Part I Mathematical Foundations

Machine learning's three core concepts: data, model, learning

1. Data
    - Goal of machine learning is to design general-purpose methodologies to extract valuable patterns from data
    - Ideally without much domain-specific expertise
2. Models: Related to the process that generates data
3. Learning: Automatically find patterns and structure in data by optimizing the parameters of the model

1.1 Finding Words for Intuition

- Algorithm

1. Predictors: "Machine learning algorithm" is a system that makes predictions based on input data
2. Training: "Machine learning algorithm" is a system taht adapts some internal parameters of the predictor so taht it performs well on future unseen input data

- In the book, consider data as vectors
- Vectors

1. Array of numbers(CS)
2. Arrow with a direction and magnitude(Physics)
3. Object that obeys addition and scaling(Math)

- Good model can be used to predict without performing real-world experiments
- Learning compoennt
    - Training a model: Use the data to optimize parameters of the model to evaluae how well the model predicts the training data
    - Interested in the model to perform well on unseen data

1.2 Two ways to read this book: Bottom-up or top-down

2 Linear Algebra

- Algebra: Construct a set of objects and a set of rules to manipulate these objects
- Linear algebra: Study of vectors and certain rules to manipulate vectors
- Vectors: speical objects that can be added together and multiplied by scalars

    1. Geometric vectors
    2. Polynomials
    3. Audio signals
    4. Elements of R^n

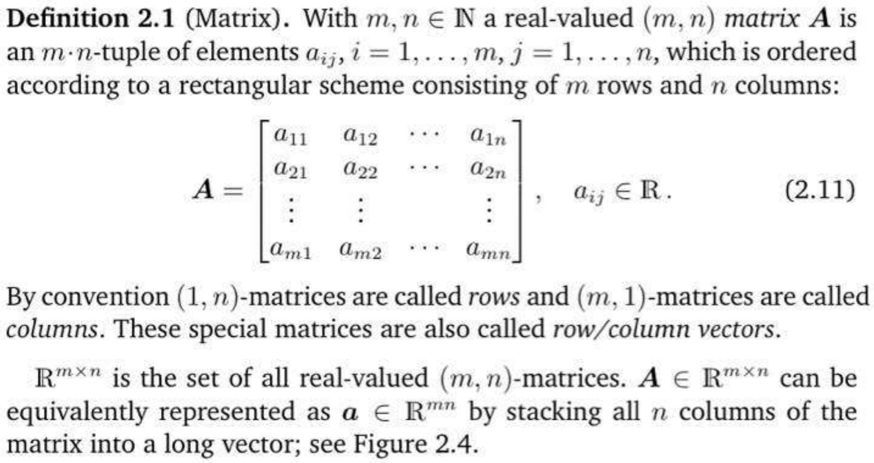
- Linear algebra focuses on the similarities between these vector concepts
- Largely focus on vectors in R^n
- Finite dimensional vector spaces
- Idea of "Closure" resulting in vector space which underlies much of machine learning

2.1 Systems of Linear Equations

- A real valued system of linear equations: No, exactly one, or infinitely many solutions
- Linear equations with two variables) each linear equation defiens a line on the x1x2-plane
    - Intersection can be a line, a, point or empty
- Three varaibles) each linear equation determines a plane in three-dimensional space
    - Solution set can be a plane, a line, a point, or empty
- Use compact notations a.k.a. matrices

2.2 Matrices

- Central role in linear algebra
- Used to compactly represent systems of linear equations, but also represent linear functions(linear mappings)
- Definition:



Definition 2.1 (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}, $i = 1, \dots, m$, $j = 1, \dots, n$, which is ordered according to a rectangular scheme consisting of m rows and n columns:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R}. \qquad (2.11)$$

By convention $(1, n)$-matrices are called *rows* and $(m, 1)$-matrices are called *columns*. These special matrices are also called *row/column vectors*.

$\mathbb{R}^{m \times n}$ is the set of all real-valued (m, n)-matrices. $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ can be equivalently represented as $\boldsymbol{a} \in \mathbb{R}^{mn}$ by stacking all n columns of the matrix into a long vector; see Figure 2.4.

