Fundamentals of Neural Networks

Pietro Marcatti

First Semester 2022/2023

1 Introduction

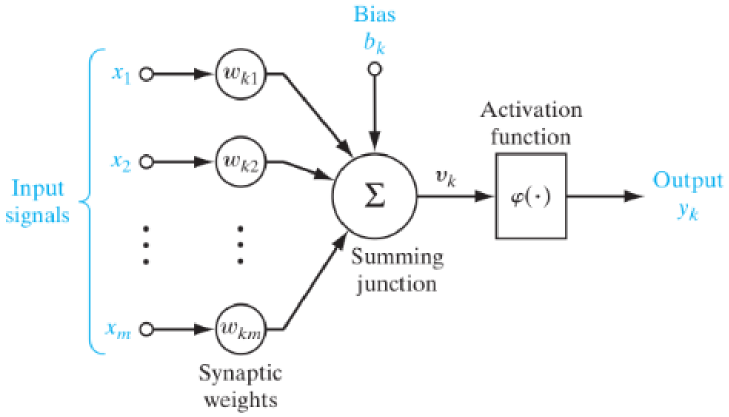
To summarize the main characteristics that artificial neural networks try to adapt from biology we can mention:

- Self-Organization and learning capability
- Generalization capability
- Fault tolerance

What, in fact, have we learned and how do we design artificial neural networks?

$$\begin{array}{rcl} \text{Many inputs} & \longrightarrow & \text{Vectors } \vec{x} \\ \text{One output} & \longrightarrow & \text{One operator } \ (\Sigma) \\ \text{Synapses} & \longrightarrow & \text{Weights } \left(\sum_{i=1} w_i x_i\right) \end{array}$$

In simple terms, the output of a single processing unit is obtained combining with some operator the weights and the inputs and putting them through a non-linear function.



Pattern Recognition

Neural networks are particularly strong at pattern which in the field is also synonimus of classification. One of the biggest application field for neural networks is computer vision were it is natural to manipulate real numbers which represent intensities in light.

The problem of recognizing the ten written digits is extremely complicated for a classic algorithm, where the only sensible approach would be to find some heuristics to classify the inputs. Immediately we would find obstacles to overcome like handling the countless variants in the input.

Ingredients and Terminology

When first approaching machine learning it can be overwhelming having to learn all the terminology involved. To start we can first list a couple characteristics that can be associated to machine learning models.

- Models can be of **supervised learning** when we are given both the training examples and the labels, or the real answers, the ground truth. They can also be **unsupervised** when we are not given the labels.
- Models can then be **regression models** when they try to learn a function so that they can predict the expected value in a real space. They can be **classifiers** and so they identify a set of input as belonging to one of a number of discrete classes.

2 Perceptron

In our digit classification problem a perceptron would constituite a single computing unit and would be capable of only recognizing one digit. The first elements of the perceptron that we encounter are

- a $28 \times 28 = 784$ real vector $\vec{x}$ describing the input or features
- an equally sized vector $\vec{w}$ describing the weight associated to every input
- a bias "weight"

The last two together are called the parameters of the model and are written as Φ. The perceptron model is governed by a parametric function of x

$$f_\Phi(x) = \begin{cases} 1 & if\ b + w \cdot x > 0 \\ 0 & otherwise \end{cases}$$

Perceptron Algorithm

If we want to introduce in some way the ground truth, the knowledge that can correct our parameters we can present the perceptron algorithm. It is important to say that, despite having many limitations, if a configuration of the parameters Φ that makes it so that all training examples are correctly labeled it will find such configuration.

- set b and all of the w to 0.
- for N iterations, or until the weights do not change
  - for each training example x^k with answer a^k
    - if $a^k - f(x^k) = 0$ continue
    - else for all weights $w_i, \Delta w_i = (a^k - f(x^k))x_i$

It is important to take a look at what the perceptron does, matemathically speaking. It determines the position of the features vector with respect to a plan. It then twiggles that hyperplane with the aim of correctly putting the features vectors that are labeled ”yes” above it and those labeled as ”no” beneath it. The learning of the model is then represented by the variation of the weights in an effort of correctly identifying all examples.

Terminology

We have now introduced more elements and we can then take a second to learn the terminology associated with them.

Hyperparameters : these are all ”second order” parameters, or more simply, all values that can be changed to impact the performance of the model but that are not the weights or the bias. An example is the number of iterations N.

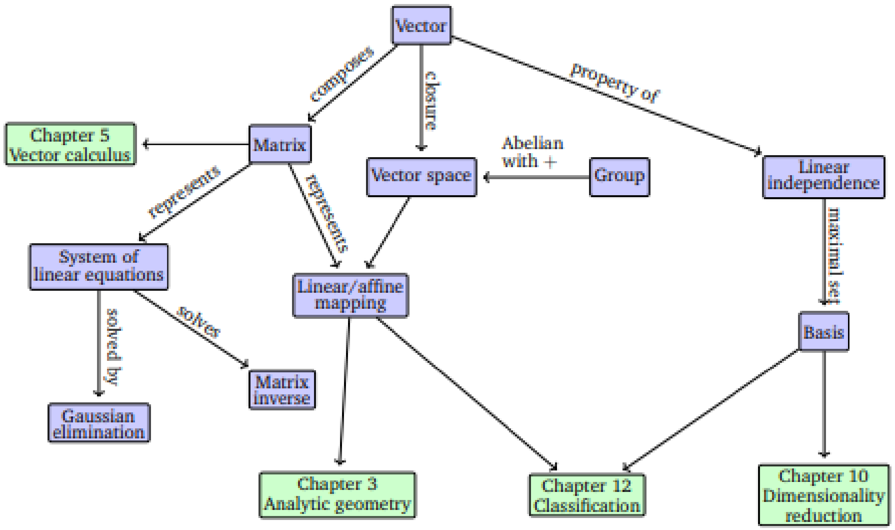
Bias : is a fixed/external knowledge, can be seen as the weight associated with a neuron (feature input) that i always firing, always set to 1.

