Tutorial: Linear Algebra

ECE421 – Introduction To Machine Learning (Fall 2022)

Stephan Rabanser University of Toronto & Vector Institute for AI stephan@cs.toronto.edu

In machine learning we typically try to analyze high-dimensional datasets consisting of a collection of numbers. As such, we routinely represent numbers in vectors and matrices. The representation and manipulation of vectors and matrices are studied in the field of linear algebra. We briefly revise some of the most relevant material for this course below.

1 Vectors & Matrices

Vector A *vector* $\boldsymbol{x} \in \mathbb{R}^n$ is a one-dimensional list of n values:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1}$$

Matrix A *matrix* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a two-dimensional grid of $m \cdot n$ values:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} \tag{2}$$

Note

Individual matrix columns can be referred to as vectors $\boldsymbol{a}_i \in \mathbb{R}^m$. Further note that vectors can be treated as single-column matrices, i.e. a vector $\boldsymbol{x} \in \mathbb{R}^n$ is can be treated as a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times 1}$. Following standard notation in ML, we denote scalar values as non-bold lower-case letters $s \in \mathbb{R}$, vectors as bold lower-case letters $\boldsymbol{x} \in \mathbb{R}^n$, and matrices as bold upper-case letters $\boldsymbol{X} \in \mathbb{R}^{m \times n}$.

Transposition The *transpose* operation swaps the row and column indices of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \qquad \boldsymbol{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nm} \end{bmatrix} \tag{3}$$

Note that $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$. Analogously, vector transposition transforms a column vector $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ into a row vector $\boldsymbol{x}^\top \in \mathbb{R}^{1\times n}$ (or vice versa).

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \tag{4}$$

Square Matrix A matrix is called a *square matrix* if both the row and the column dimensions have the same size, i.e. if $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

Symmetric Matrix A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called a *symmetric matrix* if the matrix is identical to its transpose, i.e. $\boldsymbol{A} = \boldsymbol{A}^\top$.

Identity Matrix The *identity matrix* $\boldsymbol{I}$ of size n is a matrix that contains 1 on its main diagonal[1] and 0 for all other entries:

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{5}$$

Trace The *trace* of a square matrix $\boldsymbol{A}$ is the sum of its diagonal elements:

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}. \tag{6}$$

Definitness A square and symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called *positive definite* if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ for all non-zero $\boldsymbol{x} \in \mathbb{R}^n$. The definition can be relaxed to *positive semi-definiteness* by replacing $>$ with $\geq$, i.e. $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ for all non-zero $\boldsymbol{x} \in \mathbb{R}^n$.

Rank The *rank* of a matrix A is the dimension of the vector space spanned by its columns.

Orthogonality A square matrix $\boldsymbol{A}$ is called *orthogonal* if

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}. \tag{7}$$

Note that this implies that an orthogonal matrix is always invertible since

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} \Longrightarrow \boldsymbol{A}^{-1} = \boldsymbol{A}^\top. \tag{8}$$

Diagonal Matrix A matrix $\boldsymbol{A}$ is called *diagonal* if all the entries outside of its main diagonal are 0.

[1] The main diagonal is the top-left to bottom-right diagonal.

Diagonalizability A square matrix $\boldsymbol{A}$ is called *diagonalizable* if there exists a diagonal matrix $\boldsymbol{D}$ and an invertible matrix $\boldsymbol{P}$ such that

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}. \tag{9}$$

2 Basic Operations

2.1 Arithmetic Operations

Matrix Addition & Subtraction For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$, their addition and subtraction are preformed element-wise:

$$\begin{aligned}\boldsymbol{A} + \boldsymbol{B} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}\end{aligned} \tag{10}$$

$$\begin{aligned}\boldsymbol{A} - \boldsymbol{B} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \dots & a_{1n} - b_{1n} \\ a_{21} - b_{21} & a_{22} - b_{22} & \dots & a_{2n} - b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} - b_{m1} & a_{m2} - b_{m2} & \dots & a_{mn} - b_{mn} \end{bmatrix}\end{aligned} \tag{11}$$

