Summarizing the Essentials

January 12, 2024

Contents

Summarizing the Essentials

1 Structure of Vector Space

- The fundamental notion that emerges naturally from the *axioms of vector spaces* is the concept of **linear combinations** (of vectors), from which the following fundamental relations among vectors naturally follow
  - **Linearly independent:** *"linear combinations* $= 0$ *"* has only the **trivial** solution.
  - **Linearly dependent:** *"linear combinations* $= 0$*"* has **non trivial** solution.
- A set of vectors in linearly dependent if and only if one of them can be expressed in terms of the others.
  - **Basis:** A set of linearly independent vectors such that all vectors can be written as linear combinations of them.
  - **Dimension**: The number of vectors in any basis.
- **Dimension Formula:** For U, W two subspaces of V, then $U + W$ and $U \cap W$ are subspaces and we have

$$dim\,(U + W) = dim\,U + dim\,W - dim\,(U \cap W)$$

  - This is compatible with the general **Inclusion-Exclusion Principal** (容斥原理) and can be generalized accordingly, for example

$$dim\,(U_1 + U_2 + U_3) = dim\,U_1 + dim\,U_2 + dim\,U_3 -$$

$$-dim\,(U_1 \cap U_2) - dim\,(U_1 \cap U_3) - dim\,(U_2 \cap U_3) + dim\,(U_1 \cap U_2 \cap U_3)$$

- Any $n + 1$ vectors in a n-dimensional vector space must be linear independent.
- **Basis Extension:** Any set of linearly independent vectors in V can be extended to a basis of V.
  - If V is endowed with an **inner product**, so that the notion of **orthogonality** makes sense, then we have the concept of **orthonormal basis**. Then any

linearly independent set can be transformed into an orthogonal set by by applying the **Gram-Schmidt process**.

  - In particularly, any inner product space has orthogonal basis. And any orthogonal set can be extended to a orthonormal basis.
  - A set of orthogonal vectors must be linearly independent.

- Gram-Schmidt process is based on the notion of **orthogonal projection**:

$$Proj_{\mathbf{w}}\mathbf{v} := \frac{\langle \mathbf{v},\, \mathbf{w}\rangle}{||\mathbf{w}||^2}\mathbf{w}$$

  - In general, if W is a subspace of V that has orthogonal basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$, then for any $\mathbf{v} \in V$, its orthogonal projection onto subspace W is given by

$$Proj_W\mathbf{v} = \frac{\langle \mathbf{v},\, \mathbf{w}_1\rangle}{||\mathbf{w}_1||^2}\mathbf{w}_1 + \cdots + \frac{\langle \mathbf{v},\, \mathbf{w}_k\rangle}{||\mathbf{w}_k||^2}\mathbf{w}_k$$

  - In particularly, if $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$ is **orthonormal**, then we have

$$Proj_W\mathbf{v} = \langle \mathbf{v},\, \mathbf{w}_1\rangle\mathbf{w}_1 + \cdots + \langle \mathbf{v},\, \mathbf{w}_k\rangle\mathbf{w}_k$$

- Inner product structure induces **normal (distance) function**: $||\mathbf{v}||^2 = \langle \mathbf{v},\, \mathbf{v}\rangle$. Conversely, any normal induces an inner product via **polarization identity.**

  - **Real case:**

$$\langle \mathbf{u},\, \mathbf{v}\rangle = \frac{||\mathbf{u}+\mathbf{v}||^2 - ||\mathbf{u}-\mathbf{v}||^2}{4}$$

  - **Complex case:**

$$\langle \mathbf{u},\, \mathbf{v}\rangle = \frac{||\mathbf{u}+\mathbf{v}||^2 - ||\mathbf{u}-\mathbf{v}||^2 + i||\mathbf{u}+i\mathbf{v}||^2 - i||\mathbf{u}-i\mathbf{v}||^2}{4}$$

- Norm and inner product enjoy the following properties

  - **Triangle inequality:** $||\mathbf{u}+\mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$.
  - **Pythagoras' Law:** If $\mathbf{u} \perp \mathbf{v}$, then $||\mathbf{u}+\mathbf{v}|| = ||\mathbf{u}|| + ||\mathbf{v}||$.
  - **Cauchy-Schwarz inequality:** $|\langle u,\, \mathbf{v}\rangle| \leq ||\mathbf{u}||\mathbf{v}||$.

- For complex inner product, it is **conjugate linear** in the 2^{nd} slot, i.e.,

$$\langle \mathbf{u},\, k\mathbf{v} + l\mathbf{w}\rangle = \bar{k}\langle \mathbf{u},\, \mathbf{v}\rangle + \bar{l}\langle \mathbf{u},\, \mathbf{w}\rangle$$

which is equivalent to $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$. In particularly, $\langle \mathbf{u}, \mathbf{u} \rangle \in \mathbb{R}$. So that $||\mathbf{u}||^2 = \langle \mathbf{u}, \mathbf{u} \rangle$ holds in both cases.

