Machine Learning: Mathematical Background

Linear Algebra

He Wang

Website: drhewang.com

Lecture Overview

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

Learning Outcomes for Today's Lecture

- By the end of this lecture you should be familiar with some fundamental objects in and results of **Linear Algebra**
- For the most part we will concentrate on the statement of results which will be of use in the main body of this module
- However we will not be so concerned with the proof of these results

Lecture Overview

Vector Spaces

- Setting in which **linear algebra** takes place
- A **vector space**, V, is a set, the elements of which are called **vectors**, denoted by bold lower case letters, e.g. $\mathbf{x}$, $\mathbf{y}$ etc.
- Two operations are defined on a vector space:
  - **Vector Addition**
  - **Scalar Multiplication**
- For our purposes a **scalar** is a real number, usually denoted by a lower case letter

Vector Spaces

- V must satisfy:

  1. **Additive Closure**: if $\mathbf{x}, \mathbf{y} \in V$ then $\mathbf{x} + \mathbf{y} \in V$
  2. **Scalar Closure**: if $\alpha \in \mathbb{R}$ and $\mathbf{x} \in V$ then $\alpha\mathbf{x} \in V$
  3. **Identity Element of Addition**: $\exists$ a **zero vector**, $\mathbf{0}$, such that: $\mathbf{x} + \mathbf{0} = \mathbf{x}$, $\forall \mathbf{x} \in V$
  4. **Inverse Element of Addition**: $\exists$ an **additive inverse**, $-\mathbf{x}$, for each $\mathbf{x} \in V$, such that: $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$
  5. **Identity Element of Scalar Multiplication**: $\exists$ a **multiplicative identity**, 1, such that: $1\mathbf{x} = \mathbf{x}, \quad \forall \mathbf{x} \in V$
  6. **Commutativity**: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in V$
  7. **Associativity**: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$; and $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$
  8. **Distributivity**: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$; and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$

Linear Independence, Span, Basis & Dimension

- A set of vectors $\mathbf{v}_1, ..., \mathbf{v}_n \in V$ is **linearly independent** if:

$$\alpha_1 \mathbf{v}_1 + ... + \alpha_n \mathbf{v}_n = \mathbf{0} \quad \implies \quad \alpha_1 = ... = \alpha_n = 0$$

- The **span** of $\mathbf{v}_1, ..., \mathbf{v}_n \in V$, is the set of all vectors that can be expressed as a linear combination of them:

$$\text{span}(\mathbf{v}_1, ..., \mathbf{v}_n) = \{\mathbf{v} \quad | \quad \mathbf{v} = \beta_1 \mathbf{v}_1 + ... + \beta_n \mathbf{v}_n \quad \forall \beta_1, ..., \beta_n \in \mathbb{R}\}$$

- A **basis** for V is a set of vectors which are linearly independent and which span the whole of V
  - So every linearly independent set of vectors forms a basis for its span

- A **dimension** of V, $\dim(V)$, is the number of vectors in a basis

Euclidean Space

- So far our language has been abstract, but we will be interested in a particular vector space: the **Euclidean space**, $\mathbb{R}^n$
- Here the vectors are n-tuples of real numbers defined as **column vectors**, e.g.:

$$\mathbf{x} = \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

- Similarly a **row vector** is defined as:

$$\mathbf{x}^T = [x_1, x_2, ..., x_n]$$

- Note: On occasion we use $[\mathbf{x}]_i$ as an alternative to x_i

Euclidean Space: Addition & Multiplication

- **Vector Addition**:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \cdot \\ \cdot \\ x_n + y_n \end{bmatrix}$$

- **Scalar Multiplication**:

$$\alpha \mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \cdot \\ \cdot \\ \alpha x_n \end{bmatrix}$$

Euclidean Space: The Standard Basis

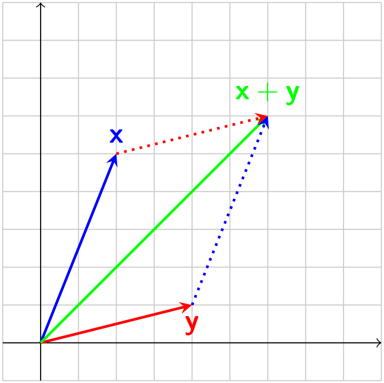
- One particular basis in $\mathbb{R}^n$ is the **standard basis**:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots \quad , \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

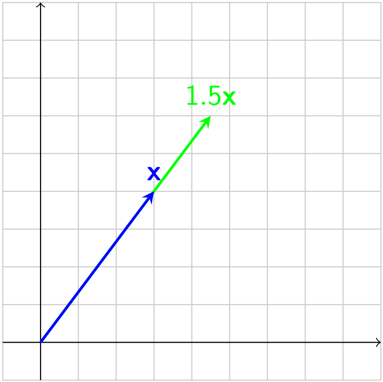
- Any vector $\mathbf{x}$ can be expressed in the standard basis:

$$\begin{aligned} \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} &= x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \ldots + x_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \\ &= x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + \ldots + x_n \mathbf{e}_n \end{aligned}$$