Matrix Multiplication For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times k}$ and a matrix $\boldsymbol{B} \in \mathbb{R}^{k \times n}$, the matrix multiplication of $\boldsymbol{A} \cdot \boldsymbol{B}$ is given as follows:

$$\begin{aligned}\boldsymbol{A} \cdot \boldsymbol{B} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mk} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{kn} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{k} a_{1i}b_{i1} & \sum_{i=1}^{k} a_{1i}b_{i2} & \dots & \sum_{i=1}^{k} a_{1i}b_{in} \\ \sum_{i=1}^{k} a_{2i}b_{i1} & \sum_{i=1}^{k} a_{2i}b_{i2} & \dots & \sum_{i=1}^{k} a_{2i}b_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{k} a_{mi}b_{i1} & \sum_{i=1}^{k} a_{mi}b_{i2} & \dots & \sum_{i=1}^{k} a_{mi}b_{in} \end{bmatrix}\end{aligned} \tag{12}$$

Caution

Matrix multiplication has a few peculiar properties differentiating it from scalar multiplcation:

- In general, $\boldsymbol{AB} \neq \boldsymbol{BA}$ (i.e. $\boldsymbol{A}$ and $\boldsymbol{B}$ do not commute).
- Matrix multiplication is not performed element-wise! However, the *Hadamard product* defines element-wise multiplication for two shape-identical matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$:

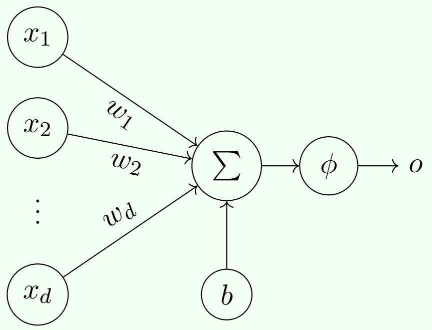
$$\begin{aligned}\boldsymbol{A} * \boldsymbol{B} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} * \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} \cdot b_{11} & a_{12} \cdot b_{12} & \dots & a_{1n} \cdot b_{1n} \\ a_{21} \cdot b_{21} & a_{22} \cdot b_{22} & \dots & a_{2n} \cdot b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} \cdot b_{m1} & a_{m2} \cdot b_{m2} & \dots & a_{mn} \cdot b_{mn} \end{bmatrix}\end{aligned}$$

Inner Product For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ we define the inner product (sometimes also called *dot product* or the *scalar product*) as follows:

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^\top \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \tag{13}$$

The dot product between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ produces a scalar $c \in \mathbb{R}$ representing the sum of all (dimension-aligned) multiplicative interactions.

Example Usage in Machine Learning



Consider the linear/logistic regression algorithm which can be used to determine the best linear fit/classification to/of the underlying data. For a given data point $\boldsymbol{x} \in \mathbb{R}^d$, we can compute it's prediction via a weighted sum (using weights $\boldsymbol{w} \in \mathbb{R}^d$) of all input features. The weighted average is typically followed by an activation function ϕ, which for linear regression corresponds to the identity map while it typically corresponds to the a softmax function for logistic regression.

$$y = \phi(\langle \boldsymbol{x}, \boldsymbol{w} \rangle + b) = \phi(\boldsymbol{x}^\top \boldsymbol{w} + b) = \phi\left(\left(\sum_{i=1}^{n} x_i w_i\right) + b\right) \tag{14}$$

2.2 ℓ_p Norms

For a real number $p \geq 1$, the ℓ_p-norm (or simply p-norm) of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$||\boldsymbol{x}||_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}. \tag{15}$$

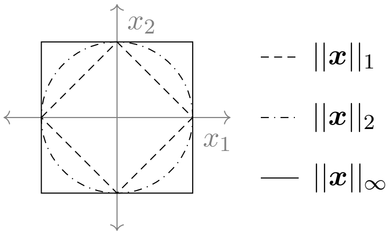


Figure 1: Illustration of unit circles for vectors $\boldsymbol{x} \in \mathbb{R}^2$ for various norms.

