Politecnico di Milano
School of Industrial and Information Engineering
Academic Year 2024-2025

# Numerical Analysis for Machine Learning

Professor: Edie Miglio

*Last updated: October 1, 2024*


Made by Roberto Benatuil Valera

# Contents

Chapter 1

# Numerical Linear Algebra tools

## 1.1 Introduction: Recap of Linear Algebra

In this section we will review some basic concepts of Linear Algebra that will be useful for the rest of the course.

### 1.1.1 Matrix-vector multiplication

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$, the matrix-vector multiplication $y = Ax$ is defined as:

$$y_i = \sum_{j=1}^{n} A_{ij} x_j \tag{1.1}$$

A matrix-vector multiplication can be considered as a linear combination of the columns of the matrix $A$. Lets see an example:

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} x_1 + \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} x_2 \tag{1.2}$$

### 1.1.2 Column space of a matrix

The column space of a matrix $A \in \mathbb{R}^{m \times n}$ is the subspace of $\mathbb{R}^m$ spanned by the columns of $A$. In other words, it is the set of all possible linear combinations of the columns of $A$. The column space of a matrix is denoted as $C(A)$.

If the columns of $A$ are linearly independent, then the column space of $A$ is the entire $\mathbb{R}^m$. If the columns of $A$ are linearly dependent, then the column space of $A$ is a subspace of $\mathbb{R}^m$ with dimension equal to the rank of $A$.

The rank of a matrix $A$ is is the size of the largest set of linearly independent columns of $A$. It is denoted as $rank(A)$. Note that $rank(A) = rank(A^T)$.

### 1.1.3 System of linear equations

A system of linear equations is a set of $m$ equations with $n$ unknowns of the form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned} \tag{1.3}$$

This system can be written in matrix form as $Ax = b$, where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

The system $Ax = b$ has a solution if and only if $b \in C(A)$. If $b \in C(A)$, then the system has a unique solution if and only if $rank(A) = n$. If $rank(A) < n$, then the system has infinitely many solutions.

### 1.1.4 CR factorization

The CR factorization of a matrix $A \in \mathbb{R}^{m \times n}$, with $m \geq n$, is a factorization of $A$ as $A = CR$, where $C \in \mathbb{R}^{m \times r}$ is a matrix with the linearly independent columns of $A$ and $R \in \mathbb{R}^{r \times n}$ is obtained by determining the coefficients of the linear combination of the columns of $C$ that give the columns of $A$. In this factorization, $r = rank(A)$.

Lets see an example:

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix} = CR \tag{1.4}$$

The matrix $C$ is also called the Row Reduced Echelon Form of $A$.

### 1.1.5 Matrix-matrix multiplication

Given two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, the matrix-matrix multiplication $C = AB$ is defined as:

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} \tag{1.5}$$

A matrix-matrix multiplication can be considered as the outer product of the columns of $A$ and the rows of $B$. Lets see an example:

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \begin{bmatrix} 3 & 4 \end{bmatrix} \tag{1.6}$$

Note that each outer product generates a matrix of the same size as the result matrix, but always with rank 1. So the matrix-matrix multiplication can be considered as a sum of rank 1 matrices, obtained by the outer products of the columns of $A$ and the rows of $B$.

### 1.1.6 Null space of a matrix

The null space of a matrix $A \in \mathbb{R}^{m \times n}$ is the set of all vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. The null space of a matrix is denoted as $N(A)$. It is also called the kernel of $A$, denoted as $ker(A)$.

Formally, we have that:

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\} \tag{1.7}$$

The null space of a matrix is a subspace of $\mathbb{R}^n$. The dimension of the null space of a matrix is called the nullity of the matrix.