2.2.1 Matrix Addition and Multiplication

- To compute element cij, we multiply the elements of the ith row of A with the jth column of B and sum them up
- Call this dot product
- Remark. Matrices can only be multiplied if their neighboring dimensions match
- Remark. Matrix multiplication is not defined as an element-wise operation on matrix elements
- Definition 2.2. Identity matrix
- Properties of matrices

2.2.2. Inverse and Transpose

- Definition 2.3 (Inverse). Consider a square matrix A ∈ R nxn. Let matrix B ∈ R nxn have the property that AB = I = BA. B is the inverse of A and denoted by A^-1
- Not every matrix A possesses an inverse A^-1
- A: regular, invertible, nonsingular
- When inverse exists, it is unique
- Definition 2.4 (Transpose). For A ∈ R mxn the matrix B ∈ R nxm with bij = aji is called the transpose of A. We write B = A^T
- Definition 2.5 (Symmetric Matrix). A matrix A ∈ R nxn is symmetric if A = A^T
- Only nxn can be symmetric, or square matrices
- If A is invertible, then so is A^T, and (A^-1)T = (A^T)-1 = A^-T

2.2.3 Multiplication by a Scalar

λ ∈ R mxn and λ ∈ R. Then λA = K, Kij = λaij

2.2.4 Compact Representation of Systems of Linear Equations

- Compactly represented as **Ax = b**
- The product Ax is a lienar combination of the columns of A

2.3 Solving Systems of Linear Equations

aij ∈ R and bi ∈ R are known and xj are unknowns

- System has 2 equations and 4 unknowns
- Solution is [42, 8, 0, 0]^T called particular solution or linear equations
- Not the only solution. To capture other solutions, need to be creative in generating 0 in a non-trivial way using the columns of th ematrix: Adding 0 to our special solution does not change the special solution
- Express the third column using the first two columns

*Remark: General approach

1. Find a particular solution to Ax = b
2. Find all solutions to Ax = 0
3. Combine the solutions from steps 1. and 2. to the general solution

- General equation systems are not of this simple form
- Fortunately, constructive algorithmic way of transforming any system of linear equations into this particularly simple form
- Gaussian elimination
- Key: Elementary transformations of systems of lienar equations

2.3.2 Elementary Transformations

- Key of solving a system: elementary transformations that keep the solution set the same, but that transform the equation system into a simpler form

- Exchange of two equations (rows in the matrix representing the system of equations)
- Multiplication of an equation (row) with a constant $\lambda \in R \{0\}$
- Addition of two equations(rows)

- Remark: Row-Echelon Form
    - A matrix is in row-echelon form if
        - All rows that contain only zeros are at the bottom of the matrix; correspondingly, 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: Basic and Free Varaibles
    - The variables corresponding to the pivots in the row-echelon form are called basic variables and others are free variables

- Remark: Obtaining a particular solution
    - Row-echelon helpful when need to determine a particular solution
    - express the right hand side of the equation system using the pivot columns

- Remark: Reduced Row Echelon Form
    - It is in row-echelon form
    - Every pivot is 1
    - The pivot is the only nonzero entry in its column
    - Allows us to determine the general solution of a system of linear equations in a straightforward way

- Remark : Gaussian Elimination(Algorithm that performs elementary transformations to bring a system of linear equations into reduced row-echelon form)

2.3.3 The Minus-1 Trick

- Practical trick for reading out the solutions x of a homogeneous system of linear equations Ax = 0, where $A \in R$ kxn, $x \in$ Rn
- To start, assume that A is in reduced row-echelon form without any rows that just contain zeros
- Extend this matrix to an nxn matrix A by adding n-k rows of the form [0 ... 0 -1 0 ... 0] so that the diagonal of tha augmented matrix A contains either 1 0r -1
- Columns of A that contain -1 as pivots are solutions of the homogeneous equation system Ax = 0
- Columns form a basis of the solution space of Ax = 0 a.k.a. kernel or null space

Calculating the Inverse

- Have to find a matrix X that satisfies AX = I, then X = A^-1
- If we bring the augmented equation system into reduced row-echelon form, we can reach out the inverse on the righ thand side of the equation system
- Determining the inverse of a matrix = solving systems of linear equations