Epoch : an entire run through the training data constituites an epoch.

When multiple computing units share the same inputs we are in multiple classification case and the neural network can be seen as a way to map, in our case, a 784-vector to a 10-vector (each unit deciding whether the input belongs to the i-th class).

3 Linear Algebra - A Mathematical Background

A mind-map summarising the key concept of this chapter and their relationship:



3.1 Systems of Linear Equations

Systems of linear equations play a central part of linear algebra. Many problems can be formulated as systems of linear equations, and linear algebra gives us the tools for solving them.
In general, for a real-valued system of linear equations we obtain either no, exactly one, or infinitely many solutions. We can introduce a useful compact notation for systems of linear equations (SLE) collecting the coefficients a_{ij} into vectors and collect the vectors into matrices.

$$\begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} x_2 + \cdots + \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} x_n = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{1}$$

$$\iff \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{2}$$

Definition 3.1 (Homogeneous SLE) *A system of linear equations is defined as homogeneous if* $\vec{b} = \vec{0}$

3.2 Matrices

Matrices play a central role in linear algebra and other than to compactly represent *SLE*s they can be used to represent linear functions (linear mappings).

Definition 3.2 (Matrix) *With $m, n \in \mathbb{N}$ a real-valued (m, n) matrix $\boldsymbol{A}$ is an $m \cdot n$-tuple of elements $a_{ij}, i = 1, \ldots, n, j = 1, \ldots, n$ which is ordere according to a rectangular scheme consisting of m rows and n columns:*

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R} \tag{3}$$

By convention $(1, n)$-matrices are called rows and $(m, 1)$-matrices are called columns. These special matrices are also called row/column vectors.
For matrices $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times k}$, the elements c_{ij} of the product $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times k}$ are computed as follows:

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}, \quad i = 1, \ldots, m \quad j = 1, \ldots, k$$

This means that the elements of the ith-row of $\mathbf{A}$ are multiplied with the elements of the jth-column of $\mathbf{B}$ and then summed together.

Definition 3.3 (Identity Matrix) *In $\mathbb{R}^{n \times n}$, we define the identity matrix as the $n \times n$ matrix containing 1 on the diagonal and 0 everywhere else.*

With the understanding of matrix multiplication, matrix addition and the identity matrix we can take a look at some properties of matrices:

Associativiy:

$$\forall \mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}, \mathbf{C} \in \mathbb{R}^{p \times q} : (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$

Distributivity

$$\begin{aligned} \forall \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}, \mathbf{C}, \mathbf{D} \in \mathbb{R}^{n \times p} : & (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} && (4) \\ & \mathbf{A}(\mathbf{C} + \mathbf{D}) = \mathbf{AC} + \mathbf{AD} && (5) \end{aligned}$$

Definition 3.4 (Inverse) *Consider a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. Let matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ have the property that $\mathbf{AB} = \mathbf{I}_n = \mathbf{BA}$. $\mathbf{B}$ is called the inverse of $\mathbf{A}$ and denoted by $\mathbf{A}^{-1}$*

Unfortunately not every matrix possesses and inverse. If this inverse does exist the matrix is called regular/invertible/nonsingular; otherwise it's called singular/noninvertible.

Definition 3.5 (Transpose) *For $\mathbf{A} \in \mathbb{R}^{m \times n}$ the matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ with $b_{ij} = a_{ji}$ is called the transpose of $\mathbf{A}$. We write it as $\mathbf{B} = \mathbf{A}^T$*

Definition 3.6 (Symmetric Matrix) *A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric if $\mathbf{A} = \mathbf{A}^T$*

3.2.1 Compact Representations of SLE

If we consider a system of linear equations and use the rules for matrix multiplication, we can write this equation system in a more compact form:

$$\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix}$$

Generally a system of linear equations can be compactly represented in their matrix form as $\mathbf{A}x = b$.

Definition 3.7 (Row-Echelon Form REF) *A matrix is in row-echelon form if:*

- *All rows that contain only zeros are at the bottom of the matrix; correspondigly, all rows that contain at least one nonzero element are on top of rows that contain only zeros.*
- *Looking at nonzero rows only, the first nonzero number from the left (also called the pivot or the leading coefficient) is always strictly to the right of the pivot of the row above it*

Remark 3.1 (Reduced Row-Echelon Form) *An equation system is in reduced row-echelon form (also row-reduces echelon form or row canonical form) if:*

- *It is in row-echelon form*
- *Every pivot is 1*
- *The pivot is the only nonzero entry in its column*

3.3 Vector Spaces

So far we have seen that systems of linear equations can be compactly represented using matrix-vector notation. In the following chapter we will have a closer look at vector spaces, i.e., a structured space in which vectors live.

3.3.1 Groups

We are ready to formalize the characteristics of vectors and scalar multiplication but we need to introduce the concept of a group. A group is a set of elements and an operation defined on these elements that keeps some structure of the set intact. Groups play an important role in computer science. Besides providing a fundamental framework for operations on sets, they are heavily used in cryptography, coding theory and graphics.

