DATA8015: Mathematical Foundation of Data Science **Fall 2025**

Part 1.1 Linear Algebra Basics

Scriber: Yingyu Liang

1 Vectors and Matrices

Vectors and matrices (and higher-order arrays) are ubiquitous in data science.

A d-dimensional vector $\boldsymbol{x}$ is a first-order array:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \cdots \\ \boldsymbol{x}_d \end{bmatrix}$$

where $\boldsymbol{x}_i$ is the i-th entry.

An m by n matrix $\boldsymbol{A}$ is a second-order array:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1n} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2n} \\ \cdots \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \cdots & \boldsymbol{A}_{mn} \end{bmatrix}$$

where $\boldsymbol{A}_{ij}$ is the (i, j)-th entry.

A d-dimensional vector can also be viewed as a d by 1 matrix (i.e., a column vector), or viewed as a 1 by d matrix (i.e., a row vector). An m by n matrix can also be viewed as stacking n column vectors $\boldsymbol{A}_{:j}(1 \le j \le n)$ together, or staking m row vectors $\boldsymbol{A}_{i:}(1 \le i \le m)$ together.

1.1 Basic Operations

Linear operations: Assuming addition/multiplication on entries are well-defined, one can perform the following basic linear operations on vectors/matrices:

- scalar multiplication: multiple each entry by the scalar
- addition: add the corresponding entries in two vectors of the same size, or two matrices of the same size

Vector products:

- dot product (which is a special kind of inner product):

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_d \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \cdots \\ \boldsymbol{y}_d \end{bmatrix} = \boldsymbol{x}_1\boldsymbol{y}_1 + \boldsymbol{x}_2\boldsymbol{y}_2 + \cdots + \boldsymbol{x}_d\boldsymbol{y}_d$$

- outer product:

$$\boldsymbol{x}\boldsymbol{y}^\top := \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \cdots \\ \boldsymbol{x}_d \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_d \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1\boldsymbol{y}_1 & \boldsymbol{x}_1\boldsymbol{y}_2 & \cdots \boldsymbol{x}_1\boldsymbol{y}_d \\ \boldsymbol{x}_2\boldsymbol{y}_1 & \boldsymbol{x}_2\boldsymbol{y}_2 & \cdots \boldsymbol{x}_2\boldsymbol{y}_d \\ \cdots \\ \boldsymbol{x}_d\boldsymbol{y}_1 & \boldsymbol{x}_d\boldsymbol{y}_2 & \cdots \boldsymbol{x}_d\boldsymbol{y}_d \end{bmatrix}$$

Matrix-Vector multiplication: For $\boldsymbol{A}$ of size $m \times n$ and $\boldsymbol{x}$ of dimension n, the product $\boldsymbol{A}\boldsymbol{x}$ is a vector of dimension m, whose i-th entry $[\boldsymbol{A}\boldsymbol{x}]_i$ is the inner product of the i-th row of $\boldsymbol{A}$ and $\boldsymbol{x}$, i.e., $\langle \boldsymbol{A}_{i:}, \boldsymbol{x} \rangle$.

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1n} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2n} \\ \cdots \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \cdots & \boldsymbol{A}_{mn} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \cdots \\ \boldsymbol{x}_n \end{bmatrix} \\ &= \begin{bmatrix} \langle \boldsymbol{A}_{1:}, \boldsymbol{x} \rangle \\ \langle \boldsymbol{A}_{2:}, \boldsymbol{x} \rangle \\ \cdots \\ \langle \boldsymbol{A}_{m:}, \boldsymbol{x} \rangle \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{x}_1 + \boldsymbol{A}_{12}\boldsymbol{x}_2 + \cdots + \boldsymbol{A}_{1n}\boldsymbol{x}_n \\ \boldsymbol{A}_{21}\boldsymbol{x}_1 + \boldsymbol{A}_{22}\boldsymbol{x}_2 + \cdots + \boldsymbol{A}_{2n}\boldsymbol{x}_n \\ \cdots \\ \boldsymbol{A}_{m1}\boldsymbol{x}_1 + \boldsymbol{A}_{m2}\boldsymbol{x}_2 + \cdots + \boldsymbol{A}_{mn}\boldsymbol{x}_n \end{bmatrix} \end{aligned}$$

Matrix-Matrix multiplication: For $\boldsymbol{A}$ of size $m \times n$ and $\boldsymbol{B}$ of size $n \times d$, the product $\boldsymbol{A}\boldsymbol{B}$ is of size $m \times d$, whose (i, j)-th entry $[\boldsymbol{A}\boldsymbol{B}]_{ij}$ is the inner product of the i-th row of $\boldsymbol{A}$ and the j-th column of $\boldsymbol{B}$, i.e., $\langle \boldsymbol{A}_{i:}, \boldsymbol{B}_{:j} \rangle$.

$$\begin{aligned} \boldsymbol{A}\boldsymbol{B} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1n} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2n} \\ \cdots \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \cdots & \boldsymbol{A}_{mn} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \cdots & \boldsymbol{B}_{1d} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \cdots & \boldsymbol{B}_{2d} \\ \cdots \\ \boldsymbol{B}_{n1} & \boldsymbol{B}_{n2} & \cdots & \boldsymbol{B}_{nd} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_{1:} \\ \boldsymbol{A}_{2:} \\ \cdots \\ \boldsymbol{A}_{m:} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{:1} & \boldsymbol{B}_{:2} & \cdots & \boldsymbol{B}_{:d} \end{bmatrix} \\ &= \begin{bmatrix} \langle \boldsymbol{A}_{1:}, \boldsymbol{B}_{:1} \rangle & \langle \boldsymbol{A}_{1:}, \boldsymbol{B}_{:2} \rangle & \cdots & \langle \boldsymbol{A}_{1:}, \boldsymbol{B}_{:d} \rangle \\ \langle \boldsymbol{A}_{2:}, \boldsymbol{B}_{:1} \rangle & \langle \boldsymbol{A}_{2:}, \boldsymbol{B}_{:2} \rangle & \cdots & \langle \boldsymbol{A}_{2:}, \boldsymbol{B}_{:d} \rangle \\ \cdots \\ \langle \boldsymbol{A}_{m:}, \boldsymbol{B}_{:1} \rangle & \langle \boldsymbol{A}_{m:}, \boldsymbol{B}_{:2} \rangle & \cdots & \langle \boldsymbol{A}_{m:}, \boldsymbol{B}_{:d} \rangle \end{bmatrix} \end{aligned}$$

It can also be viewed as multiplying $\boldsymbol{A}$ with the columns of $\boldsymbol{B}$, or multiplying the rows of $\boldsymbol{A}$ with $\boldsymbol{B}$:

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{AB}_{:1} & \boldsymbol{AB}_{:2} & \cdots & \boldsymbol{AB}_{:d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{1:}\boldsymbol{B} \\ \boldsymbol{A}_{2:}\boldsymbol{B} \\ \cdots \\ \boldsymbol{A}_{m:}\boldsymbol{B} \end{bmatrix}$$

The matrix-vector multiplication can also be viewed as a special case of matrix-matrix multiplication by viewing the vector as a matrix with a single column.