### 1.1.7 Fundamental subspaces of a matrix

Given a matrix $A \in \mathbb{R}^{m \times n}$, we can define four fundamental subspaces:

- The column space of $A$, denoted as $C(A)$
- The row space of $A$, denoted as $C(A^T)$
- The null space of $A$, denoted as $N(A)$
- The left null space of $A$, denoted as $N(A^T)$

These subspaces are related by the following properties:

$$\begin{aligned} C(A) &\perp N(A^T) \\ C(A^T) &\perp N(A) \end{aligned} \tag{1.8}$$

They also satisfy the following dimensions properties:

$$\begin{aligned} dim(C(A)) + dim(N(A)) &= n \\ dim(C(A^T)) + dim(N(A^T)) &= m \end{aligned} \tag{1.9}$$

This is known as the Rank-Nullity Theorem.

### 1.1.8 Orthogonal matrices

An orthogonal matrix is a square matrix $Q \in \mathbb{R}^{n \times n}$ such that $Q^TQ = I$, where $I$ is the identity matrix. This implies that $Q^T = Q^{-1}$.

Now, consider that $Q$ is an orthogonal matrix, and set $w = Q^Tx$. Then we have that:

$$\begin{aligned} \|w\|^2 = w^Tw &= x^TQQ^Tx \\ &= x^Tx = \|x\|^2 \end{aligned} \tag{1.10}$$

This means that the norm of a vector is preserved under an orthogonal transformation. This is called an isometry. It is a useful property for numerical algorithms, as it helps to avoid numerical instability.

There are two main types of orthogonal transformations that we are interested:

#### Rotation matrices

A rotation matrix is an orthogonal matrix that represents a rotation in $\mathbb{R}^2$ or $\mathbb{R}^3$. In $\mathbb{R}^2$, a rotation matrix is of the form:

$$Q(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{1.11}$$

#### Reflection matrices

A reflection matrix is an orthogonal matrix that represents a reflection with respect to a hyperplane. If $n$ denotes the unit normal vector to the hyperplane, then the reflection matrix is of the form:

$$Q = I - 2nn^T \tag{1.12}$$

Note that the inverse of this matrix is itself, as $Q^T = Q^{-1}$ and in this case, $Q$ is symmetric ($Q = Q^T$).

### 1.1.9 QR factorization

The QR factorization of a matrix $A \in \mathbb{R}^{m \times n}$, with $m \geq n$, is a factorization of $A$ as $A = QR$, where $Q \in \mathbb{R}^{m \times n}$ is an orthogonal matrix and $R \in \mathbb{R}^{n \times n}$ is an upper triangular matrix.

#### Gram-Schmidt process

The Gram-Schmidt process is a method to compute the QR factorization of a matrix. Given a matrix $A \in \mathbb{R}^{m \times n}$, the Gram-Schmidt process computes an orthonormal basis for the column space of $A$, as follows:

$$\begin{aligned} q_1 &= \frac{a_1}{\|a_1\|} \\ q_i &= a_i - \sum_{j=1}^{i-1} (q_j^T a_i) q_j \quad \forall i = 2, \ldots, n \end{aligned} \tag{1.13}$$

where $a_i$ denotes the $i$-th column of $A$. The matrix $Q$ is obtained by stacking the vectors $q_i$ as columns. The matrix $R$ is obtained by computing the coefficients of the linear combination of the columns of $Q$ that give the columns of $A$.

### 1.1.10 Eigenvalues and eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, a scalar $\lambda$ is called an eigenvalue of $A$ if there exists a vector $v \in \mathbb{R}^n$ such that:

$$Av = \lambda v \tag{1.14}$$

The vector $v$ is called an eigenvector of $A$ associated with the eigenvalue $\lambda$.

Let $P$ be the matrix whose columns are the eigenvectors of $A$, and $\Lambda$ be the diagonal matrix whose diagonal elements are the eigenvalues of $A$. Then we have that:

$$A = P\Lambda P^{-1} \tag{1.15}$$

This is called the eigendecomposition of $A$.