Definition 3.8 (Group) *Consider a set $\mathcal{G}$ and an operation $\otimes : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined on $\mathcal{G}$. Then $G := (\mathcal{G}, \otimes)$ is called a group if the following hold:*

1. *Closure of $\mathcal{G}$ under $\otimes : \forall x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$*

2. *Associativiy: $\forall x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$*

3. *Neutral element: $\exists e \in \mathcal{G} \forall x \in \mathcal{G} : x \otimes e = x$ and $e \otimes x = x$*

4. *Inverse element: $\forall x \in \mathcal{G} \exists y \in \mathcal{G} : x \otimes y = e$ and $y \otimes x = e$, where e is the neutral element. We oftern write x^{-1} to denote the inverse element of x.*

Remark 3.2 (Abelian Group) *If additionally $\forall x, y \in \mathcal{G} : x \otimes y = y \otimes x$, then $G = (\mathcal{G}, \otimes)$ is an Abelian group (commutative).*

3.3.2 Vector Spaces

We will now consider an extension of the definition of group that in addition to an inner operation $+$ also contain an outer operation $\cdot$, the multiplication of a vector $x \in \mathcal{G}$ by a scalar $\lambda \in \mathbb{R}$.

Definition 3.9 (Vector space) *A real-valued vector space $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations*

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$$
$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$$

where

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

2. *Distributivity*

3. *Associativiy (w.r.t. the outer operation)*

4. *Neutral element (w.r.t. the outer operation)*

The elements $x \in V$ are called vectors.

Remark 3.3 *A "vector multiplication" $\mathbf{ab}, a, b \in \mathbb{R}^n$ is not defined. We could use the matrix multiplication as previously defined however the dimensions of the vectors do not match. Only the following multiplication for vectors are defined: $\mathbf{ab}^T \in \mathbb{R}^{n \times n}$ (outer product), $\mathbf{a^T b} \in \mathbb{R}$ (inner/scalar/dot product)*

3.3.3 Vector Subspaces

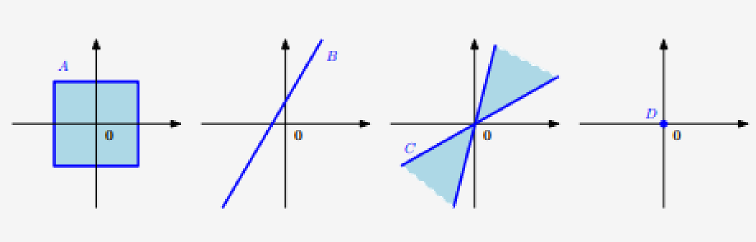
Intuitively, vector subspaces are sets contained in the original vector space with the property that when we perform space operations on elements within this subspace, we will never leave it. In this sense they are "closed". We will see how we can use vector subspaces to perform dimensionality reduction.

Definition 3.10 (Vector subspace) *Let $V = (\mathcal{V}, +, \cdot)$ be a vector space and $\mathcal{U} \subseteq \mathcal{V}, \mathcal{U} \neq \emptyset$. Then $U = (\mathcal{U}, +, \cdot)$ is called vector subspace of V (or linear subspace) if U is a vector space with the vector space operations $+$ and $\cdot$ restricted to $\mathcal{U} \times \mathcal{U}$ and $\mathbb{R} \times \mathcal{U}$. We write $U \subseteq V$ to denote a subspace U of V.*

If U is a vector subspace of V it naturally inherits many of its properties but to determine whether $(\mathcal{U},+,\cdot)$ is a subspace of V we still need to show

1. $\mathcal{U} \neq \emptyset$, in particular $\vec{0} \in \mathcal{U}$
2. Closure of U:
   (a) W.r.t. to the outer operation: $\forall \lambda \in \mathbb{R} \forall x \in \mathcal{U} : \lambda x \in \mathcal{U}$
   (b) W.r.t. to the outer operation: $\forall x, y \in \mathcal{U} : x + y \in \mathcal{U}$

Example: Only example D in the followin figureì is a subspace of R^2 (with the inner/outer operations). In A and C the closure property is violated, meanwhile B does not contain $\vec{0}$.



Remark 3.4 *Every subspace $U \subseteq (R^n,+,\cdot)$ is the solution space of a homogeneous SLE $\mathbf{A}\vec{x} = \vec{0}$ for $\vec{x} \in R^n$*

3.3.4 Linear Indipendence

In the following subsection we will take a closer look at what we can do with vectors. In particular, we can add vectors together and multiply them with scalars. The closure property guarantees that we end up with another vector in the same vector space. It is possible, we will see, to find a set of vectors with which we can represent every vector in the vector space by adding them together and scaling them. This set of vectors is a basis. Before we can explore further these concept we need to define linear combinations and linear indipendence.

Definition 3.11 (Linear combination) *Consider a vector space V and a finite number of vectors $x_1, \ldots, x_k \in V$. Then, every $v \in V$ of the form*

$$v = \lambda_1 x_1 + \cdots + \lambda_k x_k = \sum_{i=1}^{k} \lambda_i x_i \in V$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ is a linear combination of the vectors $x_1, \ldots, x_k$

The $\vec{0}$ can always be written as the linear combination of k vectors, only some of them are non-trivial. In the following we are interested in non-trivial linear combinations that represent $\vec{0}$, that is, where not all λ_i are 0.

Definition 3.12 (Linear (In)dependence) *Let us consider a vector space V with $k \in \mathbb{N}$ and $x_i, \ldots, x_k \in V$. If there is a non-trivial linear combination, such that $\vec{0} = \sum_{i=1}^{k} \lambda_i x_i$ with at least one $\lambda_i \neq 0$, the vectors $x_1, \ldots, x_k$ are linearly dependent. If only the trivial solution exists the vectors $x_1, \ldots, x_k$ are linearly independent.*

Intuitively a set of linearly independent vectors consists of vectors that have no redundancy. If we remove any of those vectors from the set, we will lose something.