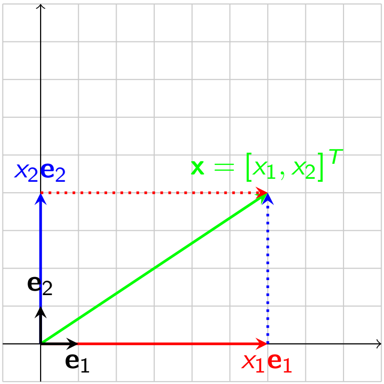
Euclidean Space: Vector Addition in $\mathbb{R}^2$



Euclidean Space: Scalar Multiplication in $\mathbb{R}^2$



Euclidean Space: The Standard Basis in $\mathbb{R}^2$



Subspaces

- So if U and W are subspaces of V then so is their sum:

$$U + W = \{\mathbf{u} + \mathbf{w} | \mathbf{u} \in U, \mathbf{w} \in W\}$$

- If $U \cap W = \{\mathbf{0}\}$ then $U + W$ is a **direct sum**, written as: $U \oplus W$

- It can be shown that:

$$\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W)$$

- And in particular:

$$\dim(U \oplus W) = \dim(U) + \dim(W)$$

Affine Subspaces: Parametric Spaces

- Affine subspaces can described by **parameters**:
- If L is a k-dimensional affine space and $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k\}$ is a basis of U, then every element $\mathbf{x} \in L$ can be described by:

$$\mathbf{x} = \mathbf{x}_0 + \lambda_1 \mathbf{x}_1 + \ldots + \lambda_k \mathbf{x}_k$$

Where $\{\lambda_i \in \mathbb{R}\}_{i=1}^{k}$ are a set **parameters**, $\mathbf{x}_0$ is called the **support point**.

Affine Subspaces: Lines, Planes & Hyperplanes

- **Lines** are 1-dimensional affine subspaces, with elements $\mathbf{y} \in \mathbb{R}^n$, described as:

  $$\mathbf{y} = \mathbf{x}_0 + \lambda \mathbf{x}_1$$

  Where $\lambda \in \mathbb{R},\ U = \text{span}(\mathbf{x}_1) \subseteq \mathbb{R}^n$

- **Planes** are 2-dimensional affine subspaces, with elements $\mathbf{y} \in \mathbb{R}^n$, described as:

  $$\mathbf{y} = \mathbf{x}_0 + \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$$

  Where $\lambda_1, \lambda_2 \in \mathbb{R},\ U = \text{span}(\mathbf{x}_1, \mathbf{x}_2) \subseteq \mathbb{R}^n$

- **Hyperplanes** are $(n-1)$-dimensional affine subspaces in $\mathbb{R}^n$, with elements $\mathbf{y} \in \mathbb{R}^n$, described as:

  $$\mathbf{y} = \mathbf{x}_0 + \sum_{i=1}^{n-1} \lambda_i \mathbf{x}_i$$

  Where $\{\lambda_i\}_{i=1}^{n-1} \in \mathbb{R},\ U = \text{span}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}) \subseteq \mathbb{R}^n$

Matrices

- Denote by a bold upper case letter, e.g. $\mathbf{A}$
- A matrix, $\mathbf{A} \in \mathbb{R}^{n \times m}$ is an $n \times m$ array of numbers with elements $\{A_{ij}\}_{i,j=1}^{n,m}$, i.e.:

$$\mathbf{A} = \underline{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ A_{21} & A_{22} & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nm} \end{bmatrix}$$

- Note: On occasion we use $[\mathbf{A}]_{ij}$ as an alternative to A_{ij}

Linear Maps

- A **linear map** is a function, $f : V \to W$ where V and W are vector spaces, that satisfies:

  1. $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in V$
  2. $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}), \quad \forall\, \mathbf{x} \in V, \forall\, \alpha \in \mathbb{R}$

Matrices as Linear Maps

- In particular, suppose V and W are finite-dimensional vector spaces with bases $\{\mathbf{v}_i\}_{i=1}^m$ and $\{\mathbf{w}_i\}_{i=1}^n$ respectively, then every matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ induces a linear map, $f : \mathbb{R}^m \to \mathbb{R}^n$ given by:

$$f(\mathbf{x}) = \mathbf{Ax}$$

  where $\mathbf{x} \in V$ and $f(\mathbf{x}) \in W$, and where:

$$[\mathbf{Ax}]_i = \sum_{j=1}^{m} A_{ij} x_j \qquad \text{for: } i = 1, ..., n$$