2.3.4 Algorithms for Solving a System of Linear Equations

- Assume that a solution exists
- Can determine the inverse A^-1 such that the solution of Ax = b is given as x = A^-1b
- However, only possible if A is a square matrix and invertible

- Disadvantage: Requires many computaitons for the matrix-matrix product and computing the inverse of ATA
- Discuss alternative approaches to solving systems of linear equations

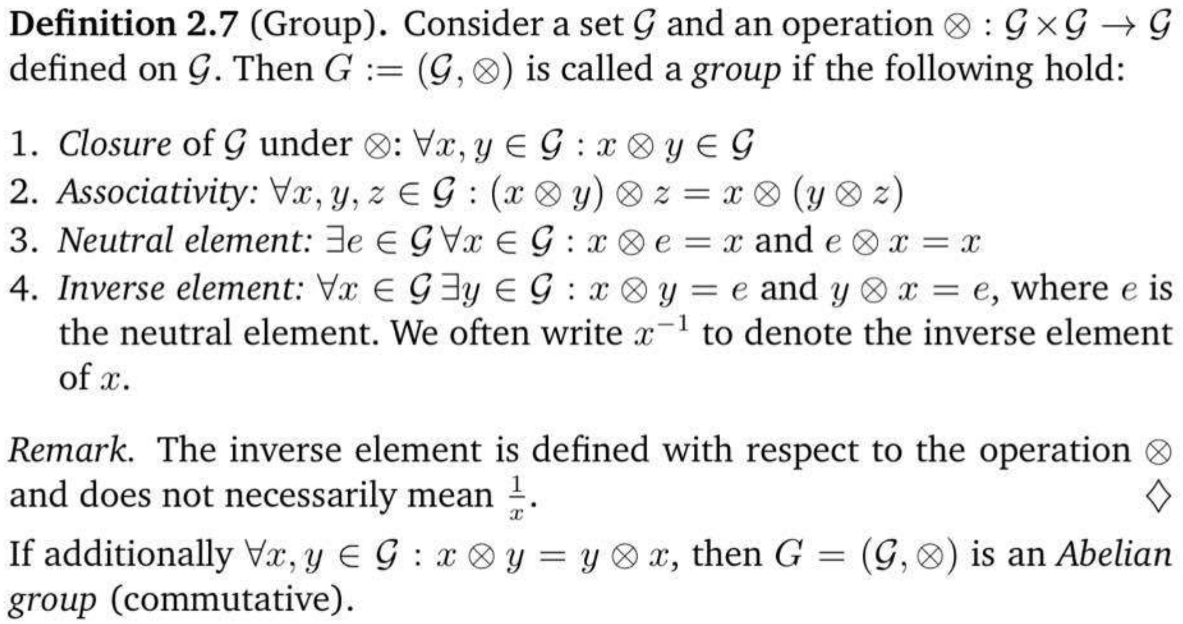
- **Gaussian elimination**
- Checking whether a set of vectors is linearly independent, computing the inverse of a matrix, computing the rank of a matrix, and determing a basis of a vector space
- Intuitive and constructive way to solve a system of linear equations with tousands of variables
- For systems with millions of variables, impractical

- In practice, systems of many linear equations are solved indirectly, by either stationary iterative methods, such as the Richardson method, the Jacobi method, and the Gauss-Seidel method, and the successive over-relaxation method, or Krylov subspace methods, generalized minimal residual, or biconjugate gradients

2.4 Vector Spaces

- Systems of linear equations can be compactly represented using matrix-vector notation
- Have a closer look at ector spaces, a structured space in which vectors live

2.4.1 Groups



Definition 2.7 (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:

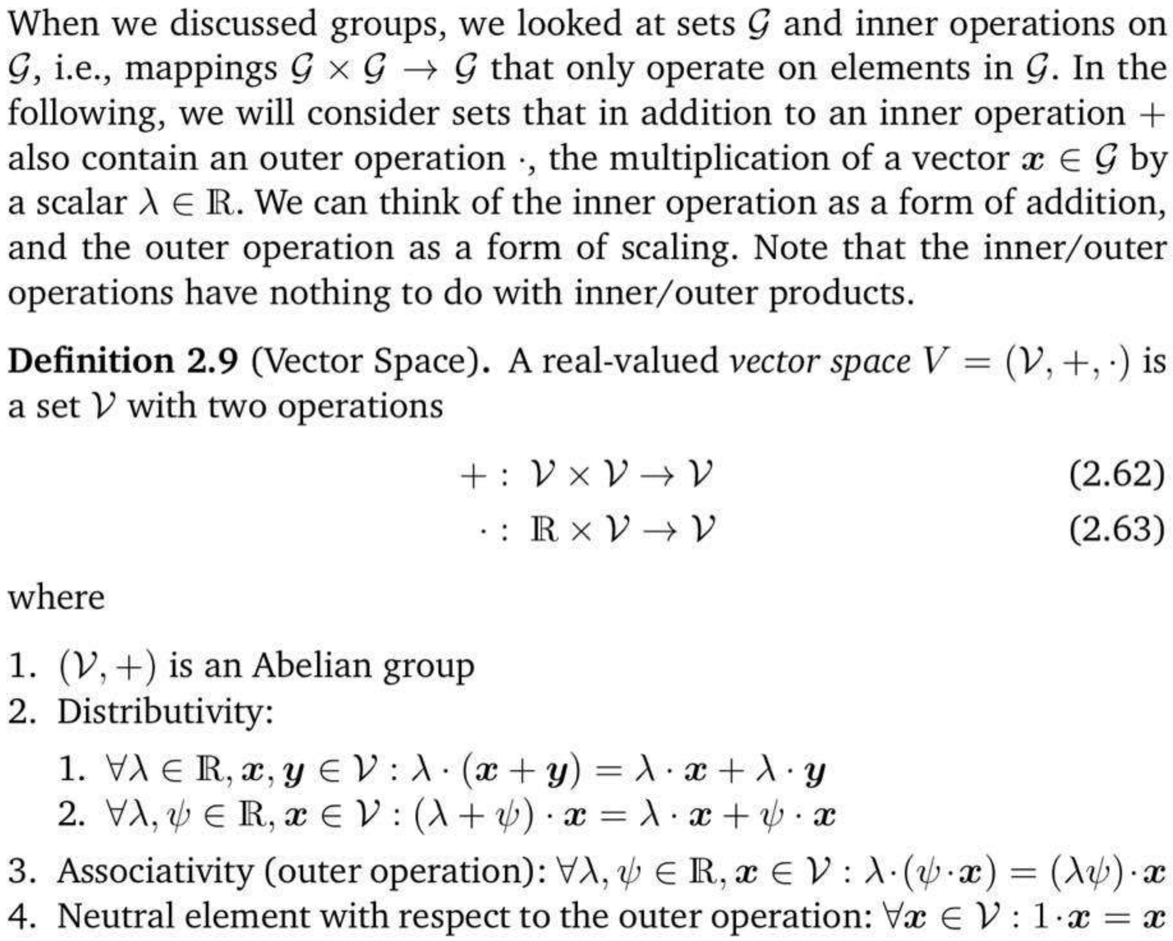
1. *Closure* of $\mathcal{G}$ under $\otimes$: $\forall x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$
2. *Associativity:* $\forall x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. *Neutral element:* $\exists e \in \mathcal{G}\, \forall x \in \mathcal{G} : x \otimes e = x$ 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 often write x^{-1} to denote the inverse element of x.

Remark. The inverse element is defined with respect to the operation $\otimes$ and does not necessarily mean $\frac{1}{x}$. ◇

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

- Definition 2.8 General Linear Group. The set of regular matrices A ∈ R nxn is a group with respect to matrix multiplication as defined in (2.13)
- General linear group GL(n, R)

2.4.2 Vector Spaces



When we discussed groups, we looked at sets $\mathcal{G}$ and inner operations on $\mathcal{G}$, i.e., mappings $\mathcal{G} \times \mathcal{G} \to \mathcal{G}$ that only operate on elements in $\mathcal{G}$. In the following, we will consider sets that in addition to an inner operation $+$ also contain an outer operation $\cdot$, the multiplication of a vector $\boldsymbol{x} \in \mathcal{G}$ by a scalar $\lambda \in \mathbb{R}$. We can think of the inner operation as a form of addition, and the outer operation as a form of scaling. Note that the inner/outer operations have nothing to do with inner/outer products.