- Let $U^{\perp}$ the **orthogonal complement** of U in V, then $dim\, U + dim\, U^{\perp} = dim\, V$.

2 Linear Transformations and Linear Operators

- A **linear transformation** T between V and W is a map $T : V \to W$ that is **linear**, i.e., *keeps the linear combinations of vectors in V and W*. That is

$$T(k\mathbf{v} + l\mathbf{w}) = kT(\mathbf{v}) + lT(\mathbf{w}) \qquad \textit{for } \mathbf{v} \in V, \mathbf{w} \in W \textit{ and } \forall\, k, l$$

  - A linear transformation form V to V is called a **linear operator**.
  - Linear transformations can be composed, $T_1 \circ T_2(\mathbf{v}) := T_1(T_2(\mathbf{v}))$. This is possible if $Rang\,(T_2) \subset Domain\,(T_1)$.
  - Linear transformation can be **inverted** only if it is one-to-one.

- Linear transformation $T : V \to W$ gives a relation between V and W.
  - If T is both *one to one* and *onto*, then we call T establishes an **isomorphism** between V and W, and denoted by $V \cong W$.
  - Isomorphic vector spaces have the same dimension. And conversely, if $dim\, V = dim\, W$, then $V \cong W$. Indeed, suppose V and W have basis given respectively by $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$. Then the linear transformation given by $T(k_1\mathbf{v}_1 + \cdots + k_n\mathbf{v}_n) := k_1\mathbf{w}_1 + \cdots + k_n\mathbf{w}_n$ establishes an isomorphism $V \cong W$.

- The extent that a general linear transformation $T : V \to W$ fails to be an isomorphism is measured by its **kernel** and **range**.
  - $ker\,(T) := \{\mathbf{v} \in V : T(\mathbf{v}) = \mathbf{0}\} \subset V$. T is one-to-one if and only if $ker\,(T) = \{\mathbf{0}\}$. So, the larger the size of *ker*, the further for T from being an **injection** (a.k.a one to one).
  - $Rang\,(T) := \{\mathbf{w} \in W : \mathbf{w} = T\,(\mathbf{v}) \textit{ for some } \mathbf{v} \in V\} \subset W$. T is onto if and only if $Range\,(T) = W$. So, the smaller the size of *range*, the further for T from being an **surjection** (a.k.a onto).

- **Fundamental identity:** $dim\, ker\,(T) + dim\, Range\,(T) = dim\,(V)$.

3 Matrix as "Coordinates"

- We can *coordinalize* a vector space V by choosing a basis $B := \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, then we have the **coordinate map** which gives the following isomporphism

$$[\ ]_B : V \xrightarrow{\cong} \mathbb{R}^n \qquad \mathbf{v} = k_1\mathbf{v}_1 + \cdots + k_n\mathbf{v}_n \longmapsto [\mathbf{v}]_B := \begin{bmatrix} k_1 \\ \vdots \\ k_n \end{bmatrix}$$

- Different basis of V are related by **invertible matrix**. If $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $B' = \{\mathbf{v}'_1, \cdots, \mathbf{v}'_n\}$ are two basis of V. Then the **transition matrix** from B to B', which is denoted by $P_{B' \leftarrow B}$, is defined to be

$$P_{B' \leftarrow B} := [[\mathbf{v}_1]_{B'} \mid \cdots \mid [\mathbf{v}_n]_{B'}] \quad \Longleftrightarrow \quad B = B' P_{B' \leftarrow B}$$

  Then for $\mathbf{v} \in V$, we have the **coordinate transformation formula**

$$[\mathbf{v}]_{B'} = P_{B' \leftarrow B}[\mathbf{v}]_B$$

- Transition matrix has the following properties
  - $P_{B \leftarrow B'} = P^{-1}_{B' \leftarrow B}$
  - $P_{B \leftarrow B'} P_{B' \leftarrow B''} = P_{B \leftarrow B''}$.
  - If S is the standard basis for $\mathbb{R}^n$, then for any basis $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, we have

$$P_{S \leftarrow B} = [\mathbf{v}_1 \mid \cdots \mid \mathbf{v}_n] =: B$$

  - For any two basis B and B' of $\mathbb{R}^n$, then above properties implies

$$P_{B' \leftarrow B} = P_{B' \leftarrow S} P_{S \leftarrow B} = P^{-1}_{S \leftarrow B'} P_{S \leftarrow B} = (B')^{-1} \cdot B$$