In particular, the ℓ_1 and ℓ_2 norms are often used in machine learning:

$$||\boldsymbol{x}||_1 = \sum_{i=1}^{n} |x_i| = |x_1| + \ldots + |x_n| \tag{16}$$

$$||\boldsymbol{x}|| = ||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + \ldots + x_n^2} \tag{17}$$

By taking $p \longrightarrow \infty$ we get the ℓ_∞ norm which is defined as the maximum over all absolute vector elements:

$$||\boldsymbol{x}||_\infty = \max_{i \in [n]} \{|x_1|, \ldots, |x_n|\} \tag{18}$$

Further note that higher-order ℓ_p norms are upper-bounded by lower-order ℓ_p norms:

$$||\boldsymbol{x}||_{p+a} \leq ||\boldsymbol{x}||_p \tag{19}$$

Example Usage in Machine Learning

Many machine learning algorithms use ℓ_p norms to either

- *measure the distance* between points in a high-dimensional data space (e.g., k-nearest neighbor classification); or to
- *bound the magnitude* of a vector to a specific value (e.g., regularization in linear regression).

Example

Consider the vector $\boldsymbol{x} = \begin{bmatrix} 5 \\ 2 \\ -3 \end{bmatrix}$. Then the ℓ_1, ℓ_2, and ℓ_∞ norms are given as follows:

$$\begin{aligned} ||\boldsymbol{x}||_1 &= |5| + |2| + |-3| = 10 \\ ||\boldsymbol{x}||_2 &= \sqrt{5^2 + 2^2 + (-3)^2} = \sqrt{25 + 4 + 9} = 6.1644 \\ ||\boldsymbol{x}||_\infty &= 5 \end{aligned} \tag{20}$$

2.3 Determinant

The *determinant* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ yields a scalar that captures a set of properties associated with the linear map represented by the matrix. For instance, an invertible matrix has a determinant not equal to 0. Moreover, the determinant can be used to define the characteristic polynomial of a matrix, and further captures the degree of volume change induced by the matrix on an n-dimensional parallelepiped.

The determinant for $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{B} \in \mathbb{R}^{3\times 3}$ matrices can be computed as follows (higher order determinants are beyond the scope of this class):

$$\det(\boldsymbol{A}) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc \tag{21}$$

$$\begin{aligned} \det(\boldsymbol{B}) &= \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \\ &= a \begin{bmatrix} e & f \\ h & i \end{bmatrix} - b \begin{bmatrix} d & f \\ g & i \end{bmatrix} + c \begin{bmatrix} d & e \\ g & h \end{bmatrix} \\ &= aei + bfg + cdh - ceg - bdi - afh \end{aligned} \tag{22}$$

Example

Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 3 & 7 \\ 1 & -4 \end{bmatrix}$. Then the determinant is given by:

$$\det(A) = 3 \cdot (-4) - 7 \cdot 1 = -19. \tag{23}$$

3 Matrix Decompositions

It is often useful to factorize a matrix into a product of matrices. In general, there exists a wide variety of decompositions[2]. We will focus on the most relevant of these factorizations for this class, namely the *Eigendecomposition* and the *Singular Value Decomposition*.

3.1 Eigendecomposition

The Eigendecomposition of a matrix factorizes a square diagonalizable matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ into a set of *eigenvalues* and *eigenvectors*.