Definition 2.9 (Vector Space)**.** A real-valued *vector space* $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations

$$+ : \ \mathcal{V} \times \mathcal{V} \to \mathcal{V} \tag{2.62}$$

$$\cdot : \ \mathbb{R} \times \mathcal{V} \to \mathcal{V} \tag{2.63}$$

where

1. $(\mathcal{V}, +)$ is an Abelian group
2. Distributivity:
   1. $\forall \lambda \in \mathbb{R}, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} : \lambda \cdot (\boldsymbol{x} + \boldsymbol{y}) = \lambda \cdot \boldsymbol{x} + \lambda \cdot \boldsymbol{y}$
   2. $\forall \lambda, \psi \in \mathbb{R}, \boldsymbol{x} \in \mathcal{V} : (\lambda + \psi) \cdot \boldsymbol{x} = \lambda \cdot \boldsymbol{x} + \psi \cdot \boldsymbol{x}$
3. Associativity (outer operation): $\forall \lambda, \psi \in \mathbb{R}, \boldsymbol{x} \in \mathcal{V} : \lambda \cdot (\psi \cdot \boldsymbol{x}) = (\lambda\psi) \cdot \boldsymbol{x}$
4. Neutral element with respect to the outer operation: $\forall \boldsymbol{x} \in \mathcal{V} : 1 \cdot \boldsymbol{x} = \boldsymbol{x}$

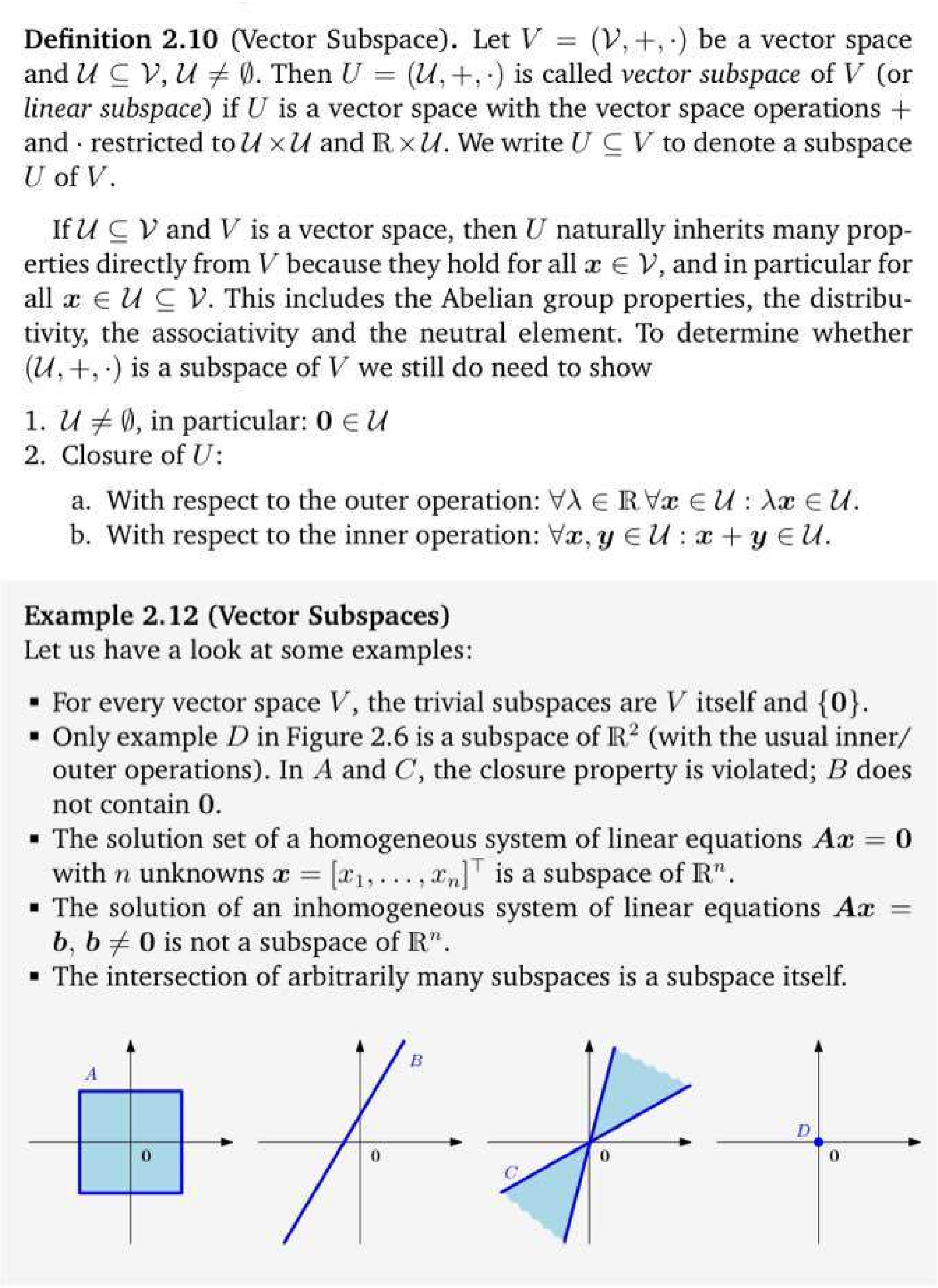
- Vectors: elements x ∊ V
- Zero vector 0 = [0,...,0]T (V, +)
- Vector addition: Inner operation +
- Scalars: Elements scalar ∊ R
- Remark. A vector multiplication ab, a, b ∊ Rn is not defined
- Array multiplication is common to many programming languages but makes matematically limited sense using the standard rules for matrix multiplication
  - By treating vectors as n X 1 matrices, can use the matrix multiplication as defined
  - However, dimensions of the vectors do not match