- Given $T : V \to W$, we can coordinalize T as follows
  - Choosing basis $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $B' = \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ of V and W respectively. Then the *matrix (representation)* for $T : V \longrightarrow W$ *relative to the basis* B and B' is defined to be

$$[T]_{B'B} = \left[\ [T(\mathbf{v}_1)]_{B'} \,\middle|\, [T(\mathbf{v}_2)]_{B'} \,\middle|\, \cdots \,\middle|\, [T(\mathbf{v}_n)]_{B'}\ \right]$$

- In terms of coordinate representations of vectors in V and W via the basis B and B' respectively, the linear transformation $T : V \longrightarrow W$ can be represented as a matrix transformation
$$[T(\mathbf{x})]_{B'} = [T]_{B'B} \cdot [\mathbf{x}]_B$$

- In other words, **at the level of coordinate representation,** $T : V \to W$ **is represented as** $T_A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$**, where** $A = [T]_{B'B}$ **is the matrix of** T **relative to the basis** B **and** B'. Graphically, we have the following

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[d, "\cong"'] & W \arrow[d, "\cong"] \\ \mathbb{R}^n \arrow[r, "T_A"'] & \mathbb{R}^m \end{tikzcd} \quad \rightsquigarrow \quad \begin{tikzcd} V \ni \mathbf{x} \arrow[r, "T"] \arrow[d, "{[\ ]_B}"'] & T(\mathbf{x}) \in W \arrow[d, "{[\ ]_{B'}}"] \\ \mathbb{R}^n \ni [\mathbf{x}]_B \arrow[r, "T_A"'] & {[T(\mathbf{x})]_{B'} \in \mathbb{R}^m} \end{tikzcd}$$

Thus, if we know the coordinate of $T(\mathbf{x})$, namely $[T(\mathbf{x})]_{B'}$, we can "recover" $T(\mathbf{x})$ from the inverse of the *coordinate map* $[\ ]_{B'}$ as follows

$$\boxed{T = [\ ]_{B'}^{-1} \circ T_A \circ [\ ]_B}$$

- If $V = W$, $T : V \to V$ is represented by a *square matrix* relative to a basis B of V. In this case, we write $[T]_B$ instead of $[T]_{BB}$.
- If $T : V \to V$ is invertible, then for any basis B of V, the matrix of T relative to it is also invertible, and $[T^{-1}]_B = [T]_B^{-1}$.
- The composition of linear transformations corresponds to the multiplication of the corresponding matrices.

$$\begin{tikzcd} V_1 \arrow[r, "T_1"] \arrow[d, "\cong"'] \arrow[rr, bend left, "T_2 \circ T_1"] & V_2 \arrow[r, "T_2"] \arrow[d, "\cong"'] & V_3 \arrow[d, "\cong"'] \\ \mathbb{R}^{dim\, V_1} \arrow[r, "T_{A_1}"] \arrow[rr, bend right, "T_{A_2} \circ T_{A_1} = T_{A_2 \circ A_1}"'] & \mathbb{R}^{dim\, V_2} \arrow[r, "T_{A_2}"] & \mathbb{R}^{dim\, V_3} \end{tikzcd}$$

- For a linear operator $T : V \to V$. Two matrix representations $[T]_B$ and $[T]_{B'}$ are related by **similarity relation**.

$$[T]_{B'} = P_{B \leftarrow B'}^{-1} [T]_B P_{B \leftarrow B'} = P_{B' \leftarrow B} [T]_B P_{B \leftarrow B'}$$

- If a real vector space V is endowed with an inner product structure, and we consider orthonormal basis with respect to it. Then the transition matrix between two orthonormal basis is **orthogonal**. i.e.,

$$A^T = A^{-1} \quad \text{or} \quad AA^T = A^T A = I.$$

  - The rows and columns of an orthogonal matrix form an orthonormal basis.
  - The product of two orthogonal matrix is orthogonal.
  - The inverse of an orthogonal matrix is orthogonal.

- If a complex vector space V is endowed with an inner product structure, and we consider orthonormal basis with respect to it. Then the transition matrix between two orthonormal basis is **unitary**. i.e.,

$$A^* := \overline{A}^T = A^{-1} \quad \text{or} \quad AA^* = A^* A = I.$$

  - The rows and columns of an unitary matrix form an orthonormal basis.
  - The product of two unitary matrix is orthogonal.
  - The inverse of an unitary matrix is unitary.

- A is *orthogonal* (*unitary*) if and only if A *keeps the dot product*, i.e.,

$$A\mathbf{v} \bullet A\mathbf{w} = \mathbf{v} \bullet \mathbf{w}$$

  if and only if A *keeps the norm*, i.e.,

$$||A\mathbf{v}|| = ||\mathbf{v}||$$

4 Similarity and Diagonalization

- For a linear operator $T : V \to V$, which is represented by $A = [T]_B$. If we have another basis $B' = \{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ such that $[T]_{B'} = D = diag\{\lambda_1, \cdots, \lambda_n\}$. Then we call the linear operator T (or the corresponding matrix A) is **diagonalizable**.