Matrix-Matrix multiplication satisfies the Associativity and Distributivity properties, but not the Commutative property.

Transposition: Transposition flips the rows and columns of a matrix. For a matrix $\boldsymbol{A}$ of size $m \times n$, its transposition $\boldsymbol{A}^\top$ is a matrix of size $n \times m$, whose (i, j)-th entry is the (j, i)-th entry of $\boldsymbol{A}$, i.e., $[\boldsymbol{A}^\top]_{ij} = \boldsymbol{A}_{ji}$.

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1n} \\ & & \cdots & \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \cdots & \boldsymbol{A}_{mn} \end{bmatrix}, \qquad \boldsymbol{A}^\top = \begin{bmatrix} \boldsymbol{A}_{11} & & \boldsymbol{A}_{m1} \\ \boldsymbol{A}_{12} & & \boldsymbol{A}_{m2} \\ \cdots & \cdots & \cdots \\ \boldsymbol{A}_{1n} & & \boldsymbol{A}_{mn} \end{bmatrix}$$

A vector is typically regarded as a column vector, and its transposition is a row vector:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \cdots \\ \boldsymbol{x}_d \end{bmatrix}, \qquad \boldsymbol{x}^\top = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_d \end{bmatrix}$$

Matrix inverse: Usually we consider matrices over real numbers. The identity matrix $\boldsymbol{I}$ is then a square matrix with 1 in the diagonal and 0 elsewhere:

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

If two matrices $\boldsymbol{A}, \boldsymbol{B}$ satisfy $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$, then $\boldsymbol{A}$ is invertible/nonsingular, and $\boldsymbol{B}$ is its inverse (denoted by $\boldsymbol{B} = \boldsymbol{A}^{-1}$). If we consider the vector-matrix product $\boldsymbol{Ax}$, then we will have $\boldsymbol{BAx} = \boldsymbol{x}$, i.e., $\boldsymbol{B}$ will inverse the effect of $\boldsymbol{A}$ on $\boldsymbol{x}$.

2 Applications in Data Science

Vectors and matrices usually serve two purposes in data science.

2.1 Representing Data

Many data types are naturally vectors or matrices. Examples:

- Information about a patient can be concatenated into a vector, such as [age, weight, height, temperature, blood pressure, ...].
- A grayscale image can be considered as a matrix I, where the (i, j)-th entry is the gray value for the (i, j)-th pixel.
- An RGB color image can be considered as a third-order tensor I, where the $(i, j, 0)$-th, $((i, j, 1)$-th, and $(i, j, 2)$-th entries are the red, green, blue values for the (i, j)-th pixel, respectively.

Many other data types are not native but converted to numeric vectors/matrices. Examples:

- A text sentence is a sequence of tokens (e.g., English words), represented as a vector of tokens $(w_1, w_2, \ldots, w_T)$ where w_i's are tokens from a vocabulary Σ. In modern applications (e.g., large language models), the tokens are typically converted into numeric vectors by an embedding $v(w) : \Sigma \to \mathbb{R}^d$. The sentence is then represented as a sequence of vectors $(v(w_1), v(w_2), \ldots, v(w_T)) \in \mathbb{R}^{T \times d}$, which can be viewed as a matrix of dimensions T by d.
- A graph $(\mathcal{V}, \mathcal{E})$ can be represented as an adjacent matrix $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where the (i, j)-th entry is set to 1 if there is an edge connecting the nodes i and j (i.e., $(i, j) \in \mathcal{E}$), and is set to 0 otherwise.

A practical reason for representing these data types as numeric vectors/matrices is such that they can be processed by various numerical tools. For example, deep learning models take a sequence of numeric vectors as input, rather than the sequence of tokens itself. The spectral clustering method operates on the Laplacian matrix of the graph (a transformation of the adjacent matrix), rather than the graph itself.

However, conceptually this is highly non-trivial: why the embedded vectors/matrices have structures that can keep the semantics of the original data and be exploited by downstream tools? The embedding thus needs to be carefully designed.

- The word embedding methods should preserve the semantics as much as possible, e.g., similar words are mapped to similar vectors. In the classic word2vec embedding method, it is observed that the embedding vectors have linear structure: $v(\text{man}) - v(\text{woman}) \approx v(\text{king}) - v(\text{queen})$.
- The Laplacian matrix of a graph is used because its eigenvectors encode the clustering structure of the graph.

In fact, the popular deep learning paradigm is also widely regarded as representation learning, which aims to learn a representation of the data that can be exploited by linear algebra. For example, a convolutional neural network trained for image classification on ImageNet can be viewed as the backbone representation (i.e., from the input to the second-to-last layer) followed by a logistic regression model (i.e., the last layer). The goal of the backbone is to get a representation function such that the representations of the images from different classes are linearly separated by the logistic regression model.

2.2 Representing Linear Functions

Vectors/matrices can also represent linear functions.

- Consider a linear function $f : \mathbb{R}^d \to \mathbb{R}$. It is defined as $f(\boldsymbol{x}) = \theta_1 \boldsymbol{x}_1 + \theta_2 \boldsymbol{x}_2 + \cdots + \theta_d \boldsymbol{x}_d$. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_d)$. Then by the definition of vector multiplication, the function can be compactly written as $f(\boldsymbol{x}) = \theta^\top \boldsymbol{x}$.
- Consider a linear function $F : \mathbb{R}^d \to \mathbb{R}^k$. It outputs a k-dimensional vector, where the i-th dimension is given by a linear function $F^i(\boldsymbol{x}) = (\boldsymbol{\theta}^i)^\top \boldsymbol{x}$ for a $\boldsymbol{\theta}^i \in \mathbb{R}^d$. Let

  $$\boldsymbol{A} = \begin{bmatrix} (\boldsymbol{\theta}^1)^\top \\ (\boldsymbol{\theta}^2)^\top \\ \cdots \\ (\boldsymbol{\theta}^k)^\top \end{bmatrix} \in \mathbb{R}^{k \times d}.$$

  Then by the definition of matrix multiplication, $F(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$.
- Many applications involve linear equation systems:

  $$\begin{aligned} \boldsymbol{A}_{11}\boldsymbol{x}_1 + \boldsymbol{A}_{12}\boldsymbol{x}_2 \cdots + \boldsymbol{A}_{1d}\boldsymbol{x}_d &= \boldsymbol{b}_1 \\ \boldsymbol{A}_{21}\boldsymbol{x}_1 + \boldsymbol{A}_{22}\boldsymbol{x}_2 \cdots + \boldsymbol{A}_{2d}\boldsymbol{x}_d &= \boldsymbol{b}_2 \\ &\cdots \\ \boldsymbol{A}_{k1}\boldsymbol{x}_1 + \boldsymbol{A}_{k2}\boldsymbol{x}_2 \cdots + \boldsymbol{A}_{kd}\boldsymbol{x}_d &= \boldsymbol{b}_k \end{aligned}$$

  This can be compactly represented as $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. This means $\boldsymbol{x}$ is in the preimage of $\boldsymbol{b}$ under the linear mapping represented by $\boldsymbol{A}$.

