# 1 Mathematics Refresher for Machine Learning

## 1.1 Sets

A set is a well-defined collection of distinct objects (possibly infinite or uncountable). E.g:

- $\{1, 2, 3\}, \{a, e, i, o, u\}, \{\pi, e\}$
- Integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
- Positive Integers $\mathbb{Z}_{++} = \{1, 2, 3, \ldots\}$
- Real Numbers $\mathbb{R}$

If x is an elements of set Z we write $x \in Z$. E.g: $x \in \mathbb{R}$ means x is a real number. Set builder notation:

- Positive reals $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$

## 1.2 Sets: empty set, cardinality, intersection, union

The empty set is the set with no elements $\emptyset = \{\}$ The cardinality of a set is the number of elements in the set. E.g: $X = \{1, 2, 3\}, |X| = \#X = 3$ The intersection of two sets is the set containing all common elements. If $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$ then the intersection $A \cap B = \{3\}$

$$\mathbb{Z} \cap \mathbb{R} = \mathbb{Z}.$$

The union of two sets is the set containing all elements that occur in either set. If $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$, then the union $A \cup B = \{1, 2, 3, 4, 5\}$

$$\mathbb{Z} \cup \mathbb{R} = \mathbb{R}.$$

## 1.3 Sets: subsets

A is a subset of B if all the elements of A are also contained in B. Written as $A \subset B$

$$\{1, 2\} \subset \{1, 2, 3\}.$$

$$\mathbb{Z} \subset \mathbb{R}.$$

$$\mathbb{Z}_{++} \subset \mathbb{Z}_{+} \subset \mathbb{Z} \subset \mathbb{R}.$$

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

## 1.4 Vectors

Vectors, in general, an abstract mathematical notation, but for the purpose of this module can be thought of as an ordered list of numbers E.g: Column vectors:

$$x = \begin{pmatrix} 1 \\ 3 \end{pmatrix} y = \begin{pmatrix} 3 \\ 1 \\ 8 \end{pmatrix}.$$

To say that a vector x is real values with D dimensions, we write $x \in \mathbb{R}^D$ E.g: $x \in \mathbb{R}^2, y \in \mathbb{R}^3$ Can write column vectors more compactly using parentheses $x = (13)$ The elements of a vector are usually denoted using subscripts E.g: if $x = (145)$ then $x_1 = 1, x_2 = 4, x + 3 = 5$ The transpose of a column vector is a row vector

$$x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} x^T = [123].$$

This gives us another way to write column vectors compactly: $x = [x_1 x_2 x_3]^T \in \mathbb{R}^3$

## 1.5 Adding and Scaling Vectors

To add two row or column vectors of the same dimension, just add their components

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{pmatrix} x + y = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_D + y_D \end{pmatrix}.$$

You cannot add a row vector to a column vector or vice versa (unless the dimension is 1) You cannot add vectors of different dimensions. To multiply a vector $x \in \mathbb{R}^D$ by a scalar $\alpha \in \mathbb{R}$, just multiply each component.

$$\alpha x = [\alpha x_1 \alpha x_2 \ldots \alpha x_D]^T.$$

## 1.6 Dot product

The dot product between two vectors $x, y \in \mathbb{R}^D$ is computed by multiplying the corresponding components of x and y and adding up the products.

$$x \cdot y = x_1 y_1 + x_2 y_2 + \ldots = \sum_{i=1}^{N} x_i y_i.$$

The dot product is also called the inner product or scalar product. Alternative notations:

- $x^T y$ (dot product as a a matrix multiplication)
- $(x, y)$ or $(x|y)$ (bracket notation, common in physics)
- $xy$ implicit notation

## 1.7 Norms

Most common norm is the Euclidean norm or $L_2$ norm, denoted $|| \, x \, ||_2$ or just $|| \, x \, ||$. Gives the length of a vector.