The eigenvalues of a matrix are the roots of the characteristic polynomial of $A$, which is defined as:

$$det(A - \lambda I) = 0 \tag{1.16}$$

### 1.1.11 Similar matrices

Two square matrices $A$ and $B$ are called similar if there exists a non-singular matrix $M$ such that:

$$B = M^{-1}AM \tag{1.17}$$

Similar matrices have the same eigenvalues, but not necessarily the same eigenvectors. Let $(\lambda, y)$ be an eigenpair of $B$, then we have:

$$By = \lambda y \Rightarrow M^{-1}AMy = \lambda y \Rightarrow A(My) = \lambda(My) \tag{1.18}$$

This means that $My$ is an eigenvector of $A$ associated with the eigenvalue $\lambda$. So, to obtain the eigenvectors of $A$ from the eigenvectors of $B$, we need to multiply the eigenvectors of $B$ by $M$.

This property can be useful. For example, if we want to compute the eigenvalues of a matrix $A$, we can find some transformation $M$ such that $M^{-1}AM = B$ is a simpler matrix to work with, usually a lower triangular matrix. Then we can compute the eigenvalues of $B$ and obtain the eigenvalues of $A$. $M$ is obtained by the permutation matrices to get from $A$ to $B$. The Givens and Householder transformations are examples of such method.

## 1.2 Power method

The power method is an iterative algorithm to compute the dominant eigenvalue of a matrix (i.e., the eigenvalue with the largest magnitude). The algorithm is as follows:

**Algorithm 1** Power method

1: Choose a random vector $x^{(0)}$, s.t. $\|x^{(0)}\| = 1$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad y^{(k)} = Ax^{(k-1)}$
4: $\quad x^{(k)} = \frac{y^{(k)}}{\|y^{(k)}\|}$
5: $\quad \lambda^{(k)} = x^{(k)T}Ax^{(k)}$
6: **end for**

The convergence rate is determined by the ratio of the largest eigenvalue to the second largest eigenvalue.

### 1.2.1 Rayleigh quotient

The Rayleigh quotient is the base of the power method algorithm. Given a matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the Rayleigh quotient is defined as:

$$R(x) = \frac{x^T A x}{x^T x} \tag{1.19}$$

For every eigenpair $(\lambda, v)$ of $A$, we have that:

$$R(v) = \frac{v^T A v}{v^T v} = \frac{v^T \lambda v}{v^T v} = \lambda \tag{1.20}$$

This means that the Rayleigh quotient is equal to the eigenvalue associated with the eigenvector $v$. This property is used in the power method to compute the dominant eigenvalue of a matrix.

### 1.2.2 Proof of convergence for the power method

The power method converges to the dominant eigenvalue of a matrix. The sketch proof is as follows:

Let $x^{(0)}$ be our initial vector, and let $\{v_1, \ldots, v_n\}$ be the eigenvectors of $A$. We can write $x^{(0)}$ as a linear combination of the eigenvectors of $A$ (since the eigenvectors of $A$ form a basis of $\mathbb{R}^n$):

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i v_i \tag{1.21}$$

Then we have that:

$$A x^{(0)} = \sum_{i=1}^{n} \alpha_i A v_i = \sum_{i=1}^{n} \alpha_i \lambda_i v_i \tag{1.22}$$

Since in every iteration $k$ of the power method we apply the matrix $A$ to the vector $x^{(k-1)}$, we have that:

$$A x^{(k-1)} = A^k x^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k v_i \tag{1.23}$$

Now, let us factorize the previous equation by the dominant eigenvalue $(\lambda_1)^k$:

$$A^k x^{(0)} = (\lambda_1)^k \left( \alpha_1 v_1 + \sum_{i=2}^{n} \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k v_i \right) \tag{1.24}$$

Note that the term inside the parenthesis converges to zero as $k \to \infty$, since the ratio of the other eigenvalues to the dominant eigenvalue is less than 1. This means that $x^(k)$ converges to the direction of the dominant eigenvector $v_1$. When it is normalized, its Rayleigh quotient converges to the dominant eigenvalue $\lambda_1$.