- Remark. Denote a vector space (V, +, .) by V when + and . are the standard vector addition and scalar multiplicaiton
- Use the notation x ∊ V for vectors in V to simplify notation

- Remark. Vector spaces Rn, Rnx1, R1xn are only different in the way we write vectors
- Write x to denote a column vector, and a row vector transpose of x

2.4.3 Vector Subspaces

- Vector subspaces: Sets contained in the original vector space with the property that when we perform vector space operations on elements within this subspce, we will never leave it
- In this sense, they are closed



Definition 2.10 (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 $\mathcal{U} \subseteq \mathcal{V}$ and V is a vector space, then U naturally inherits many properties directly from V because they hold for all $\boldsymbol{x} \in \mathcal{V}$, and in particular for all $\boldsymbol{x} \in \mathcal{U} \subseteq \mathcal{V}$. This includes the Abelian group properties, the distributivity, the associativity and the neutral element. To determine whether $(\mathcal{U}, +, \cdot)$ is a subspace of V we still do need to show

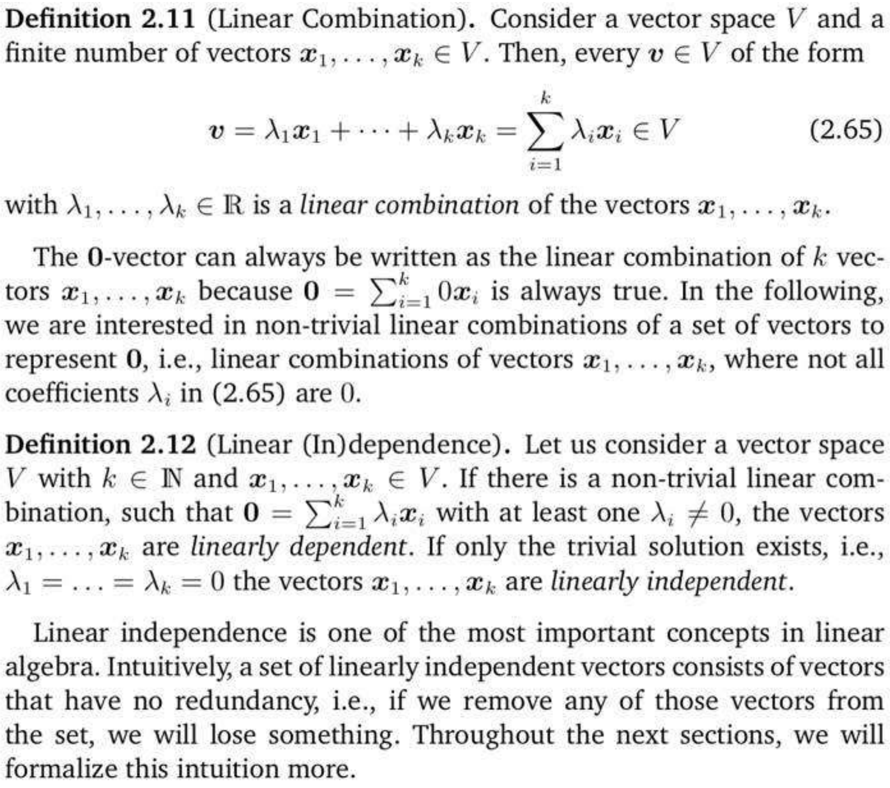
1. $\mathcal{U} \neq \emptyset$, in particular: $\mathbf{0} \in \mathcal{U}$
2. Closure of U:
   a. With respect to the outer operation: $\forall \lambda \in \mathbb{R}\, \forall \boldsymbol{x} \in \mathcal{U} : \lambda \boldsymbol{x} \in \mathcal{U}$.
   b. With respect to the inner operation: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{U} : \boldsymbol{x} + \boldsymbol{y} \in \mathcal{U}$.

Example 2.12 (Vector Subspaces)
Let us have a look at some examples:

- For every vector space V, the trivial subspaces are V itself and $\{\mathbf{0}\}$.