$$|| \, x \, || = \sqrt{x^T x} = \sqrt{\sum_i x_i^2}.$$

The squared Euclidean norm $\| x \|^2$ is often useful:

$$\| x \|^2 = x^T x = x \cdot x.$$

Another common norm is the $L_1$ norm, which is the sum of absolute values of x

$$\| x \|_1 = \sum_i | x_i | .$$

## 1.8 Properties of Norms

A norm is any function p from vectors to $\mathbb{R}$ that satisfies:

1. $p(\alpha x) = | \alpha | p(x)$ for $\alpha \in \mathbb{R}$
2. $p(x + y) \leq p(x) + p(y)$ (Triangle inequality)
3. $p(x) = 0 \iff x = 0$

We also have $p(-x) = p(x)$ E.g: $\| 5x \| = 5 \| x \|$ for any norm $\| \cdot \|$

## 1.9 Norms and Distance Metrics

Any vector space V and norm p can be used to induce a distance metric (define a metric space) by defining the distance metric to be $d(x, y) = p(x, y)$ E.g: The space of D dimensional real valued vectors $\mathbb{R}^D$ and the $L_2$ norm give the Euclidean space with metric $d(x, y) = \| x - y \|_2$ In 2D this gives the familiar Euclidean distance between two points x and y:

$$d(x, y) = \| x - y \|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

## 1.10 Matrices

A rectangular array of numbers. E.g:

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 1 & 6 \end{bmatrix} \in \mathbb{R}^{2x3}.$$

A can be thought of as a horizontal stack of M row vectors, or a vertical stack of N column vectors.

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_N \\ | & | & & | \end{bmatrix}.$$

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \dots \\ a_M^T \end{bmatrix}.$$

Beware of the overloaded notation! $a_i^T$ usually means the *ith* row rather than the *ith* column transposed.

### 1.11 Matrices: Transpose

The transpose of a matrix A, denoted $A^T$ is the same matrix with rows and columns interchanged. E.g:

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 1 & 6 \end{bmatrix} \in \mathbb{R}^{2x3}.$$

$$A^T = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{bmatrix} \in \mathbb{R}^{3x2}.$$

Note:

- $(A+B)^T = A^T + B^T$ (Transpose is linear)
- $(\alpha A)^T = \alpha(A^T) for \alpha \in \mathbb{R}$ (Transpose is linear
- $\alpha^T = \alpha for \alpha \in \mathbb{R}$ (Transpose a scalar does nothing)

### 1.12 Adding and Scaling Matrices

Matrices of the same dimenstions can be added together in the same way as with vectors: just add the corresponding components.

Matrices can be scaled by scalars by just multiplying each element by the scalar.

### 1.13 Matrix Multiplication

Two matrices can be multiplied if the number of columns in the first matrix equals the number of rows in the second matrix.

E.g it is possible to multiply A and B if $A \in \mathbb{R}^{MxN}$ and $B \in \mathbb{R}^{NxD}$. The resulting matrix AB will have dimension $MxD$. Rule: inner dimension must match, outer dimension gives dimension of result.

The $ij$ element of the product AB is equal to the inner product of row $a_i^T$ from A and column $b_j$ from B:

$$(AB)_{ij} = a_i^T b_j = \sum_{k=1}^{M} a_{ik} b_{kj}.$$

Multiplying two matrices requires performing MxD inner products

### 1.14 Matrix-Vector Multiplication

A matric-vector multiplication Ax can be performed if the number of columns in A equals the number of rows (entries) in x.

A matrix-vector product can be interpreted as a linear combination of the columns of A.

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots a_N & \\ | & | & & | \end{bmatrix}.$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}.$$

$$Ax = x_1 a_1 + x_2 a_2 + \ldots + x_N a_N.$$

I.e. each elements of x scales a colu,m of A and the result is the sum of these scaled columns.

It can also be interpreted as a set of M inner products between rows of A and x with each output being put in the resulting vector.