Remark 3.5 *Consider a vector space V with k linearly independent vectors $b_1, \ldots, b_k$ and m linear combinations*

$$x_j = \sum_{i=1}^{k} \lambda_i 1 b_i, \quad j = 1, \ldots, m$$

Defining $\mathbf{B} = [b_1, \ldots, b_k]$ as the matrix whose columns are the linearly independent vectors $b_1, \ldots, b_k$, we can write in a more compact form

$$x_j = \mathbf{B}\lambda_j, \quad \lambda_j = \begin{bmatrix} \lambda_{1j} \\ \vdots \\ \lambda_{kj} \end{bmatrix}, \quad j = 1, \ldots, m$$

A set of vectors are linearly independent if and only if no-one of the vectors can be obtained as a linear combination of the others.

3.4 Basis and Rank

In a vector space V, we are particularly interested in sets of vectors $\mathcal{A}$ that possess the property that any vector $v \in V$ can be obtained by a linear combination of the vectors in $\mathcal{A}$.

3.4.1 Generating Set and Basis

Definition 3.13 (Generating set and Span) *Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and set of vectors $\mathcal{A} = x_1, \ldots, x_k \subseteq \mathcal{V}$. If every vector $v \in \mathcal{V}$ can be expressed as a linear combination of $x_1, \ldots, x_k$, $\mathcal{A}$ is called a generating set of V. The set of all linear combinations of vectors in $\mathcal{A}$ is called the span of $\mathcal{A}$. If $\mathcal{A}$ spans the vector space V, we write $V = span[\mathcal{A}]$ or $V = span[x_1, \ldots, x_k]$*

Definition 3.14 (Basis) *Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} \subseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called minimal if there exists no smaller set $\tilde{\mathcal{A}} \subsetneq \mathcal{A} \subseteq \mathcal{V}$ that spans V. Every linearly independent generating set of V is minimal and is called a basis of V*

In our study we will only consider finite-dimensional vector spaces V. In this case, the dimension of V is the number of basis vectors of V, and we write $dim(V)$. If $U \subseteq V$ is a subspace of V, then $dim(U) \leq dim(V)$ and $dim(U) = dim(V)$ if and only if $U = V$. Intuitively, the dimension of a vector space can be thought of as the number of independent directions in this vector space. However, it is important to notice that this is not necessarily the number of elements in a basis vector but it is the number of basis vectors.

Remark 3.6 *A basis of a subspace $U = span[x_1, \dots, x_m] \subseteq R^n$ can be found by executing the followin steps:*

1. *Write the spanning vectors as columns of a matrix* $\mathbf{A}$
2. *Determine the row-echelon form of* $\mathbf{A}$
3. *The spanning vectors associtated with the pivot columns are a basis of U*

3.4.2 Rank

The number of linearly independent columns of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the rank of $\mathbf{A}$ and is denoted by $rk(\mathbf{A})$.
The columns of $\mathbf{A} \in \mathbb{R}^{m \times n}$ span a subspace $U \subseteq \mathbb{R}^m$ with $dim(U) = rk(\mathbf{A})$. Later we will call this subspace the image or range.
For all $\mathbf{A} \in \mathbb{R}^{n \times n}$ it holds that $\mathbf{A}$ is regular (invertible) if and only if $rk(\mathbf{A}) = n$
A matrix $\mathbf{A} \in R^{m \times n}$ has full rank if its rank equals the largest possible rank for a matrix of the same dimensions. This means that the rank of a full-rank matrix is the lesser of the number of rows and columns, i.e., $rk(A) = min(m, n)$. A matrix is said to be rank deficient if it does not have full rank.

3.5 Linear Mappings

In this section we will study mappings on vector spaces that preserve their structure, which will allow us to define the concept of a coordinate. In the beginning of the chapter we said that vectors are objects that can be added together and multiplied by a scalar, and the resulting object is still a vector. We wish to preserve this property when applying the mapping. Consider two real vector spaces V, W, a mapping $\Phi : V \longrightarrow W$ preserves the structure of the vector space if

$$\Phi(\vec{x} + \vec{y}) = \Phi(\vec{x}) + \Phi(\vec{y}) \tag{6}$$

$$\Phi(\lambda \vec{x}) = \lambda \Phi(\vec{x}) \tag{7}$$

for all $x, y \in V$ and $\lambda \in \mathbb{R}$. We can summarize this in the following definition.

Definition 3.15 (Linear Mapping) *For vector spaces V, W, a mapping $\Phi : V \longrightarrow W$ is called a linear mapping (or vector space homomorphism/linear transformation) if*

$$\forall x, y \in V \; \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda x + \psi y) = \lambda \Phi(x) + \psi \Phi(y)$$

It turns out that we can represent linear mappings as matrices. Recall that we can also collect a set of vectors as columns of a matrix. When working with matrices, we have to keep in mind what the matrix represents: a linear mapping or a collection of vectors.