- Only example D in Figure 2.6 is a subspace of $\mathbb{R}^2$ (with the usual inner/outer operations). In A and C, the closure property is violated; B does not contain $\mathbf{0}$.
- The solution set of a homogeneous system of linear equations $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ with n unknowns $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$ is a subspace of $\mathbb{R}^n$.
- The solution of an inhomogeneous system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{b} \neq \mathbf{0}$ is not a subspace of $\mathbb{R}^n$.
- The intersection of arbitrarily many subspaces is a subspace itself.

2.5 Linear Independence

- Concepts of lienar combinations and linear dependence



Definition 2.11 (Linear Combination)**.** Consider a vector space V and a finite number of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$. Then, every $\boldsymbol{v} \in V$ of the form

$$\boldsymbol{v} = \lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in V \tag{2.65}$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ is a *linear combination* of the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$.

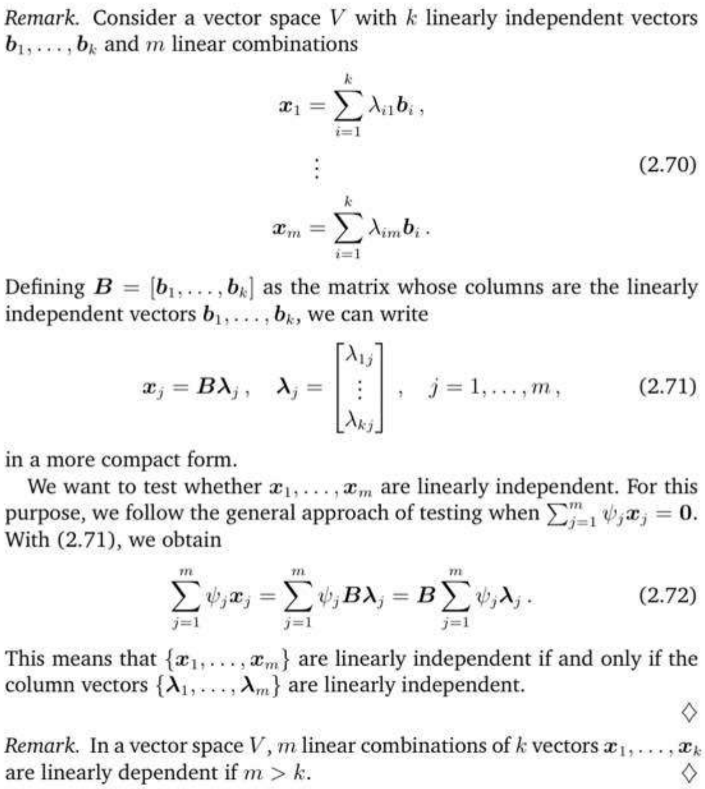
The $\boldsymbol{0}$-vector can always be written as the linear combination of k vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ because $\boldsymbol{0} = \sum_{i=1}^{k} 0\boldsymbol{x}_i$ is always true. In the following, we are interested in non-trivial linear combinations of a set of vectors to represent $\boldsymbol{0}$, i.e., linear combinations of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, where not all coefficients λ_i in (2.65) are 0.