  - This means $T\mathbf{w}_i = \lambda_i \mathbf{w}_i \quad \forall i = 1, \cdots, n$. And we have

$$D = [T]_{B'} = P_{B' \leftarrow B}[T]_B P_{B \leftarrow B'} = P_{B \leftarrow B'}^{-1} A P_{B \leftarrow B'}$$

Denote by $P = P_{B \leftarrow B'}$ the transition matrix, then the above becomes

$$D = P^{-1}AP$$

- In particularly, given $A \in M_{n\times n}$, we view it as matrix transformation $T_A : \mathbb{R}^n \to \mathbb{R}^n$. Then we see that
  - A is itself the matrix representation of T_A relative to the standard basis $S = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$.
  - If A is diagonalizable, then $\exists$ new basis $P = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for some scalars λ_i, $i = 1, \cdots, n$.
  - Denote by $P = [\mathbf{v}_1 \,|\cdots|\, \mathbf{v}_n]$, then $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ means $AP = PD$ for $D = diag\{\lambda_1, \cdots, \lambda_n\}$. That is $D = P^{-1}AP$ where $P = P_{S\leftarrow B}$.
- **Conditions of diagonalizability (for both real and complex matrices)**:
  - $A \in \mathbb{M}_{n\times n}$ is diagonalizable if and only if A has n linearly independent eigenvectors.
  - A is diagonalizable if and only if the sum of dimensions of distinct eigenspaces (a.k.a geometric multiplicities) is equal to n.
  - A is diagonalizable if and only if for each eigenvalue, its algebraic multiplicity equals its geometric multiplicity.
    - In particularly, if A has n distinct eigenvalues, i.e., if all eigenvalues of A are distinct (all eigenvalues having algebraic multiplicity 1), then A is diagonalizable.
    - Geometric multiplicity $\leq$ Algebraic multiplicity.
- **Methods of diaonalizability**:
  - First, find the eigenvalues of A by computing the *characteristic polynomial*

$$P_A(\lambda) = det\,(\lambda I - A) = 0$$

  - Then for each eigenvalue λ, find a basis for the corresponding eigenspace

$$E_\lambda = \{\mathbf{v} : \, A\mathbf{v} = \lambda\mathbf{v}\} = null\,(\lambda I - A).$$

    by solving the corresponding linear system $(\lambda I - A)\mathbf{x} = \mathbf{0}$.

- Putting all n basis vectors for all eigenspaces together to form the transition matrix P (*from the basis formed by these basis vectors to the standard basis*).
- And P can diagonalize A in the sense $P^{-1}AP = D$ where D is diagonal with diagonal entries being the eigenvalues (counted with multiplicity) that are so arranged in order of the corresponding eigenvectors in P.

- **Orthogonal and Unitary Diagonalization**

  - If we can find an *orthogonal* P such that $P^T AP = D$, we say that A is **orthogonally diagonalizable**. This amounts saying we can choose an orthonormal basis $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ for $\mathbb{R}^n$ such that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for some scalars λ_i.
    * A real matrix A is orthogonally diagonalizable if and only if A is **symmetric**, i.e., $A^T = A$.
    * For symmetric A, all its eigenvalues are real, and the eigenvectors corresponding to different eigenvalues are orthogonal.
  - If we can find an *unitary* P such that $P^*AP = D$, we say that A is **unitarily diagonalizable**. This amounts saying we can choose an orthonormal basis $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ for $\mathbb{C}^n$ such that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for some scalars λ_i.
    * A complex matrix A is unitarily diagonalizable if and only if A is **Herimitian**, i.e., $A^* = A$.
    * For Hermitian A, all its eigenvalues are real, and the eigenvectors corresponding to different eigenvalues are orthogonal.

- **Methods of Orthogonal (Unitary) Diagonalization:**

  - First, find all eigenvalues of A by computing the characteristic polynomial.
  - Find a basis for each eigenspace. It must be an orthogonal basis.
  - Putting the basis vectors for all distinct eigenspaces together, and apply the *Gram-Schmidt process* to get an orthonormal basis.
  - Using these basis as columns of an invertible P (which must be orthogonal in the real case, and unitary in the complex case).
  - P computed as above will orthogonally (unitarily) diagonalize A.

