Chapter 3

Linear Algebra

3.1 Vector Space

3.1.1 Field

Definition 80. *For 0 and 1 of a field F, the smallest n that $\sum_{i=1}^{n} 1 = 0$ is called the characteristic of F. If no such n exists, F is called characteristic zero.* □

Definition 81. *The field Z_2 has characteristic of 2 which consists of two elements 0 and 1:*

- $0 + 0 = 0$
- $0 + 1 = 1 + 0 = 1$
- $1 + 1 = 0$
- $0 \times 0 = 0$
- $0 \times 1 = 1 \times 0 = 0$
- $1 \times 1 = 1$

3.1.2 Vector

Algebra is concerned with how to manipulate symbolic combinations of object and how to equate one with another.

Definition 82. *A vector space vector space V over a field field F has two operation $\{+, \times\}$ with $\vec{0}$ and 1.* □

Definition 83. *A subspace is a subset W of vector space V that is closed under $\{+, \times\}$. When we say a subset is a subspace of a vector space, we mean it is a vector space as well.*

Theorem 91. *$\{0\}$ is a subspace of all vector space.* □

matrix is late Latin for *womb*. The idea is that a matri is a place for holding numbers.

Definition 84. *a trace of an $n \times n$ matrix M, denoted tr(M), is the sum of diagonal entries:*

$$tr(M) = \sum_{i=1}^{n} M_{ii} \tag{3.1}$$

Definition 85. *A span of a nonempty subset S of a vector space V is the set consisting of all linear combinations of the vectors in S. If span$(S) = V$, S generate (or span) V.* □

Definition 86. *The span of $\emptyset$ is $\{0\}$, not $\emptyset$.*

A span set is useful because it allow one to describe all vectors in terms of a much smaller space.

Definition 87. *A subset S of V is linearly dependent if there exist a finite number of distinct vector $u_1, u_2, \ldots, u_n$ in S and scalars $a_1, a_2, \ldots, a_n$, not all 0, that:*

$$\sum_{i=1}^{n} a_i u_i = 0 \tag{3.2}$$

S is called linearly independent if it is not linearly dependent. $\emptyset$ is linearly independent.

□

Theorem 92. *Let S be linearly independent, v is not in S. Then $S \cup v$ is linearly dependent if $v \in$ span(S).*

3.1.3 Basis

Basis tries to represent a infinite vector space using a finite set of vectors. So a complex structure could be understood using simplified structure. A linearly independent generating set has a very useful property that every vector has one and only one representation using basis.

Definition 88. *A basis β for V is a linearly independent subset of V that generate V.* □

A vector space is usually infinite. It is desirable to describe this infinite set using a finite subset, which is the role of basis.

Theorem 93. *$\emptyset$ is a basis for zero vector space $\{0\}$, so every vector space has a basis.*

Definition 89. *The standard basis for F^n is $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, $e_n = (0, 0, \ldots, 1)$.*

Definition 90. *The standard basis for $P_n(F)$ is $\{1, x, x^2, \ldots, x^n\}$.*

Theorem 94. *β is a basis of V if $\forall v \in V$, v has a unique representation as a linear combination of vectors of β.*

Theorem 95. *A finite spanning set for V can be reduced to a basis.*

Theorem 96 (Replacement Theorem)**.** *Let V be generated by a set G with n vectors. Let L be a linearly independent subset of V with m vectors. Then $m < n$ and $\exists H \subset G$ with $n - m$ vectors such that $L \cup H$ generate V.* □

Theorem 97. *Let V have a finite basis. Then every basis contains the same number of vectors. This number is an intrinsic property of V and called the dimension of V.*

Theorem 98. *Let V be a vector space with dimension n:*

- *any finite generating set for V contains at least n vectors. If they contains exactly n vectors, they are a basis.*
- *any linearly independent subset of n vectors is a basis.*
- *every linearly independent subset could be extended to a basis.*

Definition 91 (Lagrange Interpolation Formula)**.** *let $c_0, c_1, \ldots, c_n$ be distinct scalars in field F. Define $n + 1$ function $\{f_i\}$ as:*

$$f_i(x) = \prod_{k=0, k\neq i}^{n} \frac{x - c_k}{c_i - c_k} \tag{3.3}$$

then $\beta = \{f_i\}$ is a basis of $\mathbb{P}_n(F)$, where $\mathbb{P}_n(F)$ is a set of all polynomials over F. For $\forall g \in \mathbb{P}_n(F)$, we have

$$g = \sum_{i=0}^{n} g(c_i) f_i \tag{3.4}$$

To generate a function g of degree n that passes $n + 1$ points (x_i, y_i), first use $\{x_i\}$ to generate $\{f_i\}$, then $g = \sum\limits_{i=0}^{n} y_i f_i$.

Proof. since β is a basis of $\mathbb{P}_n(F)$, $\forall g \in \mathbb{P}_n(F)$,

$$g = \sum_{i=0}^{n} b_i f_i$$

it follows that

$$g(c_j) = \sum_{i=0}^{n} b_i f_i(c_j) = b_j$$

so $g = \sum\limits_{i=0}^{n} g(c_i) f_i$. □

Theorem 99. *for any two subspace W_1 and W_2 of V, their dimension has a relation:*

$$dim(W_1 + W_2) = dim(W_1) + dim(W_2) - dim(W_1 \cap W_2) \tag{3.5}$$

Definition 92. *here are the definition of common terms:*

1. *square matrix: a matrix $M_{i\times j}$ that $i = j$. It is usually denoted as M, not A.*
2. *zero vector: $\vec{0}$.*
3. *transpose: $\left(A^\top\right)_{ij} = A_{ji}$.*
4. *symmetric matrix: $A^\top = A$.*
5. *diagonal matrix: for a $n \times n$ square matrix M that $M_{ij} = 0$ if $i \neq j$.*
6. *upper triangular: $A_{ij} = 0$ if $i > j$.*

□

The following text discusses the result of infinite basis.

Definition 93. *Let F be a family of sets. A member M of F is called maximal if M is contained in no member of F other than M itself.*

Definition 94. *A collection of set C is called a chain if for each pair of sets A and B in C, either $A \subseteq B$ or $B \subseteq A$.*

Theorem 100. *Let F be a family of sets. If for each chain $C \subseteq F$, there exists a member of F that contains each member of C, then F contains a maximal member.*

Proof. use axiom of choice. Note that the maximal member may not be in C. □

Definition 95. *Let S be a subset of a vector space V. A maximal linearly independent subset of S is a subset B of S that:*

1. *B is linearly independent.*
2. *The only linearly independent subset of S that contains B is B.*

Theorem 101. *If V has a basis β, β is maximal linearly independent.*

Proof. A basis is linearly independent. Because a basis generate V, nothing could be added to it and still make it linearly independent. $\square$

Theorem 102. *Let V be a vector space and S a subset that generate V. If β is a maximal linearly independent subset of S, then β is a basis V.*

Proof. β is linearly independent, so only need to prove that β generate V. It is easy because β is maximal in S so nothing from S could be added to it. $\square$

Theorem 103. *Let S be a linearly independent subset of a vector space V. There exists a maximal linearly independent subset of V that contains S.*

Proof. Let F be a family of all linearly independent subsets of V that contains S. For a chain C in F, let U be the union of all its member. This U is linearly independent and belongs to F, so it is a maximal linearly independent subset of F, which is a basis of F. $\square$

Theorem 104. *Every vector space has a basis.*

3.2 Linear Transformation and Matrix

3.2.1 Linear Transformation

Definition 96. *A linear transformation from V to W is a function $T : V \to W$ that:*

1. $T(x+y) = T(x) + T(y)$
2. $T(cx) = cT(x)$

The two linear transformation verification criteria could be combined into one: prove that

$$T(cx+y) = cTx + Ty \tag{3.6}$$

The identity transformation $\mathrm{I}_v : V \to V$ is defined as $\mathrm{I}_v(x) = x$.
The zero transformation $T_0 : V \to W$ is defined as $T_0 = 0$.

Definition 97. *Let $T : V \to W$ be linear. the null space $\mathcal{N}(T)$ of T is the set $\{x \in V : T(x) = 0\}$. It is also called the kernel of T. It measures how much information is lost by the transformation T.*

Definition 98. *The range of T is defined as $\mathcal{R}(T) = \{T(x) : x \in V\}$. It measures how much information is retained by the transformation T.*

Theorem 105. *Let $T : V \to W$ be linear. If $\beta = \{v_i\}$ is a basis for V, then*

$$\mathcal{R}(T) = \mathbf{span}\left(T(\beta)\right) = \mathbf{span}\left(\{T(v_i)\}\right) \tag{3.7}$$

Definition 99. *Let $T : V \to W$ be linear. the nullity of T is the dimension of $\mathcal{N}(T)$. The rank of T is the dimension of $\mathcal{R}(T)$.*

Theorem 106 (Dimension Theorem)**.** *If V is finite dimensional, $T : V \to W$ is linear, then*

$$\mathbf{dim}\left(\mathcal{N}(T)\right) + \mathbf{dim}\left(\mathcal{R}(T)\right) = \mathbf{dim}\left(T\right) \tag{3.8}$$

Proof. expand nullity set to a basis and prove the image of extra parameters are independent. □

Theorem 107. *Let $V : \{v_i\}$ and $W : \{w_i\}$ be vector space over F, and their dimensions are the same. Then there exists a unique linear transformation $T : V \to W$ such that $T(v_i) = w_i$.*

Proof. For $x = \sum_{i=1}^{n} a_i v_i$, define $T : V \to W$ that $T(x) = \sum_{i=1}^{n} a_i w_i$. □

Theorem 107 is useful when proving two functions are the same.

Theorem 108. *Let $T : V \to W$ be a linear transformation. T is one-to-one if and only if $\mathcal{N}(T) = \{0\}$.*

3.2.2 Matrix Representation

Definition 100. *A ordered basis for V is a basis for V with a specific order.*

Definition 101. *$\{e_1, e_2, \ldots, e_n\}$ is the standard ordered basis for F^n. $\{1, x, \ldots, x^n\}$ is the standard ordered basis for $P_n(F)$.*

Definition 102. *Let $\beta = \{u_1, u_2, \ldots, u_n\}$ be an ordered basis for V. $\forall x \in V$, let $\{a_1, a_2, \ldots, a_n\}$ be the unique scalar such that*

$$x = \sum_{i=1}^{n} a_i u_i$$

the coordinate vector of x relative to β, is defined as

$$[x]_\beta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \tag{3.9}$$

Note that $[u_i]_\beta = e_i$.

Definition 103. *Let V with ordered basis $\beta = \{v_i\}$, W with ordered basis $\gamma : \{w_i\}$, $T : V \to W$ be linear. There exists unique scala $a_{ij} \in F$ such that*

$$T(v_j) = \sum_{i=1}^{m} a_{ij} w_j \tag{3.10}$$

The $m \times n$ matrix[1] A defined by $A_{ij} = a_{ij}$ is the matrix representation of T in the ordered basis β and γ and write $A = [T]_\beta^\gamma$. If $V = W$ and $\beta = \gamma$, we write $A = [T]_\beta$. □

Note that the j-th column of A is $\left[T(v_j)\right]_\gamma$: $[T]_\beta^\gamma = \left[\ldots, \left[T(v_j)\right]_\gamma, \ldots\right]$.

Note that T is the relationship between two basis. The value of T might be the same as basis, for example when they are operators on F^n, but T and basis are different objects. It is easy to confuse them, especially on F^n.

Theorem 109. *If $\mathrm{U}, T : V \to W$ are linear transformation that $[\mathrm{U}]_\beta^\gamma = [T]_\beta^\gamma$, then $\mathrm{U} = T$.*

Definition 104. *$\mathcal{L}(V, W)$ contains all linear transformation from V to W.*

Theorem 110. *Let T,U be linear transformation over V and W,*

1. $[T + \mathrm{U}]_\beta^\gamma = [T]_\beta^\gamma + [\mathrm{U}]_\beta^\gamma$
2. $[aT]_\beta^\gamma = a\,[T]_\beta^\gamma$ *for all scalar a*

Theorem 111. *let $T : V \to W$ and $\mathrm{U} : W \to Z$. Then $\mathrm{U}T : V \to Z$ is linear.*

Definition 105. *Let $T : V \to W$ and $\mathrm{U} : W \to Z$ be linear transformation. $A_{m\times n} = [\mathrm{U}]_\alpha^\beta$ and $B_{n\times p} = [T]_\beta^\gamma$ where $\alpha = \{v_i\}$, $\beta = \{w_i\}$, $\gamma = \{z_i\}$. Define the product of matrix AB as:*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} \tag{3.11}$$

then

$$[UT]_\alpha^\gamma = [U]_\beta^\gamma\,[T]_\alpha^\beta \tag{3.12}$$

Proof. For product $AB = [UT]_\alpha^\gamma$, we have

$$\begin{aligned}(UT)(v_j) = U(T(v_j)) &= U\left(\sum_{k=1}^{m} B_{kj} w_k\right) = \sum_{k=1}^{m} B_{kj} U(w_k) \\ &= \sum_{k=1}^{m} B_{kj}\left(\sum_{i=1}^{p} A_{ik} z_i\right) = \sum_{k=1}^{m}\left(\sum_{i=1}^{p} A_{ik} B_{kj}\right) z_i \\ &= \sum_{i=1}^{p} C_{ij} z_i\end{aligned} \tag{3.13}$$

□

Definition 106. *the Kronecker delta δ_{ij} is defined as*

$$\delta_{ij} = \begin{cases} 1 & \text{, if } i = j \\ 0 & \text{, if } i \neq j \end{cases} \tag{3.14}$$

Definition 107. *The $n \times n$ identity matrix I_n is defined as $(I_n)_{ij} = \delta_{ij}$.*

Theorem 112. *Let u_j and v_j be the jth column of AB and B, then*

1. $u_j = Av_j : AB = \left[Av_1, Av_2, \ldots, Av_j, \ldots, Av_p\right]$
2. $v_j = Be_j : B = B \times I_n$

Theorem 113. *Let $T : V \to W$ be linear, we have*

$$\left[T(u)\right]_\gamma = [T]_\beta^\gamma\,[u]_\beta \tag{3.15}$$

[1] The word matrix is Latin for womb which is the same root as matrimony. The idea is that a matrix is a receptacle for holding numbers.