## 1.15 Matrix-Matrix Multiplication

Matrix-matrix multiplication AB can be interpreted as a series of matrix vector multiplications, one for each of the colu,ms of B, with the results being stacked side-by-side in a matrix.

$$AB = \begin{pmatrix} Ab_1 & Ab_2 & \ldots & Ab_D \end{pmatrix}.$$

Rules of matrix multiplication:

1. Not commutative: in general $AB \neq BA$
2. Distributive: $A(B + C) = AB + AC$
3. Associative: $A(BC) = (AB)C = ABC$
4. Transposes: $(AB)^T = B^T A^T$

## 1.16 Square Matrices, Identities

A matrix is square if it has the same number of rows and columns. E.g $A \in \mathbb{R}^{3x3}$.

The matrix I is called the identity matrix. The 3x3 identity matrix is:

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

Multiplication by the identity matrix gives back the original matrix

- $AI = A$
- $IA = A$

## 1.17 Inverses

The square matrix B is said to be an inverse of the square matrix A if $AB = I$.

The inverse of A (if it exists) is denoted $A^{-1}$

- $AA^{-1} = I$
- $A^{-1}A = I$
- $(AB)^{-1} = B^{-1}A^{-1}$

Terminology:

- A square matrix that has an inverse is called nonsingular or invertible (also nondegenerate)
- A square matrix that has no inverse is called singular or degenerate

## 1.18 Computing Inverses

Most numerical computation packages include functions for computing inverses. The algorithm used is usually LU decomposition or Gauss-Jordan elimination.

E.g Python numpy.linlang.inv(A). MATLAB inv(A).

You can invert 2x2 matrices by hand with:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

The matrix has no inverse if the determinant $det(A) = 0$. In the 2x2 case the determinant is given by $det(A) = ad - bc$

## 1.19 System of linear equations

A system of linear equations can be solved by using the inverse. E.g. consider the linear system:

$$3x + 2y - x = 12x - 2y + 4z = -2 - x + 0.5y - z = 0.$$

This can be written in matrix form as $Ax = b$ with:

$$A = \begin{bmatrix} 3 & 2 & -1 \\ 2 & -2 & 4 \\ -1 & 0.5 & -1 \end{bmatrix} x = \begin{bmatrix} x \\ y \\ z \end{bmatrix} b = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}.$$

And can be solved by finding the inverse of A and multiplying by b:

$$x = A^{-1}b.$$

## 1.20 Solving systems of linear equations numerically

You could do the calculations in Python as follows:

```
import numpy as np
A = np.array([[3,2,-1],[2,-2,4],[-1,0.5,-1]])
b = np.array([1,-2,0])
x = np.dot(np.linlang.inv(A), b)
print(x)
```

Which gives the answer (1,-2,-2).

Note that computing the inverses is usually not the most efficient way of doing this. Just call *numpy.linlang.solve* directly.

```
x = np.linlang.solve(A,b)
```

## 1.21 Quadratic Forms

Expressions like $x^TAx$ are called quadratic forms. They are scalar quantities.

$$x^TAx = x^T(Ax) = X^Tz, z = Ax.$$

They are called quadratic since they involve powers and cross terms of x

$$x^TAx = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}x_ix_j.$$

## 1.22 Positive Definiteness

If $x^TAx > 0 \forall x \neq 0$ then the (symmetric) matrix A is said to be symmetric positive definite (SPD). Sometimes written as $A \succ 0$.

A is positive definite matrices if and only if:

1. All eigenvalues of A are positive ( $\lambda_i(A) > 0 \forall i$ )
2. All pivots of A are positive
3. All principal minors of A are positive (determinants of upper kxk submatrices)
4. $x^TAx > 0 \forall x \neq 0$

Every pos. def. matrix can be written as a product $A - L^TL$ for some matrix L (A has a cholesky decomposition)

## 1.23 Positive Semi-Definite, Negative definite, etc.

Table 1:

| | | | |
|---|---|---|---|
| Positive Definite (SPD) | $A \succ$ | $x^TAx > 0$ | $\lambda_i(A) > 0$ |
| Positive Semidefinite (PSD) | $A \succeq 0$ | $x^TAx \geq 0$ | $\lambda_i(A) \geq 0$ |
| Negative Definite | $A \prec 0$ | $x^TAx < 0$ | $\lambda_i(A) < 0$ |
| Negative Semidefinite | $A \preceq 0$ | $x^TAx \leq 0$ | $\lambda_i(A) \leq 0$ |