5 Matrix Seen In Itself

- For $A \in \mathbb{M}_{m\times n}$, the space spanned by its rows (columns) is called the *row (column) space of* A, denoted by $row\,(A)$ and $col\,(A)$ respectively.
- $dim\,row\,(A) = dim\,col\,(A) =: rank\,(A)$.
- A is invertible if and only if it is of full rank.
- Elementary row and column operations do not alter the rank of a matrix.
  - If R is a reduced row echelon form of A, then

    $$rank\,(A) = rank\,(R) = number\ of\ nonzero\ rows\ in\ R = number\ of\ leading\ 1\text{'s}\ in\ R$$

  - If $rank\,(A) = r$ then $\exists$ invertible P and Q such that $PAQ = \begin{bmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$
- $A\mathbf{x} = \mathbf{b}$ is solvable if and only if $ranl\,(A) = rank\,(A|\mathbf{b})$; if and only if $\mathbf{b} \in col\,(A)$.
- For any matrix A, $rank\,(A) = rank\,(A^T) = rank\,(AA^T) = rank\,(A^TA)$.
- $row\,(A)^\perp = null\,(A)$ and $col\,(A)^\perp = null\,(A^T)$
- $nullity\,(A) := dim\,null\,(A) = dim\,\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$.
- **Fundamental relation:** For $A \in \mathbb{M}_{m\times n}$, we have

  $$nullity\,(A) + rank\,(A) = n$$

  - Viewing A as matrix transformation $T_A : \mathbb{R}^n \to \mathbb{R}^m$, the above relation is the same as $dim\,ker\,(T_A) + dim\,Range\,(T_A) = n$.
- **QR-decomposition**: If A is an $m \times n$ matrix with linearly independent column vectors, i.e., *full column rank (列满秩)* then A can be factored as $A = QR$ where Q is an $m \times n$ matrix with **orthonormal** column vectors, and R is an $n \times n$ invertible upper triangular matrix. In particularly, every invertible matrix has a QR-decomposition.
- If $A \in \mathbb{M}_{m\times n}$, and if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of A^TA, then the numbers

  $$\sigma_1 = \sqrt{\lambda_1},\ \sigma_2 = \sqrt{\lambda_2},\ \cdots,\ \sigma_n = \sqrt{\lambda_n}$$

are called the **(singular values)** 奇异值 of A.

- **Singular Value Decomposition (SVD) Expanded Form:** If A is an $m \times n$ matrix of rank k, then A can be factored as

$$A = U\Sigma V^T = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_k \,|\, \mathbf{u}_{k+1}\ \cdots\ \mathbf{u}_m] \left[\begin{array}{cccc|c} \sigma_1 & 0 & \cdots & 0 & \\ 0 & \sigma_2 & \cdots & 0 & \mathbf{0}_{k\times(n-k)} \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & \sigma_k & \\ \hline & \mathbf{0}_{(m-k)\times k} & & & \mathbf{0}_{(m-k)\times(n-k)} \end{array}\right] \left[\begin{array}{c} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_k^T \\ \hline \mathbf{v}_{k+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{array}\right]$$

in which U, Σ and V have sizes $m \times m$, $m \times n$ and $n \times n$ respectively, and in which:

a). $V = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n]$ orthogonally diagonalizes A^TA.

b). The nonzero diagonal entries of Σ are $\sigma_1 = \sqrt{\lambda_1}$, $\sigma_2 = \sqrt{\lambda_2}$, $\cdots$, $\sigma_k = \sqrt{\lambda_k}$, where $\lambda_1, \lambda_2, \cdots, \lambda_k$ are the nonzero eigenvalues of A^TA corresponding to the column vectors of V.

c). The column vectors of V are ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$.

d). $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||} = \frac{1}{\sigma_i} A\mathbf{v}_i \quad i = 1, 2, \cdots, k$

e). $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k\}$ is an orthonormal basis for $col\,(A)$.

f). $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\}$ is an extension of $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k\}$ to an orthonormal basis for $\mathbb{R}^m$.

  - The **reduced singular value decompositon** of A reads

$$A = \underbrace{[\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_k]}_{U_1} \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{bmatrix}}_{\Sigma_1} \underbrace{\begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix}}_{V_1^T}$$

  In this form, the sizes of U_1, Σ_1 and V_1^T are $m\times k$, $k\times k$ and $k\times n$ respectively. And Σ_1 is invertible. When expanding, the above can be further written as

$$A = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T$$

which is called a **reduced singular value expansion** of A. This applies to **all** matrices, whereas the *spectral decomposition* applies only to symmetric matrices.