Proof. Fix $u \in V$, and define linear transformation $f : F \to V$ by $f(a) = au$ and $g : F \to W$ by $g(a) = aT(u)$. Let $a = \{1\}$ be the standard basis of F. Notice that $g = Tf$. we have:

$$[T(u)]_\gamma = [g(1)]_\gamma = [g]_\alpha^\gamma = [Tf]_\alpha^\gamma = [T]_\beta^\gamma [f]_\alpha^\beta = [T]_\beta^\gamma [f(1)]_\beta = [T]_\beta^\gamma [u]_\beta \tag{3.16}$$

□

Note: in the above proof, a vector could be treated as a linear transformation from a field to vector space.

Definition 108. *Let A be an $m \times n$ matrix. The mapping L_A that $L_A : F^n \to F^m$ defined by $L_A(x) = Ax$ is called left-multiplication transformation .* □

A linear transformation is different from matrix:

1. Matrix is finite dimensional, so it defines relation only in finite dimension space. A linear transformation could be of any dimension.
2. For a transformation, its matrix representation depends on the chosen basis.

Theorem 114.

$$\begin{cases} [L_A]_\alpha^\beta & = A \\ L_{[T]_\alpha^\beta} & = T \end{cases} \tag{3.17}$$

3.2.3 Inverse

Definition 109. *Let $T : V \to W$ and U $: W \to V$ be linear. U is an inverse of T if TU $= I_W$ and UT $= I_V$. If T has an inverse, T is invertable , which is denoted as T^{-1}.*

Theorem 115. $(\mathrm{UT})^{-1} = T^{-1}\mathrm{U}^{-1}$.

Definition 110. *Let A be $n \times n$ matrix. A is invertable if there is an $n \times n$ matrix B that $AB = BA = I$.*

Theorem 116. *if T is invertible,*

$$\left[T^{-1}\right]_\gamma^\beta = \left([T]_\beta^\gamma\right)^{-1}$$

Proof.

$$I_n = [I_V]_\beta = \left[T^{-1}T\right]_\beta = \left[T^{-1}\right]_\gamma^\beta [T]_\beta^\gamma$$

□

Definition 111. *V is isomorphic to W if there exists a linear transformation $T : V \to W$ that is invertible. T is called an isomorphism from V to W.*

Theorem 117. *V is isomorphic to W if* $\mathbf{dim}\,(V) = \mathbf{dim}\,(W)$.

Proof. If the dimensions are the same, choose basis β of V and γ of W and create a linear mapping $T : \beta \to \gamma$ by Theorem 107. □

Theorem 118. *Let V be a vector space over F. Then V is isomorphic to $F^n \Leftrightarrow$* $\mathbf{dim}\,(V) = n$.

Theorem 119. *The function $\Phi : \mathcal{L}(V, M) \to M_{m\times n}(F)$ defined by $\Phi(T) = [T]_\beta^\gamma$, is an isomorphism. The dimension has relation that*

$$\mathbf{dim}\left(\mathcal{L}(V, M)\right) = \mathbf{dim}\,(V) \times \mathbf{dim}\,(W) \tag{3.18}$$

3.2.4 Change of Coordinate Matrix

Theorem 120. *Let β and β' be two ordered basis of V. Let $Q = [I_V]_{\beta'}^\beta$, then*

1. *Q is invertible.*
2. *$\forall \alpha \in V$, $[\alpha]_\beta = Q\,[\alpha]_{\beta'} = [I_V]_{\beta'}^\beta\,[\alpha]_{\beta'}$.*

$Q = [I_V]_{\beta'}^\beta$, is called change of coordinate matrix that changes from β'-coordinates to β-coordinates.

Proof. $\forall \alpha \in V$, $[\alpha]_\beta = \left[I_V(\alpha)\right]_\beta = [I_V]_{\beta'}^\beta\,[\alpha]_{\beta'} = Q\,[\alpha]_{\beta'}$. □

If Q changes β'-coordinate into β-coordinate, Q^{-1} changes β-coordinate into β'-coordinate.

Definition 112. *A linear operator is a linear transformation that map from V to itself.*

Theorem 121. *If T is a linear operator on V, then*

$$[T]_{\beta'} = [I_V]_{\beta}^{\beta'} [T]_{\beta} [I_V]_{\beta'}^{\beta} = Q^{-1} [T]_{\beta} Q \tag{3.19}$$

Proof. $Q [T]_{\beta'} = [I]_{\beta'}^{\beta} [T]_{\beta'}^{\beta'} = [IT]_{\beta'}^{\beta} = [TI]_{\beta'}^{\beta} = [T]_{\beta}^{\beta} [I]_{\beta'}^{\beta} = [T]_{\beta} Q$. □

Theorem 122. *Let $A \in M_{n\times n}(F)$, and $\gamma : \{a_i\}$ is an ordered basis for F^n. Then $[L_A]_{\gamma} = Q^{-1}AQ$, where $Q = [a_1, a_2, \ldots, a_n]$.*

Proof. $[L_A]_I = A$, so

$$[L_A]_{\gamma} = [I_V]_I^{\gamma} \times [L_A]_I \times [I_V]_{\gamma}^{I} = [I_V]_I^{\gamma} \times A \times [I_V]_{\gamma}^{I}$$

A take aways is that Q is the change of coordinate matrix from γ to I. □

Theorem 123. *Let $T : V \to W$, β and β' are ordered basis of V, γ and γ' are ordered basis of W. Then*

$$[T]_{\beta'}^{\gamma'} = [I_W]_{\gamma}^{\gamma'} [T]_{\beta}^{\gamma} [I_V]_{\beta'}^{\beta} \tag{3.20}$$

Example 7. *There is an example of the usage of change of coordinate matrix: do reflection operation T against a line $y = ax$. Let β be the standard basis of R^2 and β' be the standard basis of R^2 after the rotation of $y = ax$. The operation T has a matrix representation in β'*

$$[T]_{\beta'} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Then calculate $[T]_{\beta}$ based on $[T]_{\beta'}$.

Definition 113. *B is similar to A if there is an invertible matrix Q that $B = Q^{-1}AQ$.*

Theorem 124. *If T is a linear operator on finite dimension vector space V, and if β and β' are any ordered basis of V, then $[T]_{\beta'}$ is similar to $[T]_{\beta}$.*

3.2.5 Quotient Space

Definition 114. *Let subspace $U \subset V$, The affine subset $v + U$ of V is defined as:*

$$v + U = \{v + u : u \in U\} \tag{3.21}$$

Definition 115. *Let subspace $U \subset V$. Then the quotient space V/U is defined as:*

$$V/U = \{v + U : v \in V\} \tag{3.22}$$

Definition 116. *Let subspace $U \subset V$. The quotient map $\pi : V \to V/U$ is defined as:*

$$\pi(v) = v + U \tag{3.23}$$

Theorem 125.

$$\mathbf{dim}\left(V/U\right) = \mathbf{dim}\,(V) - \mathbf{dim}\,(U) \tag{3.24}$$

Proof. Define $\pi : V \to V/U$. The null space is U. □

Theorem 126. *Define $\tilde{T} : V/\mathcal{N}(T) \to W$ by:*

$$\tilde{T}\left(v + \mathcal{N}(T)\right) = Tv$$

Then $\tilde{T}$ is an isomorphism between $V/\mathcal{N}(T)$ and T.

Proof. If $u + \mathcal{N}(T) = v + \mathcal{N}(T)$, then $u - v \in \mathcal{N}(T)$. So $T(u - v) = T(u) - T(v) = 0$ and $T(u) = T(v)$. □

3.2.6 Dual Space

Definition 117. *A linear functional is a linear transformation that map from V into F.*

Definition 118. *An i-th coordinate function f_i with respect to basis β is defined as $f_i(x) = a_i$ where*

$$[x]_\beta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_1(a) \\ f_2(a) \\ \vdots \\ f_n(a) \end{bmatrix}$$

Definition 119. *The dual space of V is the vector space $V^* = \mathcal{L}(V, F)$. The double dual space V^{**} is the dual space of V^*.*

The dimension of dual space is $\mathbf{dim}\,(V^*) = \mathbf{dim}\,\big(\mathcal{L}(V,F)\big) = \mathbf{dim}\,(V) \times \mathbf{dim}\,(F) = \mathbf{dim}\,(V)$.

Definition 120. *Let $\beta = \{x_i\}$ be an ordered basis for finite dimensional vector space V. Define $f_i(x) = a_i$ where*

$$[x]_\beta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

f_i is the i-th coordinate function with respect to basis β. let $\beta^* = \{f_i\}$. Then β^* is an ordered basis for V^*, and $\forall f \in V^*$, we have

$$f = \sum_{i=1}^{n} f(x_i) f_i \tag{3.25}$$

β^* is called the dual basis of β.

Proof. Let $g = \sum_{i=1}^{n} f(x_i) f_i$, we have

$$g(x_j) = \left(\sum_{i=1}^{n} f(x_i) f_i \right)(x_j) = \sum_{i=1}^{n} f(x_i) f_i(x_j) = \sum_{i=1}^{n} f(x_i)\delta_{ij} = f(x_j)$$

□

Theorem 127. *Let V and W be vector space over F with ordered basis β and γ. For any linear transformation $T : V \to W$, the mapping $T^t : W^* \to V^*$ defined as $T^\top(g) = gT, \forall g \in W^*$ is a linear transformation with property that $[T^\top]_{\gamma^*}^{\beta^*} = \left([T]_\beta^\gamma\right)^\top$.*

Proof. Let $\beta = \{x_i\}$ and $\gamma = \{y_i\}$ with dual basis $\beta^* = \{f_i\}$ and $\gamma^* = \{g_i\}$, $A = [T]_\beta^\gamma$. we have

$$T^\top(g_j) = g_j T = \sum_{s=1}^{n} (g_j T)(x_s) f_s$$

So the row i, column j entry of $[T^\top]_{\gamma^*}^{\beta^*}$ is

$$(g_j T)(x_i) = g_j(T(x_i)) = g_j \left(\sum_{k=1}^{m} A_{kj} y_k \right) = \sum_{k=1}^{m} A_{kj} g_j(y_k) = \sum_{k=1}^{m} A_{kj} \delta_{kj} = A_{ji}$$

Hence $[T^\top]_{\gamma^*}^{\beta^*} = A^\top$. □

Definition 121. *For $U \subset V$, the annihilator of U, denoted as U_V^0, is defined as*

$$U_V^0 = \{\phi \in V^* : \phi(u) = 0, \forall u \in U\}$$

So the annihilator map U to 0. For vectors in $V - U$, the mapping could be any result. The annihilator is a subspace.

Theorem 128.

$$\mathbf{dim}\,(U) + \mathbf{dim}\left(U_V^0\right) = \mathbf{dim}\,(V) \tag{3.26}$$

Proof. Define $i \in \mathcal{L}(U, V)$ that $i(u) = u, \forall u \in U$. $i^* \in \mathcal{L}(V^*, U^*)$. So

$$\mathbf{dim}\,\big(\mathcal{R}(i^*)\big) + \mathbf{dim}\,\big(\mathcal{N}(i^*)\big) = \mathbf{dim}\,(V^*)$$

By definition, $\mathcal{N}(i^*) = U_V^0$. Also $\mathcal{R}(i^*) = U^*$. □

Theorem 129. *Let V and W be two finite-dimentional vector space, and $T \in \mathcal{L}(V, W)$. Then*:

1. $\mathcal{N}(T^*) = (\mathcal{R}(T))^0$
2. $\mathcal{R}(T^*) = (\mathcal{N}(T))^0$
3. $\mathbf{dim}\,\big(\mathcal{R}(T^*)\big) = \mathbf{dim}\,(\textit{range } T)$
4. $\mathbf{dim}\,\big(\mathcal{N}(T^*)\big) = \mathbf{dim}\,\big(\mathcal{N}(T)\big) + \mathbf{dim}\,(W) - \mathbf{dim}\,(V)$

Proof. Suppose $\varphi \in \text{null } T^*$. Then $0 = T^*(\varphi) = \varphi T$. Then

$$0 = (\varphi T)(v) = \varphi(Tv)$$

So $\varphi \in (\text{range } T)_W^0$.

$$\begin{aligned}
\mathbf{dim}\,\big(\mathcal{R}(T^*)\big) &= \mathbf{dim}\,(W^*) - \mathbf{dim}\,\big(\mathcal{N}(T^*)\big) \\
&= \mathbf{dim}\,(W) - \mathbf{dim}\left(\mathcal{R}(T)^0\right) \\
&= \mathbf{dim}\,\big(\mathcal{R}(T)\big)
\end{aligned}$$

$$\begin{aligned}
\mathbf{dim}\,\big(\mathcal{N}(T^*)\big) &= \mathbf{dim}\left(\mathcal{R}(T)^0\right) \\
&= \mathbf{dim}\,(W) - \mathbf{dim}\,\big(\mathcal{R}(T)\big) \\
&= \mathbf{dim}\,(W) - \big(\mathbf{dim}\,(V) - \mathbf{dim}\,\big(\mathcal{N}(T)\big)\big) \\
&= \mathbf{dim}\,(W) + \mathbf{dim}\,\big(\mathcal{N}(T)\big) - \mathbf{dim}\,(V)
\end{aligned}$$

□

Definition 122. *For vector $x \in V$, define $\hat{x} : V^* \to F$ by $\hat{x}(f) = f(x)$. $\hat{x}$ is a linear functional on V^*, so $\hat{x} \in V^{**}$.*

Theorem 130. *Define $\psi : V \to V^{**}$ by $\psi(x) = \hat{X}$. Then ψ is an isomorphism.*

Theorem 131. *Let V be a finite dimension vector space with dual space V^*. Every ordered basis for V^* is the dual basis for some basis for V.*

$$\begin{tikzcd}
V \arrow[rrr, "T"] \arrow[ddr] & & & W \arrow[ddll] \\
& V^* & W^* \arrow[l, "T^\top"] & \\
& F & &
\end{tikzcd}$$

3.3 Linear Equations

3.3.1 Elementary Operations

Definition 123. *Let A be an $m \times n$ matrix. there are three elementary row operation:*