Definition 3.16 (Injective, Surjective and Bijective mappings) *Consider a mapping $\Phi : \mathcal{V} \longrightarrow \mathcal{W}$ where $\mathcal{V}, \mathcal{W}$ can be arbitrary sets. Then Φ is called*

Injective *: if $\forall x, y \in \mathcal{V} : \Phi(x) = \Phi(y) \Longrightarrow x = y$*

Surjective *:if $\Phi(\mathcal{V}) = \mathcal{W}$*

Bijective *: if Φ is both injective and surjective.*

Theorem 3.1 *Two finite-dimensional vector spaces V and W are isomorphic if and only if $dim(V) = dim(W)$*

3.5.1 Matrix Representation of Linear Mappings

From the theorem just presented we can derive that any n-dimensional vector space is isomorphic to R^n. We can consider a basis $b_1, \ldots, b_n$ of an n-dimensional vector space V. In the following subsection the order of the basis vectors will be important, therefore, we write

$$B = (b_1, \ldots, b_n)$$

and we call this n-tuple an ordered basis of V.

Remark 3.7 (Notation) *In order to keep things straight we summarise some parts of the notation here. $B = (b_1, \ldots, b_n)$ is an ordered basis, $\mathcal{B} = \{b_1, \ldots, b_n\}$ is an (unordered) basis, and $\mathbf{B} = [b_1, \ldots, b_n]$ is a matrix whose columns are the vectors $b_1, \ldots, b_n$*

Definition 3.17 (Coordinates) *Consider a vector space V and an ordere basis $B = (b_1, \ldots, b_n)$ of V. For any $x \in V$ we obtain a unique representation (linear combination) of x with respect to B*

$$x = \alpha_1 b_1 + \ldots + \alpha_n b_n$$

Then $\alpha_1, \ldots, \alpha_n$ are the coordinates of x with respect to B, and the vector

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n$$

is the coordinate vector or the coordinate representation of x with respect to the ordered basis B.

Definition 3.18 (Transformation Matrix) *Consider vector spaces V, W with corrsponding (ordered) basis $B = (b_1, \ldots, b_n)$ and $C = (c_1, \ldots, c_m)$. Moreover, we consider a linear mapping $\Phi : V \longrightarrow W$. For $j \in 1, \ldots, n$,*

$$\Phi(b_j) = \alpha_{1j}c_1 + \cdots + \alpha_{mj}c_m = \sum_{i=1}^{m} a_{ij}c_i$$

is the unique representation of $\Phi(b_j)$ with respect to C. Then, we call the $m \times n$ matrix $\mathbf{A}_\Phi$, whose elements are given by

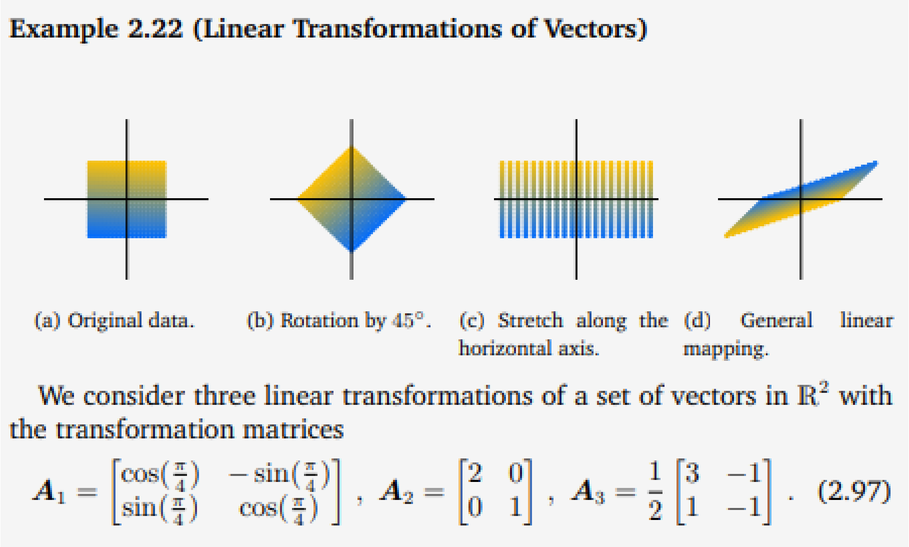
$$\mathbf{A}_\Phi(i,j) = \alpha_{i,j}$$

the transformation matrix of Φ (with respect to the ordered bases B of V and C of W)

Consider (finite-dimensional) vector spaces V, W with ordered bases B, C and a linear mapping $\Phi : V \longrightarrow W$with transformation matrix $\mathbf{A}_\Phi$. If $\hat{x}$ is the coordinate vector of $x \in V$ with respect to B and $\hat{y}$ the coordinate vector of $y = \Phi(x) \in W$ with respect to C, then

$$\hat{y} = \mathbf{A}_\Phi \hat{x}$$

This means that the transformation matrix can be used to map coordinates with respect to an ordered basis in V to coordinates with respect to an ordered basis in W



Example 2.22 (Linear Transformations of Vectors)

(a) Original data. (b) Rotation by 45°. (c) Stretch along the horizontal axis. (d) General linear mapping.

We consider three linear transformations of a set of vectors in $\mathbb{R}^2$ with the transformation matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{A}_3 = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}. \quad (2.97)$$

Questions:

3.5.2 Basis Change

In the following, we will have a closer look at ho w transformation matrices of linear mapping $\Phi : V \longrightarrow W$ change if we change the bases in V and W. Consider two ordered bases of V

$$B = (b_1, \dots, b_n) \quad \tilde{B} = (\tilde{b}_1, \dots, \tilde{b}_n)$$

and two ordered bases of W

$$C = (c_1, \dots, c_n) \quad \tilde{C} = (\tilde{c}_1, \dots, \tilde{c}_n)$$

. Moreover, $A_\Phi \in \mathbb{R}^{m \times n}$ is the transformation matrix of the linear mapping $\Phi : V \longrightarrow W$ with respect to the bases B and C, and $\tilde{A}_\Phi \in \mathbb{R}^{m \times n}$ is the corresponding transformation mapping with respect to $\tilde{B} and \tilde{C}$. In the following we will investigate how A and $\tilde{A}$ are related, for example how/whether we can transform A_Φ into $\tilde{A}_\Phi$ if we choose to perform a basis change from B, C to $\tilde{B}, \tilde{C}$.