Eigenvalues and Eigenvectors A non-zero vector $\boldsymbol{v} \in \mathbb{R}^n$ is called an eigenvector of a square diagonalizable matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ if, for a scalar $\lambda \in \mathbb{R}$ it satisfies:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{24}$$

Intuitively speaking, an eigenvector $\boldsymbol{v}$ is a vector which, under the transformation applied by $\boldsymbol{A}$, is only scaled in its magnitude. In particular, such vectors $\boldsymbol{v}$ stay on on their own span and are only elongated or shrinked. The degree of change in magnitude inflicted by $\boldsymbol{A}$ can be summarized in a single scalar λ, which is the eigenvalue corresponding to $\boldsymbol{v}$. The set of all eigenvalue-eigenvector combinations can be computed by solving for the *characteristic polynomial* yielded by $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$

[2] https://en.wikipedia.org/wiki/Matrix_decomposition

Eigen-based Decomposition For a square diagonalizable matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the eigendecomposition is given by:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} \tag{25}$$

Here, $\boldsymbol{Q}$ is also a square matrix in $\mathbb{R}^{n\times n}$ and contains the eigenvectors $\boldsymbol{q}_i \in \mathbb{R}^n$ as columns. $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries correspond to the eigenvalues of the respective eigenvectors from $\boldsymbol{Q}$. In particular $\Lambda_{ii} = \lambda_i$ is the eigenvalue associated with eigenvector $\boldsymbol{q}_i$.

Example

Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$. We first diagonalize $\boldsymbol{A}$ as follows:

$$\begin{aligned} \boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q} &= \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} &= \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \\ \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \end{aligned} \tag{26}$$

Next, we derive a separate equation for each diagonal entry:

$$\begin{aligned} \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} &= \begin{bmatrix} ax \\ cx \end{bmatrix} \\ \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} &= x \begin{bmatrix} a \\ c \end{bmatrix} \end{aligned} \tag{27}$$

$$\begin{aligned} \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} &= \begin{bmatrix} by \\ dy \end{bmatrix} \\ \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} &= y \begin{bmatrix} b \\ d \end{bmatrix} \end{aligned} \tag{28}$$

In its current form, these two equations correspond to the eigenvector problem discussed in Equation 24. Letting $\boldsymbol{v} = \begin{bmatrix} a \\ c \end{bmatrix}$ and $\boldsymbol{w} = \begin{bmatrix} b \\ d \end{bmatrix}$ we get

$$\begin{cases} \boldsymbol{A}\boldsymbol{v} = x\boldsymbol{v} \\ \boldsymbol{A}\boldsymbol{w} = y\boldsymbol{w} \end{cases} \tag{29}$$

Representing each of the above cases $\boldsymbol{A}\boldsymbol{u} = x\boldsymbol{u}$, we solve for $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}$.

$$\begin{aligned} \det(\boldsymbol{A} - \lambda\boldsymbol{I}) &= 0 \\ \det(\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}) &= 0 \\ \det(\begin{bmatrix} 2-\lambda & 3 \\ 2 & 1-\lambda \end{bmatrix}) &= 0 \end{aligned} \tag{30}$$

Solving for the determinant yields:

$$\begin{aligned} (2-\lambda)(1-\lambda) - 2\cdot 3 &= 0 \\ \lambda^2 - 3\lambda - 4 &= 0 \\ (\lambda+1)(\lambda-4) &= 0 \qquad \implies \qquad \lambda = -1 \text{ or } \lambda = 4 \end{aligned} \tag{31}$$

Substituting back into Equation 27 and Equation 28 using $x = -1$ and $y = 4$ we get:

$$\begin{cases} \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} = -1 \begin{bmatrix} a \\ c \end{bmatrix} \\ \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} = 4 \begin{bmatrix} b \\ d \end{bmatrix} \end{cases} \tag{32}$$

Solving these two equations yields:

$$\begin{cases} a + c = 0 \\ 2b - 3d = 0 \end{cases} \tag{33}$$

Thus, the matrix $\boldsymbol{Q}$ is given as:

$$\boldsymbol{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} -c & \frac{3}{2}d \\ c & d \end{bmatrix} \tag{34}$$

3.2 Singular Value Decomposition

The singular value decomposition generalizes the concepts from the Eigendecomposition to general matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{35}$$

