} & \mathbf{C} \end{pmatrix}$$

is known as having **absorbing** states, since an entity which enters a state where $a_{ii} = 1$ will never leave that state.

A transition matrix $\mathbf{A}$ is called **regular** if, for some m, $\mathbf{A}^m$ contains no zero elements. It can be shown that the limit of a sequence of powers of a regular matrix does converge; furthermore, the limit matrix $\mathbf{L}$ in question has exactly one 1-eigenvalue and the columns of $\mathbf{L}$ are all identical (they are all the eigenvector corresponding to the 1-eigenvalue, scaled to form a valid probability vector). The proof of this statement is omitted from these notes, but it relies on the following theorem:

Gershgorin's Circle Theorem

Let $\rho_i(\mathbf{A}) = \sum_{j=1}^{n} |a_{ij}|$, the sum of the magnitudes of the elements in row i of $\mathbf{A}$. The ith **Gershgorin disk** C_i is the disk in the complex plane with center a_{ii} and radius $r_i = \rho_i(\mathbf{A}) - |a_{ii}|$:

$$C_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\}.$$

Then every eigenvalue of $\mathbf{A}$ is contained in a Gershgorin disk.

6.4 Invariant Subspaces, The Cayley-Hamilton Theorem

If $\mathcal{W}$ is a subspace of $\mathcal{V}$ and $T(\mathbf{x}) \in \mathcal{W}$ for all $\mathbf{x} \in \mathcal{W}$, we call $\mathcal{W}$ **T-invariant**. The "smallest" T-invariant subspace for a vector space is, for a vector $\mathbf{v} \in \mathcal{V}$, $\mathbf{span}(\{\mathbf{v}, T(\mathbf{v}), T^2(\mathbf{v}), ...\})$. This is known as the **T-cyclic subspace generated by $\mathbf{v}$**. The characteristic polynomial of $T_{\mathcal{W}}$ neatly divides the polynomial T for the original space (this is consequence of the fact that the basis vectors for $\mathcal{W}$ can be extended to a basis for $\mathcal{V}$).

We can use this property to get more information about the characteristic polynomial for T. If $\mathcal{W}$ is the T-cyclic subspace, then $k = \mathbf{dim}(\mathcal{W})$ is the smallest k such that $\beta = \{\mathbf{v}, T(\mathbf{v}), ..., T^{k-1}(\mathbf{v})\}$ is a basis for $\mathcal{W}$. Since $\mathcal{W}$ is T-invariant, every vector $\mathbf{w}$ is a linear combination of these basis vectors, meaning $T(\mathbf{w}) = b_0 T(\mathbf{v}) + ... + b_{k-1} T^{k-1}(\mathbf{v})$. Since $T^k(\mathbf{v}) \in \mathcal{W}$, there exist scalars a_i (coefficients of the basis vectors) such that

$$T^k(\mathbf{v}) + a_0 \mathbf{I}_n(\mathbf{v}) + a_1 T(\mathbf{v}) + ... + a_{k-1} T^{k-1}(\mathbf{v}) = \mathbf{0}.$$

Since $T_{\mathcal{W}}(\beta_j) = \sum_i a_{ij} \beta_i$, the matrix form of $T_{\mathcal{W}}$ is then

$$[T_{\mathcal{W}}]_\beta = \begin{pmatrix} 0 & \dots & 0 & -a_0 \\ 1 & \dots & 0 & -a_1 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 1 & -a_{k-1} \end{pmatrix}$$

and the characteristic polynomial of $T_{\mathcal{W}}$ is

$$f(\lambda) = (-1)^k (\lambda^k + a_0 + a_1 \lambda + ... a_{k-1} \lambda^{k-1}).$$

The Cayley-Hamilton Theorem

For T, a linear operator on a finite-dimensional vector space $\mathcal{V}$, let $f(t)$ be the characteristic polynomial of T. Then $f(T) = T_0$, i.e. T satisfies its own characteristic equation.

Proof sketch: Since we know $T^k(\mathbf{v}) + a_0 \mathbf{I}_n(\mathbf{v}) + ... + a_{k-1} T^{k-1}(\mathbf{v}) = \mathbf{0}$, and the characteristic polynomial of $T_{\mathcal{W}}$ is $g(\lambda) = (-1)^k (\lambda^k + a_0 + ... + a_{k-1} \lambda^{k-1})$, substituting reveals that $g(T)(\mathbf{v}) = 0$. Since the characteristic polynomial of a T-invariant subspace neatly divides the characteristic polynomial of the original space, $\mathbf{0}$ must divide $f(T)(\mathbf{v})$ therefore $f(T) = T_0$.

Practically, the Cayley-Hamilton theorem is a powerful tool in control theory (in a sequence of infinite derivatives of the observability matrix, it lets us know after which element the remaining derivatives are no longer linearly independent). It can also be used to find the inverse of a matrix by multiplying the characteristic equation by T^{-1}, rearranging terms, and dividing by a_0.

※ Inner Product Spaces

This is arguably the single most important (and most practically applicable) portion of linear algebra, and as such it is the largest section of these notes. Many mathematical objects and sets of objects define some structure that lets us know how different two objects are. In our case, we can think of this as the distance between two vectors in a vector space. For a general vector space, we call this an **inner product**; we restrict our

attention to vector spaces over $\mathbb{R}^n$ and $\mathbb{C}^n$. An inner product on $\mathcal{V}$ is a function $\langle \mathbf{x}, \mathbf{y} \rangle$ that assigns to each pair $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ a scalar in $\mathcal{F}$ that is linear in $\mathbf{x}$, *conjugate* linear in $\mathbf{y}$, and positive semi-definite ($\langle \mathbf{x}, \mathbf{y} \rangle \geq 0$). Conjugate linearity means that $\langle \mathbf{x}, c\mathbf{y} \rangle = \overline{c}\langle \mathbf{x}, \mathbf{y} \rangle$, whereas regular linearity would mean $\langle c\mathbf{x}, \mathbf{y} \rangle = c\langle \mathbf{x}, \mathbf{y} \rangle$.

The **standard inner product** on $\mathcal{F}^n$ is defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} a_i \overline{b_i}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$, the **Frobenius inner product** is $\text{trace}(\mathbf{B}^*\mathbf{A}) = \sum\sum \overline{b}_{ij} a_{ij}$. Unless otherwise specified, these are the implicit inner products we use for $\mathcal{F}^n$ and $\mathcal{M}_{m\times n}(\mathcal{F})$. A vector space that carries with it the notion of an inner product is called an **inner product space**.

For an inner product space, a **norm** or **length** of a vector $\mathbf{x}$ is denoted $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. The **Euclidean norm** is the square root of the standard inner product on a vector space. Norms carry with them the following essential properties:

1. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$
2. $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$
3. **Cauchy-Schwarz Inequality:** $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$
4. **Triangle Inequality:** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

In the Cauchy-Schwarz inequality, $\langle \mathbf{x}, \mathbf{y} \rangle$ is less than $\|\mathbf{x}\| \cdot \|\mathbf{y}\|$ by a factor of $\cos(\theta)$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$ (so $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$). If $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ and neither vector is $\mathbf{0}$ then $\theta = \pm\pi/2$, meaning the vectors are **orthogonal**. A subset $\mathcal{S} \subseteq \mathcal{V}$ is orthogonal if any two vectors in $\mathcal{S}$ are orthogonal. If all those vectors are also **unit** vectors (have a norm of 1) we refer to $\mathcal{S}$ as **orthonormal**. We can turn any vector into a unit vector by dividing it by its norm. The majority of the remainder of this section is devoted to finding bases of orthonormal vectors.

7.1 Orthonormal Bases

An orthonormal basis is an ordered basis of unit vectors where every pair of vectors is orthogonal. These bases are especially useful because they make it easy to find the coefficients a_i of $\mathbf{y} = \sum a_i \mathbf{v}_i$. To see why this is true, assume that $\mathcal{S} = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is an orthogonal spanning subset of $\mathcal{V}$. For $\mathbf{y} \in \mathcal{V}$, we can express $\mathbf{y}$ as a weighted sum of these vectors: $\mathbf{y} = \sum a_i \mathbf{v}_i$. Then $\langle \mathbf{y}, \mathbf{v}_j \rangle = \sum a_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle = a_j \|\mathbf{v}_j\|^2$ meaning $a_j = \frac{\langle \mathbf{y}, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2}$, allowing us to determine a_j from just $\mathbf{y}$ and $\mathbf{v}_j$. If each $\mathbf{v}_j$ is also a unit vector, the term in the denominator also disappears, so $a_j = \langle \mathbf{y}, \mathbf{v}_j \rangle$.

The elements of $\mathcal{S}$ must be linearly independent by this definition, as $\mathbf{y}$ can only be zero when each a_j is zero. Therefore such a set must be an orthogonal basis. In this way, the orthogonal basis ensures that the coordinate representation of $\mathbf{y}$ preserves the linear properties of $\mathbf{y}$. Dividing by the norm of each $\mathbf{v}_j$ to retrieve an orthonormal basis is exceptionally powerful. While a coordinate representation in *any* basis will preserve

the linear properties of a vector ($c\mathbf{x} + d\mathbf{y} = [ca_1 + db_1, ca_2 + db_2, ...]$), only the coordinate representation in an orthonormal basis is *metric* preserving. This means that not only do the linear properties hold, but so do the metric (distance) ones – $\|\mathbf{x}\| = \sqrt{a_1^2 + a_2^2 + ...}$; $\langle \mathbf{x}, \mathbf{y} \rangle = a_1 b_1 + a_2 b_2 + ...$ One can easily verify that these properties do not hold in general for non-orthonormal bases.

Gram-Schmidt Orthogonalization

We can transform any linearly independent subset $\mathcal{S} = \{\mathbf{w}_1, ..., \mathbf{w}_n\}$ of $\mathcal{V}$ into an orthogonal subset $\mathcal{S}' = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ of $\mathcal{V}$ with:

$$\mathbf{v}_k = \mathbf{w}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{w}_k, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \mathbf{v}_j$$

where $\mathbf{span}\,\mathcal{S} = \mathbf{span}\,\mathcal{S}'$. Then if $\mathbf{span}\,\mathcal{S} = n$ then $\mathcal{S}'$ must be a basis. We can divide each $\mathbf{v}_k$ by $\|\mathbf{v}_k\|$ to get an orthonormal basis. Think of this as creating the next orthogonal vector by subtracting from $\mathbf{w}_k$ the projection of $\mathbf{w}_k$ onto each of the already-generated $k-1$ orthogonal vectors. If the result contains *no* components from the existing orthogonal vectors, it must be orthogonal to all of them.

Aside: Applying the Gram-Schmidt process to $P(\mathbb{R})$ yields an orthogonal basis v_k – the polynomial $(1/v_k(1))v_k$ is known as the k**th Legendre polynomial** – a system of orthogonal polynomials with numerous applications across control theory, physics, and computer science.

The set of all vectors in $\mathcal{V}$ that are orthogonal to a set $\mathcal{S}$ is called the **orthogonal complement** of $\mathcal{S}$, denoted $\mathcal{S}^\perp$. For any vector $\mathbf{v} \in \mathcal{V}$, there exist two unique vectors, $\mathbf{u} \in \mathcal{S}$ and $\mathbf{z} \in \mathcal{S}^\perp$, such that $\mathbf{v} = \mathbf{u} + \mathbf{z}$. This means that each $\mathbf{y} \in \mathcal{V}$ is uniquely characterized by a vector in $\mathcal{S}$, and a vector orthogonal to $\mathcal{S}$. The vector $\mathbf{u}$ in $\mathcal{S}$ is of particular interest – this vector is the "closest" vector to $\mathbf{y}$ that lives in $\mathcal{S}$. It is the **orthogonal projection** of $\mathbf{y}$ onto $\mathcal{S}$, and is the core of several problems in optimization theory.

Perhaps unsurprisingly, if $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_k\}$ is an orthonormal basis of $\mathcal{S}$, then it can be extended to an orthonormal basis for $\mathcal{V}$ by appending the vectors $\beta' = \{\mathbf{v}_{k+1}, ..., \mathbf{v}_n\}$. β' then (by definition) is an orthonormal basis for $\mathcal{S}^\perp$, meaning $\dim \mathcal{V} = \dim \mathcal{S} + \dim \mathcal{S}^\perp$. This can be thought of geometrically from a calculus standpoint, where we define planes by the vectors normal (orthogonal) to them – in $\mathbb{R}^3$, if $\mathcal{S}$ is a line, then its orthogonal complement is the plane defined by that line. If $\mathcal{S}$ is a plane, then its orthogonal complement is the line that defines that plane.

7.2 Adjoint Operator

Here we briefly discuss a generalization of the conjugate transpose, the **adjoint operator**, to arbitrary spaces (not to be confused with the *adjugate matrix*, or the transpose of the cofactor matrix). This general definition is especially useful in certain realms of theoretical physics.

We can represent any linear functional $g : \mathcal{V} \to \mathcal{F}$ as an inner product $g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$ where $\mathbf{y} = \sum \overline{g(\mathbf{v}_i)} \mathbf{v}_i$. Then for a linear operator $T(\mathbf{x})$, we can write $g(\mathbf{x}) = \langle T(\mathbf{x}), \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y}' \rangle$. The function $T^*(\mathbf{y}) = \mathbf{y}'$ is the **adjoint operator** for T. For $\mathbb{C}^n$, the adjoint operator is exactly the conjugate transpose.

7.3 Normal and Self-Adjoint Operators

We stated earlier that a linear operator is diagonalizable if $\mathcal{V}$ admits a basis consisting of eigenvectors of T; this let to methods to tell whether a matrix is diagonalizable (by looking at the nullity of each eigenspace $\mathcal{E}_\lambda$).

On an inner product space, we are curious if we can find an *orthonormal* basis of eigenvectors, which further simplifies computations using (potentially complicated) matrices. Our goal is to find an easily verifiable condition that tells us whether the desired basis exists. Note that this does *not* imply that an eigendecomposition (diagonalization) exists – not all upper triangular matrices are diagonalizable.

Schur Decomposition

If T is a linear operator on $\mathcal{V}$, and the characteristic polynomial of T splits over $\mathcal{F}$, then there exists an orthonormal basis γ for $\mathcal{V}$ such that $[T]_\gamma$ is upper triangular, yielding the **Schur decomposition**

$$\mathbf{A} = \mathbf{Q}\mathbf{U}\mathbf{Q}^{-1}$$

where $\mathbf{U}$ is the upper triangular matrix in question, and the columns of $\mathbf{Q}$ are the basis vectors in question. There are two principles at work here – the first is that any matrix with a splitting characteristic polynomial is similar to an upper triangular matrix (proven via induction on n), and then applying Gram-Schmidt.

If an orthonormal basis of eigenvectors β exists, then $[T]_\beta$ is a diagonal matrix; so $[T]^*_\beta$ is diagonal as well; since diagonal matrices commute, T and T^* commute (meaning $TT^* = T^*T$). Linear operators with this property are called **normal**.

In general, it is not enough to say that an operator over a real inner product space is normal for it to be diagonalizable – this is because the characteristic polynomial may not split over the reals. However, due to the fundamental theorem of algebra, every polynomial splits over $\mathbb{C}$. Schur's theorem then allows us to find an orthonormal

basis; normality allows us to prove that all vectors in this basis are eigenvectors. A brief sketch of a proof by induction: assume $\mathbf{v}_1, ..., \mathbf{v}_{k-1}$ are eigenvectors. Then if λ_j corresponds to $\mathbf{v}_{j<k}$, we have (by normality) that $T^*(\mathbf{v}_j) = \overline{\lambda}_j \mathbf{v}_j$. Then for $j \neq k$, $A_{jk} = \langle T(\mathbf{v}_k), \mathbf{v}_j \rangle = \langle \mathbf{v}_k, T^*(\mathbf{v}_j) \rangle = \lambda_j \langle \mathbf{v}_k, \mathbf{v}_j \rangle = 0$, so $A_{kk} = \lambda_k$ and $A_{jk} = 0$; so $\mathbf{v}_k$ is an eigenvector of T.

So normality is a necessary and sufficient condition for the existence of an orthonormal basis of eigenvectors (and is therefore a sufficient condition for diagonalizability) for linear operators over a complex inner product space.

For real inner product spaces, we must add the condition that $T = T^*$ (this causes all eigenvalues to be real, and hence makes the characteristic polynomial split over the reals, since $\lambda \mathbf{v} = T(\mathbf{v}) = T^*(\mathbf{v}) = \overline{\lambda} \mathbf{v}$ means $\lambda = \overline{\lambda}$). A transformation with this property is equal to its own adjoint – it is **self-adjoint** (also called **Hermitian**). Then the same logic from above follows – over real inner product spaces, an orthonormal basis of eigenvectors exists if and only if T is Hermitian. For real matrices, being Hermitian is equivalent to being symmetric; the conclusion is that every symmetric matrix over a real, finite-dimensional vector space admits an orthogonal basis composed entirely of eigenvectors.