Theorem 3.2 (Basis Change) *For a linear mapping $\Phi : V \longrightarrow W$, ordered bases of V*

$$B = (b_1, \dots, b_n) \quad \tilde{B} = (\tilde{b}_1, \dots, \tilde{b}_n)$$

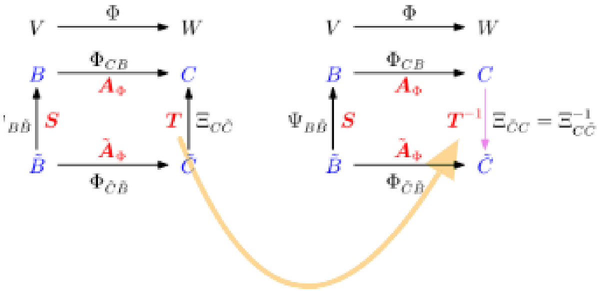
and of W

$$C = (c_1, \dots, c_n) \quad \tilde{C} = (\tilde{c}_1, \dots, \tilde{c}_n)$$

and a transformation matrix A_Φ of Φ with respect to B and C, the corresponding transformation matrix $\tilde{A}\Phi$ with respect to the bases $\tilde{B} and \tilde{C}$ is given as

$$\tilde{A}_\Phi = T^{-1} A_\Phi S$$

Here, $S \in \mathbb{R}^{n \times n}$ is the transformation matrix of id_V that maps coordinates with respect to $\tilde{B}$ onto coordinates with respect to B, and $T \in \mathbb{R}^{m \times m}$ is the transformation matrix of id_W that maps coordinates with respect to $\tilde{C}$ onto coordinates with respect to C.



3.5.3 Image and Kernel

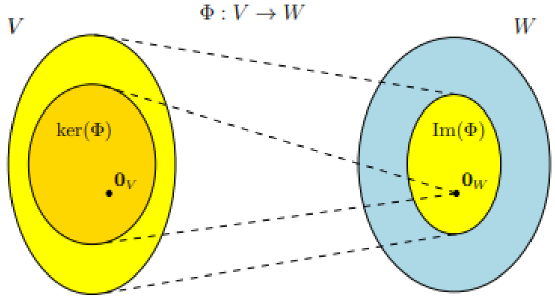
Definition 3.19 (Image and Kernel) *For $\Phi : V \to W$ we define the kernel/null space:*

$$ker(\Phi) := \Phi^{-1}(0_w) = v \in V, \Phi(v) = 0_w$$

and the image/range

$$Im(\Phi) := \Phi(V) = \{w \in W \mid \exists v \in V : \Phi(v) = w\}$$

We also call V and W the domanin and codomain of Φ respectively. Intuitively, the kernel is the set of vectors $v \in V$ that Φ maps onto the neutral element $0_W \in W$. The image is the set of vectors $w \in W$ that can be "reached" by Φ from any vector in V.



Remark 3.8 (Null Space and Column Space) *Let us consider $A \in \mathbb{R}^{m\times n}$ and a linear mapping $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \quad x \mapsto Ax$*

- *For $A = [a_1, \ldots, a_n]$, where a_i are the columns of A, we obtain*

  $$Im(\Phi) = Ax : x \in \mathbb{R}^n = \left\{ \sum_{i=1}^{n} x_i a_i : x_1, \ldots, x_n \in R \right\} = span[a_1, \ldots, a_n] \subseteq \mathbb{R}^m$$

  i.e., the image is the span of the columns of A, also called the column space. Therefore, the column space (image) is a subspace of $\mathbb{R}^m$, where m is the "height" of the matrix.

- $rk(A) = dim(Im(\Phi))$

- *The kernel/null space $ker(\Phi)$ is the general solution to the homogeneous system of linear equations $Ax = 0$ and captures all possible linear combinations of the elements in $\mathbb{R}^n$ that produce $0 \in \mathbb{R}^m$.*

- *The kernel is a subspaces of $\mathbb{R}^n$, where n is the "width" of the matrix.*

Theorem 3.3 (Rank Nullity) *For vector space V, W and a linear mapping $\Phi : V \to W$ it holds that*

$$dim(ker(\Phi)) + dim(Im(\Phi)) = dim(V)$$

The rank-nullity theorem is also referred to as the fundamental theorem of linear mappings.

The mappings that we obtain by building NN are special cases of Φ for example $R^{10\times 784}$: the strenght of this approach is that it's very simple but the cons is that it's all linear.

3.5.4 Affine Subspaces

Definition 3.20 (Affine Subspace) *Let V be a vector space, $x_0 \in V$ and $U \subseteq V$ a subspace. Then the subset:*

$$L = x_0 + U := x_0 + u : u \in U = v \in V \mid \exists u : v = x_0 + u \subseteq V$$

is called affine subspace or linear manifold of V. U is called direction or direction space, and x_0 is called support point. Note that the definition of an affine subspace excludes $\vec{0}$ if $x_0 \notin U$. Therefore, an affine subspace is not a (linear) subspace (vector subspace) of V for $x_0 \notin U$.

