Data Science for Mathematicians

Lecture 1: The Data Science Landscape & The Geometry of Data

Course Instructor

November 27, 2025

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.
- Your rigorous training in linear algebra, calculus, and probability theory is the very essence of data science.

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.
- Your rigorous training in linear algebra, calculus, and probability theory is the very essence of data science.
- **Central Idea:** We must learn to see data not as numbers in a spreadsheet, but as a geometric object.

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.
- Your rigorous training in linear algebra, calculus, and probability theory is the very essence of data science.
- **Central Idea:** We must learn to see data not as numbers in a spreadsheet, but as a geometric object.
- An observation (a row in a table) is a **point** in a high-dimensional space.

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.
- Your rigorous training in linear algebra, calculus, and probability theory is the very essence of data science.
- **Central Idea:** We must learn to see data not as numbers in a spreadsheet, but as a geometric object.
- An observation (a row in a table) is a **point** in a high-dimensional space.
- A feature (a column) corresponds to a **dimension** of that space.

The Core Conceptual Leap: Data as Geometry

- The primary objective of this course is to bridge the perceived gap between abstract mathematics and the practical field of data science.
- Your rigorous training in linear algebra, calculus, and probability theory is the very essence of data science.
- **Central Idea:** We must learn to see data not as numbers in a spreadsheet, but as a geometric object.
- An observation (a row in a table) is a **point** in a high-dimensional space.
- A feature (a column) corresponds to a **dimension** of that space.
- The entire dataset is a **cloud of points**, and its structure (shape, distances) contains the patterns we seek.

Defining Data Science

Definition (Data Science)

Data Science is an interdisciplinary field that employs scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Defining Data Science

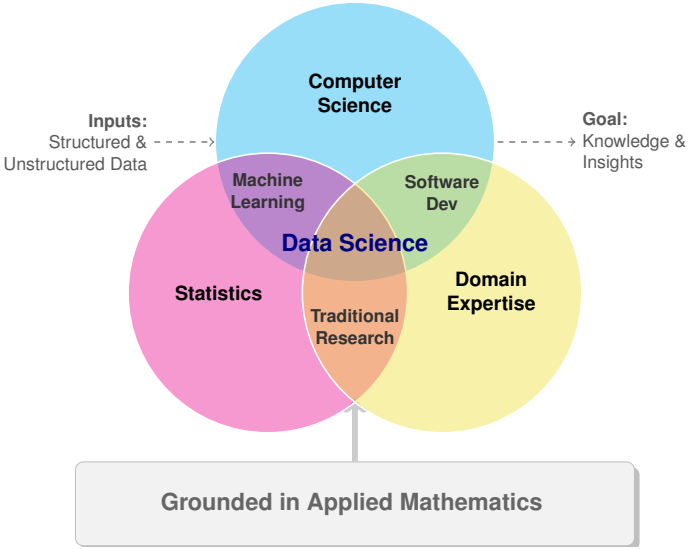
Definition (Data Science)

Data Science is an interdisciplinary field that employs scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It lies at the intersection of statistics, computer science, and domain expertise, and is fundamentally grounded in the principles of applied mathematics.

Defining Data Science

Definition (Data Science)

Data Science is an interdisciplinary field that employs scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It lies at the intersection of statistics, computer science, and domain expertise, and is fundamentally grounded in the principles of applied mathematics.


Inputs:
Structured &
Unstructured Data
Computer
Science
Goal:
Knowledge &
Insights
Machine
Learning
Software
Dev
Data Science
Statistics
Domain
Expertise
Traditional
Research
Grounded in Applied Mathematics

Data Science vs. Deterministic Calculation

The key differentiator is the use of modeling to predict or understand a complex, **uncertain (stochastic) system**.

Data Science vs. Deterministic Calculation

The key differentiator is the use of modeling to predict or understand a complex, **uncertain (stochastic) system**.

Example (Non-Data Science: Payroll Calculation)

A system calculates monthly salary (hours $\times$ rate $-$ fixed tax percentage).

- **Reason:** This is a classical **deterministic** calculation. The relationship is fixed by precise, unvarying arithmetic rules. The system is not learning or dealing with uncertainty.

Data Science vs. Deterministic Calculation

The key differentiator is the use of modeling to predict or understand a complex, **uncertain (stochastic) system**.

Example (Non-Data Science: Payroll Calculation)

A system calculates monthly salary (hours $\times$ rate $-$ fixed tax percentage).

- **Reason:** This is a classical **deterministic** calculation. The relationship is fixed by precise, unvarying arithmetic rules. The system is not learning or dealing with uncertainty.

Example (Data Science: Streaming Recommendations)

A platform uses user history and metadata to **predict** future viewing preferences.

- **Reason:** This treats user preference as a **stochastic** system and employs modern applied mathematics (e.g., matrix factorization) to build a predictive model.

The Canonical Data Science Workflow (Step-by-Step)

1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").

The Canonical Data Science Workflow (Step-by-Step)

1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").
2. **Data Acquisition and Cleaning**: Gather, handle missing values, correct errors, and reconcile inconsistencies (data wrangling).

The Canonical Data Science Workflow (Step-by-Step)

1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").
2. **Data Acquisition and Cleaning**: Gather, handle missing values, correct errors, and reconcile inconsistencies (data wrangling).
3. **Exploratory Data Analysis (EDA)**: Investigate structure through visualization and summary statistics to form initial hypotheses.

The Canonical Data Science Workflow (Step-by-Step)

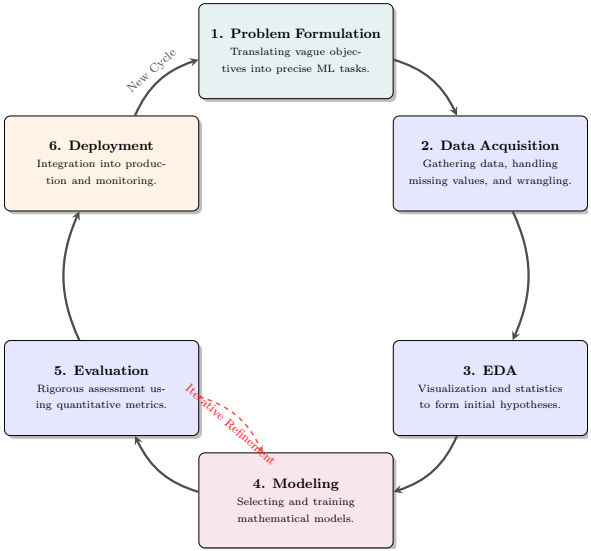
1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").
2. **Data Acquisition and Cleaning**: Gather, handle missing values, correct errors, and reconcile inconsistencies (data wrangling).
3. **Exploratory Data Analysis (EDA)**: Investigate structure through visualization and summary statistics to form initial hypotheses.
4. **Modeling**: Select and train a mathematical model (statistical or machine learning algorithm).

The Canonical Data Science Workflow (Step-by-Step)

1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").
2. **Data Acquisition and Cleaning**: Gather, handle missing values, correct errors, and reconcile inconsistencies (data wrangling).
3. **Exploratory Data Analysis (EDA)**: Investigate structure through visualization and summary statistics to form initial hypotheses.
4. **Modeling**: Select and train a mathematical model (statistical or machine learning algorithm).
5. **Evaluation**: Rigorously assess performance using quantitative metrics (how well it generalizes to new data).

The Canonical Data Science Workflow (Step-by-Step)

1. **Problem Formulation & Business Understanding**: Translate a vague goal (e.g., "reduce customer churn") into a precise, quantifiable mathematical task (e.g., "build a binary classifier").
2. **Data Acquisition and Cleaning**: Gather, handle missing values, correct errors, and reconcile inconsistencies (data wrangling).
3. **Exploratory Data Analysis (EDA)**: Investigate structure through visualization and summary statistics to form initial hypotheses.
4. **Modeling**: Select and train a mathematical model (statistical or machine learning algorithm).
5. **Evaluation**: Rigorously assess performance using quantitative metrics (how well it generalizes to new data).
6. **Deployment & Monitoring**: Integrate the model into production and continuously track its performance.


1. Problem Formulation
Translating vague objectives into precise ML tasks.
2. Data Acquisition
Gathering data, handling missing values, and wrangling.
3. EDA
Visualization and statistics to form initial hypotheses.
4. Modeling
Selecting and training mathematical models.
5. Evaluation
Rigorous assessment using quantitative metrics.
6. Deployment
Integration into production and monitoring.
Iterative Refinement
New Cycle

The Mathematician's Role: Architect of Logic

The mathematician's indispensable value is concentrated in the stages demanding abstraction and logical rigor: Formulation, Modeling, and Evaluation.

The Mathematician's Role: Architect of Logic

The mathematician's indispensable value is concentrated in the stages demanding abstraction and logical rigor: Formulation, Modeling, and Evaluation.

- **Problem Formulation as Axiomatization**
  - A practitioner might test algorithms empirically.
  - A mathematician formalizes the problem: Is it classification, regression, or clustering? What loss function captures the penalty for error? The framework choice dictates the entire logical structure.

The Mathematician's Role: Architect of Logic

The mathematician's indispensable value is concentrated in the stages demanding abstraction and logical rigor: Formulation, Modeling, and Evaluation.

- **Problem Formulation as Axiomatization**
  - A practitioner might test algorithms empirically.
  - A mathematician formalizes the problem: Is it classification, regression, or clustering? What loss function captures the penalty for error? The framework choice dictates the entire logical structure.
- **Model Selection as Theorem Proving**
  - Choosing a model is equivalent to stating a theorem: "Assuming the data follows properties A, B, and C, then algorithm X is optimal.".
  - The skill is selecting a model whose assumptions are mathematically justifiable.

The Mathematician's Role: Architect of Logic

The mathematician's indispensable value is concentrated in the stages demanding abstraction and logical rigor: Formulation, Modeling, and Evaluation.

- **Problem Formulation as Axiomatization**
  - A practitioner might test algorithms empirically.
  - A mathematician formalizes the problem: Is it classification, regression, or clustering? What loss function captures the penalty for error? The framework choice dictates the entire logical structure.
- **Model Selection as Theorem Proving**
  - Choosing a model is equivalent to stating a theorem: "Assuming the data follows properties A, B, and C, then algorithm X is optimal.".
  - The skill is selecting a model whose assumptions are mathematically justifiable.
- **Critical Analysis as Peer Review**
  - The mathematician relentlessly questions the model's assumptions (e.g., are features independent, is variance constant?).
  - This critical posture ensures the logical argument holds from premise to conclusion, preventing the misuse of black-box algorithms.

Example (Medical insurance costs)

An observation corresponds to a single individual. We consider four numerical features:

Example (Medical insurance costs)

An observation corresponds to a single individual. We consider four numerical features:

- **Age** (years)
- **BMI** (Body Mass Index)
- **Children** (number of dependents)
- **Charges** ($) (variable we wish to predict)

Example (Medical insurance costs)

An observation corresponds to a single individual. We consider four numerical features:

- **Age** (years)
- **BMI** (Body Mass Index)
- **Children** (number of dependents)
- **Charges** ($) (variable we wish to predict)

Observation	**Age**	**BMI**	**Children**	**Charges ($)**
Person 1 (x_1)	19	27.9	0	16884.92
Person 2 (x_2)	35	35.5	1	44501.40

Step 1: Observations as Vectors

The critical step is to view each row as a single, unified mathematical object: a vector.

Step 1: Observations as Vectors

The critical step is to view each row as a single, unified mathematical object: a vector.

Definition (Observation Vector)

An observation consisting of p numerical features is represented as a **vector** $x \in \mathbb{R}^p$. Each component x_j corresponds to the value of the j-th feature.

Step 1: Observations as Vectors

The critical step is to view each row as a single, unified mathematical object: a vector.

Definition (Observation Vector)

An observation consisting of p numerical features is represented as a **vector** $x \in \mathbb{R}^p$. Each component x_j corresponds to the value of the j-th feature.

Example (Person 1 as a Vector in $\mathbb{R}^4$)

The observation for Person 1 (Age=19, BMI=27.9, Children=0, Charges=16884.92) is formalized as:

$$x_1 = \begin{bmatrix} 19 \\ 27.9 \\ 0 \\ 16884.92 \end{bmatrix} \in \mathbb{R}^4.$$

Step 1: Observations as Vectors

The critical step is to view each row as a single, unified mathematical object: a vector.

Definition (Observation Vector)

An observation consisting of p numerical features is represented as a **vector** $x \in \mathbb{R}^p$. Each component x_j corresponds to the value of the j-th feature.

Example (Person 1 as a Vector in $\mathbb{R}^4$)

The observation for Person 1 (Age=19, BMI=27.9, Children=0, Charges=16884.92) is formalized as:

$$x_1 = \begin{bmatrix} 19 \\ 27.9 \\ 0 \\ 16884.92 \end{bmatrix} \in \mathbb{R}^4.$$

This vector resides in a 4-dimensional Euclidean space, where each axis corresponds to a measured feature.

Step 2: The Feature Space

Definition (Feature Space)

The p-dimensional vector space $\mathbb{R}^p$, in which each dimension corresponds to a feature and each point corresponds to a possible observation, is called the **feature space**.

Step 2: The Feature Space

Definition (Feature Space)

The p-dimensional vector space $\mathbb{R}^p$, in which each dimension corresponds to a feature and each point corresponds to a possible observation, is called the **feature space**. A dataset of n observations $\{x_1, \ldots, x_n\}$ is a **point cloud** within this space.

Step 2: The Feature Space

Definition (Feature Space)

The p-dimensional vector space $\mathbb{R}^p$, in which each dimension corresponds to a feature and each point corresponds to a possible observation, is called the **feature space**. A dataset of n observations $\{x_1, \ldots, x_n\}$ is a **point cloud** within this space.

- **Clusters** (similar customers) correspond to dense regions of points.

Step 2: The Feature Space

Definition (Feature Space)

The p-dimensional vector space $\mathbb{R}^p$, in which each dimension corresponds to a feature and each point corresponds to a possible observation, is called the **feature space**. A dataset of n observations $\{x_1, \ldots, x_n\}$ is a **point cloud** within this space.

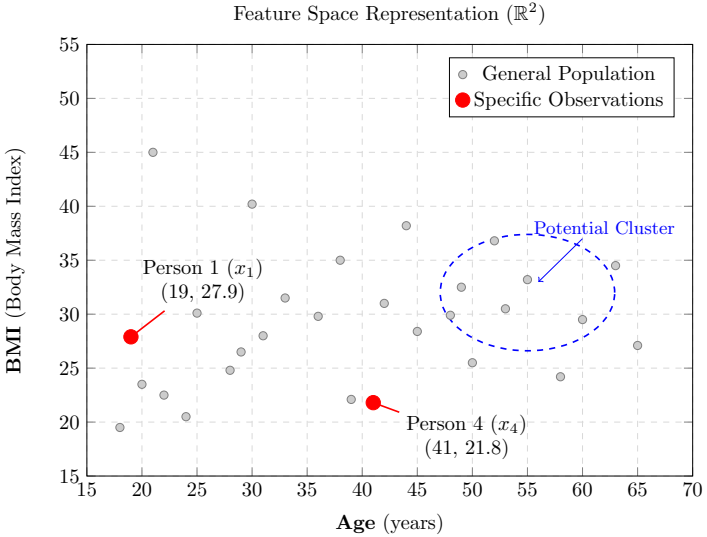
- **Clusters** (similar customers) correspond to dense regions of points.
- **Trends** (e.g., charges increasing with age) correspond to the point cloud being elongated or oriented along certain directions.

Step 2: The Feature Space

Definition (Feature Space)

The p-dimensional vector space $\mathbb{R}^p$, in which each dimension corresponds to a feature and each point corresponds to a possible observation, is called the **feature space**. A dataset of n observations $\{x_1, \ldots, x_n\}$ is a **point cloud** within this space.

- **Clusters** (similar customers) correspond to dense regions of points.
- **Trends** (e.g., charges increasing with age) correspond to the point cloud being elongated or oriented along certain directions.
- **Outliers** are points that lie far away from the main body.


Feature Space Representation ($\mathbb{R}^2$)
General Population
Specific Observations
Potential Cluster
Person 1 (x_1)
(19, 27.9)
Person 4 (x_4)
(41, 21.8)
BMI (Body Mass Index)
Age (years)
55
50
45
40
35
30
25
20
15
15
20
25
30
35
40
45
50
55
60
65
70

Step 3: The Data Matrix (Central Object)

Definition (Data Matrix)

A dataset of n observations, each with p features, is represented by a **data matrix** $X \in \mathbb{R}^{n \times p}$.

Step 3: The Data Matrix (Central Object)

Definition (Data Matrix)

A dataset of n observations, each with p features, is represented by a **data matrix** $X \in \mathbb{R}^{n \times p}$.

- Each **row** of X is the transpose of an observation vector, ${x_i}^T$.
- Each **column** of X is a vector $X_{:j} \in \mathbb{R}^n$ containing all values for a single j-th feature.

Example (Design Matrix for Medical Insurance)

For $n = 5$ observations and $p = 4$ features, the matrix $X \in \mathbb{R}^{5\times 4}$ is:

$$X = \begin{bmatrix} 19 & 27.9 & 0 & 16884.92 \\ 35 & 35.5 & 1 & 44501.40 \\ 62 & 26.3 & 0 & 27808.72 \\ 41 & 21.8 & 1 & 6272.48 \\ 25 & 42.1 & 2 & 48824.45 \end{bmatrix}$$

The entry $x_{2,2} = 35.5$ is the BMI of the second individual.

Remark: The Curse of Dimensionality

Remark

As the number of features p increases, the volume of the feature space grows exponentially.

Remark: The Curse of Dimensionality

Remark

As the number of features p increases, the volume of the feature space grows exponentially. For a fixed number of data points n, the space becomes sparse (empty). This phenomenon, known as the **curse of dimensionality**, *makes finding meaningful patterns statistically difficult, as distances between points become less informative.*

The Dot Product as a Measure of Similarity

Definition (The Dot Product)

Let $u, v \in \mathbb{R}^p$. The **dot product** (or inner product), $u \cdot v$, is defined as the sum of the products of their corresponding components:

$$u \cdot v = u^T v = \sum_{i=1}^{p} u_i v_i$$

The Dot Product as a Measure of Similarity

Definition (The Dot Product)

Let $u, v \in \mathbb{R}^p$. The **dot product** (or inner product), $u \cdot v$, is defined as the sum of the products of their corresponding components:

$$u \cdot v = u^T v = \sum_{i=1}^{p} u_i v_i$$

Example (The Dot Product as a Weighted Sum)

In linear models, the dot product aggregates contributions:

- Patient features: $x = [\text{Age}, \text{BMI}]^T = ^T$.
- Model weights: $w = [0.5, 2.0]^T$.

The Dot Product as a Measure of Similarity

Definition (The Dot Product)

Let $u, v \in \mathbb{R}^p$. The **dot product** (or inner product), $u \cdot v$, is defined as the sum of the products of their corresponding components:

$$u \cdot v = u^T v = \sum_{i=1}^{p} u_i v_i$$

Example (The Dot Product as a Weighted Sum)

In linear models, the dot product aggregates contributions:

- Patient features: $x = [\text{Age}, \text{BMI}]^T = ^T$.
- Model weights: $w = [0.5, 2.0]^T$.
- Score $= w \cdot x = (0.5)(40) + (2.0)(30) = 20 + 60 = 80$.

The dot product accumulates the contribution of each feature weighted by its importance (w).

Theorem: Geometric Interpretation of the Dot Product

Theorem (Geometric Interpretation)

Let u and v be non-zero vectors in $\mathbb{R}^p$, and let $\theta \in [0, \pi]$ be the angle between them. Then:

$$u \cdot v = \|u\|_2 \|v\|_2 \cos\theta$$

where $\|x\|_2$ is the Euclidean norm (length) of x.

Proof Sketch (using the Law of Cosines).

Step 1: Law of Cosines Consider the triangle formed by u, v, and $u - v$:

$$\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta.$$

Proof Sketch (using the Law of Cosines).

Step 1: Law of Cosines Consider the triangle formed by u, v, and $u - v$:

$$\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta.$$

Step 2: Algebraic Expansion Expand the squared norm using dot product properties:

$$\begin{aligned}\|u - v\|_2^2 &= (u - v) \cdot (u - v) \\ &= u \cdot u - 2(u \cdot v) + v \cdot v \\ &= \|u\|_2^2 - 2(u \cdot v) + \|v\|_2^2.\end{aligned}$$

Proof Sketch (using the Law of Cosines).

Step 1: Law of Cosines Consider the triangle formed by u, v, and $u - v$:

$$\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta.$$

Step 2: Algebraic Expansion Expand the squared norm using dot product properties:

$$\begin{aligned}\|u - v\|_2^2 &= (u - v) \cdot (u - v) \\ &= u \cdot u - 2(u \cdot v) + v \cdot v \\ &= \|u\|_2^2 - 2(u \cdot v) + \|v\|_2^2.\end{aligned}$$

Step 3: Equating Expressions Equate the results from Steps 1 and 2:

$$\|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta = \|u\|_2^2 - 2(u \cdot v) + \|v\|_2^2.$$

Proof Sketch (using the Law of Cosines).

Step 1: Law of Cosines Consider the triangle formed by u, v, and $u - v$:

$$\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta.$$

Step 2: Algebraic Expansion Expand the squared norm using dot product properties:

$$\begin{aligned}\|u - v\|_2^2 &= (u - v) \cdot (u - v) \\ &= u \cdot u - 2(u \cdot v) + v \cdot v \\ &= \|u\|_2^2 - 2(u \cdot v) + \|v\|_2^2.\end{aligned}$$

Step 3: Equating Expressions Equate the results from Steps 1 and 2:

$$\|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos\theta = \|u\|_2^2 - 2(u \cdot v) + \|v\|_2^2.$$

Step 4: Conclusion Subtracting like terms and dividing by -2 yields the desired result:

$$u \cdot v = \|u\|_2\|v\|_2 \cos\theta.$$

Interpreting the Dot Product Geometrically

The dot product is a measure of **alignment** between data vectors.

Interpreting the Dot Product Geometrically

The dot product is a measure of **alignment** between data vectors.

- If $u \cdot v > 0$: $\cos\theta > 0$. The angle θ is **acute** ($0 \leq \theta < \pi/2$). The vectors point in a generally similar direction (positive correlation).

Interpreting the Dot Product Geometrically

The dot product is a measure of **alignment** between data vectors.

- If $u \cdot v > 0$: $\cos\theta > 0$. The angle θ is **acute** ($0 \leq \theta < \pi/2$). The vectors point in a generally similar direction (positive correlation).
- If $u \cdot v < 0$: $\cos\theta < 0$. The angle θ is **obtuse** ($\pi/2 < \theta \leq \pi$). The vectors point in opposite directions (negative correlation).

Interpreting the Dot Product Geometrically

The dot product is a measure of **alignment** between data vectors.

- If $u \cdot v > 0$: $\cos\theta > 0$. The angle θ is **acute** ($0 \leq \theta < \pi/2$). The vectors point in a generally similar direction (positive correlation).
- If $u \cdot v < 0$: $\cos\theta < 0$. The angle θ is **obtuse** ($\pi/2 < \theta \leq \pi$). The vectors point in opposite directions (negative correlation).
- If $u \cdot v = 0$: $\cos\theta = 0$. The angle $\theta = \pi/2$. The vectors are **orthogonal** (unrelated/uncorrelated in a linear sense).

Example (Collaborative Filtering)

Consider a movie recommendation system where user preferences are represented as *mean-centered ratings* (where 0 is neutral, positive is "like", and negative is "dislike").

- **User A (Sci-Fi fan)**: Loves *Star Wars* (+2) and *Dune* (+2).
- **User B (Sci-Fi fan)**: Loves *Star Wars* (+1) and *Dune* (+3).

$$u_A = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad u_B = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

$$u_A \cdot u_B = (2)(1) + (2)(3) = 2 + 6 = 8.$$

Since $u_A \cdot u_B = 8 > 0$, the angle is acute. The vectors point in the same general direction, indicating the users share a similar taste profile. The system should recommend movies User A likes to User B.

Example (Document Similarity)

In Natural Language Processing, documents can be represented as vectors of word counts. Let our vocabulary be: {"Data", "Algorithm", "Oven", "Flour"}.

- **Document 1 (Tech Article)**: "Data and Algorithm..." $\rightarrow [1, 1, 0, 0]^T$
- **Document 2 (Recipe)**: "Oven and Flour..." $\rightarrow [0, 0, 1, 1]^T$

The dot product calculation is:

$$d_1 \cdot d_2 = (1)(0) + (1)(0) + (0)(1) + (0)(1) = 0.$$

Interpretation: Since the dot product is exactly 0, the vectors are *orthogonal*. Geometrically, they are at 90°. In a data context, this means the documents share **no common vocabulary** and are likely about completely unrelated topics.

Step 4: Centering the Data

Step 4: Centering the Data

Definition (Mean Centering)

For a feature vector $u \in \mathbb{R}^n$ with sample mean $\bar{u}$, the **mean-centered vector**

$$\tilde{u} = u - \bar{u}\mathbf{1} = \begin{pmatrix} u_1 - \bar{u} \\ \vdots \\ u_n - \bar{u} \end{pmatrix}$$

Step 4: Centering the Data

Definition (Mean Centering)

For a feature vector $u \in \mathbb{R}^n$ with sample mean $\bar{u}$, the **mean-centered vector**

$$\tilde{u} = u - \bar{u}\mathbf{1} = \begin{pmatrix} u_1 - \bar{u} \\ \vdots \\ u_n - \bar{u} \end{pmatrix}$$

- **Purpose:** Centering removes the baseline bias, showing how observations vary *relative* to the average.

Step 4: Centering the Data

Definition (Mean Centering)

For a feature vector $u \in \mathbb{R}^n$ with sample mean $\bar{u}$, the **mean-centered vector**

$$\tilde{u} = u - \bar{u}\mathbf{1} = \begin{pmatrix} u_1 - \bar{u} \\ \vdots \\ u_n - \bar{u} \end{pmatrix}$$

- **Purpose:** Centering removes the baseline bias, showing how observations vary *relative* to the average.

Example (Centering a Vector)

If $u = [2, 4, 6, 8, 10]^T$, the mean $\bar{u} = 6$.

$$\tilde{u} = [-4, -2, 0, 2, 4]^T.$$

Example (Normalizing User Bias in Recommendations)

Different users have different rating scales (biases). Some users are "easy graders" (average rating 4.5/5), while others are "harsh critics" (average rating 2.5/5). Mean centering removes this bias. Suppose a user rates four movies on a scale of 1 to 5:

$$u_{\text{ratings}} = \begin{bmatrix} 5 \\ 5 \\ 3 \\ 3 \end{bmatrix}.$$

The user's average rating is $\bar{u} = 4$. The mean-centered vector is

$$\tilde{u}_{\text{centered}} = \begin{bmatrix} 5-4 \\ 5-4 \\ 3-4 \\ 3-4 \end{bmatrix} = \begin{bmatrix} +1 \\ +1 \\ -1 \\ -1 \end{bmatrix}.$$

In the centered vector, positive values $(+1)$ now clearly indicate movies the user liked *more*

Step 5: Defining Covariance

Definition (Sample Covariance)

Given two feature vectors u and v across n observations, the **sample covariance** measures their joint variability:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Step 5: Defining Covariance

Definition (Sample Covariance)

Given two feature vectors u and v across n observations, the **sample covariance** measures their joint variability:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Geometric Interpretation of Covariance:

- When u_i and v_i are **both above** average (Top-Right) or **both below** average (Bottom-Left), their product $(u_i - \bar{u})(v_i - \bar{v})$ is **positive**.

Step 5: Defining Covariance

Definition (Sample Covariance)

Given two feature vectors u and v across n observations, the **sample covariance** measures their joint variability:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Geometric Interpretation of Covariance:

- When u_i and v_i are **both above** average (Top-Right) or **both below** average (Bottom-Left), their product $(u_i - \bar{u})(v_i - \bar{v})$ is **positive**.
- When one is above and one is below (Top-Left or Bottom-Right), the product is **negative**.

Step 5: Defining Covariance

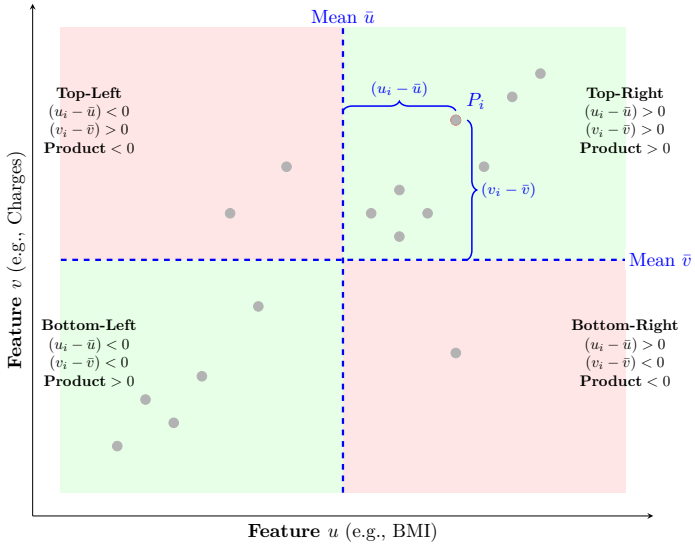
Definition (Sample Covariance)

Given two feature vectors u and v across n observations, the **sample covariance** measures their joint variability:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Geometric Interpretation of Covariance:

- When u_i and v_i are **both above** average (Top-Right) or **both below** average (Bottom-Left), their product $(u_i - \bar{u})(v_i - \bar{v})$ is **positive**.
- When one is above and one is below (Top-Left or Bottom-Right), the product is **negative**.
- Covariance is the average of these products. A positive covariance means the data mostly aligns in the positive quadrants (a positive trend).


Mean $\bar{u}$
Mean $\bar{v}$
$(u_i - \bar{u})$
P_i
$(v_i - \bar{v})$
Top-Left
$(u_i - \bar{u}) < 0$
$(v_i - \bar{v}) > 0$
Product < 0
Top-Right
$(u_i - \bar{u}) > 0$
$(v_i - \bar{v}) > 0$
Product > 0
Bottom-Left
$(u_i - \bar{u}) < 0$
$(v_i - \bar{v}) < 0$
Product > 0
Bottom-Right
$(u_i - \bar{u}) > 0$
$(v_i - \bar{v}) < 0$
Product < 0
Feature v (e.g., Charges)
Feature u (e.g., BMI)

Example

Consider a small dataset with $n = 3$ observations. Let $u = [1, 2, 3]^T$ and $v = [2, 4, 6]^T$.

1. **Compute Means:** $\bar{u} = 2$ and $\bar{v} = 4$.
2. **Compute Deviations:**
   - $u - \bar{u}\mathbf{1} = [-1, 0, 1]^T$
   - $v - \bar{v}\mathbf{1} = [-2, 0, 2]^T$
3. **Sum of Products:**

$$\sum_{i=1}^{3}(u_i - \bar{u})(v_i - \bar{v}) = (-1)(-2) + (0)(0) + (1)(2) = 2 + 0 + 2 = 4$$

4. **Normalize:**

$$\mathrm{Cov}(u, v) = \frac{1}{3-1}(4) = \frac{4}{2} = 2$$

The positive result confirms that as u increases, v also increases.

Theorem: Covariance as a Dot Product

This is a crucial unification: statistics is geometry in the centered space.

Theorem: Covariance as a Dot Product

This is a crucial unification: statistics is geometry in the centered space.

Theorem (Covariance as a Dot Product)

Let $\tilde{u}$ and $\tilde{v}$ be the mean-centered vectors for features u and v. The sample covariance is the scaled dot product of their centered vectors:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1}(\tilde{u} \cdot \tilde{v})$$

Theorem: Covariance as a Dot Product

This is a crucial unification: statistics is geometry in the centered space.

Theorem (Covariance as a Dot Product)

Let $\tilde{u}$ and $\tilde{v}$ be the mean-centered vectors for features u and v. The sample covariance is the scaled dot product of their centered vectors:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1}(\tilde{u} \cdot \tilde{v})$$

Theorem: Covariance as a Dot Product

This is a crucial unification: statistics is geometry in the centered space.

Theorem (Covariance as a Dot Product)

Let $\tilde{u}$ and $\tilde{v}$ be the mean-centered vectors for features u and v. The sample covariance is the scaled dot product of their centered vectors:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1}(\tilde{u} \cdot \tilde{v})$$

Consequence: Two variables are **uncorrelated** ($\mathrm{Cov} = 0$) if and only if their centered feature vectors are **orthogonal** ($\tilde{u} \cdot \tilde{v} = 0$).

Proof.

Step 1: Start with the Statistical Definition

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Proof.

Step 1: Start with the Statistical Definition

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Step 2: Relate Deviations to Centered Vectors By definition of mean centering, $\tilde{u}_i = u_i - \bar{u}$ and $\tilde{v}_i = v_i - \bar{v}$.

Proof.

Step 1: Start with the Statistical Definition

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Step 2: Relate Deviations to Centered Vectors By definition of mean centering, $\tilde{u}_i = u_i - \bar{u}$ and $\tilde{v}_i = v_i - \bar{v}$.

Step 3: Substitute using the Dot Product Definition:

$$\tilde{u} \cdot \tilde{v} = \sum_{i=1}^{n} \tilde{u}_i \tilde{v}_i = \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Proof.

Step 1: Start with the Statistical Definition

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Step 2: Relate Deviations to Centered Vectors By definition of mean centering, $\tilde{u}_i = u_i - \bar{u}$ and $\tilde{v}_i = v_i - \bar{v}$.

Step 3: Substitute using the Dot Product Definition:

$$\tilde{u} \cdot \tilde{v} = \sum_{i=1}^{n} \tilde{u}_i \tilde{v}_i = \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})$$

Step 4: Conclusion Substituting this back into the covariance formula yields:

$$\mathrm{Cov}(u, v) = \frac{1}{n-1} (\tilde{u} \cdot \tilde{v})$$

Definition of a Norm

Norms measure the magnitude, length, or size of a single vector.

Definition of a Norm

Norms measure the magnitude, length, or size of a single vector.

Definition (Norm)

A **norm** on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ satisfying four properties for all $x, y \in V$ and scalar α:

1. **Non-negativity:** $\|x\| \geq 0$

Definition of a Norm

Norms measure the magnitude, length, or size of a single vector.

Definition (Norm)

A **norm** on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ satisfying four properties for all $x, y \in V$ and scalar α:

1. **Non-negativity:** $\|x\| \geq 0$
2. **Positive Definiteness:** $\|x\| = 0 \iff x = \mathbf{0}$

Definition of a Norm

Norms measure the magnitude, length, or size of a single vector.

Definition (Norm)

A **norm** on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ satisfying four properties for all $x, y \in V$ and scalar α:

1. **Non-negativity:** $\|x\| \geq 0$
2. **Positive Definiteness:** $\|x\| = 0 \iff x = \mathbf{0}$
3. **Absolute Homogeneity:** $\|\alpha x\| = |\alpha|\|x\|$

Definition of a Norm

Norms measure the magnitude, length, or size of a single vector.

Definition (Norm)

A **norm** on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ satisfying four properties for all $x, y \in V$ and scalar α:

1. **Non-negativity:** $\|x\| \geq 0$
2. **Positive Definiteness:** $\|x\| = 0 \iff x = \mathbf{0}$
3. **Absolute Homogeneity:** $\|\alpha x\| = |\alpha|\|x\|$
4. **Triangle Inequality:** $\|x + y\| \leq \|x\| + \|y\|$

L_2 Norm (Euclidean Norm)

Example (L_2 Norm)

The L_2 norm of $x \in \mathbb{R}^n$ is the square root of the sum of squared components:

$$\|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2} = \sqrt{x^T x}.$$

L_2 Norm (Euclidean Norm)

Example (L_2 Norm)

The L_2 norm of $x \in \mathbb{R}^n$ is the square root of the sum of squared components:

$$\|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2} = \sqrt{x^T x}.$$

Definition (Euclidean Distance)

The Euclidean distance between two vectors x and y is given by the norm of their difference:

$$d_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}.$$

Verification of L_2 Norm Properties (1, 2 & 3).

1. **Non-negativity:** $\sum x_j^2 \geq 0$, so $\sqrt{\sum x_j^2} \geq 0$.

Verification of L_2 Norm Properties (1, 2 & 3).

1. **Non-negativity:** $\sum x_j^2 \geq 0$, so $\sqrt{\sum x_j^2} \geq 0$.
2. **Positive Definiteness:**
   - If $x = \mathbf{0}$, $\|x\|_2 = 0$.
   - If $\|x\|_2 = 0$, then $\sum x_j^2 = 0$. Since $x_j^2 \geq 0$, this requires $x_j = 0$ for all j, so $x = \mathbf{0}$.

Verification of L_2 Norm Properties (1, 2 & 3).

1. **Non-negativity:** $\sum x_j^2 \geq 0$, so $\sqrt{\sum x_j^2} \geq 0$.
2. **Positive Definiteness:**
   - If $x = \mathbf{0}$, $\|x\|_2 = 0$.
   - If $\|x\|_2 = 0$, then $\sum x_j^2 = 0$. Since $x_j^2 \geq 0$, this requires $x_j = 0$ for all j, so $x = \mathbf{0}$.
3. **Absolute Homogeneity**:

$$\begin{aligned} \|\alpha x\|_2 &= \sqrt{\sum (\alpha x_j)^2} \\ &= \sqrt{\alpha^2 \sum x_j^2} \\ &= |\alpha| \sqrt{\sum x_j^2} \\ &= |\alpha| \|x\|_2. \end{aligned}$$

Verification of L_2 Norm Properties (4).

4. Triangle Inequality (Relies on Cauchy-Schwarz Inequality):
Step 1: Square the Left Hand Side (LHS):

$$\|x + y\|_2^2 = \|x\|_2^2 + 2(x \cdot y) + \|y\|_2^2.$$

Verification of L_2 Norm Properties (4).

4. Triangle Inequality (Relies on Cauchy-Schwarz Inequality):
Step 1: Square the Left Hand Side (LHS):

$$\|x + y\|_2^2 = \|x\|_2^2 + 2(x \cdot y) + \|y\|_2^2.$$

Step 2: Apply Cauchy-Schwarz ($x \cdot y \leq \|x\|_2 \|y\|_2$):

$$\|x + y\|_2^2 \leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 = (\|x\|_2 + \|y\|_2)^2.$$

Verification of L_2 Norm Properties (4).

4. Triangle Inequality (Relies on Cauchy-Schwarz Inequality):
Step 1: Square the Left Hand Side (LHS):

$$\|x + y\|_2^2 = \|x\|_2^2 + 2(x \cdot y) + \|y\|_2^2.$$

Step 2: Apply Cauchy-Schwarz ($x \cdot y \leq \|x\|_2 \|y\|_2$):

$$\|x + y\|_2^2 \leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 = (\|x\|_2 + \|y\|_2)^2.$$

Step 3: Take the square root:

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2.$$

Example

In signal processing, the L_2 norm is often used to quantify the *energy* or strength of a discrete signal. Consider a short audio signal represented as a vector x with 3 time-steps:

$$x = \begin{bmatrix} 3 \\ -4 \\ 12 \end{bmatrix}$$

The *loudness* or magnitude of this signal is given by its Euclidean norm

$$\|x\|_2 = \sqrt{3^2 + (-4)^2 + 12^2} = \sqrt{9 + 16 + 144} = \sqrt{169} = 13.$$

If we were to normalize this signal (scale it to unit energy), we would divide the vector by this norm $\hat{x} = (1/13) \cdot x \approx [0.23, -0.31, 0.92]^T$.

Example

Suppose we are classifying fruit based on two features: $x_1 = \text{Weight (g)}$ and $x_2 = \text{Redness Index (1-10)}$.

- **Unknown Fruit (u)**: Weight = 150g, Redness = 8. $\rightarrow u = [150, 8]^T$.
- **Apple Prototype (a)**: Weight = 140g, Redness = 9. $\rightarrow a = [140, 9]^T$.
- **Banana Prototype (b)**: Weight = 100g, Redness = 1. $\rightarrow b = [100, 1]^T$.

$$
\begin{aligned}
d_2(u, a) &= \|u - a\|_2 \\
&= \sqrt{(150 - 140)^2 + (8 - 9)^2} \\
&= \sqrt{(10)^2 + (-1)^2} \\
&= \sqrt{101} \approx 10.05.
\end{aligned}
\qquad
\begin{aligned}
d_2(u, b) &= \|u - b\|_2 \\
&= \sqrt{(150 - 100)^2 + (8 - 1)^2} \\
&= \sqrt{(50)^2 + (7)^2} \\
&= \sqrt{2549} \approx 50.49.
\end{aligned}
$$

Since $d_2(u, a) < d_2(u, b)$, the algorithm classifies the unknown fruit as an Apple.

L_1 Norm (Manhattan Norm)

Example (L_1 Norm)

The L_1 norm (Manhattan/Taxicab norm) of $x \in \mathbb{R}^n$ is the sum of the absolute components:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

L_1 Norm (Manhattan Norm)

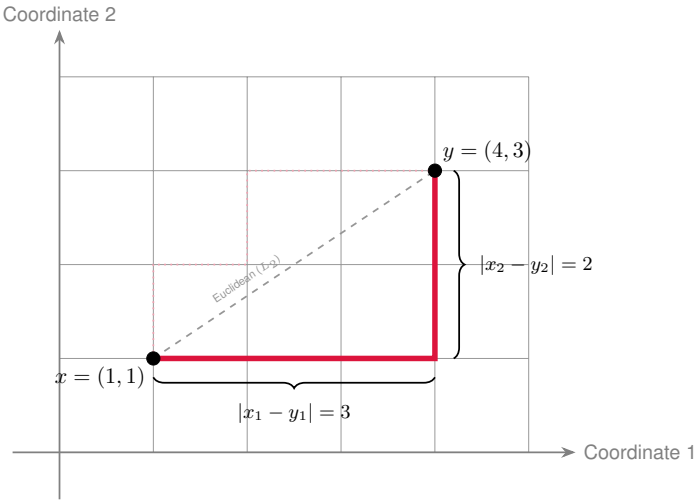
Example (L_1 Norm)

The L_1 norm (Manhattan/Taxicab norm) of $x \in \mathbb{R}^n$ is the sum of the absolute components:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

Definition (Manhattan distance)

The Manhattan distance between two vectors x and y is $d_1(x, y) = \|x - y\|_1$.


Coordinate 2
$y = (4, 3)$
$|x_2 - y_2| = 2$
Euclidean (L_2)
$x = (1, 1)$
$|x_1 - y_1| = 3$
Coordinate 1

Example

Robots in a warehouse move along a grid of aisles and shelves. They cannot move diagonally through the shelving units; they must move horizontally and vertically. Let the warehouse floor be a 2D grid.

- **Robot Position (r)**: Aisle 2, Bay 5. $\rightarrow r = [2, 5]^T$
- **Target Item (t)**: Aisle 6, Bay 1. $\rightarrow t = [6, 1]^T$

The Euclidean distance would assume a straight line through the shelves. The practical travel distance is the Manhattan distance

$$\begin{aligned} d_1(r, t) &= \|r - t\|_1 \\ &= |2 - 6| + |5 - 1| \\ &= |-4| + |4| \\ &= 4 + 4 = 8. \end{aligned}$$

The robot must travel 4 units east and 4 units south, for a total of 8 units.

The Geometry of Sparsity (Unit Balls)

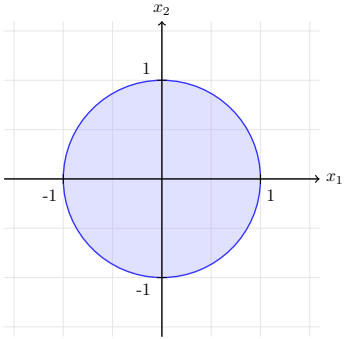
Definition (Unit Ball)

The unit ball for a norm $\|\cdot\|$ is the set $B = \{x \in V : \|x\| \leq 1\}$.

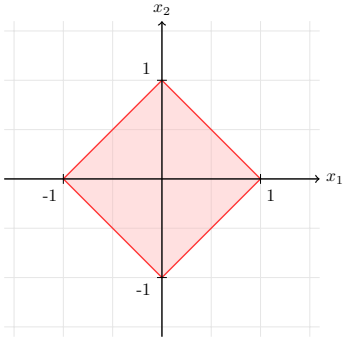
The Geometry of Sparsity (Unit Balls)

Definition (Unit Ball)

The unit ball for a norm $\|\cdot\|$ is the set $B = \{x \in V : \|x\| \leq 1\}$.



L_2 Unit Ball
(Euclidean)
$\sqrt{x_1^2 + x_2^2} \leq 1$



L_1 Unit Ball
(Manhattan)
$|x_1| + |x_2| \leq 1$

The L_p and L_∞ Norms

Example (L_p Norm)

For $p \geq 1$, the L_p norm generalizes L_1 and L_2:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The L_p and L_∞ Norms

Example (L_p Norm)

For $p \geq 1$, the L_p norm generalizes L_1 and L_2:

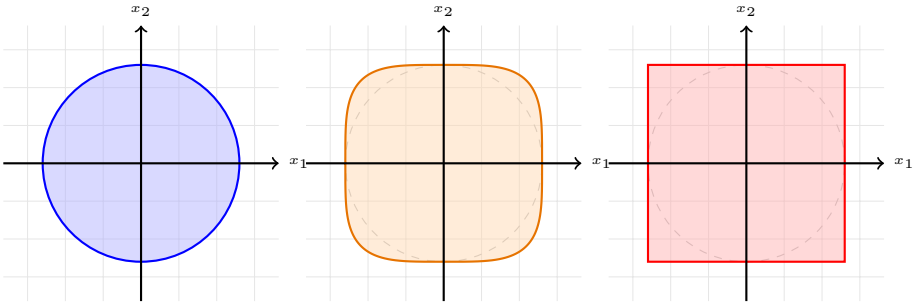
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Example (L_∞ Norm (Max Norm))

This is the limit as $p \to \infty$, defined as:

$$\|x\|_\infty = \max_{i=1,\ldots,p} |x_i|.$$

The L_∞ norm is concerned only with the single largest-magnitude feature.



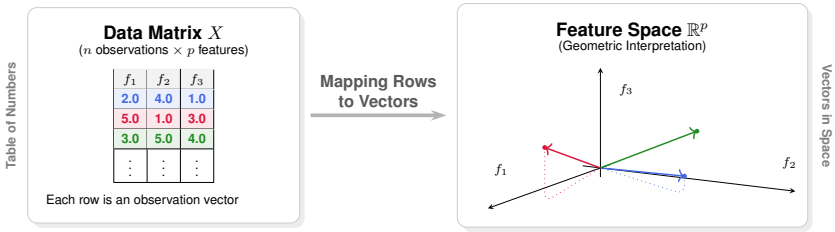
L_2 **Norm**
$\left(\sum x_i^2\right)^{1/2} \leq 1$

L_4 **Norm**
$\left(\sum x_i^4\right)^{1/4} \leq 1$

L_∞ **Norm**
$\max |x_i| \leq 1$

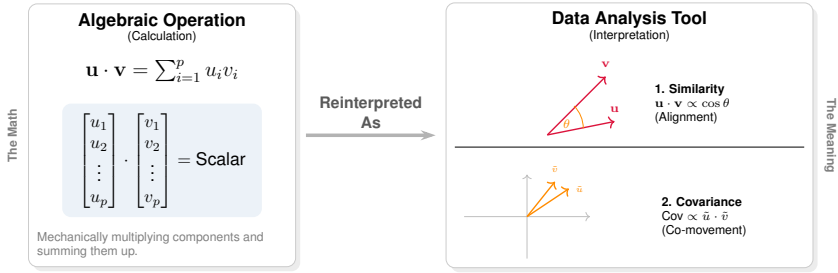
Summary of Key Conceptual Shifts

- **Data as Geometry:** A dataset is a point cloud of observation vectors in the high-dimensional feature space $\mathbb{R}^p$, represented by the data matrix X.



Summary of Key Conceptual Shifts

- **Operations as Insights:** The dot product measures alignment/similarity and is fundamentally proportional to statistical covariance in the centered feature space.



Summary of Key Conceptual Shifts

- **Norms as Metrics:** The choice of norm defines distance. L_2 is the standard distance, while L_1 induces sparsity due to the sharp corners of its unit ball geometry.