Definition 2.12 (Linear (In)dependence)**.** Let us consider a vector space V with $k \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$. If there is a non-trivial linear combination, such that $\boldsymbol{0} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i$ with at least one $\lambda_i \neq 0$, the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly dependent*. If only the trivial solution exists, i.e., $\lambda_1 = \ldots = \lambda_k = 0$ the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly independent*.

Linear independence is one of the most important concepts in linear algebra. Intuitively, a set of linearly independent vectors consists of vectors that have no redundancy, i.e., if we remove any of those vectors from the set, we will lose something. Throughout the next sections, we will formalize this intuition more.

- One of the most importanc econcepts in linear algebra
- A set of linearly independent vectors consists of vectors that have no redundancy
  - If remove any of vectors, we will lost something

- Remark. The following properties are useful to find out whether vectors are linearly independent:
  - k vectors are either linearly dependent or linearly indpendent
  - IF at least one of the vectors x1, ..., xk is 0 then they are linearly dependent. The same holds if two vectors are identical
  - The vectors { x1, ..., xk : xi != 0, i = 1, ..., k}, k>= 2, are linearly dependent if and only if (at least) one of them is a linear combination of the others
  - If one vector is a multiple of another vectors, i.e., xi = λ ∈ R then the set { x1, ..., xk : xi != 0, i = 1, ..., k} is linearly dependent
  - A practical way of checking whether vectors x1, ..., xk ∈ V are linearly independent is to use Gaussian elimination
    - Write all vectors as columns of a matrix A and perform Gaussian elimination until the matrix is in row echelon form
  - All column vectors are linearly independent if and only if all columns are pivot columns
    - If at least one non-pivot column, the columns are linearly dependent



Remark. Consider a vector space V with k linearly independent vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ and m linear combinations

$$\begin{aligned} \boldsymbol{x}_1 &= \sum_{i=1}^{k} \lambda_{i1} \boldsymbol{b}_i \,, \\ &\vdots \\ \boldsymbol{x}_m &= \sum_{i=1}^{k} \lambda_{im} \boldsymbol{b}_i \,. \end{aligned} \tag{2.70}$$

Defining $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k]$ as the matrix whose columns are the linearly independent vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$, we can write

$$\boldsymbol{x}_j = \boldsymbol{B}\boldsymbol{\lambda}_j \,, \quad \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{1j} \\ \vdots \\ \lambda_{kj} \end{bmatrix} , \quad j = 1, \ldots, m \,, \tag{2.71}$$

in a more compact form.

We want to test whether $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are linearly independent. For this purpose, we follow the general approach of testing when $\sum_{j=1}^{m} \psi_j \boldsymbol{x}_j = \boldsymbol{0}$. With (2.71), we obtain

$$\sum_{j=1}^{m} \psi_j \boldsymbol{x}_j = \sum_{j=1}^{m} \psi_j \boldsymbol{B}\boldsymbol{\lambda}_j = \boldsymbol{B} \sum_{j=1}^{m} \psi_j \boldsymbol{\lambda}_j \,. \tag{2.72}$$

This means that $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ are linearly independent if and only if the column vectors $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m\}$ are linearly independent.

◇

Remark. In a vector space V, m linear combinations of k vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are linearly dependent if $m > k$. ◇

In [ ]: