Linear Algebra

※ Vector Spaces

The fundamental structure in linear algebra is a **vector**, a collection of objects (**components**) drawn from a field $\mathcal{F}$ that is typically represented in **row form** or **column form**:

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

A set whose elements are vectors and is closed under (element-wise) addition and multiplication by a scalar is a **vector space** $\mathcal{V}$. From these definitions come other properties, such as commutativity and associativity of vector addition, and the existence of a unique zero vector.

A **matrix** $\mathcal{M}_{m\times n}(\mathcal{F})$ or $\mathbf{A}$ is a two-dimensional $m \times n$ array whose elements are also drawn from the field $\mathcal{F}$, where each element is a_{ij}, where i is the index of the row and j the column.

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

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} forms the **diagonal** of the matrix; a matrix which only has nonzero entries on the diagonal is a **diagonal matrix**, and a matrix that does not have nonzero entries *below* the diagonal is **upper triangular**. Any matrix with very few nonzero entries is **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($\mathbf{A}$).

1.1 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 $\mathcal{V}$. The intersection of subspaces of a vector space is also a subspace.

1.2 Linear Dependence

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 $\mathbf{span}(S)$. 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 $\mathbf{span}(\mathcal{S}) = \mathcal{V}$ then we say that $\mathcal{S}$ **generates** $\mathcal{V}$.

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**. By this definition, removing a vector that is the linear combination of the other vectors will not change the span of the set. A linearly independent set $\mathcal{S}$ can become linearly *dependent* by adding any vector $\mathbf{v} \in \mathbf{span}(\mathcal{S})$. No superset of a linearly dependent set can be linearly *in*dependent.

1.3 Basis, Dimension

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 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.

This “size" of the bases is known as the **dimension** of the vector space, denoted $\mathbf{dim}(\mathcal{V})$. The size of the basis is the size of the largest linearly independent subset. Conversely, any linearly independent set of size $\mathbf{dim}(\mathcal{V})$ must be a basis, as must any spanning set of size $\mathbf{dim}(\mathcal{V})$.

※ Linear Transformations, Matrices

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 $\mathbf{x}, \mathbf{y} \in \mathcal{V}$, $T(c\mathbf{x} + \mathbf{y}) = cT(\mathbf{x}) + T(\mathbf{y})$. 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. In a 2-dimensional world, this is like the difference between lines going through (0, 0) and lines with a y-intercept. Examples of such linear transformations include integration, differentiation, rotation, reflection, and projection.

Some special linear transformations: 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 $proj_{\mathcal{W}_1}(\mathbf{x}) = \mathbf{x}_1$.

2.1 Rank, Nullity

Two sets which help understand the intrinsic properties of T are:

1. The **null space** or **kernel** of T, denoted $\mathcal{N}(T)$ or $\ker(T)$, the set of vectors which T sends to zero $\{\mathbf{x} : T(\mathbf{x}) = 0\}$
2. The **range** or **image** of T, denoted $\mathcal{R}(T)$ or $\operatorname{im}(T)$, he set of vectors in $\mathcal{W}$ that result from applying T to vectors in $\mathcal{V}$, $\{\mathbf{w} \in \mathcal{W} : (\exists \mathbf{v} \in \mathcal{V} : T(\mathbf{v}) = \mathbf{w})\}$.

The null space and range are subspaces of $\mathcal{V}$ and $\mathcal{W}$, respectively. We additionally name the dimension of the kernel the **nullity** of T, and the dimension of the image the **rank**.

Rank-Nullity Theorem

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

2.2 Matrices

An **ordered basis** for a vector space $\mathcal{V}$ is a sequencing of basis vectors. Any vector $\mathbf{x} \in \mathcal{V}$ can be characterized as a linear combination of vectors in an ordered basis $\sum_i a_i \mathbf{v}_i$. The coefficients a_i again form a size-n vector, known as the **coordinate vector** $[\mathbf{x}]_\beta$.

For a linear transformation $T : \mathcal{V} \rightarrow \mathcal{W}$, suppose that $\mathcal{V}$ has a basis β and $\mathcal{W}$ has a basis γ. Then just as there is a coordinate representation for $\mathbf{x}$ according to β, there must be such a representation for T under β, γ. This representation is the coefficient matrix $\mathbf{A}$ or $[T]_{\beta\rightarrow\gamma}$ (so $T(\mathbf{v}_j) = \sum_i a_{ij}(\mathbf{w}_i)$).

Matrices preserve linearity, and are hence linear themselves, i.e. $[aT+U]_{\beta\rightarrow\gamma} = a[T]_{\beta\rightarrow\gamma} + [U]_{\beta\rightarrow\gamma}$. Therefore the set of all transformations from $\mathcal{V} \rightarrow \mathcal{W}$ is itself a vector space $\mathcal{L}(\mathcal{V}, \mathcal{W})$.

For $T : \mathcal{V} \rightarrow \mathcal{W}$ and $U : \mathcal{W} \rightarrow \mathcal{Z}$, the composition $U \circ T$ is written as $UT : \mathcal{V} \rightarrow \mathcal{Z}$ and is linear. If $T : \mathcal{V} \rightarrow \mathcal{V}$, we can compose T with itself. $T^2 = TT$, $T^3 = T^2T$, $T^k = T^{k-1}T$. If $\mathbf{A} = [U]_{\beta\rightarrow\gamma}$ and $\mathbf{B} = [T]_{\alpha\rightarrow\beta}$, then $\mathbf{AB} = [UT]_{\alpha\rightarrow\gamma}$. 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 $\mathbf{AB} \in \mathcal{F}^{m\times n}$.

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

This is **matrix multiplication**, and it is not commutative in general, but it is both associative and distributive. A matrix which, when multiplied by itself k times is zero (i.e. $\mathbf{M}$ s.t. $\mathbf{M}^k = \mathbf{0}$) is **nilpotent**.

2.2.1 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, written as T^{-1}, if $TU = I_{\mathcal{W}}$ and $UT = I_{\mathcal{V}}$. For any $\mathbf{v} \in \mathcal{V}$, $(T^{-1}T)(\mathbf{v}) = \mathbf{v}$ – and for any $\mathbf{w} \in \mathcal{W}$, $(TT^{-1})(\mathbf{w}) = \mathbf{w}$. For this to be true, T must be bijective, i.e. $\dim(\mathcal{V}) = \dim(\mathcal{W})$ (rank-nullity theorem). 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 $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$.

If a matrix's inverse is also its transpose, i.e. $\mathbf{AA}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$, then it is **orthogonal**. For a complex-valued matrix $\mathbf{A}$, the **conjugate transpose** $\mathbf{A}^*$ is $\mathbf{A}^\top$ where all elements are replaced by their complex conjugates. If $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}_n$, then we call $\mathbf{A}$ **unitary**. For $\mathbb{C}$, the conjugate transpose is the **Hermitian adjoint**.

Invertibility formalizes the notion that certain vectors "resmeble" another, like how adding two matrices in $\mathcal{M}_{2\times2}(\mathcal{F})$ looks very similar to adding two polynomials in $P_4(\mathcal{F})$. The "structure preserving" operation that relates the two spaces is an **isomorphism**, meaning there is an invertible linear transformation T between the two spaces. Two vector spaces are isomorphic if and only if they are the same dimension.

Any transformation can be described by an equivalent matrix multiplication. The **change-of-coordinate** matrix $\mathbf{Q}$, such that $\beta'_i = \sum q_{ij}\beta_j$, is an invertible matrix that lets us change a vector's basis representation.

Linear transformations that stay within $\mathcal{V}$ are known as **linear operators**, $T : \mathcal{V} \mapsto \mathcal{V}$. T in basis β can be represented in basis β' by applying consecutive change of basis transformations:

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

If there exists an $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{AQ}$, we say that $\mathbf{A}$ and $\mathbf{B}$ are **similar**.

2.3 Dual Spaces

A transformation that sends $\mathcal{V}$ to $\mathcal{F}$ is known as a **linear functional** on $\mathcal{V}$. $\mathcal{V}$ and $\mathcal{V}^*$ are isomorphic. The transformation $U : \mathcal{W}^* \to \mathcal{V}^*$ such that $(U(g))(\mathbf{v}) = (gT)(\mathbf{v})$, where $g \in \mathcal{W}^*$, is precisely the **transpose** of T. The transpose is the **algebraic adjoint** of T. The dual of the dual ($\mathcal{V}^*{}_*$) is isomorphic to $\mathcal{V}$.

※ Systems of Linear Equations

Linear algebra can be used to find solutions to sets of equations of the form $\sum_i a_ix_i - b = 0$. This is done by applying **elementary row operations** to manipulate them into forms which yield solutions. s

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 **I**, we call the result an **elementary matrix**. Consecutive row operations is equivalent to left-multiplying the corresponding elementary matrix.

3.1 Rank, Inverse

A square matrix is invertible precisely if **rankA** $= n$ (**full-rank**). Rank is preserved through multiplication with invertible matrices. Since elementary operations are invertible, they are therefore also rank-preserving. The rank of a matrix is the same as the nuber of its linearly independent rows/columns.

Any matrix can be transformed (using elementary row and column operations) into the form

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

which reveals the rank of the matrix immediately. Given two matrices **A** and **B**, we define the **augmented matrix** $\mathbf{C} = (\mathbf{A}|\mathbf{B})$ as the column-wise concatenation of **A** and **B**. If we augment **A** with the identity matrix, and perform row operations to transform **A** into **I**, we will retrieve the inverse of **A**:

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

3.2 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:

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

If $\mathbf{b} = 0$, the solution set is the null space of **A**.

We can attempt to transform the matrix into **reduced-row echelon form** in order to find the solution vector(s).

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 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** then performs operations to satisfy conditions (1) and (2) above the diagonal. This method is known as **Gaussian elimination**.

If a row in reduced-row echelon form 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.

※ Determinants

For a matrix $\mathbf{A} \in \mathcal{M}_{2\times 2}(\mathcal{F})$, the determinant is:

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

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.

The scalar $(-1)^{i+j}\det(\tilde{\mathbf{A}}_{ij})$ is called a **cofactor**. The matrix formed by all cofactors is known as the **cofactor matrix $\mathbf{C}$**; its transpose $\operatorname{adj}(\mathbf{A}) = \mathbf{C}^\top$ is the **classical adjoint** or **adjugate matrix** to $\mathbf{A}$, with the property that $\mathbf{A}\operatorname{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{I}_n$.

The determinant is impacted by performing elementary row operations on the original matrix.

1. Swapping two rows of $\mathbf{A}$ will negate $\det(\mathbf{A})$
2. Multiplying a row by a scalar results in $k\det(\mathbf{A})$
3. Adding a scalar multiple of one row to another does not hcange $\det \mathbf{A}$
4. The determinant of an upper-triangular matrix is the product of its diagonal
5. $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.
6. A matrix is invertible if and only if $\det(\mathbf{A}) \neq 0$

※ Diagonalization

Linear transformations can geometrically look like a rotation, a projection, a stretch, a shear, etc. Matrices that are **diagonalizable** admit some basis such that when they are represented in that basis, they are diagonal (only "stretching").

5.1 Eigenbasis

If $\mathbf{D} = [T]_\beta$ is a diagonal matrix, then applying $\mathbf{D}$ to a basis vector $\mathbf{v}$ is then $T(v_j) = \sum_{i=1}^n d_{ij}\mathbf{v}_i = \lambda_j\mathbf{v}_j$ where $\lambda_j = d_{jj}$. Such a vector $\mathbf{v}$ where $T(\mathbf{v}) = \lambda\mathbf{v}$ for a scalar λ is called an **eigenvector**, and the corresponding λ is called an **eigenvalue** (equivalently, $\mathbf{Av} = \lambda\mathbf{v}$). These eigenvectors form a subspace $\mathcal{E}_\lambda$.

Since $\mathbf{Av} = \lambda\mathbf{v}$, $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v} = 0$. The nullspace of this matrix is therefore nontrivial, so it is not full rank, meaning its determinant is zero. Solving $\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0$ (the **characteristic polynomial**) will yield the eigenvalues of $\mathbf{A}$ (**diagonalization**).

The characteristic polynomial has degree n, so by the fundamental theorem of algebra it has n roots (that may lie outside the field $\mathcal{F}$). For example there is no manipulation of a rotation matrix that will yield a pure stretching operation (its roots are complex). When a matrix *is* diagonalizable, we can convert $\mathbf{Av} = \lambda\mathbf{v}$ into matrix form $\mathbf{AQ} = \mathbf{DQ}$. The form $\mathbf{A} = \mathbf{QDQ}^{-1}$ ($\mathbf{D} = \mathbf{Q}^{-1}\mathbf{AQ}$) is the **eigendecomposition** of $\mathbf{A}$.

For a matrix to be diagonalizable, its characteristic polynomial must **split** over $\mathcal{F}$, i.e. it must decompose into the product of n binomial terms $c\prod(\lambda - a_i)$. A transformation is diagonalizable if and only if every eigenspace has dimension equal to the multiplicity of its corresponding eigenvalue. So if a matrix has characteristic polynomial $-(\lambda-3)^2(\lambda-4)$, the matrix $(\mathbf{A} - 3\mathbf{I})$ must have nullity 2 (meaning the eigenspace corresponding to $\lambda = 3$ must have dimension 2). This is not always true – if λ has multiplicity m, its corresponding eigenspace can have any dimension up to m.

5.2 Markov Chains

Long-running systems, which can be modeled as a series of repeated, identical matrix multiplications, can be made inexpensive with diagonal matrices. So the limit of a sequence of matrices can be represented as:

$$\lim_{m\to\infty} \mathbf{A}^m = \lim_{m\to\infty} (\mathbf{QDQ}^{-1})^m = \lim_{m\to\infty} \mathbf{QDQ}^{-1}\mathbf{QDQ}^{-1}... = \lim_{m\to\infty} \mathbf{QD}^m\mathbf{Q}^{-1}$$

If $\mathbf{D}$ is the eigenvalue matrix, the limit can only exist if $|\lambda_{max}| \leq 1$ (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, it is a **transition matrix**, where entry (i, j) is the probability of a state change from state i to state j.