1. *interchange any two row of A.*
2. *multiply any row of A by nonzero scalar.*
3. *add any scalar multiple of a row of A to another row.*

Definition 124. *An $n \times n$ elementary matrix is a matrix obtained by performing* one *elementary operation on I_n.*

Definition 125. *The rank of $A_{m\times n}$, denoted rank(A), is the rank*[2] *of linear transformation $L_A : F^n \to F^m$.*

Theorem 132. *the rank of a matrix equals the maximum number of linearly independent columns.*

Proof. For any $A \in M_{m\times n}(F)$,

$$\begin{aligned} \text{rank}(A) &= \textbf{rank}(L_A) = \textbf{dim}\left(R(L_A)\right) = \textbf{span}\left(L_A(\beta)\right) \\ &= \textbf{span}\left(\{L_A(e_1), L_A(e_2), \ldots, L_A(e_n)\}\right) \end{aligned}$$

we have $L_A(e_j) = Ae_j = a_j$ where a_j is the jth column of A. Hence

$$R(L_A) = \textbf{span}\left(\{a_1, a_2, \ldots, a_n\}\right)$$

□

Theorem 133. *Let $A_{m\times n}$ has rank r. Then there exist invertible matrix $B_{m\times m}$ and $C_{n\times n}$ that $D = BAC$, where:*

$$D = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$

Theorem 134. *Every invertible matrix is a product of elementary matrices.*

Definition 126. *For system $Ax = b$, the matrix $(A|b)$ is the augmented matrix.*

Theorem 135. *If A is an invertible matrix, it is possible to transform augmented matrix $(A|I_n)$ into matrix $(I_n|A^{-1})$ by means of a finite number of elementary row operations.*

3.3.2 System of Equations

Definition 127. *A system $A_{m\times n}x = b$ of m linear equation in n unknowns is homogeneous if $b = 0$. Otherwise the system is nonhomogeneous.*

Definition 128. *A system is consistent if its solution set is not empty. otherwise it is called inconsistent.*

Theorem 136. *Let K be the set of all solutions for $Ax = 0$. Then $K = \mathcal{N}(L_A)$ has dimension of $n - \textbf{rank}(L_A) = n - \textbf{rank}(A)$.*

Theorem 137. *if $m < n$, the system $Ax = 0$ has nonzero solution.*

Proof. $\textbf{rank}(A) \le m < n$, so $\mathcal{N}(A) = n - \textbf{rank}(A) > 0$. □

Theorem 138. *Let K be the solution set of $Ax = b$, K_H be the solution set of $Ax = 0$. Then for all solution s to $Ax = b$,*

$$K = \{s\} + K_H = \{s + k : k \in K_H\} \tag{3.27}$$

Theorem 139. *Let $A_{n\times n}x = b$ be a system of equations. If A is invertible, the solution is $A^{-1}b$. Conversely, if the system has exactly one solution, A is invertible.*

Theorem 140. *Let $Ax = b$ be a system of linear equations. the system is consistent $\Leftrightarrow$ rank(A) = rank$(A|b)$.*

Proof. $R(L_A) = \textbf{span}\left(\{a_1, a_2, \ldots, a_n\}\right)$. Since $b \in R(L_A)$, the extended span is the same. □

Definition 129. *A matrix is in reduced row echelon form if:*

1. *any row containing a nonzero entry precedes any row in which all the entries are zero.*
2. *the first nonzero entry in each row is the only nonzero entry in its column.*

[2]The rank of a linear transformation is defined in Definition (99) on page 33.

3. *the first nonzero entry in each row is* 1 *and it occurs in a column to the right of the first nonzero entry in the preceding row.*

Theorem 141. *For* $A_{m\times n}$ *and* $B_{n\times p}$*, we have*:

$$\mathbf{rank}(AB) = \mathbf{rank}(B) - \mathbf{dim}\left(\mathcal{N}(A)\cap\mathcal{R}(B)\right) \tag{3.28}$$

Proof. Let β_i be the basis of $\mathcal{N}(A)\cap\mathcal{R}(B)$, expand to the basis $\beta\cup\alpha$ of B. Prove α is a basis of $\mathcal{R}(AB)$. □

Theorem 142. *For* $A_{m\times n}$*, we have*

1. $\mathbf{rank}(A^\top A) = \mathbf{rank}(A) = \mathbf{rank}(AA^\top)$.
2. $\mathcal{R}(A^\top A) = \mathcal{R}(A^\top)$.
3. $\mathcal{N}(A^\top A) = \mathcal{N}(A)$.

$A^\top$ *could be replaced by* A^* *in* C*.*

Proof. If $\exists x \neq 0\left(x\in\mathcal{N}(A^\top)\cap\mathcal{R}(A)\right)$. Then $(A^\top x = 0)\wedge\left(\exists y(x = Ay)\right)$. So $x^\top x = y^\top A^\top x = y^\top(A^\top x) = 0$ and then $x = 0$. According to Theorem 141, $\mathbf{rank}(A^\top A) = \mathbf{rank}(A^\top) - \mathbf{dim}\Big(\mathcal{N}(A^\top)\cap\mathcal{R}(A)\Big) = \mathbf{rank}(A)$. □

Theorem 143. *For a system of linear equation* $Ax = b$*, the associated system of normal equations is defined as* $n\times n$ *system*

$$A^\top Ax = A^\top b \tag{3.29}$$

$A^\top Ax = A^\top b$ is always consistent and has unique solution when $\mathbf{rank}(A) = n$*. If* $Ax = b$ *is consistent, two solutions are the same.* □

3.4 Determinants

Definition 130. *Let $A \in M_{n\times n}(F)$. If $n = 1$, let $A = (A_{11})$ and we define det$(A) = A_{11}$. For $n \geq 2$, det(A) (or $|A|$) is defined as*

$$|A| = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} \times \left|\tilde{A}_{ij}\right| \tag{3.30}$$

where $\tilde{A}_{ij}$ is obtained from A by deleting row i and column j. This is called Laplace expansion. □

Theorem 144. *A function $\delta : M_{n\times n}(F) \to F$ is the same as $|A|$ if it satisfies the following 3 properties:*

1. *It is n-linear function: for a scalar k,*

$$\left|\begin{bmatrix} a_1 \\ \vdots \\ u + kv \\ \vdots \\ a_n \end{bmatrix}\right| = \left|\begin{bmatrix} a_1 \\ \vdots \\ u \\ \vdots \\ a_n \end{bmatrix}\right| + k \left|\begin{bmatrix} a_1 \\ \vdots \\ v \\ \vdots \\ a_n \end{bmatrix}\right| \tag{3.31}$$

2. *It is alternating: $\delta(A) = 0$ if any two adjacent rows are identical.*
3. *$\delta(I) = 1$.*

The determinate is linear on each row when the remaining rows are held fixed. □

Theorem 145. *The effect of elementary row operation on the determinant of a matrix A is:*

1. *interchange any two rows: $|B| = -|A|$.*
2. *multiply a row: $|B| = k|A|$.*
3. *add a multiple of a row to another: $|B| = |A|$.*

Theorem 146. *If* **rank**$(A_{n\times n}) < n$*, then* $|A| = 0$.

Proof. If **rank**$(A_{n\times n}) < n$, one row is a linear combination of all other rows. □

Theorem 147.

$$|AB| = |A| \times |B| \tag{3.32}$$

Theorem 148. *A matrix $A \in M_{n\times n}(F)$ is invertible $\Leftrightarrow |A| \neq 0$. If it is invertible, $\left|A^{-1}\right| = \dfrac{1}{|A|}$.*

Definition 131. *The cofactor of A is defined as*

$$\textbf{cof } A_{ij} = (-1)^{i+j} \left|\tilde{A}_{ij}\right| \tag{3.33}$$

□

If the determinate is calculated using cofactor operation, the performance is $n!$ multiplication. However if it is calculated using elementary row operation, the performance is $\dfrac{n^3 + 2n - 3}{3}$ multiplication.

Definition 132. *The adjugate of A is defined as*

$$\textbf{adj } A = (\textbf{cof } A)^{\top} \tag{3.34}$$

Theorem 149. *The inverse of invertible square matrix A is:*

$$A^{-1} = \frac{1}{|A|} \textbf{adj } A$$

Theorem 150 (Cramer's Rule)**.** *Let $Ax = b$ be a system of n equation with n unknowns. If $|A| \neq 0$, the system has a unique solution:*

$$x_k = \frac{|M_k|}{|A|} \tag{3.35}$$

where M_k is a $n \times n$ matrix obtained from A by replacing column k of A by b.

Proof. Let a_k be the kth column of A and X_k denote the matrix obtained from replacing the column k of identity matrix I_n by x. Then $AX_k = M_k$:

$$\begin{aligned} AX_k &= A \begin{bmatrix} 1 & & & x & & \\ & 1 & & x & & \\ & & \ddots & \vdots & & \\ & & & x & & \\ & & & \vdots & \ddots & \\ & & & x & & 1 \end{bmatrix} \\ &= \begin{bmatrix} Ae_1, Ae_2, \ldots, Ax, \ldots, Ae_n \end{bmatrix} \\ &= \begin{bmatrix} a_1, a_2, \ldots, b, \ldots, a_n \end{bmatrix} \\ &= M_k \end{aligned}$$

Evaluate X_k by cofactor expansion along row k produces

$$|X_k| = x_k \times |I_{n-1}| = x_k$$

Hence

$$|M_k| = |AX_k| = |A| \times |X_k| = |A| \times x_k$$

Therefore

$$x_k = \frac{|M_k|}{|A|}$$

□

Note: Cramer's Rule is too slow for real world calculation.

Theorem 151. *In geometry, for a square matrix $A \in M_{n\times n}(F)$, $|\mathbf{det}\, A|$ is the n-dimensional volume of the parallelepiped having vector $A_{i,\cdot}$ as adjacent sides.*

3.5 Diagonalization

There are two questions for a linear operator T:

1. Is there an ordered basis β that $[T]_\beta$ is a diagonal matrix?
2. If such basis exists, how can it be found?

3.5.1 Eigenvalue and Eigenvectors

Definition 133. *A linear operator T on V is diagonalizable if there is an ordered basis β of V that $[T]_\beta$ is a diagonal matrix. A matrix is diagonalizable if L_A is diagonalizable.*

If an operator T is diagonalizable, for $\beta = \{v_i\}$, we have

$$T(v_j) = \sum_{i=1}^{n} D_{ij} v_j = D_{jj} v_j = \lambda_j v_j$$

So to prove a linear operator T is diagnolizable is to find a basis $\beta = \{v_i\}$ and $\{\lambda_j\}$ that $T(v_i) = \lambda_i v_i$. □

Definition 134. *A* non-zero *vector $v \in V$ is called an eigenvector of linear operator T if $\exists \lambda : T(v) = \lambda v$. λ is called eigenvalue corresponding to eigenvector v. Eigenvector is also called characteristic vector. Eigenvalue is also called characteristic value.*

A eigenvalue could be 0, but eigenvector could not be $\vec{0}$. An eigenvector is an invariant subspace of dimension 1.

Theorem 152. *A linear operator T is diagonalizable if there exists an ordered basis consisting of eigenvectors of T.*

Theorem 153. *λ is an eigenvalue of $A \iff |A - \lambda I_n| = 0$.*

Proof. If λ is an eigenvalue of A, $\exists v \in F^n, v \neq 0$ that $Av = \lambda v$, which is $(A - \lambda I_n)(v) = 0$, which means $A - \lambda I_n$ is not invertible because $v \neq 0$, so $|A - \lambda I_n| = 0$. □

Theorem 154. *Every eigenvalue has at least one eigenvector.*

Proof. Since $|A - \lambda I_n| = 0$, $(A - \lambda I_n)x = 0$ is a homogeneous equation with $\mathbf{dim}\,(A - \lambda I_n) < n$. □

Definition 135. *For $A = [T]_\beta$ the polynomial $f_A(t) = |A - tI_n|$ is called the characteristic polynomial of A and T.*

Theorem 155. *For all eigenvalues λ_i of A, define*

$$S_k(A) = \sum_{1 \le j_1 \le j_2 \le \cdots \le j_k} \prod_{j=1}^{k} \lambda_{i_j} \tag{3.36}$$

that is $S_k(A)$ is the sum of the product of all k eigenvalues, which is the coefficient of characteristic polynomial of $f_A(t)$:

$$f_A(t) = (-1)^n t^n + (-1)^{n-1} S_1(\lambda) t^{n-1} + \cdots + (-1)^{n-k} S_k t^{n-1} + \cdots + S_n \tag{3.37}$$

Define the sum of all[3] *principal minor of size k of A as $E_k(A)$. We have*

$$E_k(A) = S_k(A) \tag{3.38}$$

So

$$trA = \sum \lambda_i \tag{3.39}$$

and

$$|A| = \prod \lambda_i \tag{3.40}$$

Proof. calculate the coefficient by $\left.\dfrac{1}{k!}\dfrac{\mathrm{d}^k f_A(t)}{\mathrm{d}t^k}\right|_{t=0}$ □

Theorem 156. *The choice of basis β did not change the eigenvalue of T.*

Proof.

$$\left|[T]_\beta - \lambda I\right| = \left|Q^{-1}\left([T]_\alpha - \lambda I\right)Q\right| = \left|Q^{-1}\right| \times \left|[T]_\alpha - \lambda I\right| \times |Q| = \left|[T]_\alpha - \lambda I\right|$$

□

[3]There are $\binom{n}{k}$ of them.

Theorem 157. *Similar matrices have the same characteristic function.*

Proof. Assume A is similar to B: $A = P^{-1}BP$. We have

$$f_A(\lambda) = |Ax - \lambda I| = \left|P^{-1}BP - \lambda P^{-1}P\right| = \left|P^{-1}\right| \times |B - \lambda I| \times |P| = |B - \lambda I| = f_B(\lambda)$$

□

Theorem 158. *if Q is a matrix with columns of eigenvectors of β, then according to Theorem 123 , $Q^{-1}AQ$ is a diagonal matrix with eigenvalue.*

3.5.2 Diagonalizability

Theorem 159. *Let λ_i be distinct eigenvalue of T. If $\{v_i\}$ are eigenvector that corresponding to λ_i, then $\{v_i\}$ is* linearly independent.

Proof. suppose it works for $k - 1 \geq 1$ and we have k eigenvector $\{v_i\}$. Suppose

$$a_1v_1 + a_2v_2 + \cdots + a_kv_k = 0$$

multiply $T - \lambda_k I$ to both sides, we have

$$a_1(\lambda_1 - \lambda_k)v_1 + a_1(\lambda_2 - \lambda_k)v_2 + \cdots + a_1(\lambda_{k-1} - \lambda_k)v_{k-1} + = 0$$

because $\{v_1, v_2, \ldots, v_{k-1}\}$ are linearly independent, we have

$$a_1(\lambda_1 - \lambda_k) = a_1(\lambda_2 - \lambda_k) = a_1(\lambda_{k-1} - \lambda_k) = 0$$

because λ_i are different, we have $a_i = 0$. □

Theorem 160. *if T has n distinct eigenvalues, then T is diagonalizable. If T is diagonalizable, it may not have n distinct eigenvalues, for example the identity matrix I_V.*

Definition 136. *A polynomial $f(t)$ in $P(F)$ split over F if there are scalars $c, a_1, \ldots, a_n$ (not necessarily distinct) in F that*

$$f(t) = c(t - a_1)(t - a_2)\ldots(t - a_n)$$

the multiplicity of λ is the largest positive integer k for which $(t - \lambda)^k$ is a factor of $f(t)$.

Theorem 161. *the characteristic polynomial of any diagonalizable linear operator splits.*

Proof. choose a basis β of eigenvectors. $[\mathrm{T}]_\beta$ is a diagonal matrix D. The characteristic polynomial of T is $|D - tI|$ splits. □

Be careful that the characteristic polynomial splits does not mean the matrix is diagonalizable. The eigenvectors need to form a basis.

Definition 137. *let λ be an eigenvalue of T. Let $E_\lambda = \mathcal{N}(T - \lambda I_V)$. the set E_λ is called the eigenspace of T corresponding to eigenvalue λ. So is it for matrix.*

Theorem 162. *let λ be an eigenvalue of T having multiplicity m. then $1 \leq$ **dim** $(E_\lambda) \leq m$.*

Proof. choose ordered basis $\{v_1, v_2, \ldots, v_p\}$ for E_λ, and extend it to ordered basis $\beta = \{v_1, v_2, \ldots, v_p, v_{p+1}, \ldots, v_n\}$ for V, and let $A = [T]_\beta$. let $v_i(1 \leq i \leq q)$ be an eigenvector of T corresponding to λ, we have

$$A = \begin{pmatrix} \lambda I_p & B \\ 0 & C \end{pmatrix}$$

so

$$\begin{aligned} f(t) &= |A - tI_n| \\ &= \left|\begin{bmatrix} (\lambda - t)I_p & B \\ 0 & C - tI_{n-p} \end{bmatrix}\right| \\ &= \left|(\lambda - t)I_p\right| \times \left|C - tI_{n-p}\right| \\ &= (\lambda - t)^p g(t) \end{aligned}$$

So $(\lambda - t)^p$ is a factor of $f(t)$, and the multiplicity of λ is at least $p = \dim(E_\lambda)$, so $\dim(E_\lambda) \leq m$ □

Theorem 163. *let $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ be distinct eigenvalue of T. let S_i be a finite linearly independent subset of eigenspace E_{λ_i}. then $S_1 \cup S_2 \cup \dots \cup S_k$ is a linearly independent subset of V.*

Theorem 164. *let $\lambda_1, \lambda_2, \dots, \lambda_k$ be distinct eigenvalue of T, then*

1. *T is diagonalizable $\iff$ the multiplicity of λ_i is equal to* **dim** *(E_{λ_i}) for all i.*
2. *If T is diagonalizable and β_i is an ordered basis for E_{λ_i} for each i, then $\beta = \beta_1 \cup \beta_2 \cup \dots \cup \beta_k$ is an ordered basis for V consisting of eigenvectors of T.*

Theorem 165. *T is diagonalizable $\iff$ both of the following holds:*

1. *the characteristic polynomial of T splits.*
2. *for each eigenvalue λ of T, the multiplicity of λ equals $n -$ **rank**$(T - \lambda I)$.*

Definition 138. *Let W_i be subspaces of a vector space V. The sum of these subspaces is defined as:*

$$\sum_{i=1}^{k} W_i = \{v_1 + v_2 + \dots + v_k : v_i \in W_i \text{ for } 1 \leq i \leq k\} \tag{3.41}$$

Definition 139. *let W_i be subspace of V. V is the direct sum of subspace $\{W_1, W_2, \dots, W_k\}$, or $V = W_1 \oplus W_2 \oplus \dots \oplus W_k$ if*

$$V = \sum_{i=1}^{k} W_i$$

and

$$W_j \cap \sum_{i \neq j} W_i = \emptyset, (1 \leq j \leq k)$$

Theorem 166. *T is diagonalizable $\iff$ V is the direct sum of eigenspaces of T.*

3.5.3 Invariant Subspaces

Definition 140. *A subspace W of V is T-invariant subspace of V if $T(W) \subseteq W$. Common T-invariant subspaces are: $\emptyset$, V, $R(T)$, $N(T)$.* □

Theorem 167. *A subspace W with basis $\alpha = \{v_1, v_2, \dots, v_k\}$ is T-invariant. Let $\beta = \alpha \cup \gamma$ as the expanded basis of V. Then*

$$[T]_\beta = \begin{bmatrix} A_{k\times k} & B \\ 0 & C \end{bmatrix} \tag{3.42}$$

The reverse is true. If $[T]_\beta$ has such representation, the first k basis of β is T-invariant.

Definition 141. *A T-cyclic subspace of V generated by x is defined as $W = span\left(\{x, T(x), T^2(x), \dots\}\right)$.*

Theorem 168. *Let T be a linear operator on finite-dimensional vector space V, and let W be a T-invariant subspace of V. Then the characteristic polynomial of T_W divides the characteristic polynomial of T.*

Proof. Choose ordered basis γ for W and expand it to β for V. Calculate $[T]_\beta$ and $[T]_\gamma$. □

Theorem 169. *Let T be a linear operator on finiate-dimensional vector space V, and let W be a T-cyclic subspace of V generated by nonzero vector $v \in V$. Let $k =$ **dim** (W). Then:*

1. *$\{v, T(v), T^2(v), \dots, T^{k-1}(v)\}$ is a basis for W.*
2. *If $a_0 v + a_1 T(v) + a_2 T^2(v) + \dots + a_{k-1} T^{k-1}(v) + T^k(v) = 0$, then the characteristic polynomial of T_W is $f(t) = (-1)^k \left(a_0 + a_1 t + \dots + a_{k-1} t^{k-1} + t^k\right)$.*

Proof. Let $\beta = \{v, T(v), T^2(v), \dots, T^{k-1}(v)\}$, and let a_i be the scalars that

$$a_0 v + a_1 T(v) + a_2 T^2(v) + \dots + a_{k-1} T^{k-1}(v) + T^k(v) = 0$$

Fors basis $\{v, T(v), T^2(v), \dots, T^{k-1}(v)\}$, $[T(v)]_\beta = [0, 1, \dots, 0]$, $T\left(T(v)\right)_\beta = [0, 0, 1, \dots, 0]$, etc, we have:

$$[T_W]_\beta = \begin{bmatrix} 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & \dots & 0 & -a_1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -a_{k-1} \end{bmatrix}$$

which has characteristic polynomial

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

□

Theorem 170 (Cayley-Hamilton)**.** *Let T be linear operator on a finite-dimensional vector space V, and let $f(t)$ be the characteristic polynomial of T. Then $f(T) = 0$.*

Proof. Suppose $v \neq 0$. Let W be the T-cyclic subspace generated by v, and suppose the $\mathbf{dim}(W) = k$. So there exists scalars $\{a_i\}$ that

$$a_0 v + a_1 T(v) + a_2 T^2(v) + \cdots + a_{k-1}T^{k-1}(v) + T^k(v) = 0$$

which implies the characteristic polynomial of T_W is

$$g(t) = (-1)^k \left(a_0 + a_1 t + \cdots + a_{k-1}t^{k-1} + t^k\right)$$

We have

$$g(T)(v) = (-1)^k \left(a_0 I + a_1 T + \cdots + a_{k-1}T^{k-1} + T^k\right)(v) = 0$$

Because $g(t)$ divides $f(t)$, $\exists q(t)$ that $f(t) = g(t)q(t)$. So

$$f(T)(v) = q(T)g(T)(v) = q(T)\left(g(T)(v)\right) = q(T)(0) = 0$$

□

Definition 142. *Let $B_1 \in M_{m\times m}(F)$, and $B_2 \in M_{n\times n}(F)$. The direct sum of B_1 and B_2, denoted as $B_1 \oplus B_2$, as the $(m+n)\times(m+n)$ matrix A that*

$$A = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}$$

Theorem 171. *Suppose $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$, where W_i is a T-invariant subspace of V. Suppose $f_i(t)$ is the characteristic polynomial of T_{W_i}, Then $\prod_{i=1}^{k} f_i$ is the characteristic polynomial of T. Let β_i be an ordered basis for W_i, and let $\beta = \bigcup_{i=1}^{k} \beta_i$. Let $A = [T]_\beta$, and $B_i = [T_{W_i}]_\beta$. Then $A = B_1 \oplus B_2 \oplus \cdots \oplus B_k$.*

3.5.4 Limit of Markov Chain Matrix

Definition 143. *A sequence $\{A_1, A_2, \dots\}$ converge to limit L if $\lim_{m\to\infty}(A_m)_{ij} = L_{ij}$.*

Theorem 172. *If $A_i \to L$, them for any P and Q, $\lim_{m\to\infty} PA_m = PL$ and $\lim_{m\to\infty} A_m Q = LQ$.*

Theorem 173. *Let Q be invertible and $A_i \to L$. Then $\lim_{m\to\infty}(QAQ^{-1})^m = QAQ^{-1}$.*

Definition 144. *Define a set S which consists of the interior of unit disk and 1:*

$$S = \left\{\lambda \in C : |\lambda| < 1 \vee \lambda = 1\right\} \tag{3.43}$$

Theorem 174. *Let A be square matrix in C. $\lim_{m\to\infty} A^m$ exists if and only if:*

1. *Every eigenvalue of A is in S.*
2. *If 1 is an eigenvalue of A, then the dimension of its eigenspace equals its multiplicity.*

Proof. use Jordan canonical form. □

Theorem 175. *For square matrix A in C, if*

1. *Every eigenvalue of A is in S.*
2. *A is diagonalizable.*

Then $\lim_{m\to\infty} A^m$ exists.

Proof. Since A is diagonalizable, $\exists Q : A = QDQ^{-1}$. So $A^m = QD^mQ^{-1}$. This is used to calculate A^m. □

Definition 145. *transition matrix or stochastic matrix is a square matrix A that $A_{ij} \geq 0 \wedge \forall j \left(\sum_i A_{ij} = 1\right)$.*

Definition 146. *P is a probability vector if its entries are all non-negative and sum to 1.*

Definition 147. *$\vec{1_n}$ is a column vector that each coordinate is 1.*

Theorem 176. *Let M be a square matrix with non-negative real entries, and v a column vector with real non-negative coordinates. Then*

1. *M is a transition matrix if and only if $M^\top \vec{1_n} = \vec{1_n}$.*
2. *v is a probability vector if and only if $\vec{1_n}^\top v = 1$.*
3. *The product of two transition matrix is transition matrix.*
4. *The product of a transition matrix and probability vector is a probability vector.*

Definition 148. *A transition matrix is regular if some power of the matrix contains only positive entries. It may contain zero entries.*

Definition 149. *For square matrix A, define $\rho_i(A) = \sum_j |A_{ij}|$ and $v_j(A) = \sum_i |A_{ij}|$. The row sum $\rho(A) = \max \rho_i$ and column sum $v(A) = \max v_j$.*

Definition 150. *For square matrix $A_{n\times n}$, the Gerschgorin disk C_i is defined as:*

$$C_i = \left\{ z \in C : |z - A_{ii}| < \rho_i(A) - |A_{ii}| \right\} \tag{3.44}$$

So the disk center is the diagonal entry, and the radius is the sum of the absolute values of all rest row entries.

Theorem 177. *Every eigenvalue of A is contained in a Gerschgorin disk.*

Proof. Let λ be a eigenvalue with eigenvector v. So $\sum_{j=1}^n A_{ij} v_j = \lambda v_i$. Assume v_k is the coordinate of v that has the largest absolute value. Then $v_k \neq 0$ because $v \neq 0$. We have

$$|\lambda v_k - A_{kk} v_k| = \left| \sum_{j=1}^n A_k j v_j - A_{kk} v_k \right| = \left| \sum_{j\neq k} A_{kj} v_j \right| \leq \sum_{j\neq k} |A_{kj}||v_j| \leq \sum_{j\neq k} |A_{kj}||v_k| = |v_k| \left(\rho_i(A) - |A_{kk}|\right)$$

So $|v_k| \times |\lambda - A_{kk}| \leq |v_k| \left(\rho_i(A) - |A_{kk}|\right)$ and $|\lambda - A_{kk}| \leq \left(\rho_i(A) - |A_{kk}|\right)$. □

Theorem 178. *Let λ be any eigenvalue of A. Then $|\lambda| \leq \rho(A)$.*

Proof. $|\lambda| = \left|(\lambda - A_{kk}) + A_{kk}\right| \leq |\lambda - A_{kk}| + |A_{kk}| \leq \rho_i(A) - |A_{kk}| + |A_{kk}| = \rho_i(A)$ □

Theorem 179. *Let λ be any eigenvalue of A. Then $|\lambda| \leq \min\left\{\rho(A), v(A)\right\}$.*

Proof. λ is an eigenvalue of $A^\top$. □

Theorem 180. *If λ is an eigenvalue of transition matrix, then $|\lambda| \leq 1$.*

Theorem 181. *Every transition matrix has 1 as eigenvalue.*

Proof. $A^\top \times \vec{1_n} = \vec{1_n}$. □

Theorem 182. *Let A be a matrix with positive entries, and let λ be an eigenvalue of A that $|\lambda| = \rho(A)$. Then $\lambda = \rho(A)$ and $\vec{1_n}$ is a basis for E_λ.*

Proof. Let v be an eigenvector for λ, and v_k is the coordinate that has the largest absolute value $b = |v_k|$. Then

$$|\lambda|\, b = |\lambda v_k| = \left| \sum_{j=1}^n A_{kj} v_j \right| \leq \sum_{j=1}^n |A_{kj} v_j| = \sum_{j=1}^n |A_{kj}||v_j| \leq \sum_{j=1}^n |A_{kj}|\, b = \rho_k(A) b \leq \rho(A) b$$

Since $|\lambda| = \rho(A)$, all inequalities are equalities, so

1. $\left| \sum_{j=1}^n A_{kj} v_j \right| = \sum_{j=1}^n |A_{kj} v_j|$
2. $|A_{kj}||v_j| = \sum_{j=1}^n |A_{kj}|\, b$
3. $\rho_k(A) \leq \rho(A)$

For Item 1 to hold, $A_{kj}v_j$ are non-negative multiples of a common complex number z. Assume $|z| = 1$. Then $\left(\exists \left\{c_j\right\} \subset R^+\right)(A_{kj}v_j = c_j z)$.