Many data science tools use linear functions as parts. For example, each layer in a standard feedforward network consists of a linear function followed by a nonlinear activation function. To understand the tools, it is then crucial to understand the linear functions, which requires diving into linear spaces and related concepts in linear algebra.

3 Linear Space and Basis

Here we will begin with abstract linear spaces and based on that introduce notions such as coordinates, matrices, matrix-vector multiplications, and their properties. This perspective provides deeper understanding about those basic notions.

3.1 Linear Space

Definition 1 (Linear Space)**.** A linear space (or vector space) $V = (\mathcal{V}, +, \cdot)$ over a field K is a set $\mathcal{V}$ with two operations:

$$\begin{aligned} + &: \mathcal{V} \times \mathcal{V} \to \mathcal{V} \\ \cdot &: K \times \mathcal{V} \to \mathcal{V} \end{aligned}$$

where

1. $(\mathcal{V}, +)$ is an Abelian group

2. $\cdot$ satisfy Associativity:

$$\forall \lambda, \psi \in K, \boldsymbol{x} \in \mathcal{V} : \lambda \cdot (\psi \cdot \boldsymbol{x}) = (\lambda\psi) \cdot \boldsymbol{x}$$

3. Neural element of $\cdot$ is the unit $1 \in K$: $\forall \boldsymbol{x} \in \mathcal{V}, 1 \cdot \boldsymbol{x} = \boldsymbol{x}$

4. $+$ and $\cdot$ satisfy Distributivity: $\forall \lambda, \psi \in K, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$,

$$\lambda \cdot (\boldsymbol{x} + \boldsymbol{y}) = \lambda \cdot \boldsymbol{x} + \lambda \cdot \boldsymbol{y}$$
$$(\lambda + \psi) \cdot \boldsymbol{x} = \lambda \cdot \boldsymbol{x} + \psi \cdot \boldsymbol{x}$$

Elements in K are called scalars. Typically, we consider $K = \mathbb{R}$, the real numbers; another common field used is $K = \mathbb{C}$, the complex numbers. We usually hide the notation $\cdot$, e.g., write $\lambda\boldsymbol{x}$ instead of $\lambda \cdot \boldsymbol{x}$.

Definition 2 (Linear Subspace)**.** For a linear space $V = (\mathcal{V}, +, \cdot)$, a subset $\mathcal{U}$ of $\mathcal{V}$ is called a subspace, if the outputs of operations $+, \cdot$ on $\mathcal{U}$ still belong to $\mathcal{U}$.

That is, the subspace is closed under the operations $+, \cdot$.

3.2 Linear Independence

To connect general linear subspaces and linear mappings on them with vectors and matrices, we need the concept of basis. Before introducing basis, we first introduce the concepts of linear combinations and linear independence.

Definition 3 (Linear Combination)**.** A linear combination of k vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$ of a linear space is a vector of the form

$$\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2 + \cdots + \lambda_k\boldsymbol{x}_k, \quad \lambda_1, \cdots, \lambda_k \in K.$$

Note that all the linear combinations of $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$ is a subspace and is the smallest subspace containing $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$. It is called the subspace spanned by $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$, or the span of $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$.

Definition 4 (Linear (In)Dependence)**.** The vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$ are called linearly dependent if there is a nontrivial linear relation between them, that is,

$$\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2 + \cdots + \lambda_k\boldsymbol{x}_k = 0$$

where not all $\lambda_1, \cdots, \lambda_k$ are zero. Otherwise, they are called linearly independent.

3.3 Basis, Dimension, and Coordinates

Definition 5 (Basis)**.** A finite set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$ from $\mathcal{V}$ is called a basis of V, if (1) they are linearly independent, (2) every vector in $\mathcal{V}$ can be represented as their linear combination. $\boldsymbol{x}_i$'s are called the basis vectors in this basis.

So the basis can generate the whole space while being linearly independent. The linear independence makes sure that it is minimal: no proper subset of the basis can generate the whole space.

A linear space may have many bases. However, the following theorem guarantees that all of them have the same number of basis vectors. Then we can use this number to measure the "size" of the space, i.e., its dimension.

Theorem 6 (Dimension)**.** If a linear space V has a basis, then all its bases contain the same number of vectors. This number is called the dimension of V and denoted as $\dim(V)$.

The theorem can be proved by the following lemma.

Lemma 7. Suppose (1) every vector in $\mathcal{V}$ can be represented as a linear combination of $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$; (2) $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, y_k$ in $\mathcal{V}$ are linearly independent. Then $k \leq n$.

Proof. Given assumption (1), for some $\lambda_1, \cdots, \lambda_n$,

$$\boldsymbol{y}_1 = \lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_n \boldsymbol{x}_n.$$

Given assumption (2), $\boldsymbol{y}_1 \neq 0$ (otherwise $\boldsymbol{y}_j$'s must be linearly dependent). Then not all λ_j's are 0. Say $\lambda_i \neq 0$. Then $\boldsymbol{x}_i$ can be represented as a linear combination of $\boldsymbol{y}_1$ and the remaining $\boldsymbol{x}_j$'s. So the set consisting of $\boldsymbol{x}_j$'s with $\boldsymbol{x}_i$ replaced by $\boldsymbol{y}_1$ can generate the whole space.

Suppose $k > n$. Then repeat this step n times and conclude that $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ can generate the whole space. In particular, $\boldsymbol{y}_{n+1}$ is a linear combination of $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$, which contradicts the assumption (2). Therefore, $k \leq n$. □

With the concept of basis $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n\}$, we can represent elements in general linear spaces as arrays of scalars (i.e., numeric vectors we are familiar with). Since the order of the basis vectors will be important for this, we will order the basis vectors and get an ordered basis $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$.

Definition 8 (Coordinates)**.** Consider a linear space V and an ordered basis $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$ of V. For any $\boldsymbol{x} \in \mathcal{V}$, there is a unique representation (linear combination)

$$\boldsymbol{x} = \alpha_1 \boldsymbol{b}_1 + \cdots + \alpha_n \boldsymbol{b}_n$$

of $\boldsymbol{x}$ with respect to B. Then α_1, α_n are the coordinates of $\boldsymbol{x}$ with respect to B, and the vector

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_n \end{bmatrix} \in K^n$$

is the coordinate vector/coordinate representation of $\boldsymbol{x}$ with respect to the ordered basis B.

So a basis effectively defines a coordinate system. Note that the representation is unique given the linear independence of the basis vectors.

3.4 Linear Mappings and Their Matrix Representation

Here we consider mappings on linear spaces that preserve their structure.

Definition 9 (Linear Mappings)**.** For vector spaces V and W (over the same field K), a mapping $\Phi : V \to W$ is called a linear mapping (or vector space homomorphism/linear transformation) if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}, \forall \lambda, \psi \in K : \Phi(\lambda \boldsymbol{x} + \psi \boldsymbol{y}) = \lambda \Phi(\boldsymbol{x}) + \psi \Phi(\boldsymbol{y}).$$

Now we are ready to make an explicit connection between matrices and linear mappings between finite-dimensional linear spaces.