Remark 3.9 *For inhomogeneous systems of linear equations and affine subspaces: the solution of the system is either the empty set or an affine subspace of $\mathbb{R}^n$ of dimension $n - rk(A)$. Every k-dimensional affine subspace is the solution of an inhomogeneous system of linear equations.*

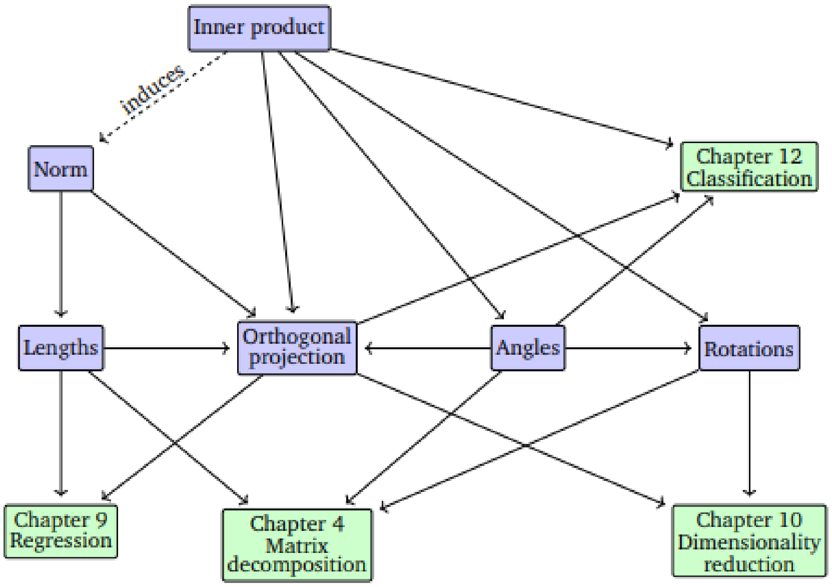
Definition 3.21 (Affine Mappings) *For two vector spaces V, W a linear mapping $\Phi : V \longrightarrow W$ and $a \in W$ the mapping*

$$\begin{aligned} \phi :& V \longrightarrow W \\ & x \mapsto a + \Phi(x) \end{aligned}$$

is an affine mapping from V to W. The vector a is called the translation vector of Φ. Every affine mapping $\phi : V \longrightarrow W$ is also the composition of a linear mapping $\Phi : V \longrightarrow W$ and a translation $\tau : W \longrightarrow W$ such that $\phi = \tau \circ \Phi$. The mappings Φ and τ are uniquely determined.

4 Analytic Geometry

In this chapter we will add some geometric interpretation and intuition to all of these concepts. In particular, we will look at geometric vectors and comput their lengths and distances or angle between two vectors. To be able to do this, we equip the vector space with an inner product that induces the geometry of the vector space.



4.1 Norms

Definition 4.1 (Norm) *A norm on a vector space V is a function*

$$\|\cdot\| : V \longrightarrow \mathbb{R} \qquad\qquad x \mapsto \|x\|$$

which assigns each vector x its length $\|x\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $x, y \in V$ the following hold:

- *Absolutely homogeneous:* $\|\lambda x\| = |\lambda| \, \|x\|$
- *Triangle inequality:* $\|x + y\| \leq \|x\| + \|y\|$
- *Positive definite:* $\|x\| \geq 0$ *an* $\|x\| = 0 \iff x = 0$

The norm we will be using throughout these notes is going to be the Euclidean norm by default if not stated otherwise.

Definition 4.2 (Euclidean Norm) *The Euclidean norm of $x \in \mathbb{R}^n$ is defined as:*

$$\|x\|_2 = \sqrt[2]{\sum_{i=1}^{n} x^2} = \sqrt{x^T x}$$

and it computes the Euclidean distance from x to the origin. The Euclidean norm is also called the ℓ_2 norm.

4.2 Inner Products

Recall the linear mappings where we can rearrange the mapping with respect to addition and multiplication with a scalar. A bilinear mapping Ω is a mapping with two arguments, and it is linear in each argument, i.e., when we look at a vector space V then it holds that for all $x, y, z \in V, \lambda, \psi \in \mathbb{R}$ that

$$\begin{aligned}\Omega(\lambda x + \psi y, z) &= \lambda\Omega(x, y) + \psi\Omega(y, z)\\ \Omega(x, \lambda y + \psi z) &= \lambda\Omega(x, y) + \psi\Omega(x, z)\end{aligned}$$

A bilinear mappings doesnt work on a single vector space but on the cartesian product of two vector spaces. We want to focus on biliear mappings that are:

- Symmetric: $\Omega(x, y) = \omega(y, x), \quad \forall x, y \in V$
- Positive definite:

$$\forall x \in V \setminus \{0\} : \Omega(x, x) > 0, \quad \Omega(\vec{0}, \vec{0}) = 0$$

Definition 4.3 (Inner product) *Let V be a vector space and $\Omega : V \times V \longrightarrow \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then*

- *A positive definite, symmetric bilinear mapping $\Omega : V \times V \longrightarrow \mathbb{R}$ is called an inner product on V. We typically write $\langle x, y \rangle$ instead of $\Omega(x, y)$.*
- *The pair $(V, \langle\cdot,\cdot\rangle)$ is called an inner product space or (real) vector space with inner product. If we use the dot product we call $(V, \langle\cdot,\cdot\rangle)$ a Euclidean vector space. We will refer to these spaces as inner product spaces.*