## 1.24 More Notation

Set of symmetric matrices:

$$\S^n = \{X \in \mathbb{R}^n xn : X = X^T\}.$$

Set of symmetric positive definite matrices:

$$\S^n_{++} = \{X \in \S^n : X \succ_0\}.$$

Set of symmetric positive semidefinite matrices:

$$\S^n_{+} = \{X \in \S^n : X \succeq_0\}.$$

## 1.25 Eigenvalues and Eigenvectors

For an nxn matrix A, scalars $\lambda$ and vectors $x \neq 0$ satisfying

$$Ax = \lambda x.$$

are called eigenvalues and eigenvextors of A.

The set of distinct eigenvalues, denoted by $\lambda(A)$, is called the spectrum of A.

Interpretation: Eigenvectors are directions in which A behaves as if it were a scalar. The coresponding eigenvalue is the amount by which the eigenvector is scaled in that direction.

Eigenvectors and eigenvalues are tycalculated numerically (e.g. using numpy.linalg.eig)

Notation:

- $\lambda_i(A)$ is the i-th eigenvalue of A
- $\lambda_{max}(A)$ is the largest eiigenvalue of A
- $\lambda_{min}(A)$ is the smallest eigenvalue of A

Note:

- If $\lambda_i(A) = 0$ for any i then A is singular
- $tr(A) = \sum_{i=1}^{n} \lambda_i(A)$
- $det(A) = \prod_{i=1}^{n} \lambda_i(A)$

## 1.26 Eigenvalue Decomposition

If $A \in \mathbb{R}^{nxn}$ has n distinct eigenvalues then we can write down n eigenvalue equations:

$$\begin{aligned} Ax_1 &= \lambda_1 x_1. \\ Ax_2 &= \lambda_2 x_2. \\ &\vdots \\ Ax_n &= \lambda_n x_n. \end{aligned}$$

These can be grouped into a single matrix equation:

$$AV = V\Lambda.$$

with

$$V = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \Lambda = diag \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_n \end{pmatrix}.$$
$$AV = V\Lambda.$$

Left multiplication with $V^{-1}$ shows that V diagonalizes A:

$$V^{-1}AV = \Lambda.$$

Right multiplication by $V^{-1}$ shows that A can be decomposed into a product of the matrices of eiginvectors and eigenvalues:

$$A = V\Lambda V^{-1}.$$

This is called the eigenvalue decomposition of A (it is not always possible)

### 1.27 The spectral theorem for symmetric matrices

If $A \in \S^n$ then:

1. All eigenvalues are real:
$$\lambda_i(A) \in \mathbb{R}, i = 1, \ldots, n.$$
2. Eigenvectors corresponding to distinct eigenvalues are orthogonal
$$\lambda_i \neq \lambda_j \implies x_i^T x_j = 0.$$
3. A is orthogonally diagonalizable. There exists is a diagonal matrix D and a orthogonal matrix Q such that $A = QDQ^T$

### 1.28 The Singular Value Decomposition (SVD)

Every matrix $A \in R^{mxn}$ can be factored as follows:

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

Where U and V are orthogonal matrices and $\Sigma$ is a diagonal matrix with non-negative entries $\Sigma = diag\left(\sigma_1, \quad \ldots, \quad \sigma_n\right)$. This can also be written as:

$$A = \sum_{n=k} \sigma_k u_k v_k^T.$$

i.e. Every matrix can be written as a linear combination of rank one matrices. Compute with svd(A) (MATLAB) or numpy.linalg.svd(A) (Python)

### 1.29 SVD

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

- U contain the left singular vectors
- V contain the right singular vectors
- $\sigma_1, \ldots, \sigma_n$ are the singular values

By convention $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$

### 1.30 Eckart-Young Theorem