Definition 10 (Transformation Matrix)**.** Consider linear spaces V, W with corresponding ordered bases $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$ and $C = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_m)$, and a linear mapping $\Phi : V \to W$. For $j \in \{1, \cdots, n\}$,

$$\Phi(\boldsymbol{b}_j) = \sum_{i=1}^{m} \alpha_{ij} \boldsymbol{c}_i$$

is the unique representation of $\Phi(\boldsymbol{b}_j)$ with respect to C. Then the $m \times n$ matrix $\boldsymbol{A}_\Phi$ with entries $[\boldsymbol{A}_\Phi]_{ij} = \alpha_{ij}$ is called the transformation matrix of Φ (with respect to the ordered bases B of V and C of W).

$\boldsymbol{A}_\Phi$ is the matrix presentation of Φ: $\boldsymbol{A}_\Phi$ and matrix-vector multiplication are defined in the particular way such that the linear mapping Φ is represented by multiplying $\boldsymbol{A}_\Phi$ with the coordinate vector of the input. More precisely, for $\boldsymbol{y} = \Phi(\boldsymbol{x})$, let $\widehat{\boldsymbol{x}}$ be the coordinate vector of $\boldsymbol{x}$ w.r.t. B and $\widehat{\boldsymbol{y}}$ be the coordinate vector of $\boldsymbol{y}$ w.r.t. C, then

$$\widehat{\boldsymbol{y}} = \boldsymbol{A}_\Phi \widehat{\boldsymbol{x}}.$$

To see this, by the definition of the linear mappings,

$$\begin{aligned}
\boldsymbol{y} &= \Phi(\boldsymbol{x}) \\
&= \Phi\left(\sum_{j=1}^{n} \widehat{\boldsymbol{x}}_j \boldsymbol{b}_j\right) \\
&= \sum_{j=1}^{n} \widehat{\boldsymbol{x}}_j \Phi\left(\boldsymbol{b}_j\right) \\
&= \sum_{j=1}^{n} \widehat{\boldsymbol{x}}_j \sum_{i=1}^{m} \alpha_{ij} \boldsymbol{c}_i \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_{ij} \widehat{\boldsymbol{x}}_j \boldsymbol{c}_i \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} \widehat{\boldsymbol{x}}_j\right) \boldsymbol{c}_i
\end{aligned}$$

which means

$$\widehat{\boldsymbol{y}}_i = \sum_{j=1}^{n} \alpha_{ij} \widehat{\boldsymbol{x}}_j = [\boldsymbol{A}_\Phi \widehat{\boldsymbol{x}}]_i$$

and thus $\widehat{\boldsymbol{y}} = \boldsymbol{A}_\Phi \widehat{\boldsymbol{x}}$.

This means the transformation matrix (working together with coordinates with respect to an ordered basis in V and coordinates with respect to an ordered basis in W) represents the linear mapping from V to W. Therefore, the study of linear mappings corresponds to the study of matrices, and vice versa.

3.5 Basis Change

The coordinate vectors and matrix representation (i.e., the transformation matrix) depend on the ordered bases used. It is then possible to choose the bases such that the transformation matrix has a particular simple form, and thus reveal interesting properties. In the following, we first consider how the transformation matrix changes when the bases change.

Consider linear spaces V, W with ordered bases $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$ and $C = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_m)$, respectively. Consider a linear mapping $\Phi : V \to W$ and its transformation matrix $\boldsymbol{A}_\Phi$ with respect to the ordered bases B and C. Given new bases $\widetilde{B} = (\widetilde{\boldsymbol{b}}_1, \cdots, \widetilde{\boldsymbol{b}}_n)$ for V and $\widetilde{C} = (\widetilde{\boldsymbol{c}}_1, \cdots, \widetilde{\boldsymbol{c}}_m)$ for W, we would like to find the transformation matrix $\widetilde{\boldsymbol{A}}_\Phi$ with respect to the new ordered bases $\widetilde{B}$ and $\widetilde{C}$.

The new bases are connected to the old bases as follows.

$$\widetilde{\boldsymbol{b}}_j = \sum_{i=1}^{n} s_{ij} \boldsymbol{b}_i, \quad j = 1, \cdots, n.$$

$$\widetilde{\boldsymbol{c}}_k = \sum_{l=1}^{m} t_{lk} \boldsymbol{c}_l, \quad k = 1, \cdots, m.$$

Define a matrix $\boldsymbol{S}$ of size $n \times n$ with entries $\boldsymbol{S}_{ij} = s_{ij}$, and define a matrix $\boldsymbol{T}$ of size $m \times m$ with entries $\boldsymbol{T}_{lk} = t_{lk}$.

Theorem 11 (Basis Change)**.** Suppose $V, W, B, C, \widetilde{B}, \widetilde{C}, \boldsymbol{A}_\Phi, \widetilde{\boldsymbol{A}}_\Phi, \boldsymbol{S}, \boldsymbol{T}$ are defined as above. Then $\widetilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}$.

Proof. One can prove the theorem by showing that $\boldsymbol{S}$ is the transformation matrix of the identity mapping id_V on V that maps the coordinates w.r.t. $\widetilde{B}$ onto coordinates w.r.t. B. Similarly, $\boldsymbol{T}$ is the transformation matrix of the identity mapping id_W on W that maps the coordinates w.r.t. $\widetilde{C}$ onto coordinates w.r.t. C. Then $\boldsymbol{T}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}$ is the matrix representation of $\mathrm{id}_W \circ \Phi \circ \mathrm{id}_V = \Phi$ mapping coordinates w.r.t. $\widetilde{B}$ to coordinates w.r.t. $\widetilde{C}$, which is exactly $\widetilde{\boldsymbol{A}}_\Phi$.

Here we present a direct proof by looking at $\Phi(\widetilde{\boldsymbol{b}}_j)$ from two perspectives. First, for any $j = 1, \cdots, n$,

$$\Phi(\widetilde{\boldsymbol{b}}_j) = \sum_{k=1}^{m} [\widetilde{\boldsymbol{A}}_\Phi]_{kj} \widetilde{\boldsymbol{c}}_k = \sum_{k=1}^{m} [\widetilde{\boldsymbol{A}}_\Phi]_{kj} \sum_{l=1}^{m} t_{lk} \boldsymbol{c}_l = \sum_{l=1}^{m} \left(\sum_{k=1}^{m} [\widetilde{\boldsymbol{A}}_\Phi]_{kj} t_{lk} \right) \boldsymbol{c}_l.$$

Alternatively,

$$\Phi(\widetilde{\boldsymbol{b}}_j) = \Phi(\sum_{i=1}^{n} s_{ij}\boldsymbol{b}_i) = \sum_{i=1}^{n} s_{ij}\Phi(\boldsymbol{b}_i) = \sum_{i=1}^{n} s_{ij}\sum_{l=1}^{m}[\boldsymbol{A}]_{li}\boldsymbol{c}_l = \sum_{l=1}^{m}\left(\sum_{i=1}^{n} s_{ij}[\boldsymbol{A}]_{li}\right)\boldsymbol{c}_l.$$

Comparing the two equations, and noting that $\boldsymbol{c}_l$'s are linearly independent, we have

$$\sum_{k=1}^{m}[\widetilde{\boldsymbol{A}}_\Phi]_{kj}t_{lk} = \sum_{i=1}^{n} s_{ij}[\boldsymbol{A}]_{li}, \quad \forall j = 1, \cdots, n \text{ and } l = 1, \cdots, m$$

and thus

$$\boldsymbol{T}\widetilde{\boldsymbol{A}}_\Phi = \boldsymbol{A}_\Phi\boldsymbol{S}$$

which leads to the theorem statement. (The proof that $\boldsymbol{T}$ is invertible is ignored here.) □

3.6 Transposition

Consider linear spaces V, W with corresponding ordered bases $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$ and $C = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_m)$, and a linear mapping $\Phi : V \to W$.