Markov Chains

A system where elements belong to states that switch probabilistically over time is a **stochastic process**. If the probability of of transitioning from $i \rightarrow j$ 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 $\mathbf{A}$ is called **regular** if, for some m, $\mathbf{A}^m$ contains no zero elements. A state that can be entered but never left (i.e. with probability zero) is an **absorbing state**.

5.3 Invariant Subspaces, The Cayley-Hamilton Theorem

If $T(\mathbf{v})$ stays in the subspace $\mathcal{W}$ for all $\mathbf{v}$, then $\mathcal{W}$ is T**-invariant**. The "smallest" T-invariant subspace for a vector space is **span**$(\{\mathbf{v}, T(\mathbf{v}), T^2(\mathbf{v}), ...\})$., the T**-cyclic subspace generated by v**. The characteristic polynomial of T_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 $T(\mathbf{w}) = b_0T(\mathbf{v}) + ... + b_{k-1}T^{k-1}(\mathbf{v})$.

Since $T^k(\mathbf{v}) \in \mathcal{W}$, there exist scalars a_i (coefficients of the basis vectors) such that

$$T^k(\mathbf{v}) + a_0\mathbf{I}_n(\mathbf{v}) + a_1T(\mathbf{v}) + \ldots + a_{k-1}T^{k-1}(\mathbf{v}) = \mathbf{0}.$$

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

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

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 $T^k(\mathbf{v}) + a_0\mathbf{I}_n(\mathbf{v}) + \ldots + a_{k-1}T^{k-1}(\mathbf{v}) = \mathbf{0}$, and the characteristic polynomial of $T_{\mathcal{W}}$ is $g(\lambda) = (-1)^k(\lambda^k + a_0 + \ldots + a_{k-1}\lambda^{k-1})$, 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

An **inner product** $\langle\cdot,\cdot\rangle : \mathcal{V} \times \mathcal{V} \mapsto \mathcal{F}$ is an object that lets us identify similarity between two vectors. It is linear in $\mathbf{x}$, conjugate linear in $\mathbf{y}$ ($\langle \mathbf{x}, c\mathbf{y}\rangle = \overline{c}\langle \mathbf{x}, \mathbf{y}\rangle$), and always non-negative. The **standard inner product** on $\mathcal{F}^n$ is $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^n a_i\overline{b_i}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$, the **Frobenius inner product** is trace($\mathbf{B}^*\mathbf{A}$) $= \sum\sum \overline{b}_{ij}a_{ij}$. These are the implicit inner products for these spaces.

The **norm** or **length** of a vector $\mathbf{x}$ is $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$. 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**. Any vector can be made into a unit vector by dividing it by its norm.

6.1 Orthonormal Bases

An orthonormal basis is an ordered basis of unit vectors where every pair of vectors is orthogonal. Orthogonal bases make it easier to invert matrices and the normalization helps remove scaling terms from matrix-vector equations. Orthonormal bases are in this way metric-preserving, i.e. they preserve the length and angle of vectors.

Gram-Schmidt Orthogonalization

Any linearly independent set of vectors $\mathcal{S} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ can be made an orthogonal subset by iteratively subtracting orthogonal components from each vector:

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

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 $\mathcal{S}^\perp$. Any vector in $\mathcal{V}$ can be uniquely characterized as the sum of 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 $\mathcal{S}$ (**orthogonal projection** of $\mathbf{y}$ onto $\mathcal{S}$).

6.2 Normal and Self-Adjoint Operators

A linear operator is diagonalizable if $\mathcal{V}$ admits a basis consisting of eigenvectors of T. As a consequence we can determine diagonalizability by looking at the nullity of each eigenspace $\mathcal{E}_\lambda$. Similarly, **normality** is a necessary and sufficient condition for whether an inner product space admits an orthonormal basis of eigenvectors.

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.

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**.

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.

6.3 Unitary and Orthogonal Operators

Transformations that are **length-preserving**, i.e. $\langle T(\mathbf{v}), T(\mathbf{v})\rangle = \langle \mathbf{v}, \mathbf{v}\rangle$ are **unitary** (over $\mathbb{C}$) or **orthogonal** (over $\mathbb{R}$).

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**).

6.4 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. $T_iT_j = \delta_{ij}T_i$;
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 ($\mathbf{A} = \mathbf{PDP}^\top$).

6.5 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 $\mathbf{dim}(\mathcal{V}) = n$ 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 $L : \ker(T)^\perp \mapsto \operatorname{im}(T)$. The matrix $T^\dagger = L^{-1}(y)$ for $y \in \operatorname{im}(T)$ (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 $\Sigma^\dagger_{ii} = 1/\sigma_i$.

※ 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.

7.1 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

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

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 $\mathcal{V}$.

7.2 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 $\prod_i (t - \lambda)_i^{p_i}$.