For item 2, since $b = \max|v_j|, |v_j| = b$. So $b = |v_j| = \left|\dfrac{c_j}{A_{kj}}z\right| = \dfrac{c_j}{A_{kj}}$, and $v_j = \dfrac{c_j}{A_{kj}}z = bz$, and $v = bz\vec{1_n}$.

Since A and $\vec{1_n}$ are all positive, $A\vec{1_n} = \lambda\vec{1_n}$, so $\lambda > 0$. □

Theorem 183. *Let A be a transition matrix that each entry is positive, and let λ be any eigenvalue of A other than 1. Then $|\lambda| < 1$. Moreover, the eigenspace of eigenvalue 1 has dimension 1.*

Theorem 184. *Let A be a regular transition matrix, and λ be one of its eigenvalue, then*

1. $|\lambda| \leq 1$.
2. *If $|\lambda| = 1$, then $\lambda = 1$ and* $\mathbf{dim}\left(E_\lambda\right) = 1$.

Theorem 185. *Let A be a disagonalizable regular transition matrix, then $\lim\limits_{m\to\infty} A^m$ exists.*

Theorem 186. *Let A be a regular transition matrix, then*

1. *the multiplicity of eigenvalue 1 is 1.*
2. $\lim\limits_{m\to\infty} A^m$ *exists.*
3. $L = \lim\limits_{m\to\infty} A^m$ *is a transition matrix.*
4. $AL = LA = L$.
5. *The column of L are identical vector v which is the probability vector in E_1.*
6. *For any probability vector w, $\lim\limits_{m\to\infty} A^m w = v$.*

Proof. Since $AL = L$, L are columns of eigenvector for eigenvalue 1. Let $y = \lim\limits_{m\to\infty} A^m w = Lw$, $Ay = ALw = Lw = y$. So y is an eigenvector for eigenvalue 1, and $y = v$. □

3.6 Inner Product Space

3.6.1 Inner Product and Norm

Definition 151. *An inner product on V is a function $V \to V \to F$ (F is either C or R) that $\forall x, y, z \in V$ and $\forall c \in F$ that:*

1. $\langle x+z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$
2. $\langle cx, y\rangle = c\langle x, y\rangle$
3. $\overline{\langle x, y\rangle} = \langle y, x\rangle$
4. $\langle x, x\rangle > 0$ *if* $x \neq 0$

Item (1) and (2) means the inner product is linear in first component. *Please be noted that the result of inner product could be a complex value, but the result of $\langle x, x\rangle$ is a non-negative real number.* □

Theorem 187. *properties of inner product:*

1. $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$
2. $\langle x, cy\rangle = \bar{c}\langle x, y\rangle$
3. $\langle x, x\rangle = 0$ *if and only if* $x = 0$.
4. *If* $\langle x, y\rangle = \langle x, z\rangle$ *for all* $x \in V$, *then* $y = z$.

Item (1) and (2) means the inner product is conjugate linear in second component.

Definition 152. *the standard inner product on F^n for $x = [a_1, a_2, \ldots, a_n]$ and $y = [b_1, b_2, \ldots, b_n]$ is:*

$$\langle x, y\rangle = \sum_{i=1}^{n} a_i \overline{b_i} \tag{3.45}$$

when $F = R$, it is usually called dot product and denoted as $x \cdot y$.

Definition 153. *For $A \in M_{m\times n}(F)$, the conjugate transpose or adjoint of A is $A^* \in M_{n\times m}(F)$ that $(A^*)_{ij} = \overline{A_{ji}}$. If A is complex, $A^* = \overline{A^\top}$. If A is real, A^* is $A^\top$.*

Definition 154 (Forbenius Inner Product). *Let $V = M_{n\times n}(F)$, the Forbenius Inner Product is defined as:*

$$\langle A, B\rangle = \mathbf{tr}(B^* A) \tag{3.46}$$

Theorem 188. *For square matrix $A_{n\times n}$, we have*

$$\langle A, A\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \left|A_{ij}\right|^2 \geq 0 \tag{3.47}$$

Definition 155. *The continuous complex-valued function on interval $[0, 2\pi]$ is a inner product space H:*

$$\langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}dt \tag{3.48}$$

Definition 156. *the norm or length of x is:*

$$\|x\| = \sqrt{\langle x, x\rangle} \tag{3.49}$$

Theorem 189. *the property of norm:*

- $\|cx\| = |c| \cdot \|x\|$
- $\|x\| = 0 \iff x = 0$
- *Cauchy-Schwarz Inequality* $\left|\langle x, y\rangle\right| \leq \|x\| \cdot \|y\|$
- *Triangle Inequality* $\|x+y\| \leq \|x\| + \|y\|$

Theorem 190. *If $\forall x \in C, \langle T(x), x\rangle = 0$. Then $T = 0$.*[4]

Proof.

$$\begin{aligned} \langle T(x+y), x+y\rangle = \langle T(x), y\rangle + \langle T(y), x\rangle &= 0 \\ \langle T(x+iy), x+iy\rangle = \langle T(x), y\rangle - \langle T(y), x\rangle &= 0 \end{aligned}$$

So $\forall y \in V, T(x) = 0$. So $\forall x \in V, T(x) = 0$ and $T = 0$. □

Theorem 191.

$$\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right) \tag{3.50}$$

[4]For it to work in all V, T needs to be self-adjoint. See Theorem 227 on page 56.

3.6.2 Orthogonal and Gram-Schmidt Process

Definition 157. *x and y are orthogonal if $\langle x, y\rangle = 0$. A subset S of V is orthogonal if any two vectors in S are orthogonal. A subset S of V is orthonormal if S is orthogonal and consists entirely of unit vectors.*

Definition 158.

$$\langle x, y\rangle = \|x\| \cdot \|y\| \cos(\theta) \tag{3.51}$$

Definition 159. *A vector is unit vector if $\|x\| = 1$. A normalizing to non-zero x is $\frac{1}{\|x\|}x$.*

Theorem 192. *Let $f_n(t) = e^{int}$ where $0 \le t \le 2\pi$. All f_i are orthogonal.*

Proof.

$$\begin{aligned}
\langle f_m, f_n\rangle &= \frac{1}{2\pi}\int_0^{2\pi} e^{imt}\overline{e^{int}}\,\mathrm{d}t \\
&= \frac{1}{2\pi}\int_0^{2\pi} e^{i(m-n)t}\,\mathrm{d}t \\
&= \frac{1}{2\pi(m-n)} e^{i(m-n)t}\Big|_0^{2\pi} \\
&= 0
\end{aligned} \tag{3.52}$$

□

Theorem 193 (Pythagorean Theorem)**.** *Suppose u and v are orthogonal in V, then*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2 \tag{3.53}$$

Theorem 194. *For a finite dimensional subspace U of V, we have*

$$V = U \oplus U^{\perp} \tag{3.54}$$

Definition 160. *A orthonormal basis for V is an ordered basis that is orthonormal.*

Theorem 195. *Let $S = \{v_1, v_2, \ldots, v_k\}$ be an orthogonal subset of V consisting of non-zero vectors. If $y \in$ **span** (S), then*

$$y = \sum_{i=1}^{k} \frac{\langle y, v_i\rangle}{\|v_i\|^2} v_i \tag{3.55}$$

Define the projection of vector a onto vector u as $\mathbf{proj}_u a = \frac{\langle a, u\rangle}{\|u\|^2}$. So

$$y = \sum_{i=1}^{k} \left(\mathbf{proj}_{v_i} y\right) v_i \tag{3.56}$$

If S is orthonormal, then

$$y = \sum_{i=1}^{k} \langle y, v_i\rangle\, v_i \tag{3.57}$$

Proof. let $y = \sum_{i=1}^{k} a_i v_i$. we have

$$\langle y, v_i\rangle = \left\langle \sum_{i=1}^{k} a_i v_i, v_j \right\rangle = \sum_{i=1}^{k} a_i \langle v_i, v_j\rangle = a_j \|v_j\|^2$$

So $a_j = \frac{\langle y, v_j\rangle}{\|v_j\|^2}$.

□

Theorem 196. *An orthogonal subset of V is linearly independent.*

Definition 161 (Gram-Schmidt process). *Let $S = \{w_1, w_2, \ldots, w_n\}$ be linearly independent subset of V. Define $S' = \{v_1, v_2, \ldots, v_n\}$, where $v_1 = w_1$ and*

$$v_k = w_k - \sum_{j=1}^{k-1} \frac{\langle w_k, v_j \rangle}{\|v_j\|^2} v_j \tag{3.58}$$

then S' is an orthogonal set of non-zero vectors that $\mathbf{span}(S') = \mathbf{span}(S)$. The process is that for the k-th basis w_k, first project it on top of the $k-1$ orthogonal vectors $\sum_{j=1}^{k-1} \frac{\langle w_k, v_j \rangle}{\|v_j\|^2} v_j$, and calculate the reciprocal vector $w_k - \sum_{j=1}^{k-1} \frac{\langle w_k, v_j \rangle}{\|v_j\|^2} v_j$. □

Theorem 197 (QR Decomposition). *Let $A_{m\times n} = [a_1, a_2, \ldots, a_n]$ with $\mathbf{rank}(A) = n$, so $\{a_i\}$ is linearly independent. Use Gram-Schmidt process to form n orthonomal basis:*

$$\begin{aligned}
u_1 &= a_1 && , e_1 = \frac{u_1}{\|u_1\|} \\
u_2 &= a_2 - \mathbf{proj}_{u_1} a_2 && , e_2 = \frac{u_2}{\|u_2\|} \\
\ldots \\
u_n &= a_n - \sum_{j=1}^{n-1} \mathbf{proj}_{u_j} a_n && , e_n = \frac{u_n}{\|u_n\|}
\end{aligned}$$

Then $\forall k$, $a_k = \sum_{j=1}^{k} \langle a_k, e_k \rangle e_k$. So

$$A = QR = [e_1, e_2, \ldots, e_n] \times \begin{bmatrix} \langle a_1, e_1 \rangle & \langle a_2, e_1 \rangle & \langle a_3, e_1 \rangle & \cdots & \langle a_n, e_1 \rangle \\ 0 & \langle a_2, e_2 \rangle & \langle a_3, e_2 \rangle & \cdots & \langle a_n, e_2 \rangle \\ 0 & 0 & \langle a_3, e_3 \rangle & \cdots & \langle a_n, e_3 \rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \langle a_n, e_n \rangle \end{bmatrix} \tag{3.59}$$

The Q is an orthonormal matrix. R could be calculated by:

$$R = Q^\top QR = Q^\top A \tag{3.60}$$

Theorem 198. *If V has an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$, then $\forall x \in V$,*

$$x = \sum_{i=1}^{n} \langle x, v_j \rangle v_i \tag{3.61}$$

Definition 162. *Let β be an orthonormal subset (not basis) of V. For $x \in V$, the Fourier coefficients of x relative to β are $\langle x, y_i \rangle$ for all $y_i \in \beta$.*

Theorem 199. *Let V with an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$. T is a linear operator on V and let $A = [T]_\beta$. then $A_{ij} = \langle T(v_j), v_i \rangle$.*

Proof. From Theorem 198 we have

$$T(v_j) = \sum_{i=1}^{n} \langle T(v_j), v_i \rangle v_i$$

□

Definition 163. *Let S be nonempty subset of V. The orthogonal complement of S is $S^\perp$ that $\forall x \in S, \forall y \in S^\perp, \langle x, y \rangle =$* 0.

Theorem 200. *Let W be a subspace of V. For $y \in V$, there is* unique *$u \in W$ and $z \in W^{\perp}$ that $y = u + z$. u is the* orthogonal projection *of y on W. If $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis of W, then*

$$\begin{aligned} u &= \sum_{i=1}^{k} \langle y, v_i \rangle v_i \\ z &= y - \sum_{i=1}^{k} \langle y, v_i \rangle v_i \end{aligned} \tag{3.62}$$

Theorem 201. *For $S = \{v_1, v_2, \ldots, v_k\}$ be an orthogonal subset of V. For $\forall y \in V$, the orthogonal projection of y on S is $u = \sum_{i=1}^{k} \frac{\langle y, v_i \rangle}{\|v_i\|^2} v_i$. If S are orthonormal, $u = \sum_{i=1}^{k} \langle y, v_i \rangle v_i$. If y is in span of S, then $y = u$.*

Theorem 202. *Let y,u,z as defined in Theorem 200. u is the closest vector in W to y that is $\forall x \in W \left(\|y - x\| \geq \|y - u\|\right)$.*

Proof.

$$\|y - x\|^2 = \|u + z - x\|^2 = \left\|(u - x) + z\right\|^2 = \|u - x\|^2 + \|z\|^2 \geq \|z\|^2 = \|y - u\|^2$$

□

3.6.3 Adjoint of Linear Operator

Theorem 203 (Riesz Representation Theorem)**.** *Let $g : V \to F$ be a linear transformation. Then there exist a unique $y \in V$ that $\forall x \in V$, $g(x) = \langle x, y \rangle$. The y is*

$$y = \sum_{i=1}^{n} \overline{g(v_i)} v_i \tag{3.63}$$

So every vector in the dual space[5] can be represented by an inner product.

Proof. Define $h(x) = \langle x, y \rangle$ with y defined above. So

$$h(v_j) = \langle v_j, y \rangle = \left\langle v_j, \sum_{i=1}^{n} \overline{g(v_i)} v_i \right\rangle = \sum_{i=1}^{n} \left\langle v_j, \overline{g(v_i)} v_i \right\rangle = \sum_{i=1}^{n} g(v_i) \langle v_j, v_i \rangle = g(v_j)$$

□

Theorem 204. *Let T be a linear operator on V. Then there existing a unique linear operator $T^* : V \to V$ that $\langle T(x), y \rangle = \langle x, T^*(y) \rangle$ for all $x, y \in V$. T^* is called the* adjoint *of T.*

Proof. For each y, $\langle T(x), y \rangle$ is a linear operator from V to F, so by Theorem 203, $\exists y'$ that $\langle T(x), y \rangle = \langle x, y' \rangle$. Define T^* as $T^*(y) = y'$. □

Theorem 205.

$$\begin{aligned} \langle T(x), y \rangle &= \langle x, T^*(y) \rangle \\ \langle x, T(y) \rangle &= \langle T^*(x), y \rangle \end{aligned} \tag{3.64}$$

So * is added to T when change the location of T.

Proof.

$$\langle x, T(y) \rangle = \overline{\langle T(y), x \rangle} = \overline{\langle y, T^*(x) \rangle} = \langle T^*(x), y \rangle$$

□

Theorem 206. *Let β be a orthonormal basis for V. If T is a linear operation on V then*

$$[T^*]_\beta = \left([T]_\beta\right)^* \tag{3.65}$$

Let A be an $n \times n$ matrix. Then

$$L_{A^*} = (L_A)^* \tag{3.66}$$

[5] Defined in Theorem 119 on page 37.

Proof. Let $A = [T]_\beta$, $B = [T^*]_\beta$, and $\beta = \{v_1, v_2, \ldots, v_n\}$. Then according to Theorem 199:

$$B_{ij} = \langle T^*(v_j), v_i \rangle = \overline{\langle v_i, T^*(v_j) \rangle} = \overline{\langle T(v_i), v_j \rangle} = \overline{A_{ji}} = (A^*)_{ij}$$

□

Theorem 207. *Let T and U be linear operator on V, then*

1. $(aT + bU)^* = \overline{a}T^* + \overline{b}U^*$
2. $(UT)^* = T^*U^*$
3. $T^{**} = T$

Definition 164. *Let $T : V \to W$ be a linear transformation where V and W are finite dimensional inner product space with inner product $\langle \cdot, \cdot \rangle_V$ and $\langle \cdot, \cdot \rangle_W$. A function $T^* : W \to V$ is called adjoint of T if $\langle T(x), y \rangle_W = \langle x, T^*(y) \rangle_V$.*

Theorem 208. *Let T^* be an adjoint of $T : V \to W$. If β and γ are orthonormal basis for V and W, then*

$$[T^*]^\alpha_\beta = \left([T]^\alpha_\beta\right)^* \tag{3.67}$$

Theorem 209. *Let T^* be an adjoint of $T : V \to W$, we have:*

$$\langle T^*(x), y \rangle_V = \langle x, T(y) \rangle_W \tag{3.68}$$

Theorem 210. *If V is finite dimentional, let T be a linear operator on V, then*

$$\begin{aligned}
\mathcal{R}(T^*)^\perp &= \mathcal{N}(T) \\
\mathcal{R}(T^*) &= \mathcal{N}(T)^\perp \\
\mathcal{R}(T)^\perp &= \mathcal{N}(T^*) \\
\mathcal{R}(T) &= \mathcal{N}(T^*)^\perp
\end{aligned}$$

So $\mathcal{R}(T^*) \perp \mathcal{N}(T)$.

Proof. If $m \in R(T^*)^\perp$, $\forall x \in V$, $0 = \langle m, T^*x \rangle = \langle T(m), x \rangle$, so $m \in N(T)$. □

3.6.4 Examples in Statistics

The following two examples show that for linear equation $Ax - y = 0$,

1. if it is consistent, that is there is solution, we want to find the solution with minimal norm.
2. If it is inconsistent, that is no solution, we want a result that has the least norm.

The same topic is discussed in pseudo inverse.

3.6.4.1 Least Square Approximation

Definition 165. *The Least Square Approximation is a problem that for* $A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$, *find* $x_0 = \begin{bmatrix} c \\ d \end{bmatrix}$ *that minimize* $\|Ax - y\|$.

Definition 166. *For $x, y \in F^n$, define $\langle x, y \rangle_n = y^* \times x$.*

Theorem 211. *Let $A \in M_{m\times n}(F)$, $x \in F^n$, $y \in F^m$, then*

$$\langle Ax, y \rangle_m = \langle x, A^*y \rangle_n \tag{3.69}$$

Proof. $\langle Ax, y \rangle_m = y^* \times (Ax) = (y^* \times A)x = (A^*y)^*x = \langle x, A^*y \rangle_n$ □

Theorem 212. *Let $A \in M_{m\times n}(F)$. Then*[6]

$$\mathbf{rank}(A^*A) = \mathbf{rank}(A) \tag{3.70}$$

So if $\mathbf{rank}(A) = n$, A^*A is invertible.

[6] See Theorem 142 for another proof.

Proof. For equation $A^*Ax = 0$ and $Ax = 0$. $Ax = 0$ implies that $A^*Ax = 0$. Then assume $A^*Ax = 0$, then

$$0 = \langle 0, x\rangle_n = \langle A^*Ax, x\rangle_n = \langle Ax, A^{**}x\rangle_m = \langle Ax, Ax\rangle_m$$

$\square$

Theorem 213. *Let $A \in M_{m\times n}(F)$, $y \in F^m$. Then there exists $x_0 \in F^n$ that $(A^*A)x_0 = A^*y$ and $\forall x \in F^n$, $\|Ax_0 - y\| \le \|Ax - y\|$. If* $\textbf{rank}(A) = n$*, then $x_0 = (A^*A)^{-1}A^*y$.*

Proof. Define $W = \mathcal{R}(L_A)$. There exists a x_0 that is closest to y that $Ax_0 - y \in W^\perp$, so $\langle Ax, Ax_0 - y\rangle_m = 0$. So $\left\langle x, A^*(Ax_0 - y)\right\rangle_n = 0$, so $A^*(Ax_0 - y) = 0$ and $(A^*A)x_0 = A^*y$. $\square$

3.6.4.2 Minimal Solution to Linear Equations

Definition 167. *A solution s is minimal solution of $Ax = b$ if $\|s\| \le \|u\|$ for any solution u.*

Theorem 214. *Let $A \in M_{m\times n}(F)$, $y \in F^m$. Suppose $Ax = y$ is consistent. Then there exists unique minimal solution $s \in R(L_{A^*})$ of $Ax = y$. And s is the only solution in $R(L_{A^*})$. If u is a solution to $(AA^*)u = y$, then $s = A^*u$.*

Proof. By Theorem 210 define $W = R(L_{A^*})$ and $W^\perp = N(L_A)$. $\forall x$ that $Ax = y$, we have $s \in W$ and $t \in W^\perp$ that $x = s + t$. So $y = Ax = A(s + t) = As + At = As$. So s is a solution to $Ax = y$. From Theorem 138, all solution to $Ax = y$ has the form $x' = s + t'$ where $t' \in W^\perp$. And $\left\|x'\right\|^2 = \left\|s + t'\right\|^2 = \|s\|^2 + \left\|t'\right\|^2 \ge \|s\|^2$. $\square$

3.7 Operator

3.7.1 Normal

Theorem 215. *If T has eigenvector, then T^* has eigenvector.*

Proof. $0 = \langle 0, x\rangle = \big\langle (T-\lambda I)(v), x\big\rangle = \big\langle v, (T-\lambda I)^*(x)\big\rangle = \Big\langle v, (T^* - \overline{\lambda} I)(x)\Big\rangle$. Since $v \neq 0$ is reciprocal to the range of $T^* - \overline{\lambda} I$, $v \notin \mathcal{R}(T^* - \overline{\lambda} I)$, so $\mathcal{N}(T^* - \overline{\lambda} I) \neq \{0\}$. $\square$

Theorem 216 (Schur)**.** *Suppose the characteristic polynomial of T splits. Then there exists an orthonormal basis β for V that the $[T]_\beta$ is upper trianglar. Note:*

1. *β does* not *need to be eigenvectors of T.*
2. *It works in $\mathcal{R}$ as long as T splits.*

Proof. Use induction. Since T splits, it has a eigenvector. By Theorem 215 T^* has eigenvector, and make it a unit eigenvector z. Let $W = \text{span}\{z\}$. Then prove $W^\perp$ is T-invariant: for $\forall y \in W^\perp$ and $x = cz \in W$:

$$\big\langle T(y), x\big\rangle = \big\langle T(y), cz\big\rangle = \big\langle y, T^*(cz)\big\rangle = \big\langle y, cT^*(z)\big\rangle = \langle y, c\lambda z\rangle = \overline{c\lambda}\,\langle y, z\rangle = 0$$

According to induction, **dim** $\left(W^\perp\right) = n-1$ and there exists an orthonormal basis γ that $[T_{W^\perp}]_\gamma$ is upper triangular. Take $\gamma \cup \{z\}$. $\square$

Theorem 217. *If β is an orthonormal basis and $[T]_\beta$ is a diagonal matrix, $[T^*]_\beta = \left([T]_\beta\right)^*$ is also a diagonal matrix.*

Theorem 218. *If an operator T has orthogonal eigenvectors β that are basis of the inner product space, then $[T]_\beta$ is a diagonal matrix.*

Definition 168. *T is normal if $TT^* = T^*T$. A square matrix A is normal if $AA^* = A^*A$.*

Theorem 219. *T is normal if and only of $[T]_\beta$ is normal under orthonormal basis β.*

Theorem 220. *Properties of normal operator T on V:*

1. $\forall x \in V, \big\|T(x)\big\| = \big\|T^*(x)\big\|$
2. *$\forall c \in F$, $T - cI$ is normal.*
3. *If x is a eigenvector of eigenvalue λ for T, $T^*(x) = \overline{\lambda} x$, so x is also an eigenvector of eigenvalue $\overline{\lambda}$ for T^*.*
4. *If x_1 and x_2 are for eigenvalues λ_1 and λ_2, $\langle x_1, x_2\rangle = 0$*

Proof.

$$\big\|T(x)\big\|^2 = \big\langle T(x), T(x)\big\rangle = \big\langle T^*T(x), x\big\rangle = \big\langle TT^*(x), x\big\rangle = \big\langle T^*(x), T^*(x)\big\rangle = \Big\|T^*(x)^2\Big\|$$

$$0 = \big\|(T-\lambda I)(x)\big\| = \big\|(T-\lambda I)^*(x)\big\| = \Big\|(T^* - \overline{\lambda} I)(x)\Big\|$$

$$\lambda_1 \langle x_1, x_2\rangle = \langle \lambda x_1, x_2\rangle = \big\langle T(x_1), x_2\big\rangle = \big\langle x_1, T^*(x_2)\big\rangle = \Big\langle x_1, \overline{\lambda_2} x_2\Big\rangle = \lambda_2 \langle x_1, x_2\rangle$$

So $(\lambda_1 - \lambda_2)\langle x_1, x_2\rangle = 0$. Since $\lambda_1 \neq \lambda_2$, $\langle x_1, x_2\rangle = 0$ $\square$

Theorem 221. *If T is normal, $\mathcal{N}(T) = \mathcal{N}(T^*)$ and $\mathcal{R}(T) = \mathcal{R}(T^*)$. So being normal will refine Theorem 210.*

Proof. If $x \in \mathcal{N}(T), \big\|T(x)\big\| = \|T^*\| = 0$, so $T^*(x) = 0$ and $x \in \mathcal{N}(T^*)$. $\square$

Theorem 222. *In $\mathcal{C}$, let V be finite dimensional inner product space. T is normal if and only if there exists an orthonormal basis for V consisting of eigenvectors of T.*

Proof. in C the polynomial always splits. According to Theorem 216 there exists a orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ that $[T]_\beta = A$ is upper triangular. v_1 is an eigenvector because $T(v_1) = A_{1,1}v_1$. Assuming $v_1, v_2, \ldots, v_{k-1}$ are eigenvector of T, we prove that v_k is also an eigenvector of T. Because A is upper triangular,

$$T(v_k) = A_{1,k}v_1 + A_{2,k}v_2 + \cdots + A_{j,k}v_j + \cdots + A_{k,k}v_k$$

Because $\forall j < k$, $A_{j,k} = \big\langle T(v_k, v_j\big\rangle = \big\langle v_k, T^*(v_j)\big\rangle = \Big\langle v_k, \overline{\lambda} v_j\Big\rangle = \lambda_j \big\langle v_k, v_j\big\rangle = 0$, we have $T(v_k) = A_{k,k}v_k$, so v_k is an eigenvector of T.

btw, it does not work in infinite dimensional complex inner product space. $\square$

3.7.2 Hermitian

Definition 169. *T is self-adjoint (Hermitian) if $T = T^*$, or $A = A^*$. For real matrix, it means A is symmetric.*

Theorem 223. *Let T be a linear operator on complex inner product space. Then T is self-adjoint if and only if $\forall x \in V, \langle T(x), x \rangle \in \mathcal{R}$.*

Proof. If T is self-adjoint, $\overline{\langle T(x), x \rangle} = \langle x, T(x) \rangle = \langle T^*(x), x \rangle = \langle T(x), x \rangle$. So $\langle T(x), x \rangle \in \mathcal{R}$.

If $\langle T(x), x \rangle \in \mathcal{R}$, $\langle T(x), x \rangle = \overline{\langle T(x), x \rangle} = \langle x, T(x) \rangle = \langle T^*(x), x \rangle$. So $\forall x \in V$, $\langle (T - T^*)(x), x \rangle = 0$. According to Theorem (190), $T - T^* = 0$. □

Theorem 224. *Let T be a self-adjoint operator on finite dimensional inner product space V. Then:*

1. *every eigenvalue is real.*
2. *If V is a real inner product space, the characteristic polynomial for T splits.*

Proof. Because T is self-adjoint, T is also normal. So according to Theorem 220 if λ is an eigenvalue of T, $\overline{\lambda}$ is an eigenvalue of T^*. So:

$$\lambda x = T(x) = T^*(x) = \overline{\lambda} x$$

So $\lambda = \overline{\lambda}$, and λ is real.

For a orthonormal basis β, $A = [T]_\beta$ is self-adjoint because $A^* = ([T]_\beta)^* = [T^*]_\beta = [T]_\beta = A$. Define $L_A(x) = Ax$ in $\mathcal{C}^n$. Here we create a function in $\mathcal{C}^n$ from a function in $\mathcal{R}^n$. Let γ be the standard basis for $\mathcal{C}$ which is orthonormal. $[L_A]_\gamma = A$ is self-adjoint, so L_A is self-adjoint in $\mathcal{C}^n$. The characteristic polynomial of L_A splits. Since L_A is self-adjoint, all eigenvalues are real, so the polynomial split in $\mathcal{R}$. But L_A, A and T has the same characteristic polynomial. □

Theorem 225. *Let T be a linear operator on finite dimensional real inner product space. T is self-adjoint if and only if there exists an orthonormal basis β for V consisting of eigenvectors of T.*

Proof. By Theorem 216 there exists orthonormal basis β for V that $A = [T]_\beta$ is upper triangular. Because $A^* = ([T]_\beta)^* = [T^*]_\beta = [T]_\beta = A$, A is diagonal matrix. □

Theorem 226. *For the orthonormal basis of eigenvector T problem we have:*

1. *If T splits, we have orthonormal basis that make T upper triangular in $\mathcal{R}$ or $\mathcal{C}$. This basis may not be eigenvectors, or T may not have eigenvectors.*
2. *T is complex normal.*
3. *T is real symmetric.*

Theorem 227. *Let T be self-adjoint operator. If $\forall x \in V, \langle T(x), x \rangle = 0$. Then $T = 0$.*[7]

Proof. Choose orthonormal basis β that consist of eigenvector of T. For $x \in \beta$, $T(x) = \lambda x$. So

$$0 = \langle x, T(x) \rangle = \langle x, \lambda x \rangle = \overline{\lambda} \langle x, x \rangle$$

Hence $\overline{\lambda} = 0$ and $\forall x \in \beta, T(x) = 0$. □

3.7.3 Positive Operator

Definition 170. *An operator T is called positive operator if T is self-adjoint and $\forall x \in V$:*

$$\langle Tx, x \rangle \geq 0 \tag{3.71}$$

Definition 171. *An Operator R is called a square root of an operator T if*

$$R^2 = T \tag{3.72}$$

Theorem 228. *All the following are equivalent:*

1. *T is positive.*
2. *T is self-adjoint and all eigenvalue of T are non-negative.*
3. *T has positive square root.*
4. *T has self-adjoint square root.*
5. *$\exists R : T = R^* R$*

[7] Self-adjoint is not needed of $V = \mathcal{C}$. See Theorem 190 on page 49.

Proof. For 2, if T is positive, $0 \le \langle Tv, v\rangle = \langle \lambda v, v\rangle = \lambda \langle v, v\rangle$, so $\lambda \ge 0$.

For 3, if T is self-adjoint, by Theorem 225 there are orthonormal basis $\beta = \{v_i\}$ with eigenvalue λ_i. Define $R(v_i) = \sqrt{\lambda_i} v_i$. Then $\forall v_i \in \beta, R^2(v_i) = T(v_i)$.

For 1, $\langle Tv, v\rangle = \langle R^* Rv, v\rangle = \langle Rv, Rv\rangle \ge 0$. □

Theorem 229. *A positive operator has a unique positive square root.*

Definition 172. *If T is a positive operator, $\sqrt{T}$ is its positive square root.*

3.7.4 Isometry

Definition 173. *Let T be a linear operator on finite dimensional inner product space V over F. If $\forall x \in V, \|T(x)\| = \|x\|$, we call T unitary operator if $F = \mathcal{C}$ or orthogonal operator if $F = \mathcal{R}$. Unitary and orthogonal are also called isometry.*

Definition 174. *A square matrix A is called unitary matrix if $AA^* = A^*A = I$ and orthogonal matrix if $AA^\top = A^\top A = I$.*

Theorem 230. *Let T be an linear operator. Then the following are equivalent:*

1. *$TT^* = T^*T = I$.*
2. *$\langle T(x), T(y)\rangle = \langle x, y\rangle$.*
3. *If β is an orthonormal basis for V. Then $T(\beta)$ is an orthonormal basis.*
4. *$\|T(x)\| = \|x\|$.*

So unitary or orthogonal operator preserve inner product and norm.

Proof. $\langle x, y\rangle = \langle T^*Tx, y\rangle = \langle T(x), T(y)\rangle$.

If $\beta = \{v_1, v_2, \ldots, v_n\}$ is an orthonormal basis. $\langle T(v_i), T(v_j)\rangle = \langle v_i, v_j\rangle = 0$.

If β and $T(\beta)$ are both orthonormal basis, expand $\|T(x)\|$ and $\|x\|$ to prove they are equal.

$\langle x, x\rangle = \|x\|^2 = \|T(x)\|^2 = \langle T(x), T(x)\rangle = \langle x, T^*Tx\rangle$. So $\forall x \in V, \langle x, (I - T^*T)(x)\rangle = 0$. $I - T^*T$ is normal, so according to Theorem 227, $I - T^*T = 0$. □

Theorem 231. *Unitary operator is normal.*

Proof. See Theorem 230 property (1). □

Theorem 232. *Let T be a linear operator on* real *inner product space V. V has an orthonormal basis of eigenvectors of T with absolute value of all eigenvalues equal to 1 if and only if T is self-adjoint and orthogonal.*

Proof. If T is self-adjoint, there is orthonormal basis β of eigenvectors. If T is orthogonal, $\forall v_i \in \beta, |\lambda_i| \times \|v_i\| = \|\lambda_i v_i\| = \|T(v_i)\| = \|v_i\|$, so $|\lambda_i| = 1$.

If V has orthonormal basis β of eigenvectors, T is self-adjoint. $\forall v_i \in \beta$, we have $TT^*(v_i) = T(\lambda_i v_i) = \lambda_i T(v_i) = \lambda_i^2 v_i$. If $|\lambda_i| = 1$, $TT^* = I$. □

Theorem 233. *Let T be a linear operator on* complex *inner product space V. V has an orthonormal basis of eigenvectors of T with absolute value of all eigenvalues equal to 1 if and only if T is unitary.*

Proof. If T is unitary, it is normal, so there is orthonormal basis β of eigenvectors. If T is unitary, $\forall v_i \in \beta$, $|\lambda_i| \times \|v_i\| = \|\lambda_i v_i\| = \|T(v_i)\| = \|v_i\|$, so $|\lambda_i| = 1$.

If V has orthonormal basis β of eigenvectors, T is normal. If $|\lambda_i| = 1$, $\forall v_i \in \beta, |\lambda_i| \times \|v_i\| = \|\lambda_i v_i\| = \|T(v_i)\| = \|v_i\|$, so $\|T(v_i)\| = \|v_i\|$ and it is unitary. □

Theorem 234. *T is isometry if $[T]_\beta$ is isometry for a orthonormal basis β of V.*

Definition 175. *A is unitarily equivalent or orthogonally equivalent to D if and only if there exists a unitary or orthogonal matrix P that $A = P^*DP$.*

Theorem 235. *Let A be a complex square matrix. A is normal if and only if it is unitarily equivalent to a diagonal matrix.*

Theorem 236. *Let A be a real square matrix. A is symmetric if and only if it is orthogonally equivalent to a diagonal matrix.*

3.7.5 Rigid motion

Definition 176. *Let V be real inner product space. $f : V \to V$ is a rigid motion if*

$$\left\|f(x) - f(y)\right\| = \|x - y\| \tag{3.73}$$

Definition 177. *Let V be real inner product space. $g : V \to V$ is a translation by $v_0 \in V$ if*

$$\exists v_0 \forall x \in V \left(g(x) = x + v_0\right) \tag{3.74}$$

Theorem 237. *A translation is a rigid motion. And a composite of rigid motion is rigid motion.*

Theorem 238. *Let f be a rigid motion. Then there exists a unique orthogonal operator T and unique translation g that $f = g \circ T$.*

Proof. Define $T(x) = f(x) - f(0)$. T is a composite of rigid motion, so it is a rigid motion. Therefore $\left\|T(x)\right\| = \left\|f(x) - f(0)\right\| = \|x - 0\| = \|x\|$. Since

$$\begin{aligned} \left\|T(x) - T(y)\right\|^2 &= \|x\|^2 - 2\left\langle T(x), T(y)\right\rangle + \|y\|^2 \\ \|x - y\|^2 &= \|x\|^2 - 2\langle x, y\rangle + \|y\|^2 \\ \left\|T(x) - T(y)\right\|^2 &= \|x - y\|^2 \end{aligned}$$

We have $\left\langle T(x), T(y)\right\rangle = \langle x, y\rangle$.

Then $\left\|T(ax + y) - aT(x) - T(y)\right\|^2 = 0$ after expansion, T is linear. So T is an orthogonal operator. So we have unique T and g that

$$\begin{aligned} T(x) = f(x) &\quad -f(0) \\ g(x) = x &\quad +f(0) \end{aligned} \tag{3.75}$$

□

Theorem 239. *Let T be an orthogonal operator on R^2, and let $A = [T]_\beta$ where β is the standard basis of R^2. Then one of the following is satisfied:*

1. *T is a rotation, so $|T| = 1$.*
2. *T is a reflection about a line through the origin, so $|T| = -1$.*

Proof. Because T is orthogonal, $T(\beta) = \left\{T(e_1), T(e_2)\right\}$ is an orthonormal basis of R^1. Since $T(e_1)$ is an unit vector, it has the form $T(e_1) = (\cos\theta, \sin\theta)$. Since $T(e_2)$ is orthogonal to $T(e_1)$, it has the form $T(e_2) = (-\sin\theta, \cos\theta)$ or $T(e_2) = (\sin\theta, -\cos\theta)$. □

Theorem 240. *For expression $f(x, y) = ax^2 + 2bxy + cy^2$, let $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $X = \begin{pmatrix} x \\ y \end{pmatrix}$, the formula is $f(X) = X^\top AX = \langle AX, X\rangle$. Since A is symmetric, there is an orthogonal matrix P and diagonal matrix D that $A = P^\top DP$. Define $X_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ that $X = PX_0$. We have $f(X) = X^\top AX = (PX_0)^\top A(PX_0) = X_0^\top DX_0 = \lambda_1 x_1^2 + \lambda_2 x_2^2$. So the xy term could be removed by rotation.*

3.7.6 Spectral Theorem

Definition 178. *Let $V = W_1 \oplus W_2$. T is a projection on W_1 along W_2 if $\forall x = x_1 + x_2$ that $x_1 \in W_1$ and $x_2 \in W_2$, $T(x) = x_1$.*

Theorem 241. *T is a projection if and only if $T^2 = T$.*

Definition 179. *T is an orthogonal projection if $\mathcal{R}(T)^\perp = \mathcal{N}(T)$ and $\mathcal{R}(T) = \mathcal{N}(T)^\perp$[8].*

Theorem 242. *T is an orthogonal projection if and only if T has an adjoint T^* that $T^2 = T = T^*$.*

Proof. $T^2 = T$ because T is a projection. Let $x = x_1 + x + 2$ and $y = y_1 + y_2$ where $x_1, y_1 \in \mathcal{R}(T)$ and $x_2, y_2 \in \mathcal{N}(T)$. So

$$\begin{aligned} \left\langle x, T(y)\right\rangle &= \langle x_1 + x_2, y_1\rangle = \langle x_1, y_1\rangle \\ \left\langle T(x), y\right\rangle &= \langle x_1, y_1 + y_2\rangle = \langle x_1, y_1\rangle \end{aligned}$$

So $T = T^*$ and $T^2 = T = T^*$.

For the reverse side, prove that $\mathcal{R}(T)^\perp = \mathcal{N}(T)$ and $\mathcal{R}(T) = \mathcal{N}(T)^\perp$. □

[8] In finite dimensional space V, $\mathcal{R}(T)^\perp = \mathcal{N}(T) \leftrightarrow \mathcal{R}(T) = \mathcal{N}(T)^\perp$

Theorem 243 (Spectral Theorem)**.** *Let T be real symmetric or complex normal with distinct eigenvalue λ_i and its corresponding eigenspace W_i. Let T_i be the orthogonal projection on W_i. We have:*

1. $T_iT_j = \delta_{ij}T_i$
2. $I = \sum_{i=1}^{k} T_i$
3. $T = \sum_{i=1}^{k} \lambda_i T_i$

λ_i is the spectrum of T. I is the resolution of the identity operator induced by T. $T = \sum_{i=1}^{k} \lambda_i T_i$ is the spectral decomposition of T.

Proof. Let $x = \sum_{i=1}^{k} x_i$ where $x_i \in W_i$. Then

$$T(x) = \sum_{i=1}^{k} T(x_i) = \sum_{i=1}^{k} \lambda_i x_i = \sum_{i=1}^{k} \lambda_i T_i(x_i) = \sum_{i=1}^{k} \lambda_i T_i(x) = \left(\sum_{i=1}^{k} \lambda_i T_i \right) x$$

□

Theorem 244. *Let $F = \mathcal{C}$. T is normal if and only if $\exists g \in P, T^* = g(T)$.*

Proof. Let $T = \sum_{i=1}^{k} \lambda_i T_i$ be the spectral decomposition of T. Take the adjoint of both side and we have

$$T^* = \sum_{i=1}^{k} \overline{\lambda_i} T_i^* \tag{3.76}$$

According to Lagrange formula[9] , $\exists g$, $g(\lambda_i) = \overline{\lambda_i}$. So $g(T) = T^*$. The reverse is easy to prove. □

Theorem 245. *Let $F = \mathcal{C}$. T is unitary if and only if T is normal and $|\lambda| = 1$ for all eigenvalue λ of T.*

Proof. Let $T = \sum_{i=1}^{k} \lambda_i T_i$ be the spectral decomposition of T. We have

$$TT^* = \left(\sum_{i=1}^{k} \lambda_i T_i \right) \times \left(\sum_{i=1}^{k} \overline{\lambda_i} T_i \right) = \sum_{i=1}^{k} |\lambda_i|^2 T_i^2 = \sum_{i=1}^{k} |\lambda_i|^2 T_i = \sum_{i=1}^{k} T_i = I$$

□

Theorem 246. *Let $F = \mathcal{C}$ and T normal. T is self-adjoint if and only if every eigenvalue of T is real.*

Proof. $T^* = \sum_{i=1}^{k} \overline{\lambda_i} T_i = \sum_{i=1}^{k} \lambda_i T_i = T$, so $\overline{\lambda_i} = \lambda_i$. □

3.7.7 Single Value Decomposition

Theorem 247. *Let $T : V \to W$ be a linear transformation with rank r. Then there exists orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ for V and $\gamma = \{u_1, u_2, \ldots, u_m\}$ for W and positive scalars singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ such that*

$$T(v_i) = \begin{cases} \sigma_i u_i & \text{if } 1 \leq i \leq r \\ 0 & \text{if } i > r \end{cases} \tag{3.77}$$

Conversely, for $1 \leq i \leq n$, v_i is an eigenvector of T^*T with corresponding eigenvalue σ_i^2 if $1 \leq i \leq r$ and 0 if $i > r$.

[9]Theorem (91) on page 31.

Proof. T^*T has rank r according to Theorem 142, and positive semidefinite by Theorem 228. So there is an orthonormal basis v_i for V consisting of eigenvectors of T^*T with corresponding eigenvalues λ_i where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and $\lambda_i = 0$ for $i > r$. For $1 \leq i \leq r$, define $\sigma_i = \sqrt{\lambda_i}$ and $u_i = \frac{1}{\sigma_i}T(v_i)$. We have:

$$\langle u_i, u_j \rangle = \left\langle \frac{1}{\sigma_i}T(v_i), \frac{1}{\sigma_j}T(v_j) \right\rangle = \frac{1}{\sigma_i\sigma_j}\langle T^*T(v_i), v_j \rangle = \frac{1}{\sigma_i\sigma_j}\langle \lambda_i v_i, v_j \rangle = \frac{\sigma_i^2}{\sigma_i\sigma_j}\langle v_i, v_j \rangle = \delta_{ij}$$

So $\{u_1, u_2, \ldots, u_r\}$ are orthogonal. Because the choice of $\sqrt{\lambda_i}$, they are unitary and therefore orthonormal. Extend it to an orthonormal basis $\{u_1, u_2, \ldots, u_m\}$. □

Definition 180. *The singular values of A is the singular value of L_A.*

Theorem 248 (Singular Value Decomposition Theorem)**.** *Let $A_{m\times n}$ be of rank r with positive singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, and let $\Sigma_{m\times n}$ be*

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 \end{cases} \tag{3.78}$$

Then there exists singular value decomposition that with $U_{m\times m}$ and $V_{n\times n}$, we have

$$A = U\Sigma V^* \tag{3.79}$$

The process to find singular value decomposition is:

1. *find singular value of A by calculating the eigenvalue of A^*A.*
2. *sort the singular value from big to small.*
3. *for non-zero singular value σ_i, put $\sqrt{\sigma_i}$ to the i-th diagonal of Σ.*
4. *form U of normalized eigenvector of A^*A.*
5. *for non-zero singular value σ_i, calculate orthonormal vector $u_i = \frac{1}{\sigma_i}L_A(v_i)$.*
6. *expand the u_i to orthonormal basis and form V.*

3.7.8 Polar Decomposition

Theorem 249 (Polar Decomposition)**.** *Any square matrix A, there exists a Polar Decomposition using unitary matrix W and a positive semidefinite matrix P that*

$$A = WP \tag{3.80}$$

If A is invertible, the Polar Decomposition is unique.

Proof. Use singular value decomposition on A and we get $A = U\Sigma V^* = UV^*V\Sigma V^* = (UV^*)(V\Sigma V^*) = WP$. So let $W = UV^*$ and $P = V\Sigma V^*$. □

3.7.9 Pseudoinverse

Definition 181. *Let $T : V \to W$ be a linear transformation. Let $L : \mathcal{N}(T)^\perp \to \mathcal{R}(T)$ be a linear transformation that $\forall x \in \mathcal{N}(T)^\perp$, $L(x) = T(x)$. The pseudoinverse (or Moore-Penrose generalised inverse) of T is a unique linear transformation from W to V that*

$$T^\dagger(y) = \begin{cases} L^{-1}(y) & \text{for } y \in \mathcal{R}(T) \\ 0 & \text{for } y \in \mathcal{R}(T)^\perp \end{cases} \tag{3.81}$$

Let $\{v_1, v_2, \ldots, v_r\}$ be a basis for $\mathcal{N}(T)^\perp$, $\{v_{r+1}, v_{r+2}, \ldots, v_n\}$ be a basis for $\mathcal{N}(T)$, $\{u_1, u_2, \ldots, u_r\}$ be basis for $\mathcal{R}(T)$, $\{u_{r_1}, u_{r+2}, \ldots, u_m\}$ be a basis for $\mathcal{R}(T)^\perp$, then:

$$T^\dagger(u_i) = \begin{cases} \frac{1}{\sigma_i}v_i & \text{if } 1 \leq i \leq r \\ 0 \end{cases}$$

So although not all T has inverse, the restriction $T|_{\mathcal{N}(T)^\perp}$ could have proper inverse.

Theorem 250. *Let $A_{m\times n}$ be a square matrix of rank r with singular value decomposition $A = U\Sigma V^*$ and non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. Let $\Sigma^\dagger_{m\times n}$ be a matrix that*

$$\Sigma^\dagger_{ij} = \begin{cases} \frac{1}{\sigma_i} & \text{if } i = j \leq r \\ 0 \end{cases} \tag{3.82}$$

Then $A^\dagger = V\Sigma^\dagger U^*$ is a singular value decomposition of A.

Theorem 251. *Let $T : V \to W$ be a linear transformation, then*

1. *$T^\dagger T$ is the orthogonal projection of V on $\mathcal{N}(T)^\perp$.*
2. *$TT^\dagger$ is the orthogonal projection of W on $\mathcal{R}(T)$.*

Proof. Define $L : \mathcal{N}(T)^\perp \to W$ by $L(x) = T(x)$. If $x \in \mathcal{N}(T)^\perp$, then $T^\dagger T(x) = L^{-1}L(x) = x$. If $x \in \mathcal{N}(T)$, then $T^\dagger T(x) = T^\dagger(0) = 0$. □

Theorem 252. *For a system of linear equations $Ax = b$. If $z = A^\dagger b$, then*

1. *If $Ax = b$ is consistent, then z is the unique solution with minimal norm.*
2. *If $Ax = b$ is inconsistent, then z is the best approximation: $\forall y, \|Ax - b\| \leq \|Ay - b\|$. Also if $Az = Ay$, then $\|z\| \leq \|y\|$.*

$A^\dagger b$ is the optimal solution discussed in section 3.6.4 on page 53.

Proof. Let $z = A^\dagger b$. If the equation is consistent, then $b \in \mathcal{R}(T)$, then $Az = AA^\dagger b = TT^\dagger(b) = b$ because $TT^\dagger$ is a orthogonal projection, so z is a solution to the linear system.

If y is any solution, then $T^\dagger T(y) = A^\dagger Ay = A^\dagger b = z$. So z is a orthogonal projection of y on $\mathcal{N}(T)^\perp$. So $\|z\| \leq \|y\|$.

If the equation is inconsistent, then $Az = AA^\dagger b$ is the orthogonal projection of b on $\mathcal{R}(T)$, so Az is the nearest vector to b. □

3.7.10 Conditioning

Definition 182. *For $Ax = b$, if a small change to A and b cause small change to x, the property is called well-conditioned. Otherwise the system is ill-conditioned.*

Definition 183. *The relative change in b is $\dfrac{\|db\|}{\|b\|}$ with $\|\cdot\|$ be the standard norm on $\mathcal{C}^n$.*

Definition 184. *The Euclidean norm of square matrix A is*

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{3.83}$$

Definition 185. *Let B be a self-adjoint matrix. The Rayleigh quotient for $x \neq 0$ is $R(x) = \dfrac{\langle Bx, x\rangle}{\|x\|^2}$*

Theorem 253. *For a self-adjoint matrix B, the $\max\limits_{x\neq 0} R(x)$ is the largest eigenvalue of B and $\min\limits_{x \neq 0} R(x)$ is the smallest eigenvalue of B.*

Proof. Choose the orthonormal basis v_i of B such that $Bv_i = \lambda_i v_i$ where $\lambda_1 \geq \lambda_2 \geq \lambda_n$. $\forall x \in F^n$, $\exists a_i$ that $x = \sum\limits_{i=1}^{n} a_i v_i$. So

$$R(x) = \frac{\langle Bx, x\rangle}{\|x\|^2} = \frac{\left\langle \sum\limits_{i=1}^{n} a_i \lambda_i v_i, \sum\limits_{j=1}^{n} a_j v_j \right\rangle}{\|x\|^2} = \frac{\sum_{i=1}^{n} \lambda_i |a_i|^2}{\|x\|^2} \leq \frac{\lambda_1 \sum_{i=1}^{n} |a_i|^2}{\|x\|^2} = \frac{\lambda_1 \|x\|^2}{\|x\|^2} = \lambda_1$$

□

Theorem 254. *$\|A\| = \sqrt{\lambda}$ where λ is the largest eigenvalue of A^*A.*

Theorem 255. *λ is an eigenvalue of A^*A if and only if λ is an eigenvalue of AA^*.*

Theorem 256. *Let A be invertible matrix. Then $\|A^{-1}\| = \dfrac{1}{\sqrt{\lambda}}$ where λ is the smallest eigenvalue of A^*A.*

Definition 186. *$\|A\| \times \|A^{-1}\|$ is the condition number of A and denoted as $cond(A)$.*

Theorem 257. *For system $Ax = b$ where A is invertible and $b \neq 0$, we have:*

1. *For any norm $\|\cdot\|$, we have $\dfrac{1}{cond(A)} \dfrac{\|db\|}{\|b\|} \leq \dfrac{\|dx\|}{\|x\|} \leq cond(A) \dfrac{\|db\|}{\|b\|}$.*
2. *If $\|\cdot\|$ is the Euclidean norm, then $cond(A) = \sqrt{\dfrac{\lambda_1}{\lambda_n}}$ where λ_1 and λ_n are the largest and smallest eigenvalue of A^*A.*

So when cond$(b) \geq 1$. *If cond*(b) *is close to* 1, *the relative error in* x *is small when relative error of* b *is small. However when cond*(b) *is large, the relative error in* x *could be large or small.*

cond(x) *is seldom calculated because when calculating* A^{-1} *in computer, there are rounding errors which is related to cond*(A).

3.8 Matrix Calculus

3.8.1 Layout

There are two different layout:

- numerator layout:

$$\begin{bmatrix} \nabla f \\ \nabla g \end{bmatrix} \tag{3.84}$$

- denominator layout:

$$[\nabla f, \nabla g] \tag{3.85}$$

numerator layout is preferred.

3.8.2 Jacobian Matrix

for $\mathbf{y}_{1\times m} = \mathbf{f}(\mathbf{x}_{1\times n})$, its Jacobian matrix is:

$$\nabla_{\mathbf{x}}\mathbf{y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_1(\mathbf{x}) \\ \vdots \\ \nabla f_m(\mathbf{x}) \end{bmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \vdots \\ \frac{\partial f_m}{\partial x} \end{pmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{x_1} & \frac{\partial f_1(\mathbf{x})}{x_2} & \dots & \frac{\partial f_1(\mathbf{x})}{x_n} \\ \frac{\partial f_2(\mathbf{x})}{x_1} & \frac{\partial f_2(\mathbf{x})}{x_2} & \dots & \frac{\partial f_2(\mathbf{x})}{x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{x_1} & \frac{\partial f_m(\mathbf{x})}{x_2} & \dots & \frac{\partial f_m(\mathbf{x})}{x_n} \end{bmatrix} \tag{3.86}$$

3.8.3 Element-wise binary operator

for element-wise binary operator

$$\mathbf{y} = \mathbf{f}(\mathbf{w}) \bigcirc \mathbf{g}(\mathbf{x}) \tag{3.87}$$

$\bigcirc$ could be $+, -, \times$[10]$, \div, max$. The gradient is:

$$\nabla_{\mathbf{x}}\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x}) \\ f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x}) \end{bmatrix} \tag{3.88}$$

The expanded matrix could be differentiated using Jacobian matrix.

3.8.4 Vector Sum

Vector sum operation sum could be expressed as

$$y = \text{sum}\Big(\mathbf{f}(\mathbf{x})\Big) = \sum_{i=1}^{n} f_i(\mathbf{x}) \tag{3.89}$$

$\nabla \mathbf{y}$ could be calculated as usual.

3.8.5 Chain Rules

In machine learning there are two ways of taking chain rules:

- forward differentiation: $\frac{dy}{dx} = \frac{du}{dx} \times \frac{dy}{du}$
- backward differentiation: $\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$

Backward differentiation is preferred for matrix operation.
The full expression of $\mathbf{y} = \mathbf{f}(\mathbf{g}(\mathbf{x}))$ is:

[10] called *hadamard product*

$$
\begin{aligned}
\nabla_{\mathbf{x}} f &= \frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} \\
&= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \times \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \\
&= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{g_1} & \frac{\partial f_1(\mathbf{x})}{g_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{g_n} \\ \frac{\partial f_2(\mathbf{x})}{g_1} & \frac{\partial f_2(\mathbf{x})}{g_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{g_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{g_1} & \frac{\partial f_m(\mathbf{x})}{g_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{g_n} \end{bmatrix}_{m \times n} \times \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{x_1} & \frac{\partial g_1(\mathbf{x})}{x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{x_r} \\ \frac{\partial g_2(\mathbf{x})}{x_1} & \frac{\partial g_2(\mathbf{x})}{x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{x_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n(\mathbf{x})}{x_1} & \frac{\partial g_n(\mathbf{x})}{x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{x_r} \end{bmatrix}_{n \times r}
\end{aligned} \tag{3.90}
$$