Consider any linear mapping $\ell : W \to K$ from W to the field K. It induces a mapping $r = \ell \circ \Phi$ from V to K, which is also linear. We denote the mapping from ℓ to r as Φ' and call it the transpose of Φ. One can show that the space of all linear mappings on W (or V) is a linear space, which is called the dual space of W and denoted as W' (or V'). Furthermore, Φ' is also a linear mapping (from the linear space W' to V')!

Let $\boldsymbol{A}_\Phi = [\alpha_{ij}]$ denote the matrix representation of Φ. The matrix representation of ℓ is a vector

$$\boldsymbol{v}^\ell = [\ell(\boldsymbol{c}_1), \ldots, \ell(\boldsymbol{c}_m)]$$

such that for any $\boldsymbol{y} \in W$ with coordinates $\widehat{\boldsymbol{y}}$, $\ell(\boldsymbol{y}) = \ell(\sum_{i=1}^m \widehat{\boldsymbol{y}}_i\boldsymbol{c}_i) = \sum_{i=1}^m \widehat{\boldsymbol{y}}_i\ell(\boldsymbol{c}_i) = \langle\boldsymbol{v}^\ell, \widehat{\boldsymbol{y}}\rangle$.

Now consider the matrix representation of $r = \Phi'\ell$, which is a vector

$$\boldsymbol{v}^r = [r(\boldsymbol{b}_1), \cdots, r(\boldsymbol{b}_n)], \text{ where } r(\boldsymbol{b}_j) = \ell \circ \Phi(\boldsymbol{b}_j) = \ell\left(\sum_{i=1}^{m}\alpha_{ij}\boldsymbol{c}_i\right) = \sum_{i=1}^{m}\alpha_{ij}\ell(\boldsymbol{c}_i).$$

Now it can be verified that for $r = \Phi'\ell$, their representations satisfy

$$\boldsymbol{v}^r = [\boldsymbol{A}_\Phi]^\top\boldsymbol{v}^\ell$$

where $[\boldsymbol{A}_\Phi]^\top$ is the matrix transposition of $\boldsymbol{A}_\Phi$. In other words, matrix transposition is defined in such a way as to represent the linear mapping transposition.

3.7 More on Dimensions and Ranks

Isomorphism. Intuitively, the dimension of a linear space measures the "size" of the space. Actually, it is more fundamental: two linear spaces over the same field and of the same dimension can be viewed as the same.

Definition 12 (Isomorphism)**.** A one-to-one correspondence between two linear spaces over the same field that maps sums into sums and scalar multiples into scalar multiples is called an isomorphism.

That is, isomorphic linear spaces are indistinguishable by means of operations available in linear spaces.

Theorem 13 (Dimension and Isomorphism)**.** Finite-dimensional linear spaces V and W are isomorphic if and only if $\dim(V) = \dim(W)$.

Proof. If V, W are isomorphic, clearly the two have the same dimension. For the other direction, let n denote the dimension and K denote the field. Let K^n denote the linear space formed by the set of all row vectors: $(a_1, a_2, \cdots, a_n), a_j \in K$ with addition, multiplication defined componentwise. It can be shown that any n-dimensional linear space over K is isomorphic to K^n. Then V, W are isomorphic. □

It also means that one can view all linear subspaces of dimension n as K^n.

Range and Nullspace. The following is a fundamental result about linear maps.

Definition 14 (Range and Nullspace)**.** Consider a linear mapping $\Phi : V \to W$. The range of Φ, denoted as R_Φ, is the image of V under Φ. The nullspace (or kernel) of Φ, denoted as N_Φ, is the subset of V mapped into 0 by Φ.

Theorem 15 (Dimensions in Linear Mappings)**.** Let $\Phi : V \to W$ be a linear map. Then

$$\dim(N_\Phi) + \dim(R_\Phi) = \dim(V).$$

Proofsketch. Define Φ acting on the quotient space V/N_Φ by setting $\Phi\{\boldsymbol{x}\} := \Phi(\boldsymbol{x})$. This is an isomophism and thus $\dim(V/N_\Phi) = \dim(R_\Phi)$. The theorem then follows from the property of quotient space that $\dim(V/N_\Phi) = \dim(V) - \dim(N_\Phi)$. □

Rank. The dimension of the range R_Φ is also called the rank of Φ. For a matrix $\boldsymbol{A}$, the number of linearly independent columns is called the column rank, and the number of linearly independent rows is called the row rank. It is easy to see that the rank of Φ is the same as the column rank of its matrix representation $\boldsymbol{A}_\Phi$.

A nontrivial fact is that the row rank is the same as the column rank, so all three rank notions are the same. To see this, first note that the row rank of $\boldsymbol{A}_\Phi$ is the column rank of its transpose $[\boldsymbol{A}_\Phi]^\top$, which is the rank of the transpose Φ' (since $[\boldsymbol{A}_\Phi]^\top$ is the representation of Φ'). Also note that the column rank of $\boldsymbol{A}_\Phi$ is the rank of Φ. So it is sufficient to show that the rank of a linear mapping equals that of its transpose.

Theorem 16 (Dimension of Transposition)**.** For a linear mapping Φ and its transpose Φ', $\dim(R_\Phi) = \dim(R_{\Phi'})$.

Proofsketch. Suppose $\Phi : V \to W$. Let $R_\Phi^\perp$ denote the annihilator of the range of Φ (i.e., all the linear functions that annihilate R_Φ). We note the following key properties about annihilator without proofs:

$$\dim(R_\Phi^\perp) + \dim(R_\Phi) = \dim(W).$$

and also

$$R_\Phi^\perp = N_{\Phi'}.$$

On the other hand, by the fundamental result Theorem 15 about the dimensions in linear maps,

$$\dim(N'_\Phi) + \dim(R'_\Phi) = \dim(W').$$

Furthermore,

$$\dim(W') = \dim(W).$$

These equations lead to the theorem. □

4 Determinant and Trace

Here, we consider square matrices (matrices with the same number of rows and columns), which can be viewed as representations of linear mappings from a linear space to itself. We introduce two important scalar-valued functions of square matrices. They remain the same under basis change of the linear space and thus are regarded as inherent properties of a linear mapping. They are also related to and thus useful for testing linear dependence and doing eigendecomposition.

Determinant. The determinant can be introduced in different ways and here we choose to use axioms inspired by the intuitive properties of volume in geometry. The geometry of linear space is discussed more later, but it turns out that the volume notion is somehow independent of that.