### 1.2.3 Inverse power method

The inverse power method is an iterative algorithm to compute the eigenvalue with the smalles magnitude of a matrix. Note that the smallest eigenvalue of a matrix is the largest eigenvalue of its inverse, since:

$$Ax = \lambda x \Rightarrow A^{-1}x = \frac{1}{\lambda}x \tag{1.25}$$

The algorithm is similar to the power method, but instead of applying the matrix $A$ to the vector $x$, we apply the inverse of the matrix $A$:

**Algorithm 2** Inverse power method

1: Choose a random vector $x^{(0)}$, s.t. $\|x^{(0)}\| = 1$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad y^{(k)} = A^{-1}x^{(k-1)}$
4: $\quad x^{(k)} = \frac{y^{(k)}}{\|y^{(k)}\|}$
5: $\quad \lambda^{(k)} = x^{(k)T}Ax^{(k)}$
6: **end for**

In practice, we normally don't compute the inverse of the matrix $A$, but instead solve the linear system $Ax = y^{(k)}$ in each iteration.

### 1.2.4 Shifted inverse power method

The shifted inverse power method is an iterative algorithm to compute the eigenvalue with the smalles magnitude of a matrix, but with a shift $\mu$ added to the matrix $A$. The algorithm is as follows:

**Algorithm 3** Shifted inverse power method

1: Choose a random vector $x^{(0)}$, s.t. $\|x^{(0)}\| = 1$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad y^{(k)} = (A - \mu I)^{-1}x^{(k-1)}$
4: $\quad x^{(k)} = \frac{y^{(k)}}{\|y^{(k)}\|}$
5: $\quad \lambda^{(k)} = x^{(k)T}Ax^{(k)}$
6: **end for**

Note that with this algorithm, we are computing the eigenvalue of $A$ that is closest to the shift $\mu$. This can be useful to compute the eigenvalues of a matrix that are close to a given value.

## 1.3 Symmetric matrices

A matrix $A \in \mathbb{R}^{n\times n}$ is called symmetric if $A = A^T$. Symmetric matrices have important properties:

- The eigenvectors of a symmetric matrix form an orthonormal basis of $\mathbb{R}^n$.

- All the eigenvalues of a symmetric matrix are real.

Let us prove the first property:

Let $A$ be a symmetric matrix, and let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues. Let $v_1, \ldots, v_n$ be the eigenvectors associated with the eigenvalues $\lambda_1, \ldots, \lambda_n$. Let us take two eigenvectors $v_i$ and $v_j$, such that $i \neq j$. Then we have that:

$$Av_i = \lambda_i v_i \quad \text{and} \quad Av_j = \lambda_j v_j \tag{1.26}$$

Then we have that:

$$(A - \lambda_i I)v_i = 0 \quad \text{and} \quad (A - \lambda_i I)v_j = (\lambda_j - \lambda_i)v_j$$

$$\Rightarrow v_i \in N(A - \lambda_i I) \quad \text{and} \quad v_j \in C(A - \lambda_i I) \tag{1.27}$$

Since $A$ is symmetric, we have that $A - \lambda_i I$ is also symmetric. Then we have that:

$$N(A - \lambda_i I) = N((A - \lambda_i)^T) \perp C(A - \lambda_i I) \tag{1.28}$$

Concluding that $v_i$ and $v_j$ are orthogonal. Since this holds for all pairs of eigenvectors, we have that the eigenvectors of a symmetric matrix form an orthonormal basis of $\mathbb{R}^n$.