Recall that any vectors $x, y \in V$ can be written as linear combinations of the basis vectors so that $x = \sum_{i=1}^{n} \psi_i b_i \in V$ and $y = \sum_{j=1}^{n} \lambda_j b_j \in V$ for suitable $\psi_i, \lambda_j \in \mathbb{R}$. Due to the bilinearity of the inner product, it holds for all $x, y \in V$ that

$$\langle x, y \rangle = \left\langle \sum_{i=1}^{n} \psi_i b_i, \sum_{j=1}^{n} \lambda_j b_j \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_i \langle b_i, b_j \rangle \lambda_j = \hat{x} A \hat{y}$$

where $A_{i,j} := \langle b_i, b_j \rangle$ and $\hat{x}, \hat{y}$ are the coordinates of x and y with respect to the basis B. This implies that the inner product $\langle\cdot,\cdot\rangle$ is uniquely determined through A. The symmetry of the inner product also means that A is symmetric. Furthermore, the positive definiteness of the inner product implies that

$$\forall x \in V \setminus \{0\} : x^T A x > 0$$

Definition 4.4 (Symmetric, Positive Definite Matrix) *A symmetric matrix $A \in \mathbb{R}^{n \times n}$ that satisfies*

$$\forall x \in V \setminus \{0\} : x^T A x > 0$$

is called symmetric, positive definite or just positive definite.

4.3 Lengths and Distances

Remark 4.1 (Cauchy-Schwarz inequality) *For an inner product vector space $(V, \langle \cdot, \cdot \rangle)$ the induced norm $\|\cdot\|$ satisfies the Cauchy-Schwarz inequality*

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|$$

Definition 4.5 (Distance and Metric) *Consider an inner product space $(V, \langle \cdot, \cdot \rangle)$. Then*

$$d(x, y) := \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

is called the distance between x and y for $x, y \in V$. If we use the dot product as the inner product, then the distance is called Euclidean distance.
The mapping

$$\begin{aligned} d : & V \times V \longrightarrow \mathbb{R} \\ & (x, y) \mapsto d(x, y) \end{aligned}$$

A metric d satisfies the following:

- d is positive definite, i.e., $d(x, y) \geq 0 \; \forall x, y \in V$ and $d(x, y) = 0 \iff x = y$
- d is symmetric, i.e., $d(x, y) = d(x, y) \; \forall x, y \in V$
- Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z) \; \forall x, y, z \in V$

Remark 4.2 *At first glance, the lists of properties of inner products and metrics look very similar. However, we observe that $\langle x, y \rangle$ and $d(x, y)$ behave in opposite directions. Very similar x and y will result in a large value for the inner product and a small value for the metric.*

4.4 Angles and Orthogonality

In addition to enabling the definition of lenghts of vectors, as well as the distance between two vectors, inner products also capture the geometry of a vector space by defining the angle ω between two vectors. We use the Cauchy-Schwarz inequality to define angles ω in inner product spaces between two vectors x,y, and this notation coincides with our intuition in $\mathbb{R}^2$ and $\mathbb{R}^3$. Assume that $x \neq 0, y \neq 0$. Then

$$-1 \leq \frac{\langle x, y \rangle}{\|x\| \, \|y\|} \leq 1$$

Therefore, there exists a unique $\omega \in [0, \pi]$ with

$$\cos \omega = \frac{\langle x, y \rangle}{\|x\| \, \|y\|}$$

Intuitively, the angle between two vectors tells us how similar their orientations are.

Definition 4.6 (Orthogonality) *Two vectors x and y are orthogonal if and only if $\langle x, y \rangle = 0$ and we write $x \perp y$. If additionally $\|x\| = 1 = \|y\|$, the vectors x and y are orthonormal.*

Definition 4.7 (Orthogonal Matrix) *A square matrix $A \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if and only if its columns are orthonormal so that*

$$AA^T = I = A^T A$$

which implies that

$$A^{-1} = A^T$$

4.5 Orthonormal Basis

Definition 4.8 (Orthonormal Basis) *Consider an n-dimensional vector space V and a basis $b_1, \ldots, b_n$ of V. If*

$$\begin{aligned} \langle b_i, b_j \rangle = 0 \; for \; i \neq j \\ \langle b_i, b_j \rangle = 1 \end{aligned}$$

for all $i, j = 1, \ldots, n$ then the basis is called an orthonormal basis (ONB).

5 Tensor Flow

Let's follow the computation of our program to recognise the 10 digits, our input matrix is (1 x 784), our weights matrix W (784 x 10) our bias matrix (10 x 1). Matrix representation of neural networks:

$$L = XW + B$$

L = logits. These are the steps:

$$Pr(A(x)) = \sigma(xW + b)$$

$$L(x) = -log(Pr(A(x) = a)) \quad crossentrpylossfunction$$

$$\nabla_l X(\Phi) = \left(\frac{\partial X(\Phi)}{\partial l_1}, \ldots, \frac{\partial X(\Phi)}{\partial l_m} \right)$$

$$\Delta W = -\mathcal{L} X^T \nabla_l X(\Phi)$$

How to create the nabla in matrix form??
Important we want to manipulate input in batches. Now the input matrix is now (m x 784), m batch size. When adding bias you might have to adjust the size of the matrix to work with the m rows of the batch input matrix. Tensor flow is the programming language and Python is the environment. Tensorflow plays with tensors (often typed). Vectors in Tensorflow only have one dimension! In the environment we set-up our nn-model by defining parameters and an architecture. We go trhough 3 stages:

- create the tensors
- Turn it into a variable
- Initialize it

A tensorflow program:

- Load data
- set-up the model (batch-size)
- learn the variables (parameters)

Along the way of training the model we must keep in mind to compute the accuracy.
Rule of thumb: the smaller the batch size the smaller the learning rate.
We can see with simple linear algebra matrix multiplication properties that adding a new layer doesn't add anything so we must add some non linearity. This non linearity is expressed in the form of activation function.