A key reason for introducing the determinant is to test whether an $n \times n$ square matrix is of full rank (i.e., rank n), equivalently, to test if the columns/rows are linearly independent. Viewing the column vectors as n points, they form a simplex together with the origin. Intuitively, if they are dependent, then the simplex falls into a subspace and has 0 volume. This thus inspires using the volume as a test.

The determinant is related to the signed volume of the simplex up to a scaling factor $n!$ (where the sign depends on the order of the columns/vertices). Rather than start with a formula for the determinant, here we will deduce it from the geometric properties of signed volume.

Let $a_1, \cdots, a_n$ denote the columns of a matrix $\boldsymbol{A}$. Let $\det(\boldsymbol{A}) = D(a_1, \cdots, a_n)$ denote the determinant. It should have the following properties:

Property (i). $D(a_1, \cdots, a_n) = 0$ if $a_i = a_j, i \neq j$.

Property (ii). $D(a_1, \cdots, a_n)$ is a multilinear function of its arguments, in the sense that if all $a_i, i \neq j$ are fixed, D is a linear function of the remaining argument a_j.

Property (iii). Normalization: $D(e_1, \cdots, e_n) = 1$, where e_i is the vector with 1 on the i-th dimension and 0 elsewhere.

These three properties can then deduce all remaining properties. In particular, it is not difficult to prove that if $a_1, \cdots, a_n$ are linearly dependent, then $D(a_1, \cdots, a_n) = 0$. In fact, the three properties completely characterize the determinant.

Theorem 17 (Determinant)**.** Properties (i),(ii), and (iii) uniquely determine the determinant as a function of $a_1, \cdots, a_n$. In particular,

$$\det(\boldsymbol{A}) := D(a_1, \cdots, a_n) = \sum_{p\text{: all permutations}} \sigma(p) a_{p_1 1} \cdots a_{p_n n}$$

where $\sigma(p) = (-1)^k$ is the signature of the permutation p, and k is the number of transpositions needed to form p.

Proofsketch. Representing $a_j = \sum_i a_{ji} e_i$ and using Property (ii), we have

$$D(a_1, \cdots, a_n) = \sum_f a_{f_1 1} a_{f_2 2} \cdots a_{f_n n} D(e_{f_1}, e_{f_2}, \cdots, e_{f_n})$$

where the summation is over all functions mapping $\{1, \cdots, n\}$ into $\{1, \cdots, n\}$. By Property (i), this reduces to a summation over all permutations.

It can be shown that a single transposition changes the value of D by a factor of -1 (e.g., $D(e_2, e_1, e_3, \cdots, e_n) = (-1)D(e_1, e_2, e_3, \cdots, e_n) = -1$). Then k transpositions change it by a factor $(-1)^k$, and $D(e_{p_1}, e_{p_2}, \cdots, e_{p_n}) = \sigma(p) D(e_1, e_2, \cdots, e_n) = \sigma(p)$. This leads to the theorem. □

Here is another important property.

Theorem 18 (Determinant of Product)**.** For all pairs of $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\det(\boldsymbol{BA}) = \det(\boldsymbol{A})\det(\boldsymbol{B}).$$

Proofsketch. Again, let $a_1, \cdots, a_n$ denote the columns of $\boldsymbol{A}$.

First, consider $\det(\boldsymbol{B}) \neq 0$. Define the function C as follows:

$$C(a_1, \cdots, a_n) := \frac{\det(\boldsymbol{BA})}{\det(\boldsymbol{B})} = \frac{D(\boldsymbol{B}a_1, \cdots, \boldsymbol{B}a_n)}{\det(\boldsymbol{B})}.$$

It satisfies the Properties (i)(ii) and (iii), so it equals $\det(\boldsymbol{A})$. Therefore, $\det(\boldsymbol{BA}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.

When $\det(\boldsymbol{B}) = 0$, we define $\boldsymbol{B}(t) = B + tI$. Clearly $\boldsymbol{B}(0) = \boldsymbol{B}$. $D(\boldsymbol{B}(t))$ is a polynomial of degree n with the coefficient of t^n being 1, so $D(\boldsymbol{B}(t)) \neq 0$ for all t near zero but not equal to 0. According to the above, $\det(\boldsymbol{BA}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$ for all such t. The theorem then follows by letting $t \to 0$ and noting the continuity of polynomials. □

Corollary 19 (Determinant and Inverse)**.** An $n \times n$ matrix $\boldsymbol{A}$ is invertible iff $\det(\boldsymbol{A}) \neq 0$.

Proofsketch. Suppose $\boldsymbol{A}$ is not invertible. One can prove the property that a linear mapping from a linear space to itself is invertible iff its representation matrix is invertible. This means the range of $\boldsymbol{A}$ is a proper subspace, and thus the columns are linearly dependent. Then $\det(\boldsymbol{A}) = 0$.

Suppose $\boldsymbol{A}$ is invertible, i.e., there exists $\boldsymbol{B}$ with $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$. By the above theorem, $\det(\boldsymbol{B})\det(\boldsymbol{A}) = \det(\boldsymbol{I})$. Since $\det(\boldsymbol{I}) = 1$, we know $\det(\boldsymbol{A}) \neq 0$. □

In summary, by using Properties (i)(ii) and (iii) to define the determinant, we achieve our goal: an $n \times n$ matrix $\boldsymbol{A}$ is of full rank (i.e., the columns/rows of $\boldsymbol{A}$ are independent) iff $\boldsymbol{A}$ is invertible iff $\det(\boldsymbol{A}) \neq 0$.

Trace. Another important function of square matrices is the trace.

Definition 20 (Trace)**.** The trace of a square matrix $\boldsymbol{A}$, denoted as $\mathrm{tr}(\boldsymbol{A})$, is the sum of the entries on its diagonal:

$$\mathrm{tr}(\boldsymbol{A}) = \sum_i [\boldsymbol{A}]_{ii}.$$

Similarity. The similarity notion is related to basis change and defines an equivalence relation.

Definition 21 (Similarity)**.** The matrix $\boldsymbol{A}$ is called similar to the matrix $\boldsymbol{B}$ if there is an invertible matrix $\boldsymbol{S}$ such that

$$\boldsymbol{A} = \boldsymbol{SBS}^{-1}.$$

For any linear mapping of a linear space into itself, two different representation matrices from two different bases are similar. The determinant and trace are the same for similar matrices, so we can define the determinant and trace of a linear mapping as those of any matrix representing it.

Theorem 22 (Determinant/Trace and Similarity)**.** Similar Matrices have the same determinant and the same trace.

5 Spectral Theory and Eigendecomposition

Spectral theory analyzes linear mappings of a linear space into itself by decomposing them into their basic constituents formed by eigenvectors. Typically, we consider linear spaces over the field K, where K is the field of complex numbers or that of real numbers.

Definition 23 (Eignevalues and Eigenvectors)**.** Consider a square matrix $\boldsymbol{A} \in K^{n\times n}$. If

$$\boldsymbol{Ax} = \lambda \boldsymbol{x}, \text{ where } \lambda \in K, \boldsymbol{x} \in K^n, \boldsymbol{x} \neq 0,$$

then λ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x}$ is the corresponding eigenvector.
The set of all eigenvectors of $\boldsymbol{A}$ associated with an eigenvalue λ spans a subspace, which is called the eigenspace of $\boldsymbol{A}$ with respect to λ. The set of all eigenvalues of $\boldsymbol{A}$ is called the eigenspectrum of $\boldsymbol{A}$.