* **Spectral decomposition for symmetric matrix:** Given $A \in \mathbb{M}_{n\times n}(\mathbb{R})$ that is symmetric, we know that $\exists$ *orthogonal* $P = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_n]$ that diagonalizes A, i.e.,

$$A = P^T DP = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_n] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix} =$$

$$= [\lambda_1 \mathbf{u}_1 \ \cdots \ \lambda_n \mathbf{u}_n] \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \cdots + \lambda_n \mathbf{u}_n \mathbf{u}_n^T$$

6 Optimization by Least Square Method

The **least squares problem:** $\min_{\forall\, \mathbf{x}\in\mathbb{R}^n}\{||\mathbf{b} - A\mathbf{x}||\}$ admits solutions that are given by $\hat{\mathbf{x}}$ such that

$$A\hat{\mathbf{x}} = proj_{col\,(A)}\mathbf{b}. \qquad \Longrightarrow \qquad ||\mathbf{b} - A\hat{\mathbf{x}}|| = \min_{\forall\, \mathbf{x}\in\mathbb{R}^n}\{||\mathbf{b} - A\mathbf{x}||\}$$

Notice that such $\hat{\mathbf{x}}$, when exists, may not be unique, and any such solution is called a **least squares solution** of $A\mathbf{x} = \mathbf{b}$. Each such solution $\hat{\mathbf{x}}$ has the same error vector and thus the same least squares error.

Solving the above least squares problem is equivalent to solving the so called *normal equation:* $A^T A\mathbf{x} = A^T\mathbf{b}$. Notice that the normal equation is always consistent even if $A\mathbf{x} = \mathbf{b}$ is inconsistent. Indeed, this establishes an *one to one* correspondence between the following solution sets

$$\left\{ \begin{array}{c} \textit{Solutions to} \\ A\mathbf{x} = proj_{col\,(A)}\mathbf{b} \end{array} \right\} \overset{1-1}{\longleftrightarrow} \left\{ \begin{array}{c} \textit{Solutions to} \\ A^T A\mathbf{x} = A^T\mathbf{b} \end{array} \right\}$$

And the **uniqueness** of least square solutions is equivalent to the **uniqueness** of solutions to the associated normal system.

- When A is of full column rank, i.e., the column vectors of A are linearly independent, then A^TA is invertible, and the normal equation $A^TA\mathbf{x} = A^T\mathbf{b}$ has **unique solution** $\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$, thus the least squares problem admits unique solution. In particular, the orthogonal projection $proj_{col(A)}\mathbf{b}$ is given by $A\hat{\mathbf{x}} = A(A^TA)^{-1}A^T\mathbf{b}$. In general, if $W \subseteq V$ is a subspace that can be described by $col\,(A)$ for some matrix A, then orthogonal projection $proj_W$ is given by $proj_W = A(A^TA)^{-1}A^T$
- When A is of full column rank, by applying Gram-Schmidt process to the set of column vectors followed by normalization, we will get the QR decomposition of A, i.e., $A = QR$, where Q is orthogonal and R is upper triangular, then for each $\mathbf{b} \in \mathbb{R}^m$ the system $A\mathbf{x} = \mathbf{b}$ has a unique least squares solution given by $\hat{\mathbf{x}} = R^{-1}Q^T\mathbf{b}$

Least Squares Lines of Best Fit: Suppose we want to fit a straight line $y = a + bx$ to the experimentally determined points

$$(x_1,\, y_1),\; (x_2,\, y_2),\, \cdots, (x_n,\, y_n)$$

Optimization Problem: $\min_{a,b}\left\{\sum_{i=1}^{n}[y_i - (a + bx_i)]^2 = \sum_{i=1}^{n} d_i^2\right\}$ where $d_i = |y_i - (a+bx_i)|$ measures the *error in* y_i *at the point* x_i, which are called the *residuals (残余)*.

To solve this problem, we first turn it into a linear algebra problem. If there's no error, we would have

$$\begin{cases} y_1 = a + bx_1 \\ y_2 = a + bx_2 \\ \qquad \vdots \\ y_n = a + bx_n \end{cases}$$

Denote by $\mathbf{y} = [y_1\; y_2\; \cdots\; y_n]^T,\; \mathbf{v} = [a\; b]^T$ and $M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ then the above system can be written as $M\mathbf{v} = \mathbf{y}$.

When the data $\{(x_i,\, y_i)\}$ are not co-linear (共线), then there is **NO** $\mathbf{v} = [a\; b]^T$ that solves $M\mathbf{v} = \mathbf{y}$.