Now, let us prove the second property:

Since $A$ is symmetric, we have that $A = A^T$. Then we have that:

$$\begin{aligned} Ax &= \lambda x \\ A\bar{x} &= \bar{\lambda}\bar{x} \end{aligned} \tag{1.29}$$

Then we have that:

$$\begin{aligned} \bar{x}^T Ax &= \lambda x^T x = \lambda \|x\|^2 \\ x^T A\bar{x} &= \bar{\lambda} x^T \bar{x} = \bar{\lambda}\|x\|^2 \end{aligned} \tag{1.30}$$

Since $A = A^T$, we have that:

$$\bar{x}^T Ax = (Ax)^T \bar{x} = x^T A^T \bar{x} = x^T A\bar{x} \tag{1.31}$$

Then we have that:

$$\lambda\|x\|^2 = \bar{\lambda}\|x\|^2 \Rightarrow \lambda = \bar{\lambda} \tag{1.32}$$

This means that the eigenvalues of a symmetric matrix are real.

### 1.3.1 Symmetric positive definite matrices

A matrix $A \in \mathbb{R}^{n \times n}$ is called symmetric positive definite if it is symmetric and if for every vector $x \in \mathbb{R}^n$ we have that:

$$x^T A x > 0 \quad \forall x \neq 0 \tag{1.33}$$

Symmetric positive definite matrices have important properties:

- All the eigenvalues of a symmetric positive definite matrix are positive.
- The Cholesky factorization of a symmetric positive definite matrix exists and is unique:

$$A = L^T L \tag{1.34}$$

In fact, a symmetric matrix is positive definite if and only if all its eigenvalues are positive, so:

$$x^T A x > 0 \quad \forall x \neq 0 \quad \Leftrightarrow \quad \lambda_i > 0 \quad \forall i \tag{1.35}$$

Note that the quantity $x^T A x$ is called the energy of the vector $x$ with respect to the matrix $A$. This quantity is always positive for a symmetric positive definite matrix.

## 1.4 Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is a factorization of a matrix $A \in \mathbb{R}^{m \times n}$ as:

$$A = U \Sigma V^T \tag{1.36}$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times n}$ is a quasi-diagonal matrix with the singular values of $A$. The singular values of $A$ are the square roots of the eigenvalues of $A^T A$.

Note that if $rank(A) = r$, then the matrix $\Sigma$ has $r$ non-zero singular values, and the remaining singular values are zero. Assume that the singular values of $A$ are $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$. Then we have that:

$$\begin{aligned} A v_i &= \sigma_i u_i \quad \forall i = 1, \ldots, r \\ A v_i &= 0 \quad \forall i = r+1, \ldots, n \end{aligned} \tag{1.37}$$

where $u_i$ and $v_i$ are the columns of $U$ and $V$, respectively. The vectors $u_i$ and $v_i$ are called the left and right singular vectors of $A$, respectively.

The SVD can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{1.38}$$

### 1.4.1 Economy SVD

The economy SVD is a factorization of a matrix $A \in \mathbb{R}^{m \times n}$ as:

$$A = U_r \Sigma_r V_r^T \tag{1.39}$$

where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $\Sigma \in \mathbb{R}^{r \times r}$, with $r = rank(A)$.

The economy SVD is useful when we are only interested in the first $r$ singular values of $A$, which are the non-zero singular values.

### 1.4.2 Low-rank approximation

The SVD can be used to compute a low-rank approximation of a matrix $A \in \mathbb{R}^{m \times n}$. Given a rank $k$, with $k < r = rank(A)$, the low-rank approximation of $A$ is given by:

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T \tag{1.40}$$

where $U_k \in \mathbb{R}^{m \times k}$, $V_k \in \mathbb{R}^{n \times k}$ and $\Sigma_k \in \mathbb{R}^{k \times k}$, with $k < r = rank(A)$.

Because the singular values of $A$ are sorted in decreasing order, the low-rank approximation only considers the first $k$ singular values of $A$, as they are the components of $A$ with the largest contribution.

The low-rank approximation of a matrix is useful for data compression, as it allows to represent a matrix with a smaller number of parameters. Note that the low-rank approximation of a matrix is the best rank-$k$ approximation of the matrix in the Frobenius norm. This is called the Eckart-Young theorem.