Note that similar matrices have the same eigenvalues, so eigenvalues are also inherent properties of a linear mapping that are independent of the choice of the basis.

When there are n linearly independent eigenvectors $\boldsymbol{x}_i$ (corresponding to eigenvalues λ_i), they form a basis for the whole linear space. Using the ordered basis $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, the matrix representation of the linear mapping is then a diagonal matrix with λ_i's on the diagonal:

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

which is in a simple form. This is the diagonalization of the matrix $\boldsymbol{A}$.

Unfortunately, there may not always be n linearly independent eigenvectors. Below we investigate when there are.

5.1 General Square Matrices

The determinant can be used to compute the eigenvalues/eigenvectors.

Theorem 24 (Characteristic Polynomial)**.** λ is an eigenvalue of $\boldsymbol{A}$ iff λ is a root of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ of $\boldsymbol{A}$, which is defined as:[1]

$$p_{\boldsymbol{A}}(\lambda) := \det(\lambda \boldsymbol{I} - \boldsymbol{A}).$$

Proof. By definition,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \text{ and } \boldsymbol{x} \neq 0 &\iff (\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = 0 \text{ and } \boldsymbol{x} \neq 0 \\ &\iff \lambda \boldsymbol{I} - \boldsymbol{A} \text{ is not invertible} \\ &\iff \det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0. \end{aligned}$$

□

The characteristic polynomial is a polynomial of degree n, and its coefficient of the highest power λ^n is 1.

According to the fundamental theorem of algebra, a polynomial of degree n with complex coefficients has n complex roots (some may be multiple), which are the eigenvalues of $\boldsymbol{A}$. So for linear spaces over the field of complex numbers, the characteristic polynomials have a full set of roots. However, even over the field of complex numbers, some root λ_i may be multiple and the corresponding eigenspace has dimension (called the geometric multiplicity of λ_i) less than the root's multiplicity (called the algebraic multiplicity of λ_i). In this case, the eigenvectors cannot form a basis covering the whole space.

The good news is that the eigenvectors can form a basis if there are n distinct eigenvalues.

Theorem 25 (Distinct Eigenvalues)**.** Eigenvectors of a matrix $\boldsymbol{A}$ corresponding to distinct eigenvalues are linearly independent.

[1] Some textbooks use $p_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda \boldsymbol{I})$ which differs by a multiplicative factor $(-1)^n$.

Proof. Let λ_i's be a set of distinct eigenvalues, and $\boldsymbol{x}_i$'s be the corresponding eigenvectors. Suppose they are linearly independent, i.e., there are nontrivial linear relations among the $\boldsymbol{x}_i$'s. Among them, there is one involving the least number m of eigenvectors (w.l.o.g., $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$):

$$\sum_{j=1}^{m} b_j \boldsymbol{x}_j = 0, \quad b_j \neq 0, j = 1, \cdots, m. \tag{1}$$

Note that $m \leq 2$ since $\boldsymbol{x}_j \neq 0$. Applying $\boldsymbol{A}$ leads to

$$\sum_{j=1}^{m} b_j \boldsymbol{A} \boldsymbol{x}_j = \sum_{j=1}^{m} b_j \lambda_j \boldsymbol{x}_j = 0$$

Multiply (1) by λ_m and subtract from the above:

$$\sum_{j=1}^{m} (b_j \lambda_j - b_j \lambda_m) \boldsymbol{x}_j = 0.$$

The coefficient of $\boldsymbol{x}_m$ is 0 and none of the others is 0, so this is a linear relation involving only $m-1$ eigenvectors, which is a contradiction. □

For general square matrices on the complex field, one can introduce generalized eigenvectors; together with the genuine eigenvectors, they can cover the whole space.

5.2 Real Symmetric Matrices

Below is one of the basic theorems of linear algebra (and of mathematics itself). To state the theorem, note that a matrix $\boldsymbol{A}$ is symmetric if $\boldsymbol{A}^\top = \boldsymbol{A}$. A set of vectors $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^n$ form an orthonormal basis if $\|\boldsymbol{x}\|^2 := \langle \boldsymbol{x}, \boldsymbol{x} \rangle := \boldsymbol{x}_i^\top \boldsymbol{x}_i = 1$ and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}_i^\top \boldsymbol{x}_j = 0, \forall i \neq j$.[2]

Theorem 26 (Spectral Theorem)**.** If $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ is symmetric, then it has real eigenvalues and has a set of corresponding eigenvectors that form an orthonormal basis.

Proof. For $\boldsymbol{x} \in \mathbb{R}^n$, let $q(\boldsymbol{x}) := \langle \boldsymbol{x}, \boldsymbol{H}\boldsymbol{x} \rangle$, $p(\boldsymbol{x}) := \langle \boldsymbol{x}, \boldsymbol{x} \rangle$. Define the Rayleigh quotient of $\boldsymbol{H}$:

$$R(\boldsymbol{x}) := \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} = \frac{\langle \boldsymbol{x}, \boldsymbol{H}\boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}.$$

$R(\boldsymbol{x})$ is a homogeneous real function of degree zero, i.e., $R(k\boldsymbol{x}) = R(\boldsymbol{x})$ for every scalar $k \neq 0$. Then in seeking its minimum, it suffices to consider the unit sphere $\|\boldsymbol{x}\| = 1$. This is a compact set, so $R(\boldsymbol{x})$ takes on its minimum at some point of the unit sphere; call this point $\boldsymbol{f}$. Let $\boldsymbol{g}$ be any other vector and t be a real variable. Since $\boldsymbol{H}$ is symmetric,

$$R(\boldsymbol{f} + t\boldsymbol{g}) = \frac{\langle \boldsymbol{f}, \boldsymbol{H}\boldsymbol{f} \rangle + 2t\langle \boldsymbol{g}, \boldsymbol{H}\boldsymbol{f} \rangle + t^2 \langle \boldsymbol{g}, \boldsymbol{H}\boldsymbol{g} \rangle}{\langle \boldsymbol{f}, \boldsymbol{f} \rangle + 2t\langle \boldsymbol{g}, \boldsymbol{f} \rangle + t^2 \langle \boldsymbol{g}, \boldsymbol{g} \rangle} =: \frac{q(t)}{p(t)}.$$

[2] The theorem can be generalized to self-adjoint mappings over complex Euclidean space into itself (real symmetric generalized to self-adjoint, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ generalized to conjugate-symmetric scalar product), with only slight modifications to the proof.

It achieves its minimum at $t = 0$, so

$$\frac{d}{dt} R(\boldsymbol{f} + t\boldsymbol{g})\Big|_{t=0} = \frac{\dot{q}(0)p(0) - q(0)\dot{p}(0)}{p(0)^2} = 0.$$

$p(0) = 1$ since $\|\boldsymbol{f}\| = 1$; denoting $R(\boldsymbol{f}) = \min R$ by λ, the above can be rewritten as:

$$\dot{q}(0) - \lambda \dot{p}(0) = 0.$$

Plugging in $\dot{q}(0) = 2\langle \boldsymbol{g}, \boldsymbol{H}\boldsymbol{f}\rangle$ and $\dot{p}(0) = 2\langle \boldsymbol{g}, \boldsymbol{f}\rangle$,

$$\langle \boldsymbol{g}, \boldsymbol{H}\boldsymbol{f} - \lambda \boldsymbol{f}\rangle = 0.$$

This holds for all vectors $\boldsymbol{g}$, so

$$\boldsymbol{H}\boldsymbol{f} - \lambda \boldsymbol{f} = 0,$$

that is, λ is an eigenvalue and $\boldsymbol{f}$ is an eigenvector of $\boldsymbol{H}$.

Now proceed by recursion. Consider the orthogonal complement $\boldsymbol{X}_1$ of , that is, the set of $\boldsymbol{x}$ with $\langle \boldsymbol{x}, \boldsymbol{f}\rangle = 0$. Clearly, $\boldsymbol{X}_1$ is a subspace of dimension $n - 1$. Furthermore, $\boldsymbol{H}$ maps $\boldsymbol{X}_1$ into itself, that is, $\langle \boldsymbol{H}\boldsymbol{x}, \boldsymbol{f}\rangle = 0$ for any $\boldsymbol{x} \in \boldsymbol{X}_1$. This is because

$$\langle \boldsymbol{H}\boldsymbol{x}, \boldsymbol{f}\rangle = \langle \boldsymbol{x}, \boldsymbol{H}\boldsymbol{f}\rangle = \langle \boldsymbol{x}, \lambda\boldsymbol{f}\rangle = \lambda\langle \boldsymbol{x}, \boldsymbol{f}\rangle = 0$$

where the first step follows from $\boldsymbol{H}$ is symmetric. Therefore, we can consider the minimum of $R(\boldsymbol{x})$ over all nonzero vectors in the subspace $\boldsymbol{X}_1$ and repeat the argument to get the second smallest eigenvalue and the corresponding eigenvector of $\boldsymbol{H}$. In this fashion we produce successively n real eigenvalues in increasing order and a set of corresponding eigenvectors that form an orthonormal basis. □

The theorem means that any real symmetric matrix $\boldsymbol{H}$ has the eigendecomposition:

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{H}$ on the diagonal, the columns of $\boldsymbol{U}$ are the corresponding eigenvectors of $\boldsymbol{H}$, and $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$.

5.3 Singular Vector Decomposition

The singular value decomposition (SVD) is also a fundamental theorem of linear algebra and a central matrix decomposition method. It can be deduced based on the spectrum theorem, but it can be applied to all matrices, and it always exists.

Theorem 27 (Singular Value Decomposition)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a rectangular matrix of rank $r \in [0, \min(m, n)]$. The singular value decomposition (SVD) of $\boldsymbol{A}$ is a decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m\times m}, \boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal (i.e., $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$), and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ satisfies $\boldsymbol{\Sigma}_{ii} \geq 0$ and $\boldsymbol{\Sigma}_{ij} = 0, i \neq j$.

The diagonal entries σ_i's of $\boldsymbol{\Sigma}$ are called the singular values. The columns of $\boldsymbol{U}$ are called the left-singular vectors, and the columns of $\boldsymbol{V}$ are called the right-singular vectors. By convention, the singular values are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.

For intuition, given the SVD of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we can see that $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top$, which is exactly the eigendecomposition of the matrix $\boldsymbol{A}^\top\boldsymbol{A}$ (σ_i^2's are the eigenvalues and the columns of $\boldsymbol{V}$ are the eigenvectors). Similarly, $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$ is the eigendecomposition of the matrix $\boldsymbol{A}\boldsymbol{A}^\top$. In fact, we can prove the theorem by constructing SVD in the reverse direction: eigendecompose $\boldsymbol{A}^\top\boldsymbol{A}$ to construct $\boldsymbol{V}$, then construct $\boldsymbol{U}$, and connect them together.

Proofsketch. Suppose W.L.O.G. $n \geq m$. Consider the symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{n\times n}$. By the Spectral Theorem for real symmetric matrices, it has real eigenvalues λ_i's and has a set of corresponding eigenvectors $\boldsymbol{v}_i$'s that form an orthonormal basis; suppose λ_i's are sorted in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots$. Construct $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ by stacking $\boldsymbol{v}_i$'s as the columns; note that $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$. Construct a diagonal matrix $\boldsymbol{D}$ with λ_i's on the diagonal. Then with $\boldsymbol{V}$ as the basis, we can diagonalize the matrix $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top.$$

Furthermore, $\boldsymbol{A}^\top\boldsymbol{A}$ is positive semidefinite, i.e., $\boldsymbol{x}^\top(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{A}\boldsymbol{x}) \geq 0$ for any vector $\boldsymbol{x} \in \mathbb{R}^n$. This implies $\lambda_i \geq 0$.

Now construct $\boldsymbol{u}_i$ as follows. Note that there is the rank r such that $\lambda_i > 0$ for $i \leq r$ and $\lambda_i = 0$ for $i > r$. Then for $i \leq r$,

$$\boldsymbol{u}_i := \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i.$$

We claim that $\boldsymbol{u}_i^\top\boldsymbol{u}_i = 1$ and $\boldsymbol{u}_i^\top\boldsymbol{u}_j = 0$ for $i \neq j \leq r$, since

$$\begin{aligned}
\boldsymbol{u}_i^\top\boldsymbol{u}_i &= \frac{1}{\lambda_i}(\boldsymbol{A}\boldsymbol{v}_i)^\top\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\lambda_i}\boldsymbol{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_i = \frac{1}{\lambda_i}\boldsymbol{v}_i^\top\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{v}_i = 1,\\
\boldsymbol{u}_i^\top\boldsymbol{u}_j &= \frac{1}{\sqrt{\lambda_i\lambda_j}}(\boldsymbol{A}\boldsymbol{v}_i)^\top\boldsymbol{A}\boldsymbol{v}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}}\boldsymbol{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}}\boldsymbol{v}_i^\top\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{v}_j = 0.
\end{aligned}$$

For $i > r$, one can sequentially add a new vector to existing $\boldsymbol{u}_i$'s to make an orthonormal basis. Construct $\boldsymbol{U}$ by stacking $\boldsymbol{u}_i$'s as the columns; note that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$.

Finally, let $\sigma_i := \sqrt{\lambda_i}$ for $i = 1, \ldots, r$. Then we have

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i, \quad i = 1, \cdots, r.$$

This yields

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}.$$

Multiplying $\boldsymbol{V}^\top$ on both sides gives the SVD decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. □

6 Further Reading

The lecture note is partially based on two resources [1, 2]. [1] is a good resource for reviewing the mathematical tools for machine learning; see its website `https://mml-book.github.io` for details. [2] presents an in-depth treatment of linear algebra, while this lecture note only covers a small fraction.

References

[1] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[2] Peter D Lax. *Linear algebra and its applications*. John Wiley & Sons, 2007.