Principles of Machine Learning

Lecture 1: Introduction and Course Overview

Sharif University of Technology
Dept. of Aerospace Engineering

February 13, 2025

Table of Contents

Course Overview

- **Course Title:** Principles of Machine Learning
- **Target Audience:** Undergraduate and graduate students from diverse academic backgrounds
- **Focus:** Balancing theoretical foundations with practical implementations
- **Key Tools:** Python, NumPy, Pandas, TensorFlow/Keras
- **Outcome:** Develop skills to understand, implement, and evaluate machine learning solutions

Course Objectives

By the end of this course, students will be able to:

1. Understand core machine learning concepts and algorithms
2. Design and implement ML solutions using Python
3. Select appropriate algorithms for different problems
4. Preprocess and prepare data for ML tasks
5. Evaluate and validate ML models
6. Create end-to-end ML solutions for real-world problems

Course Structure

- **Week 1:** Introduction to Machine Learning
- **Weeks 1-3:** Mathematical Foundations
- **Weeks 4-5:** Supervised Learning - Regression
- **Weeks 5-6:** Supervised Learning - Classification
- **Week 7:** Model Evaluation and Validation
- **Week 8:** Practical Considerations
- **Weeks 9-10:** Unsupervised Learning
- **Week 11:** Neural Networks Fundamentals
- **Weeks 12-13:** Deep Learning
- **Week 14:** Reinforcement Learning
- **Week 15:** Advanced Topics

Week 1: Introduction to Machine Learning

- What is machine learning?
- Types of machine learning problems
- Python programming basics (in Lab)
- Introduction to scientific computing (NumPy, Pandas) (in Lab)
- Case studies: ML applications in aerospace industry

Mathematical Foundations (Weeks 1-3)

- Linear algebra review
- Probability and statistics
- Calculus concepts
- Optimization basics

Supervised Learning (Weeks 4-6)

- **Regression:** Linear regression, polynomial regression, regularization techniques, gradient descent
- **Classification:** Logistic regression, SVMs, decision trees, random forests, k-Nearest Neighbors
- Applications: System modeling, fault detection, flight phase classification

Model Evaluation and Validation (Week 7)

- Cross-validation techniques
- Performance metrics (accuracy, precision, recall, F1-score)
- Bias-variance tradeoff
- Model selection and hyperparameter tuning

Practical Considerations (Week 8)

- Data cleaning, normalization, and transformation
- Feature selection and extraction techniques
- Industry case studies: Predictive maintenance, flight data analysis, aerospace design optimization
- Model deployment basics

Unsupervised Learning (Weeks 9-10)

- Clustering algorithms (K-means, Hierarchical, DBSCAN)
- Dimensionality reduction (PCA, t-SNE)
- Anomaly detection
- Applications: Process monitoring, pattern discovery, flight data clustering

Neural Networks and Deep Learning (Weeks 11-12)

- Neural Networks Fundamentals: Architecture, backpropagation, activation functions
- Deep Learning: CNNs, RNNs, time series data, modern architectures (Transformers), transfer learning
- Applications: Signal processing, computer vision, object detection in satellite imagery

Reinforcement Learning (Week 13)

- Markov Decision Processes
- Q-learning
- Policy Gradient methods
- Applications: Control systems, resource optimization

Advanced Topics (Week 15)

- Transfer learning
- Ensemble methods
- Explainable AI
- Few-shot and zero-shot learning

What is Machine Learning

Machine Learning

- The science (and art) of programming computers to *learn from data*
- "To give computers the ability to learn without explicit programming"
  - *Arthur Samuel*, 1959
- "To learn from Experience E w.r.t some Task T and performance measure P, if its performance on T, as measured by P, improves with E"
  - *Tom Mitchell*, 1997

What is Machine Learning

There are three concepts at the core of ML [1]:

Data

- Training data
- The model input
- Stores valuable information

Model

- The input-output mapping
- Parametric
- Non-parametric

Learning

- Determining model's parameters
- Optimization
- Generalization

Why ML?

Traditional approach vs. ML approach

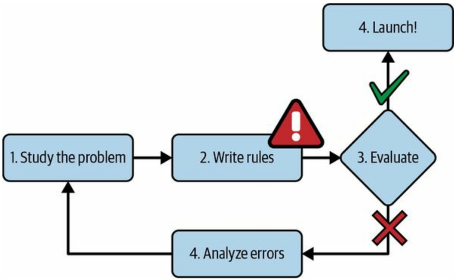


Figure: The traditional approach

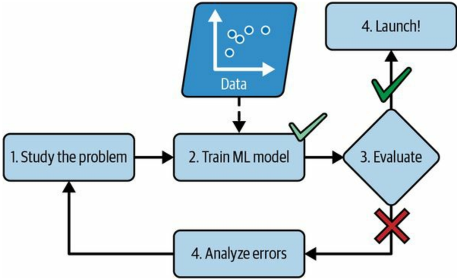


Figure: The machine learning approach

What is ML

Traditional vs. Machine Learning-based solutions

Traditional approach:

- Expert knowledge
- High interpretability
- Well-known and typical scenarios
- Lacking access to sufficient amount of data

Machine learning-based approach:

- Data-rich scenarios
- Unseen settings
- Adaptability for various conditions
- Difficulty of deriving a governing model

Why ML?

Automatic adaptation for complex problems

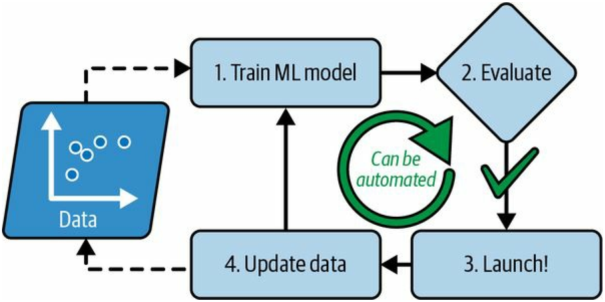


Figure: Automatically adapting to change

Why ML?

Inspect ML models to see what they have learned (*not always possible though*)

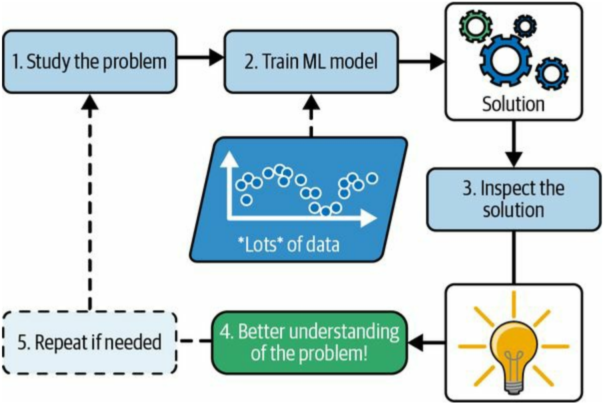


Figure: ML can help humans learn

What is Machine Learning

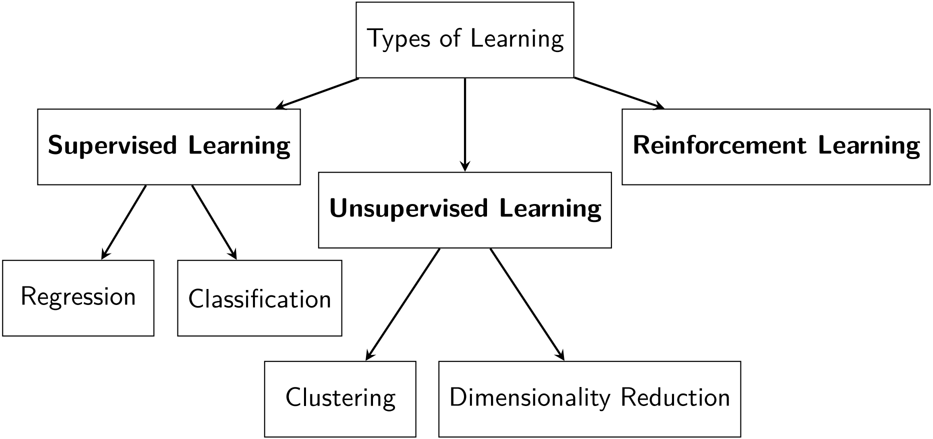
When ML?

- Problems with the traditional solutions being a long list of rules
- Complex problems with the traditional solutions being no good
- Fluctuating environments
- Revealing insights about the data

Examples of Applications

- Detecting tumors in brain scans
  (*CNNs/Transformers*)
- Forecasting a company's revenue next year
  (*Regression, Random forest, Artificial neural network*)
- Representing a complex, high-dimensional dataset in a clear and insightful diagram
  (*Dimensionality reduction*)
- Building an intelligent bot for a game
  (*RL*)

What is Machine Learning



* Some modern learning paradigms, such as semi-supervised learning, are not mentioned above.
* Deep learning, Deep Generative Modelling, etc. are not explicitly discussed in the domain of ML.

Supervised Learning

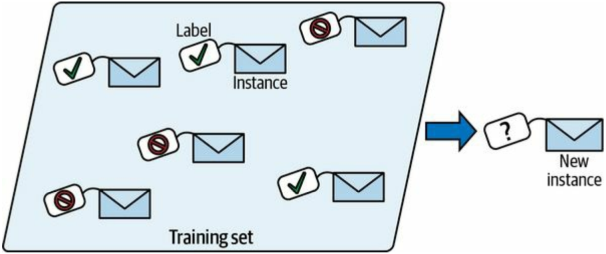


Figure: A labeled training set for spam classification (an example of supervised learning)

Supervised Learning - An Example

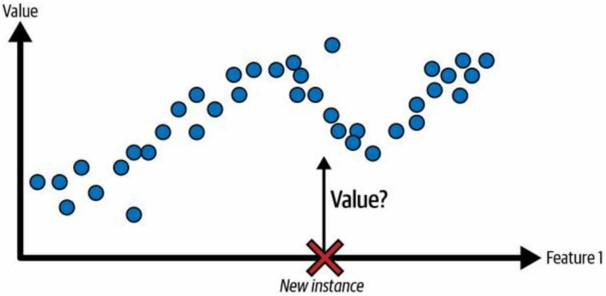


Figure: A regression problem: predict a value (or multiple values), given an input feature (or multiple input features)

Unsupervised Learning

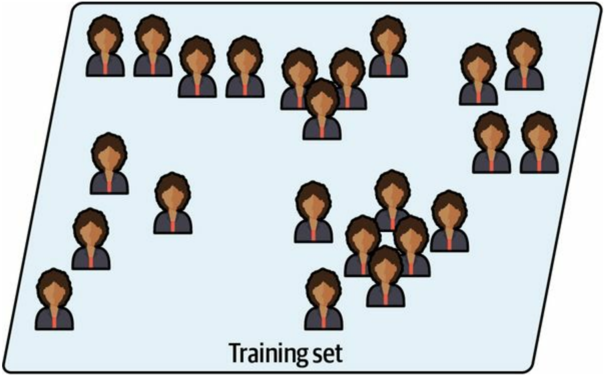


Figure: An unlabeled training set for unsupervised learning

Unsupervised Learning - An Example

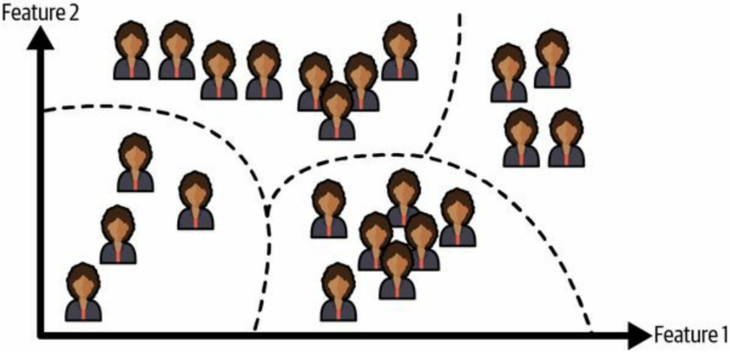


Figure: Clustering

Unsupervised Learning - An Example

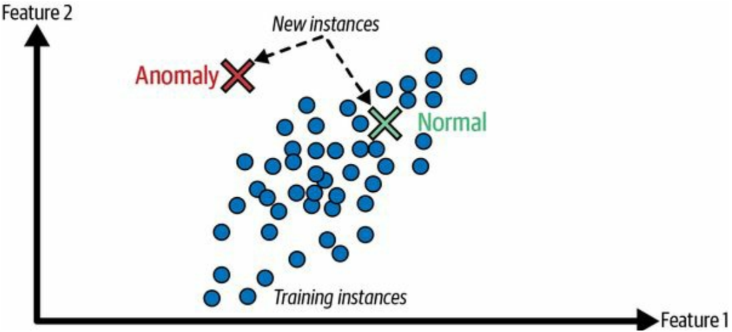


Figure: Anomaly detection

Semi-supervised learning

Combines unsupervised and supervised learning

- leverages both unlabeled and partially labeled data

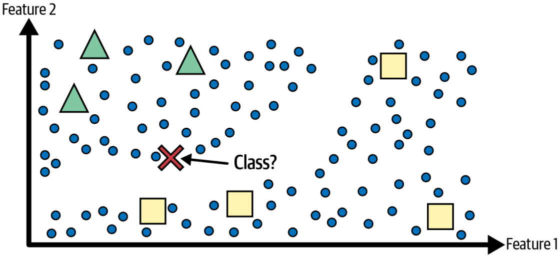


Figure: Semi-supervised learning with two classes (triangles and squares): the unlabeled examples (circles) help classify a new instance (the cross) into the triangle class rather than the square class, even though it is closer to the labeled squares

What is Machine Learning

Self-Supervised Learning

Generating a labeled dataset from a fully unlabeled dataset

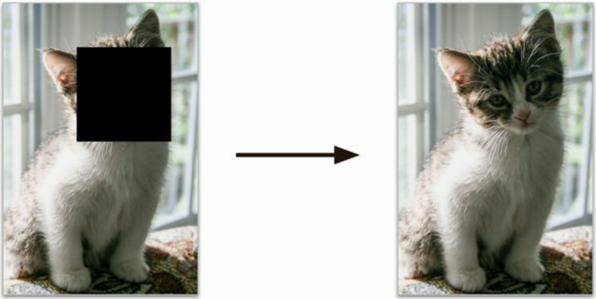


Figure: Self-supervised learning example: input (left) and target (right)

Reinforcement Learning

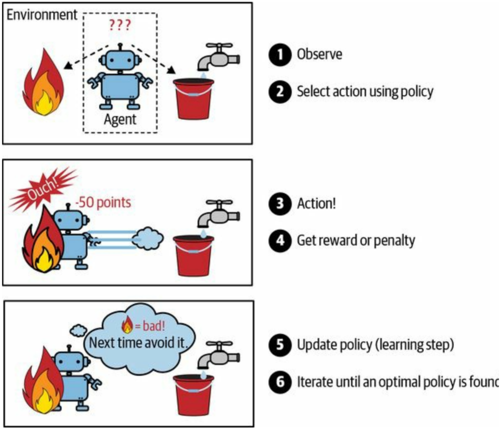


Figure: Reinforcement learning

Batch Learning vs. Online Learning

Batch Learning (Offline Learning)

- incapable of learning incrementally
- train using all available data
- training before launching into production
- susceptible to *model rot* or *data drift*
- a lot of computing resources (CPU, memory space, etc.)

Batch Learning vs. Online Learning

Online Learning

- capable of learning incrementally
- train using mini-batches
- capturing new data on the fly
- out-of-core learning
- adapt quickly to changing data

Batch Learning vs. Online Learning

Online Learning

- A model is trained, deployed, and continuously updated with new data.

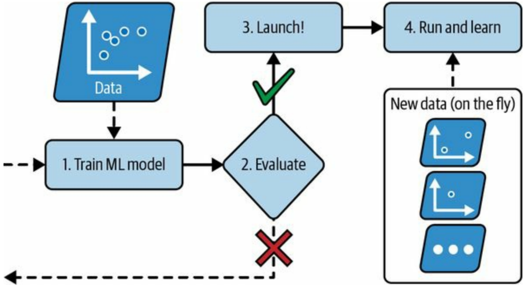


Figure: In online learning, a model is trained and launched into production, and then it keeps learning as new data comes in

Batch Learning vs. Online Learning

Online Learning

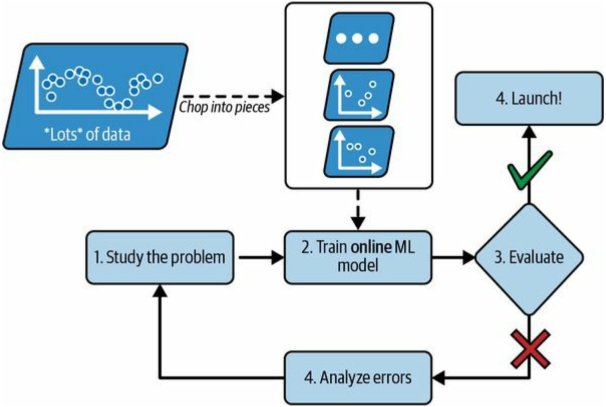


Figure: Using online learning to handle huge datasets

Instance-Based Learning vs. Model-Based Learning

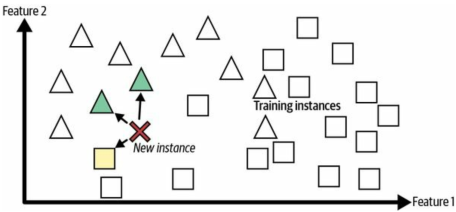


Figure: Instance-based learning

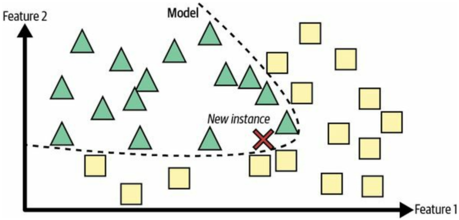


Figure: Model-based learning

Main Challenges of Machine Learning

1. Insufficient Quantity of Training Data

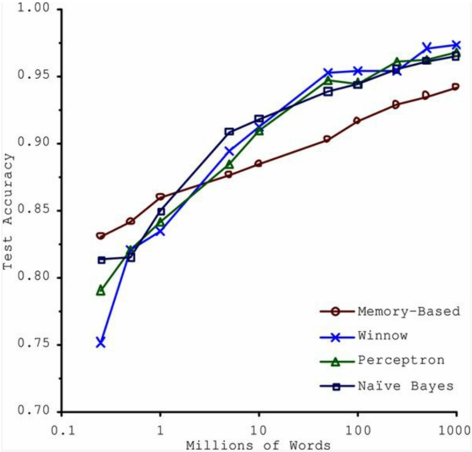


Figure: The importance of data versus algorithms

Main Challenges of Machine Learning

2. Nonrepresentative Training Data

Training data should be representative of the new cases to generalize to

- too small samples $\longrightarrow$ sampling noise
- very large samples $\longrightarrow$ sampling bias

Main Challenges of Machine Learning

3. Poor-Quality Data

Errors, outliers, noise, etc.

- Discard the outliers or fix the errors manually
- Handling missing features (ignore feature, ignore data, or fill in the missing values)

Main Challenges of Machine Learning

4. Irrelevant Features

Feature engineering:

- *Feature selection* (selecting the most useful features)
- *Feature extraction* (combining existing features to produce a useful one)
- Gathering new data $\longrightarrow$ Creating new features

Main Challenges of Machine Learning

5. Overfitting the Training Data

Performing well on training data but not generalizing well
Complex relative to the amount and noisiness of the training data
Possible solutions:

- Simplifying or constraining the model
- Gather more training data
- Reduce the noise in the training data

Main Challenges of Machine Learning

6. Underfitting the Training Data

Too simple to learn the underlying structure of data
Possible solutions:

- Select a model with more parameters
- Feed better features (feature engineering)
- Reduce the constraints

Testing and Validating

Hyperparameter Tuning and Model Selection

- Train the model on the *training set*
- Test the model on the *test set*
- *Generalization error*: The error rate on new cases
- Evaluate the model on the test set $\longrightarrow$ Estimating the generalization error
- The value above indicates how good your model performs on unseen data
- Common 80-20 splitting

But, what if the model is the best model only for *that particular set?*

Testing and Validation

Hyperparameter Tuning and Model Selection

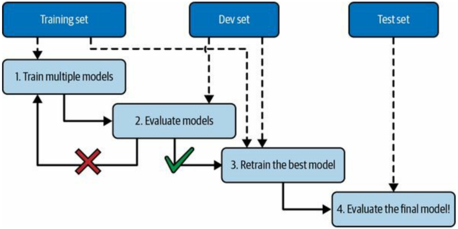


Figure: Model selection using holdout validation

Validation set, too small or too large $\longrightarrow$ cross validation

Data Mismatch

- Representativity of the test and validation (development) sets
- How to decide the poor performance is due to model overfitting or data mismatch $\longrightarrow$ using train-dev set

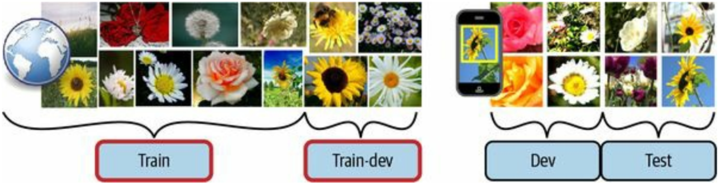


Table of Contents

- **Linear Algebra**
- Probability & Statistics
- Calculus
- Optimization

Mathematical Foundations: Linear Algebra

- Linear Algebra: The study of vectors and their manipulations
- Any mathematical object that satisfies the following two properties can be considered a **vector**:
  - $\mathbf{x} + \mathbf{y} = \mathbf{z}$
  - $\lambda \mathbf{x} = \mathbf{w}$, $\lambda \in \mathbb{R}$
- Examples of vectors: polynomials, audio signals, elements of $\mathbb{R}^n$

Definition (Matrix)

Matrix **A** is an $m \cdot n$-tuple of elements

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

Definition (Matrix Addition and Multiplication)

The sum of two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$:

$$\mathbf{A} + \mathbf{B} := \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

For matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times k}$, the elements of the product $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times k}$ are:

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

Definition (Associativity)

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

Definition (Distributivity)

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

Definition (Inverse)

For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ is the *inverse* of $\mathbf{A}$ and denoted as $\mathbf{A}^{-1}$ if:

$$\mathbf{AB} = \mathbf{I}_n = \mathbf{BA}$$

The inverse does exist $\longrightarrow$ $\mathbf{A}$ is *invertible/nonsingular*

Definition (Transpose)

For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ is the *transpose* of $\mathbf{A}$ and denoted as $\mathbf{A}^\top$ if:

$$b_{ij} = a_{ji}$$

If $\mathbf{A} = \mathbf{A}^\top$ $\longrightarrow$ $\mathbf{A}$ is *symmetric*

A system of linear equations as below

$$a_{11}x_1 + \cdots + a_{1n}x_n = b_1$$
$$\vdots$$
$$a_{m1}x_1 + \cdots + a_{mn}x_n = b_m$$

can be formulated using matrices notation as

$$\mathbf{Ax} = \mathbf{b}.$$

How to solve them?

Example 1. For $a \in \mathbb{R}$, we seek all solutions of the following system of equations:

$$
\begin{array}{rcrcrcrcrcr}
-2x_1 & + & 4x_2 & - & 2x_3 & - & x_4 & + & 4x_5 & = & -3 \\
4x_1 & - & 8x_2 & + & 3x_3 & - & 3x_4 & + & x_5 & = & 2 \\
x_1 & - & 2x_2 & + & x_3 & - & x_4 & + & x_5 & = & 0 \\
x_1 & - & -2x_2 & & & - & 3x_4 & + & 4x_5 & = & a
\end{array}
$$

Example 1.

$$\mathbf{Ax} = \mathbf{b} \xrightarrow{\textit{augmented matrix}} [\mathbf{A} \mid \mathbf{b}]:$$

$$\left[\begin{array}{ccccc|c} -2 & 4 & -2 & -1 & 4 & -3 \\ 4 & -8 & 3 & -3 & 1 & 2 \\ 1 & -2 & 1 & -1 & 1 & 0 \\ 1 & -2 & 0 & -3 & 4 & a \end{array}\right]$$

After applying some *elementary transformations*:

$$\rightsquigarrow \left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 3 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & a+1 \end{array}\right]$$

Known as *row-echelon form*.

Only for $a = 1$ this system can be solved.

Example 1.

A *particular solution* is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}.$$

The *general solution* is

$$\left\{ \mathbf{x} \in \mathbb{R}^5 : \mathbf{x} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ -1 \\ 2 \\ 1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R} \right\}.$$

Definition (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*) is always strictly to the right of the pivot of the row above it.

Definition (Reduced Row-Echelon Form)

An equation system is in *rref* or *row canonical form* if

- It is in row-echelon form.
- Every pivot is 1.
- The pivot is the only non-zero entry in its column.

Definition

Gaussian elimination is an algorithm that performs elementary transformations to bring a system of linear equations into reduced row-echelon form.

Example 2. Calculating an Inverse Matrix by Gaussian Elimination:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\xrightarrow{\textit{augmented matrix}}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right]$$

Example 2.

After using Gaussian Elimination to reach rref:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -1 & 2 & -2 & 2 \\ 0 & 1 & 0 & 0 & 1 & -1 & 2 & -2 \\ 0 & 0 & 1 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 2 \end{array}\right]$$

Thus,

$$\mathbf{A}^{-1} = \begin{bmatrix} -1 & 2 & -2 & 2 \\ 1 & -1 & 2 & -2 \\ 1 & -1 & 1 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

Check this by $\mathbf{AA}^{-1} = \mathbf{I}$

Definition (Group)

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

1. *Closure* of C under $\otimes$: $\forall x, y \in \mathbb{C} \ : \ x \otimes y \in \mathbb{C}$
2. *Associativity*: $\forall x, y, z \in \mathbb{C} \ : \ (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. *Neutral element*: $\exists e \in \mathbb{C} \, \forall x \in \mathbb{C} \ : \ x \otimes e = x$ and $e \otimes x = x$
4. *Inverse element*: $\forall x \in \mathbb{C} \, \exists y \in \mathbb{C} \ : \ x \otimes y = e$ and $y \otimes x = e$

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

Definition (Vector Space)

A real-valued *vector space* $\mathrm{V} = (\mathcal{V}, +, \cdot)$ 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

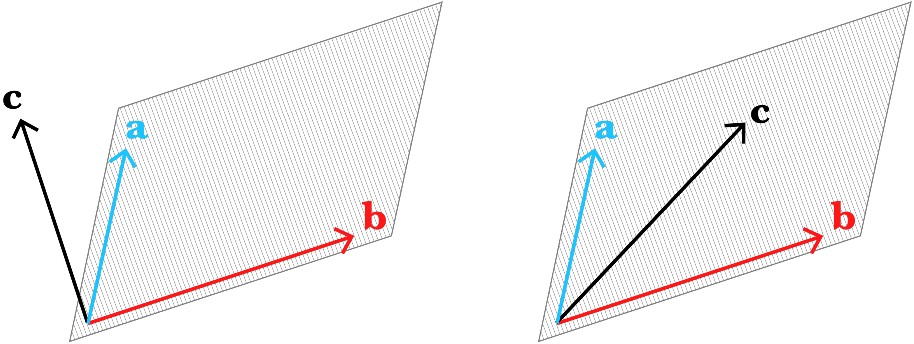
1. $(\mathcal{V}, +)$ is an Abelian Group.
2. Distributivity for **x**, **y** $\in \mathcal{V}$
3. Associativity (outer operation)
4. Neutral element with respect to the outer operation

Definition (Linear (In)dependence)

Consider a vector space V with $k \in \mathbb{N}$ and $\mathbf{x_1}, \ldots, \mathbf{x_k} \in \mathrm{V}$. If there is a non-trivial linear combination, such that $0 = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i$ with at least one $\lambda_i \neq 0$, the vectors $\mathbf{x_1}, \ldots, \mathbf{x_k}$ are *linearly dependent*.
If only the trivial solution exists, i.e., $\lambda_1 = \cdots = \lambda_k = 0$ the vectors $\mathbf{x_1}, \ldots, \mathbf{x_k}$ are *linearly independent*.

- To investigate linear independency of n vectors $\longrightarrow$ solve a homogenuous linear system of n equations.

Graphical interpretation of "Linear Independence"



Definition (Generating Set and Span)

For $\mathrm{V} = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subseteq \mathcal{V}$, $\mathcal{A}$ is a *generating set* of V if for every $\mathbf{v} \in \mathrm{V}$:

$$\mathbf{v} = \lambda_1 \mathbf{x}_1 + \cdots + \lambda_k \mathbf{x}_k.$$

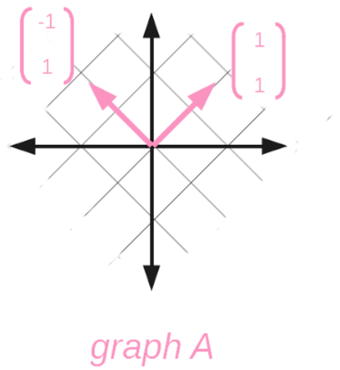
The set of all linear combinations of vectors in $\mathcal{A}$ is the *span* of the $\mathcal{A}$ and $\mathrm{V} = \mathrm{span}[\mathcal{A}]$ if $\mathcal{A}$ spans V.

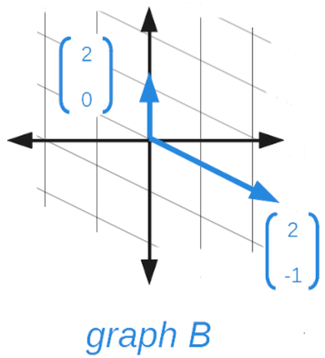
Definition (Basis)

For $\mathrm{V} = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subseteq \mathcal{V}$, a generating set $\mathcal{A}$ of V is *minimal* if there exists no smaller set $\tilde{\mathcal{A}} \not\subseteq \mathcal{A} \subseteq \mathcal{V}$ that spans V.

Every linearly independent generating set of V that is minimal, is a *basis* of V.

Graphical interpretation of “Basis Vectors”





Example 3. The first two sets are both bases in $\mathbb{R}^3$, however the third set is not a base in $\mathbb{R}^4$. (Why?)

$$\mathcal{B}_1 = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

$$\mathcal{B}_2 = \left\{ \begin{bmatrix} 0.5 \\ 0.8 \\ 0.4 \end{bmatrix}, \begin{bmatrix} 1.8 \\ 0.3 \\ 0.3 \end{bmatrix}, \begin{bmatrix} -2.2 \\ -1.3 \\ 3.5 \end{bmatrix} \right\}$$

$$\mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ -4 \end{bmatrix} \right\}$$

- No unique basis
- All bases have the same number of elements, called the *basis vector*
- *Dimension* of V, dim(V): The number of basis vectors of V
- $\mathrm{U} \subseteq \mathrm{V} \longrightarrow \dim(\mathrm{U}) \leq \dim(\mathrm{V})$ & $\mathrm{U} = \mathrm{V} \longrightarrow \dim(\mathrm{U}) = \dim(\mathrm{V})$
- Intuitively, dim(V) is the number of independent directions in V.

Definition (Rank)

The *rank* of $\mathbf{A}$ is the number of linearly independent columns (or rows) of $\mathbf{A} \in \mathbb{R}^{m \times n}$.

- $\text{rk}(\mathbf{A}) = \text{rk}(\mathbf{A}^\top)$
- The columns of $\mathbf{A}$ span a subspace $U \subseteq \mathbb{R}^m$ with $\dim(U) = \text{rk}(\mathbf{A})$
- The rows of $\mathbf{A}$ span a subspace $W \subseteq \mathbb{R}^n$ with $\dim(W) = \text{rk}(\mathbf{A})$
- $\mathbf{A} \in \mathbb{R}^{n \times n}$ is regular $\iff$ $\text{rk}(\mathbf{A}) = n$
- $\mathbf{Ax} = \mathbf{b}$ can be solved $\iff$ $\text{rk}(\mathbf{A}) = \text{rk}(\mathbf{A}|\mathbf{b})$
- Subspace of solutions for $\mathbf{Ax} = \mathbf{0}$ (*kernel* or *null space*) have dimension $n - \text{rk}(\mathbf{A})$
- $\mathbf{A}$ has *full rank* if $\text{rk}(\mathbf{A}) = \min(m, n)$, otherwise has *rank deficiency*

Definition (Linear Mapping)

For vector spaces V, W, a mapping $\Phi : \mathrm{V} \to \mathrm{W}$ is a *linear mapping/ vector space homomorphism/linear transformation* if

$$\Phi(\lambda \mathbf{x} + \psi \mathbf{y}) = \lambda \Phi(\mathbf{x}) + \psi \Phi(\mathbf{x})$$

Definition (Injective, Surjective, Bijective)

For sets $\mathcal{V}, \mathcal{W}$, a mapping $\Phi : \mathcal{V} \to \mathcal{W}$ is

- *Injective* if $\Phi(\mathbf{x}) = \Phi(\mathbf{y}) \implies \mathbf{x} = \mathbf{y}$
- *Surjective* if $\Phi(\mathcal{V}) = \mathcal{W}$
- *Bijective* if satisfies both of above

- *Isomorphism*: $\Phi : \mathrm{V} \to \mathrm{W}$ linear & bijective
- *Endomorphism*: $\Phi : \mathrm{V} \to \mathrm{V}$ linear
- *Automorphism*: $\Phi : \mathrm{V} \to \mathrm{V}$ linear & bijective
- $\mathrm{id}_{\mathrm{V}} : \mathrm{V} \to \mathrm{V}, \mathbf{x} \mapsto \mathbf{x}$ as the *identity mapping* in V

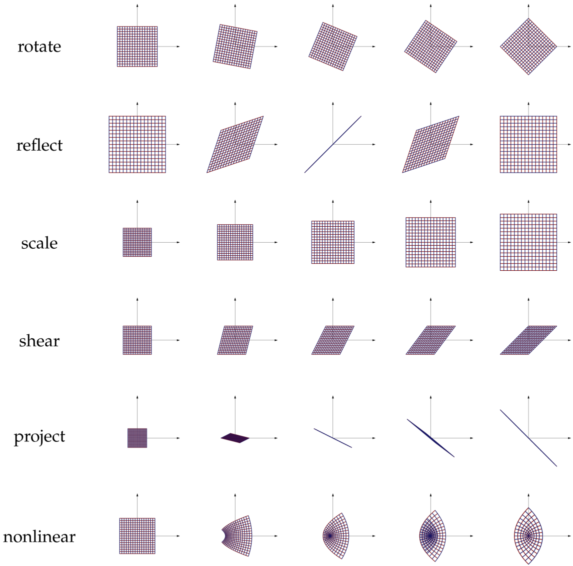
Definition (Transformation Matrix)

Consider vector spaces V, W with corresponding (ordered) bases $\mathrm{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ and $\mathrm{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_m)$. Also, we consider a *linear mapping* $\Phi : \mathrm{V} \to \mathrm{W}$. For $j = \{1, \ldots, n\}$, $\Phi(\mathbf{b}_j) = \alpha_{1j}\mathbf{c}_1 + \cdots + \alpha_{mj}\mathbf{c}_m = \sum_{i=1}^{m} \alpha_{ij}\mathbf{c}_i$ is the unique representation of $\Phi(\mathbf{b}_j)$ with respect to C. Then, we call the $m \times n-$matrix $\mathbf{A}_\Phi$, whose elements are given by

$$A_\Phi(i,j) = \alpha_{ij},$$

the *transformation matrix* of Φ w.r.t. the ordered bases of B of V and C of W.

Examples of Transformation of Vectors



Theorem (Basis Change)

For a linear mapping $\Phi : \mathrm{V} \to \mathrm{W}$, *ordered bases*

$$\mathrm{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_\mathrm{n}), \quad \widetilde{\mathrm{B}} = (\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_\mathrm{n})$$

of V *and*

$$\mathrm{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{m}), \quad \widetilde{\mathrm{C}} = (\widetilde{\mathbf{c}}_1, \ldots, \widetilde{\mathbf{c}}_\mathrm{m})$$

of **W**, *transformation matrices w.r.t. to the preceding ordered bases are given as:*

$$\widetilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1}\mathbf{A}_\Phi\mathbf{S},$$

where $\mathbf{S} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ *and* $\mathbf{T} \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$ *are the transformation matrices of* id_V *and* id_W

Definition (Equivalence)

Two matrices $\mathbf{A}$, $\widetilde{\mathbf{A}} \in \mathcal{R}^{m\times n}$ are *equivalent* if there exist regular matrices $\mathbf{S} \in \mathbb{R}^{n\times n}$ and $\mathbf{T} \in \mathbb{R}^{m\times m}$, s.t.

$$\widetilde{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{S}.$$

Definition (Similarity)

Two matrices $\mathbf{A}$, $\widetilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ are *similar* if there exists a regular matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ with

$$\widetilde{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}.$$

Example 4. For a linear mapping $\Phi : \mathbb{R}^3 \to \mathbb{R}^4$ and the given transformation matrix in bases B, C, find the same matrix w.r.t bases $\widetilde{\mathrm{B}}, \widetilde{\mathrm{C}}$

$$\mathbf{A}_\Phi = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ -1 & 2 & 4 \end{bmatrix}$$

$$\mathrm{B} = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right), \quad \mathrm{C} = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right)$$

$$\widetilde{\mathrm{B}} = \left(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right), \quad \widetilde{\mathrm{C}} = \left(\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right)$$

Example 4.

$$\implies \quad \mathbf{S} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\widetilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1}\mathbf{A}_\Phi\mathbf{S} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}\begin{bmatrix} 3 & 2 & 1 \\ 0 & 4 & 2 \\ 10 & 8 & 4 \\ 1 & 6 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} -4 & -4 & -2 \\ 6 & 0 & 0 \\ 4 & 8 & 4 \\ 1 & 6 & 3 \end{bmatrix}.$$

Definition (Image & Kernel)

For $\Phi : \mathrm{V} \to \mathrm{W}$, the *kernel/null space* is:

$$\ker(\Phi) := \Phi^{-1}(\mathbf{0}_{\mathrm{W}}) = \{\mathbf{v} \in \mathrm{V} : \Phi(v) = \mathbf{0}_{\mathrm{W}}\}$$

and the *image/range* is:

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

V and W $\longrightarrow$ *domain* and *codomain* of Φ

Graphical interpretation of "Image" and "Kernel"

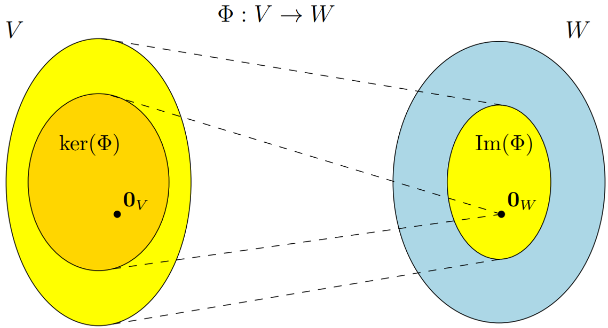


Figure: Kernel and image of a linear mapping $\Phi : \mathrm{V} \to \mathrm{W}$.

Theorem (Rank-Nullity / Fundamental Theorem of Linear Mappings)

For vector spaces V, W *and a linear mapping* $\Phi : V \rightarrow W$*:*

$$\dim(\ker(\Phi)) + \dim(\mathrm{Im}(\Phi)) = \dim(V)$$

- If $\dim(\ker(\Phi)) < \dim(V)$
  - $\ker(\Phi)$ is non-trivial $\longrightarrow$ kernel contains more than $\mathbf{0}_V$ and $\dim(\ker(\Phi)) \geq 1$
  - $\mathbf{A}_\Phi \mathbf{x} = \mathbf{0}$ has infinite number of solutions

- If $\dim(V) = \dim(W)$
  - Φ is injective
  - Φ is surjective
  - Φ is bijective

Definition (Norm)

A *norm* on a vector space V is a function

$$\begin{aligned} \|\cdot\| : \mathrm{V} &\to \mathbb{R}, \\ \mathbf{x} &\mapsto \|\mathbf{x}\|, \end{aligned}$$

which assigns each vector $\mathbf{x}$ its *length* $\|\mathbf{x}\| \in \mathbb{R}$, such that:

- $\|\lambda \mathbf{x}\| = |\lambda| \, \|\mathbf{x}\|$
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
- $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$

Example 5. Manhattan & Euclidean distance

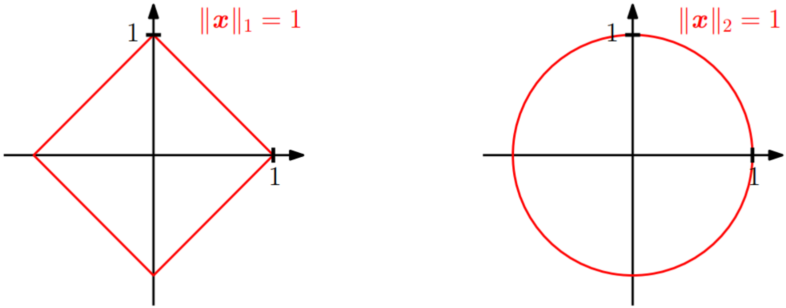


Figure: The red lines indicate the set of vectors with norm 1. Left: Manhattan norm (ℓ_1); Right: Euclidean norm (ℓ_2)

Dot product

A particular type of *Inner Product* is *dot product*:

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} \mathrm{x_i y_i}$$

Definition (bilinear mapping)

A mapping Ω is a *bilinear mapping* with two arguments and is linear in each argument:

$$\Omega(\lambda \mathbf{x} + \psi \mathbf{y}, \mathbf{z}) = \lambda \Omega(\mathbf{x}, \mathbf{z}) + \psi \Omega(\mathbf{y}, \mathbf{z})$$
$$\Omega(\mathbf{x}, \lambda \mathbf{y} + \psi \mathbf{z}) = \lambda \Omega(\mathbf{x}, \mathbf{y}) + \psi \Omega(\mathbf{x}, \mathbf{z})$$

Definition (Inner Product)

For a vector space V and a bilinear mapping $\Omega : \mathrm{V} \times \mathrm{V} \rightarrow \mathbb{R}$,

- Ω is *symmetric* if $\Omega(\mathbf{x}, \mathbf{y}) = \Omega(\mathbf{y}, \mathbf{x})$
- Ω is *positive definite* if

$$\forall \mathbf{x} \in \mathrm{V} \backslash \{\mathbf{0}\} : \Omega(\mathbf{x}, \mathbf{x}) > 0, \quad \Omega(\mathbf{0}, \mathbf{0}) = 0$$

- *Inner product* $\langle \mathbf{x}, \mathbf{y} \rangle$: A positive, symmetric bilinear mapping $\Omega : \mathrm{V} \times \mathrm{V} \rightarrow \mathbb{R}$
- *Inner product space*: The pair $(\mathrm{V}, \langle \cdot, \cdot \rangle)$
  - For dot product, $(\mathrm{V}, \langle \cdot, \cdot \rangle)$ is a *Euclidean vector space*

Due to the bilinearity of the inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{i=1}^{n} \psi_i \mathbf{b}_j, \sum_{j=1}^{n} \lambda_i \mathbf{b}_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_i \langle \mathbf{b}_i, \mathbf{b}_j \rangle \lambda_j = \hat{\mathbf{x}}^\top \mathbf{A} \hat{\mathbf{y}},$$

where $A_{ij} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle$ and $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ are the coordinates of $\mathbf{x}$ and $\mathbf{y}$ w.r.t. the ordered basis B,

$$\forall \mathbf{x} \in V \backslash \{\mathbf{0}\} : \mathbf{x}^\top \mathbf{A} \mathbf{x} > 0 \qquad (*)$$

Definition (Symmetric, Positive Definite Matrix)

$\mathbf{A}$ that satisfies $(*)$ $\longrightarrow$ *symmetric, positive definite*

$\mathbf{A}$ that satisfies $(*)$ only if $\geq$ holds $\longrightarrow$ *symmetric, positive semidefinite*

Example 6. Symmetric, Positive Definite Matrices

$$\mathbf{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix}$$

$$\begin{aligned} \mathbf{x}^\top \mathbf{A}_1 \mathbf{x} &= \begin{bmatrix} \mathrm{x}_1 & \mathrm{x}_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \end{bmatrix} \\ &= 9\mathrm{x}_1^2 + 12\mathrm{x}_1\mathrm{x}_2 + 5\mathrm{x}_2^2 = (3\mathrm{x}_1 + 2\mathrm{x}_2)^2 + \mathrm{x}_2^2 > 0 \end{aligned}$$

$\Longrightarrow$ $\mathbf{A}_1$ is positive definite

However, $\mathbf{A}_2$ is symmetric but not positive definite (why?)

Theorem

For a real-valued, finite-dimensional vector space V*,* $\langle \cdot, \cdot \rangle : \mathrm{V} \times \mathrm{V} \to \mathbb{R}$ *is an inner product if and only if there exists a symmetric, positive definite matrix* $\mathbf{A} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ *with*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \hat{\mathbf{x}}^\top \mathbf{A} \hat{\mathbf{y}}$$

Thus, if $\mathbf{A} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ is symmetric and positive definite:

- The null space (kernel) of $\mathbf{A}$ consists only of $\mathbf{0} \implies \mathbf{A}\mathbf{x} \neq \mathbf{0}$ if $\mathbf{x} \neq \mathbf{0}$
- Diagonal elements a_{ii} of $\mathbf{A}$ are positive

Definition (Distance and Metric)

For an inner product space $(\mathrm{V}, \langle\cdot,\cdot\rangle)$,

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

is the *distance* between $\mathbf{x}$ and $\mathbf{y}$

The mapping

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

is a *metric*

A metric d satisfies:

- d is positive definite
- d is symmetric
- $\mathrm{d}(\mathbf{x},\mathbf{z}) \leq \mathrm{d}(\mathbf{x},\mathbf{y}) + \mathrm{d}(\mathbf{y},\mathbf{z})$

Angle between vectors
Example 7.

The angle between $\mathbf{x} = \begin{bmatrix}1 & 1\end{bmatrix}^\top \in \mathbb{R}^2$ and $\mathbf{y} = \begin{bmatrix}1 & 2\end{bmatrix}^\top \in \mathbb{R}^2$?

$$\cos\omega = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle\langle \mathbf{y}, \mathbf{y}\rangle}} = \frac{\mathbf{x}^\top\mathbf{y}}{\sqrt{\mathbf{x}^\top\mathbf{x}\mathbf{y}^\top\mathbf{y}}} = \frac{3}{\sqrt{10}}$$

$$\Longrightarrow \arccos\left(\frac{3}{\sqrt{10}}\right) \approx 0.32\,\text{rad}$$

Definition (Orthogonality)

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* ($\mathbf{x} \perp \mathbf{y}$) if and only if

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0$$

If additionally $\|\mathbf{x}\| = 1 = \|\mathbf{y}\|$, i.e., the vectors are unit vectors, then $\mathbf{x}$ and $\mathbf{y}$ are *orthonormal*

Definition (Orthogonal Matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is an *orthogonal matrix* if and only if its columns are orthonormal ($\mathbf{A}\mathbf{A}^\top = \mathbf{I} = \mathbf{A}^\top \mathbf{A}$), implying that

$$\mathbf{A}^{-1} = \mathbf{A}^\top$$

Transformations by Orthogonal Matrices

- Length of a vector $\mathbf{x}$ is not changed

$$\|\mathbf{A}\mathbf{x}\|^2 = (\mathbf{A}\mathbf{x})^\top(\mathbf{A}\mathbf{x}) = \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{I}\mathbf{x} = \mathbf{x}^\top\mathbf{x} = \|\mathbf{x}\|^2$$

- The angle between any two vectors $\mathbf{x}, \mathbf{y}$ is unchanged

$$\cos\omega = \frac{(\mathbf{A}\mathbf{x})^\top(\mathbf{A}\mathbf{y})}{\|\mathbf{A}\mathbf{x}\|\,\|\mathbf{A}\mathbf{y}\|} = \frac{\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y}}{\sqrt{\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}\mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y}}} = \frac{\mathbf{x}^\top\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$$

Definition (Orthonormal Basis)

For an *n*-dimensional vector space V and a basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ of V, the basis is an *orthonormal basis* if

$$\langle \mathbf{b}_\mathrm{i}, \mathbf{b}_\mathrm{j} \rangle = 0 \quad \text{for } \mathrm{i} \neq \mathrm{j}$$
$$\langle \mathbf{b}_\mathrm{i}, \mathbf{b}_\mathrm{i} \rangle = 1$$

Example 8. Orthonormal Basis

The canonical/standard basis for a Euclidean vector space $\mathbb{R}^n$ is an orthonormal basis, where the inner product is the dot product of vectors. In $\mathbb{R}^2$, the vectors

$$\mathbf{b}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \quad \mathbf{b}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$$

form an orthonormal basis. (Why?)

Orthogonal Complement

- Generally used to describe hyperplanes in n-dimensional vector and affine spaces (important in *linear dimensionality reduction*
- For a D-dimensional vector space V and an M-dimensional subspace $U \subseteq V$, its *orthogonal complement* $U^{\perp}$ is a $(D-M)$-dimensional subspace of V
- Contains all vectors in V that are orthogonal to every vector in U
- Since $U \cap U^{\perp} = \{\mathbf{0}\}$, any vector $\mathbf{x} \in V$ can be uniquely decomposed into

$$\mathbf{x} = \sum_{m=1}^{M} \lambda_m \mathbf{b}_m + \sum_{j}^{D-M} \psi_j \mathbf{b}_j^{\perp}$$

- The vector ω with $\|\omega\| = 1$, which is orthogonal to a 2D subspace U, is the basis vector of $U^{\perp}$ (see next slide)
- The vector ω is the *normal vector* of U

Mathematical Foundations: Linear Algebra

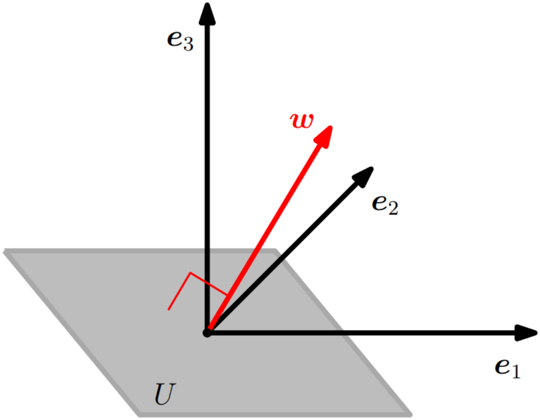


Figure: A plane U in a three-dimensional vector space can be described by its normal vector, which spans its orthogonal complement $\mathrm{U}^{\perp}$

Proejections

- An important class of linear transformations
- Used to represent the original high-dimensional data onto a lower-dimensional feature space
- A fundamental mathematical tool in *data compression* tasks
- To retain as much information as possible is to minimize the difference/error between the original high-dimensional data and the projected lower-dimensional subspace (illustrated in the next slide)

Mathematical Foundations: Linear Algebra

An example of an **Orthogonal Projection**

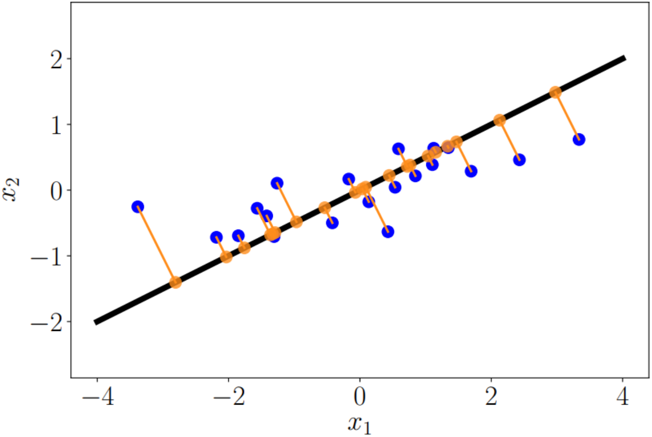


Figure: Orthogonal projection (orange dots) of a two-dimensional dataset (blue dots) onto a one-dimensional subspace (straight line)

Definition (Projection)

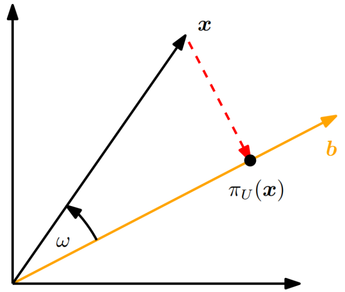
For a vector space V and $\mathrm{U} \subseteq \mathrm{V}$ a subspace of V, a linear mapping $\pi : \mathrm{V} \to \mathrm{U}$ is called a *projection* if

$$\pi^2 = \pi \circ \pi = \pi$$

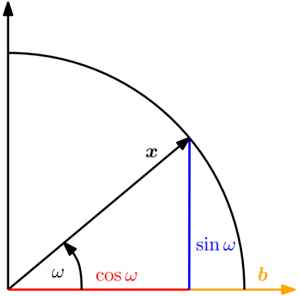
The preceding definition applies to a special kind of transformation matrices, the *projection matrices* $\mathbf{P}_\pi$, which exhibits the property that $\mathbf{P}_\pi^2 = \mathbf{P}_\pi$.

Projection onto 1D Subspaces (Lines)

By projecting $\mathbf{x} \in \mathbb{R}^n$ onto U, the vector $\pi_{\mathrm{U}}(\mathbf{x}) \in \mathrm{U}$ that is closest to $\mathbf{x}$ is sought.



(a) Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis vector $\boldsymbol{b}$.



(b) Projection of a two-dimensional vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ onto a one-dimensional subspace spanned by $\boldsymbol{b}$.

Figure: Examples of projections onto 1D subspaces

Projection onto 1D Subspaces (Lines)

1. Finding the coordinate λ (note that $\pi_{\mathrm{U}}(\mathbf{x}) = \lambda\mathbf{b}$)

$$\lambda = \frac{\mathbf{b}^\top \mathbf{x}}{\mathbf{b}^\top \mathbf{b}} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2}$$

2. Finding the projection point $\pi_{\mathrm{U}}(\mathbf{x}) \in \mathrm{U}$

$$\pi_{\mathrm{U}}(\mathbf{x}) = \lambda\mathbf{b} = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\|\mathbf{b}\|^2}\mathbf{b} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2}\mathbf{b},$$

$$\|\pi_{\mathrm{U}}(\mathbf{x})\| = |\cos\omega|\,\|\mathbf{x}\|$$

3. Finding the projection matrix $\mathbf{P}_\pi$

$$\mathbf{P}_\pi = \frac{\mathbf{b}\mathbf{b}^\top}{\|\mathbf{b}\|^2}$$

Example 9. Projection onto a Line
Find the projection matrix onto the line through the origin spanned by $\mathbf{b} = \begin{bmatrix}1 & 2 & 2\end{bmatrix}^\top$.

$$\mathbf{P}_\pi = \frac{\mathbf{b}\mathbf{b}^\top}{\mathbf{b}^\top\mathbf{b}} = \frac{1}{9}\begin{bmatrix}1\\2\\2\end{bmatrix}\begin{bmatrix}1 & 2 & 2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}$$

$$\pi_{\mathrm{U}}(\mathbf{x}) = \mathbf{P}_\pi\mathbf{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in \mathrm{span}[\begin{bmatrix}1\\2\\2\end{bmatrix}]$$

We can further show that $\pi_{\mathrm{U}}(\mathbf{x})$ is an *eigenvector* of $\mathbf{P}_\pi$, and the corresponding *eigenvaue* is 1

Projection onto a 2D subspace

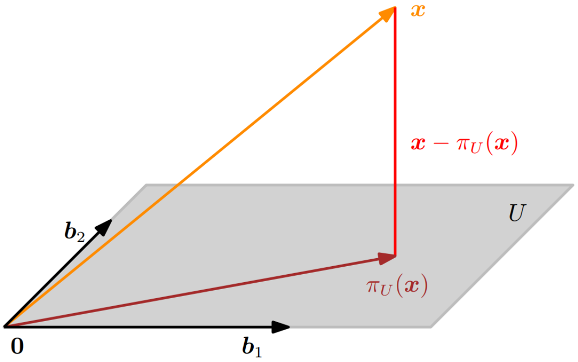


Figure: Projection onto a 2D subspace U with basis $\mathbf{b}_1$ and $\mathbf{b}_2$. The projection $\pi_{\mathrm{U}}(\mathbf{x})$ of $\mathbf{x} \in \mathbb{R}^3$ onto U can be expressed as a linear combination of $\mathbf{b}_1$, $\mathbf{b}_2$ and the displacement vector $\mathbf{x} - \pi_{\mathrm{U}}(\mathbf{x}$ is orthogonal to both $\mathbf{b}_1$ and $\mathbf{b}_2$

Projection onto General Subspaces

1. Finding the coordinate λ (note that $\pi_{\mathrm{U}}(\mathbf{x}) = \lambda \mathbf{b}$)

$$\lambda = \frac{\mathbf{b}^\top \mathbf{x}}{\mathbf{b}^\top \mathbf{b}} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2}$$

2. Finding the projection point $\pi_{\mathrm{U}}(\mathbf{x}) \in \mathrm{U}$

$$\pi_{\mathrm{U}}(\mathbf{x}) = \lambda \mathbf{b} = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\|\mathbf{b}\|^2} \mathbf{b} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2} \mathbf{b},$$

$$\|\pi_{\mathrm{U}}(\mathbf{x})\| = |\cos \omega| \, \|\mathbf{x}\|$$

3. Finding the projection matrix $\mathbf{P}_\pi$

$$\mathbf{P}_\pi = \frac{\mathbf{b}\mathbf{b}^\top}{\|\mathbf{b}\|^2}$$

Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong.
Mathematics for machine learning.
Cambridge University Press, Cambridge and New York NY, 2020.