Finally, some common types of self-adjoint matrices: if $\langle T(\mathbf{x}), \mathbf{x} \rangle > 0$ (equivalently, $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$, or $\lambda > 0$ for all λ), we call T **positive definite**. If we relax the strict inequality to a non-strict inequality, we call it **positive semi-definite**. **Negative-definiteness** and **negative-semidefiniteness** are defined similarly.

7.4 Unitary and Orthogonal Operators

We have previously discussed “structure-preserving" transformations such as linear operators, which preserve the properties of vector addition and scalar multiplication, and isomorphisms, which preserve all structures of a vector space. With our introduction if inner products, which can be taken to represent an idea of “length," we may also wonder about operations which are length-preserving; i.e. transformations T such that $\langle T(\mathbf{x}), T(\mathbf{x}) \rangle = \langle \mathbf{x}, \mathbf{x} \rangle$ (implying $T^*T = TT^* = I$).

Such a T is called **unitary** if $\mathcal{F}$ is $\mathbb{C}$, or **orthogonal** if $\mathcal{F}$ is $\mathbb{R}$. By definition, unitary/orthogonal operators are normal. We say that such operators **preserve inner product**.

Since $T(\mathbf{x}) = \lambda \mathbf{x}$ for appropriate eigenvalues/eigenvectors λ and $\mathbf{x}$, and since $\langle T(\mathbf{x}), T(\mathbf{x}) \rangle = \lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle$, we must have $|\lambda| = 1$. In fact $\mathcal{V}$ has an orthonormal basis of eigenvectors with all $|\lambda| = 1$ if and only if there exists some T that is unitary – if $\mathcal{F} = \mathbb{R}$ we require the stronger condition that T is also self-adjoint (such an operator, where $T = T^{-1}$, is called **involutory**).

A complex normal/real symmetric matrix $\mathbf{A}$ admits an orthonormal basis consisting of eigenvectors. Therefore, for the corresponding diagonal matrix $\mathbf{D} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$, each $\mathbf{Q}$ must be a unitary/orthogonal matrix – we say then that $\mathbf{A}$ is **unitarily/orthogonally equivalent** to $\mathbf{D}$ (the resulting decomposition, $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^*$, is known as the **spectral decomposition**).

7.5 Spectral Theorem

Recall that, for a subspace described by a direct sum $\mathcal{W} = \mathcal{W}_1 \oplus \mathcal{W}_2$, the linear transformation $T(\mathbf{w}) = \mathbf{w}_1$ is the projection of $\mathbf{w}$ onto $\mathcal{W}_1$. While there are multiple ways to perform a projection onto $\mathbf{W}_1$, the one we are most interested in is the **orthogonal projection** – the projection in which we map a vector onto the *closest* (defined in terms of inner product, which, again, represents "distance") vector in $\mathcal{W}$. A projection T is an orthogonal projection if $\text{im}(T)^\perp = \ker(T)$ ($\ker(T)^\perp = \text{im}(T)$). By the nature of being a projection, such a T must have $T^2 = T$ (it is **idempotent**). Additionally, since $\ker(T) = \text{im}(T)^\perp$, we may determine that $T = T^*$ (T is normal/self-adjoint).

Spectral Theorem

Let T is a linear operator on a finite-dimensional inner product space $\mathcal{V}$ over $\mathcal{F}$ with k distinct eigenvalues, and T is either normal if $\mathcal{F} = \mathbb{C}$ or self-adjoint if $\mathcal{F} = \mathbb{R}$. Then suppose $\mathcal{W}_i$ is the eigenspace corresponding to λ_i (recall $\mathcal{E}_\lambda = \ker(\mathbf{A} - \lambda\mathbf{I})$). Additionally, let T_i be the orthogonal projection of $\mathcal{V}$ onto $\mathcal{W}_i$. Then:

1. $\mathcal{V} = \bigoplus \mathcal{W}_i$;
2. $\mathcal{W}^\perp = \bigoplus_{j \neq i} \mathcal{W}_i$;
3. $T_i T_j = \delta_{ij} T_i$;
4. $I = \sum T_i$;
5. $T = \sum \lambda_i T_i$.