However, we can seek the **least squares solutions** that always exist. The *least squares error* in this case is given by

$$e = ||\mathbf{y} - M\mathbf{v}|| = \left\| \begin{bmatrix} y_1 - (a + bx_1) \\ y_2 - (a + bx_2) \\ \vdots \\ y_n - (a + bx_n) \end{bmatrix} \right\| = \sqrt{\textit{sum of the squares of the data errors}}$$

Thus, the *least squares solution* of $M\mathbf{v} = \mathbf{y}$ will yield the *least squares line of best fit*. We consider the *normal equation* associated to $M\mathbf{v} = \mathbf{y}$.

$$M^T M\mathbf{v} = M^T\mathbf{y}$$

$$M^T M = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

So $det(M^T M) \neq 0$ when $\sigma^2 \neq 0$, i.e., there is error. In this case, the normal equation has unique solution, thus we have **unique** line of best fit given by

$$y = a^* + b^* x \qquad \textit{where} \qquad \mathbf{v}^* = \begin{bmatrix} a^* \\ b^* \end{bmatrix} = (M^T M)^{-1} M^T \mathbf{y}.$$

The above can be generalized to the case when we need to find a polynomial $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$ that best fit the data $\{(x_1, y_1), (x_2, y_n), \cdots, (x_n, y_n)\}$. That is, find $a_0, a_1, \cdots, a_m$ such that the following quantity is minimized

$$erro = \left\| \begin{bmatrix} y_1 - a_0 - a_1 x_1 - \cdots - a_m x_1^m \\ y_2 - a_0 - a_1 x_2 - \cdots - a_m x_2^m \\ \vdots \\ y_n - a_0 - a_1 x_n - \cdots - a_m x_n^m \end{bmatrix} \right\|$$

That is we need to find the *least squares solutions* to the following system

$$\begin{cases} a_0 + a_1 x_1 + \cdots + a_m x_1^m = y_1 \\ a_0 + a_1 x_2 + \cdots + a_m x_2^m = y_2 \\ \qquad\qquad \vdots \\ a_0 + a_1 x_n + \cdots + a_m x_n^m = y_n \end{cases} \iff \underbrace{\begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & \cdots & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}}$$

which is equivalent to solving the corresponding *normal equation*: $M^T M\mathbf{x} = M^T\mathbf{y}$.

7 SVD and the "Big Picture"

Given $A \in \mathbb{M}_{m\times n}(\mathbb{R})$. A can be viewed as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ by left multiplication; similarly A^T can be viewed as a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$.

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "T_A"] \arrow[d, equal] & \mathbb{R}^m \arrow[d, equal] \\ \mathbb{R}^n & \mathbb{R}^m \arrow[l, "T_{A^T}"] \end{tikzcd}$$

We know that the fundamental facts about the pair of maps T_A, T_{A^T} are given as follows:

$$dim\, ker\,(T_A) + dim\, R(T_A) = dim\,(\mathbb{R}^n) = n$$

$$dim\, ker\,(T_{A^T}) + dim\, R(T_{A^T}) = dim\,(\mathbb{R}^m) = m$$

In terms of the language of matrix, the above become

$$nullity\,(A) + rank\,(A) = n; \quad nullity\,(A^T) + rank\,(A^T) = m \quad (*)$$

Of course we know that $rank\,(A) = rank\,(A^T)$ (i.e., the row rank equals the column rank), that the above two equations can be combined into one identity (recall that we have called it the *"baby index theorem"*)

$$nullity\,(A) - nullity\,(A^T) = n - m$$

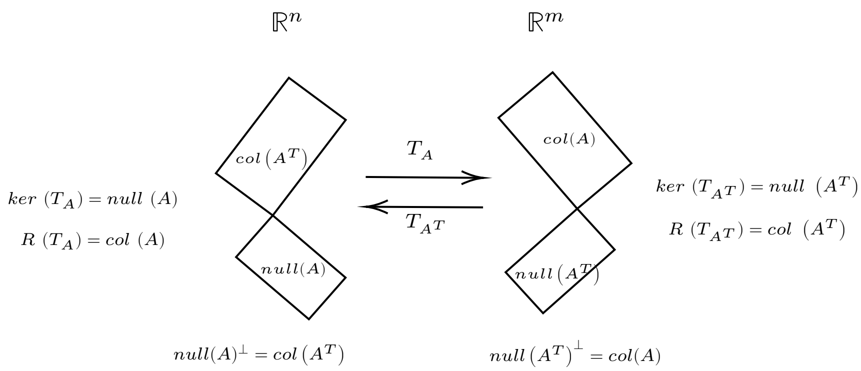
which gives the full information about the solvability of the linear system $A\mathbf{x} = \mathbf{b}$.

Next, we endow $\mathbb{R}^n$ and $\mathbb{R}^m$ with the standard inner product structure, then we have the notion of *orthogonal complement* of a subspace. And we know the following important facts:

$$null\,(A)^{\perp} = col\,(A^T); \qquad null\,(A^T)^{\perp} = col\,(A)$$

As we know that if $U \subseteq V$, then $dim\,U + dim\,U^{\perp} = dim\,V$, thus by applying "dim" on both sides of the above orthogonal relation, we get the fundamental relations $(*)$.

The above discussion can be "encoded" into the following picture, which we call "big picture".



In this framework, SVD for A turns to be an algorithm for computing all ingredients contained in the above "big picture".

$$A = U\Sigma V^T = [\mathbf{u}_1 \;\cdots\; \mathbf{u}_k \,|\, \mathbf{u}_{k+1} \;\cdots\; \mathbf{u}_m] \left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0}_{k\times(n-k)} \\ 0 & \cdots & \sigma_k & \\ \hline & \mathbf{0}_{(m-k)\times k} & & \mathbf{0}_{(m-k)\times(n-k)} \end{array}\right] \left[\begin{array}{c} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \\ \hline \mathbf{v}_{k+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{array}\right]$$

Then we claim that

- $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is an orthonormal basis for $col(A)$.
- $\{\mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\}$ is an orthonormal basis for $col(A)^{\perp} = null(A)$.
- $\{\mathbf{v}_{k+1}, \cdots, \mathbf{v}_n\}$ is an orthonormal basis for $null(A)$.
- $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is an orthonormal basis for $null(A)^{\perp} = col(A^T)$.

Explanation: Recall that $\{\mathbf{v}_1, \cdots, \mathbf{v}_k, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_n\}$ and $\mathbb{R}^n$ is an orthonormal basis for $\mathbb{R}^n$, and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\}$ is an orthonormal basis for $\mathbb{R}^m$ such that

$$A^T A\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i, i = 1, \cdots, k; \quad A^T A\mathbf{v}_j = \mathbf{0}, \; j = k+1, \cdots, n$$

As $null(A^T A) = null(A)$, we also have $A\mathbf{v}_j = \mathbf{0}, \; j = k+1, \cdots, n$.

Also recall that $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||} = \frac{A\mathbf{v}_i}{\sigma_i}, \; i = 1, \cdots, k$, and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k; \mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\}$ is an extension to an orthonormal basis for $\mathbb{R}^m$.

In particularly, we have $A\mathbf{v}_i = \sigma_i \mathbf{u}_i, \; i = 1, \cdots, k$. As $\sigma_j = 0$ for $j = k+1, \cdots, n$, and $A\mathbf{v}_j = \mathbf{0}$, $j = k+1, \cdots, n$, we can also write $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$, for all i.

Besides one can prove that $AA^T\mathbf{u}_i = \sigma_i^2 \mathbf{u}_i$, $\forall i$. Indeed, we have

$$AA^T\mathbf{u}_i = AA^T\frac{A\mathbf{v}_i}{\sigma_i} = \frac{1}{\sigma_i}A(A^T A\mathbf{v}_i) = \frac{1}{\sigma_i}A(\sigma_i^2\mathbf{v}_i) = \sigma_i^2\frac{A\mathbf{v_i}}{\sigma_i} = \sigma_i^2\mathbf{u}_i, \; i = 1, \cdots, k$$

For $j = k+1, \cdots, n$, $\sigma_j^2 = 0$, and by SVD algorithm, we know that

$$span\{\mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\} = col(A)^\perp = null(A^T)$$

That is $A^T\mathbf{u}_j = \mathbf{0}$, and consequently $AA^T\mathbf{u}_j = \mathbf{0}, j = k+1, \cdots, m$. Thus we see that $AA^T\mathbf{u}_i = \sigma_i^2\mathbf{u}_i$ holds for all i.

In conclusion, we see that $\mathbf{v}_i's$ are eigenvectors of $A^T A$ (with corresponding non zero eigenvalues given by $\sigma_1^2 \geq \cdots \geq \sigma_k^2 > 0$); while $\mathbf{u}_i's$ are eigenvectors of AA^T (with the same set of eigenvalues as that of $A^T A$). Recall that AA^T and $A^T A$ have the same set of eigenvalues.

With the above preparations, SVD for A follows easily. Indeed, we have

$$AV = A[\mathbf{v}_1, \cdots, \mathbf{v}_k, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_n] = [A\mathbf{v}_1, \cdots, A\mathbf{v}_k, \mathbf{0}, \cdots, \mathbf{0}] = [\sigma_1\mathbf{u}_1, \cdots, \sigma_k\mathbf{u}_k, \mathbf{0}, \cdots, \mathbf{0}]$$

$$= \underbrace{[\mathbf{u}_1, \cdots, \mathbf{u}_k; \mathbf{u}_{k+1}, \cdots, \mathbf{u}_m]}_{U} \underbrace{\left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0}_{k\times(n-k)} \\ 0 & \cdots & \sigma_k & \\ \hline \multicolumn{3}{c|}{\mathbf{0}_{(m-k)\times k}} & \mathbf{0}_{(m-k)\times(n-k)} \end{array}\right]}_{\Sigma}$$

Multiplying both sides by V^T from the right, and using the orthogonal property

$VV^T = I_n$, we get finally that

$$A = AVV^T = U\Sigma V^T$$