Here, $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix, $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ is a diagonal matrix, and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix. The SVD can be derived by

1. Computing $\boldsymbol{A}^\top\boldsymbol{A}$, yielding a square symmetric matrix.
2. Finding the eigenvalues for $\boldsymbol{A}^\top\boldsymbol{A}$.
3. Finding the eigenvectors for $\boldsymbol{A}^\top\boldsymbol{A}$.
4. Normalizing the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ to get $\boldsymbol{V}$.
5. Finding $\boldsymbol{U}$ using the normalized eigenvectors of $\boldsymbol{V}$ where $\boldsymbol{u}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i$.

Example Usage in Machine Learning

The singular value decomposition is frequently used as part of unsupervised learning algorithms. In particular, the SVD forms the basis of an algorithm called *principal components analysis (PCA)*, which reduces data dimensions such that the reduced data still contains the maximum amount of variability given the number of reduced dimensions. The reduction axes picked by PCA directly correspond to the singular vectors identified by the singular value decomposition.

Example

Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$. We start by computing $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix} \tag{36}$$

We derive the characteristic polynomial as

$$\det(\boldsymbol{A}^\top \boldsymbol{A} - \lambda \boldsymbol{I}) = -\lambda(\lambda^2 - 34\lambda + 225) = -\lambda(\lambda - 25)(\lambda - 9) \tag{37}$$

The singular values are hence given by $\sigma_1 = \sqrt{25} = 5$ and $\sigma_2 = \sqrt{9} = 3$. Next, we find the orthonormal (orthogonal & normalized) set of the eigenvectors in $\boldsymbol{A}^\top \boldsymbol{A}$, which will form our columns in $\boldsymbol{V}$. The eigenvalues for $\boldsymbol{A}^\top \boldsymbol{A}$ are given by 25, 9, and 0.
For $\lambda = 25$, we get

$$\boldsymbol{A}^\top \boldsymbol{A} - 25\boldsymbol{I} = \begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix} \xRightarrow[\text{reduction}]{\text{row}} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \xRightarrow[\text{vector}]{\text{unit-length}} \boldsymbol{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \tag{38}$$

For $\lambda = 9$, we get

$$\boldsymbol{A}^\top \boldsymbol{A} - 25\boldsymbol{I} = \begin{bmatrix} 4 & 12 & 2 \\ 12 & 4 & -2 \\ 2 & -2 & -1 \end{bmatrix} \xRightarrow[\text{reduction}]{\text{row}} \begin{bmatrix} 1 & 0 & -\frac{1}{4} \\ 0 & 1 & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix} \xRightarrow[\text{vector}]{\text{unit-length}} \boldsymbol{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{18}} \\ -\frac{1}{\sqrt{18}} \\ \frac{4}{\sqrt{18}} \end{bmatrix} \tag{39}$$

For $\lambda = 0$, we find a unit-vector perpendicular to $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. A perpendicular vector to $\boldsymbol{v}_1 = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ requires $-a = b$ and an 0 inner-product with $\boldsymbol{v}_2$, i.e. $\frac{2a}{\sqrt{18}} + \frac{4c}{\sqrt{18}} = 0$ yielding $-a = 2c$. Hence, $\boldsymbol{v}_3 = \begin{bmatrix} a \\ -a \\ -\frac{a}{2} \end{bmatrix}$. Normalizing the vector requires $a = \frac{2}{3}$ and results in $\boldsymbol{v}_3 = \begin{bmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}$.
Having obtained all entries in $\boldsymbol{V}$, we can now compute the values in $\boldsymbol{U}$ using $\boldsymbol{u}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i$.
This yields the final singular value decomposition:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{18}} & -\frac{1}{\sqrt{18}} & \frac{4}{\sqrt{18}} \\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix} \tag{40}$$

More details can be found in `https://www.d.umn.edu/~mhampton/m4326svd_example.pdf`.