We call the eigenvalues the **spectrum** of T, and equality (4) is known as the **resolution of the identity operator**. The final statement (5) is more broadly referred to as the **spectral decomposition**.

Broadly, we conclude that if T is a normal/self-adjoint operator, then its eigenvectors form an orthonormal basis for $\mathcal{V}$, and T can be described as the weighted (by the eigenvalues) sum of projections onto those eigenvectors. An even more straightforward consequence: if $\mathbf{A}$ is a real symmetric matrix, then it is orthogonally diagonalizable ($\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$).

7.6 Singular Value Decomposition

We previously established the relationship between the existence of an orthonormal basis of eigenvectors and the property of being normal or self-adjoint. Here, we propose

a more general theorem that extends to all linear transformations on complex and real (finite-dimensional) inner product spaces.

Singular Value Theorem

For a rank r linear transformation $T : \mathcal{V} \to \mathcal{W}$, where $\dim(\mathcal{V}) = n$ and $\dim(\mathcal{W}) = m$, there exist orthonormal bases $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ for $\mathcal{V}$ and $\{\mathbf{u}_1, ..., \mathbf{u}_m\}$ for $\mathcal{U}$ and a set of positive scalars $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$ such that $T(\mathbf{v}_i) = \sigma_i \mathbf{u}_i$, where $\sigma_i = 0$ if $i > r$. The σ_i in question are the **singular values** of T. In fact, σ_i^2 is the eigenvalue of T corresponding to eigenvector $\mathbf{v}_i$, meaning that the singular values are uniquely determined by T (though the vectors are not).

The singular value theorem states that, by choosing the appropriate bases for $\mathcal{V}$ and $\mathcal{W}$, *any* linear transformation can be expressed as a diagonal matrix (stretching operation). What's more, these bases are guaranteed to be orthonormal. Intuitively, a sphere in $\mathcal{V}$ will get stretched (by the singular values) and rotated (by the change of basis) into an ellipsoid in $\mathcal{W}$, where the dimensions of the two objects before and after need not be the same. We can think of $\mathbf{V}^*$ as rotating the original ellipsoid so that it aligns with the standard basis vectors; of Σ as performing a stretch in that alignment; and $\mathbf{U}$ as finally re-rotating the newly stretched ellipsoid into its final orientation.

For an $m \times n$ matrix $\mathbf{A}$ of rank r, let $\Sigma \in \mathcal{M}_{m \times n}(\mathcal{F})$ such that $\Sigma_{ii} = \sigma_i$ for $i < r$ and 0 otherwise. Then there exist unitary matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{AV} = \Sigma\mathbf{U}$; or more commonly, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$. This is known as the **singular value decomposition (SVD)**.

The singular value decomposition helps us partially extend the concept of an inverse to nonsquare, noninvertible matrices. We achieve this by inverting the "part" of the transformation that *is* invertible. Namely, we restrict such a transformation T exclusively to $\ker(T)^\perp$, to yield an invertible transformation $L : \ker(T)^\perp \mapsto \text{im}(T)$. The matrix $T^\dagger = L^{-1}(y)$ for $y \in \text{im}(T)$ (and 0 otherwise) is known as the **Moore-Penrose pseudoinverse** of T. If $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$, then $\mathbf{A}^\dagger = \mathbf{V}\Sigma^\dagger\mathbf{U}^*$, where $\Sigma_{ii}^\dagger = 1/\sigma_i$.

※ Canonical Forms

Diagonalizable operators are immensely useful, but not all matrices are diagonalizable. Here, we introduce analogues, termed **canonical forms**, which introduce a similar representation that makes it easier to reason about some properties of matrices – this practice is especially useful in the study of dynamical systems. In particular, we are concerned with the **Jordan canonical form**, which only requires that the characteristic polynomial splits, which it does over any algebraically closed field.

8.1 Jordan Form

Succinctly, we can find a union of ordered bases β such that

$$[T]_\beta = \begin{pmatrix} \mathbf{A}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{A}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_k \end{pmatrix}$$

where each $\mathbf{A}_i$ has form

$$\mathbf{A}_i = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix}$$

Such an $\mathbf{A}_i$ is known as a **Jordan block** and β is the **Jordan canonical basis**.

Take the Jordan block $\mathbf{A}_i$ above. For basis vectors $\mathbf{v}_1, ..., \mathbf{v}_k$, $\mathbf{v}_1$ must be an eigenvector. Then $T(\mathbf{v}_2) = \mathbf{v}_1 + \lambda\mathbf{v}_2$, meaning $(T - \lambda I)\mathbf{v}_2 = \mathbf{v}_1$, and $(T - \lambda I)\mathbf{v}_3 = \mathbf{v}_2$ and so on. Since $\mathbf{v}_1 \in \ker(\mathbf{T} - \lambda I)$, we necessarily have for all appropriate $\mathbf{v}_i$ that $(T - \lambda I)^p \mathbf{v}_i = 0$, where p is the size of the corresponding Jordan block. Such a vector $\mathbf{x}$ such that $(T - \lambda I)^p(\mathbf{x}) = 0$ for some positive integer p is called a **generalized eigenvector** corresponding to λ. Analogously, the subspace $\mathcal{K}_\lambda = \{\mathbf{x} \in \ker((T - \lambda_I)^p) : p \in \mathbb{Z}^+\}$ is the **generalized eigenspace** corresponding to λ.

Each $\mathcal{K}_\lambda$ is a T-invariant subspace that contains $\mathcal{E}_\lambda$, and each $\mathcal{K}_{\lambda_1}$ is mutually exclusive with each $\mathcal{K}_{\lambda_2}$. If T splits, $\mathcal{K}_\lambda$ is exactly equal to $\ker((T - \lambda)^m)$ where m is the multiplicity of λ (this is the consequence of the Cayley-Hamilton theorem, though we omit the proof). The key consequence: any vector $\mathbf{x} \in \mathcal{V}$ can be expressed as a sum of vectors in $\mathcal{K}_{\lambda_i}$. More generally, if T splits, and if β_i is an ordered basis for $\mathcal{K}_{\lambda_i}$, then $\beta_i \cap \beta_j = \emptyset$ and $\bigcup \beta_i$ is a basis for $\mathcal{V}$ (meaning $\dim(\mathcal{K}_{\lambda_i}) = m$). Notice that if $\mathcal{K}_\lambda = \mathcal{E}_\lambda$, this is equivalent to saying that T is diagonalizable.

The problem now becomes how to select bases for $\mathcal{K}_\lambda$ such that the resulting union is the Jordan canonical basis. In the example above, note that there is some *cyclical* relationship between the generalized eigenvectors corresponding to λ. For p, the smallest integer such that $(T - \lambda I)^p(\mathbf{x}) = 0$ (for any generalized eigenvector $\mathbf{x}$), define the set

$$\{(T - \lambda I)^{p-1}(\mathbf{x}), ..., (T - \lambda I)(\mathbf{x}), \mathbf{x}\},$$

the **cycle of generalized eigenvectors** (note the similarity to §6.4). Then if β is the *union* of the cycles of generalized eigenvectors of T, then for each cycle γ in β, $\mathcal{W} = \mathbf{span}(\gamma)$ is T-invariant and $[T_\mathcal{W}]_\gamma$ is a Jordan block; the basis β is a Jordan canonical basis for $\mathcal{V}$.

8.2 The Minimal Polynomial

The Cayley-Hamilton theorem tells us that for every T there exists an f such that $f(T) = T_0$, and we saw that the characteristic polynomial fulfills this property. In fact, each linear operator admits a *unique* polynomial p with smallest degree, known as the **minimal polynomial**, which has some applications in field theory (so it is only mentioned briefly here).

This $p(t)$ divides every other polynomial f, since $f(t) = q(t)p(t) + r(t) \implies T_0 = f(T) = q(T)p(T) + r(T) = q(T)T_0 + T_0$ (since $r(t)$ must have lower degree than p, but since p is minimal, $r(t)$ can only equal T_0). From this property, and from the fact that the characteristic polynomial satisfies the Cayley-Hamilton theorem, we may deduce that λ is an eigenvalue of T if and only if $p(\lambda) = 0$, so p is of form $\prod_i (t - \lambda_i)^{m_i}$ for some powers m_i. Furthermore T is diagonalizable if and only if $p = \prod(t - \lambda_i)$. There is a relationship between the minimal polynomial and the Jordan form – if p_i is the number of rows in the largest Jordan block corresponding to eigenvalue λ_i, the minimal polynomial of T is $\prod_i (t - \lambda)_i^{p_i}$.