- For $n = 3, m = 2$:

$$f(\mathbf{x}) = \mathbf{Ax} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 \\ A_{21}x_1 + A_{22}x_2 \\ A_{31}x_1 + A_{32}x_2 \end{bmatrix}$$

Matrix Addition

- We can define a **matrix addition** operation for $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times m}$ such that:

$$\mathbf{C} = \mathbf{A} + \mathbf{B}$$

  By:

$$C_{ij} = A_{ij} + B_{ij}$$

- For $n = 3, m = 2$:

$$\begin{aligned}\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \\ C_{31} & C_{32} \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \\ A_{31} + B_{31} & A_{32} + B_{32} \end{bmatrix}\end{aligned}$$

Matrix Addition

- This definition implies that for $\mathbf{x} \in \mathbb{R}^m$:

$$
\begin{aligned}
[\mathbf{C}\mathbf{x}]_i = \sum_{j=1}^{m} C_{ij} x_j &= \sum_{j=1}^{m} (A_{ij} + B_{ij}) x_j \\
&= \sum_{j=1}^{m} A_{ij} x_j + \sum_{j=1}^{m} B_{ij} x_j \\
&= [\mathbf{A}\mathbf{x}]_i + [\mathbf{B}\mathbf{x}]_i
\end{aligned}
$$

- Hence the definition implies that:

$$
\mathbf{C}\mathbf{x} = (\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}
$$

- Thus matrix addition offers a more efficient mechanism for adding $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$

Matrix Multiplication

- We can define a **matrix multiplication** operation for $\mathbf{A} \in \mathbb{R}^{n \times l}$, $\mathbf{B} \in \mathbb{R}^{l \times m}$, $\mathbf{C} \in \mathbb{R}^{n \times m}$ such that:

$$\mathbf{C} = \mathbf{AB}$$

  By:

$$C_{ik} = \sum_{k=1}^{l} A_{ik} B_{kj}$$

- For $n = 3, m = 2, l = 3$:

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \\ C_{31} & C_{32} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \times \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{bmatrix}$$
$$= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} + A_{13}B_{31} & A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32} \\ A_{21}B_{11} + A_{22}B_{21} + A_{23}B_{31} & A_{21}B_{12} + A_{22}B_{22} + A_{23}B_{32} \\ A_{31}B_{11} + A_{32}B_{21} + A_{33}B_{31} & A_{31}B_{12} + A_{32}B_{22} + A_{33}B_{32} \end{bmatrix}$$

- And in general $\mathbf{BA} \neq \mathbf{AB}$

Matrix Multiplication

- This definition implies that for $\mathbf{x} \in \mathbb{R}^m$:

$$
\begin{aligned}
[\mathbf{C}\mathbf{x}]_i = \sum_{j=1}^{m} C_{ij} x_j = \sum_{j=1}^{m} \sum_{k=1}^{l} A_{ik} B_{kj} x_j &= \sum_{k=1}^{l} A_{ik} \sum_{j=1}^{m} B_{kj} x_j \\
&= \sum_{k=1}^{l} A_{ik} [\mathbf{y}]_k \\
&= [\mathbf{z}]_i
\end{aligned}
$$

  Where: $\mathbf{y} = \mathbf{B}\mathbf{x}$ and $\mathbf{z} = \mathbf{A}\mathbf{y}$

- Hence the definition implies that:

$$\mathbf{C}\mathbf{x} = \mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{x})$$

- Thus matrix multiplication offers a mechanism for performing the operations $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^l$ followed by $\mathbf{A} : \mathbb{R}^l \to \mathbb{R}^n$

Scalar Multiplication

- We can define a **scalar multiplication** operation for $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{n \times m}$ and $\alpha \in \mathbb{R}$ such that:

$$\mathbf{C} = \alpha \mathbf{A}$$

  By:

$$C_{ij} = \alpha A_{ij}$$

- For $n = 3, m = 3$:

$$\alpha \mathbf{A} = \begin{bmatrix} \alpha A_{11} & \alpha A_{12} & \alpha A_{13} \\ \alpha A_{21} & \alpha A_{22} & \alpha A_{23} \\ \alpha A_{31} & \alpha A_{32} & \alpha A_{33} \end{bmatrix}$$

Scalar Multiplication

- This definition implies that for $\mathbf{x} \in \mathbb{R}^m$:

$$[\mathbf{Cx}]_i = \sum_{j=1}^{m} C_{ij} x_j = \sum_{j=1}^{m} \alpha A_{ij} x_j = \alpha \sum_{j=1}^{m} A_{ij} x_j$$
$$= \alpha [\mathbf{Ax}]_i$$

- Hence the definition implies that:

$$(\alpha \mathbf{A})\mathbf{x} = \alpha(\mathbf{Ax})$$

Matrix Transpose

- If $\mathbf{A} \in \mathbb{R}^{n \times m}$ then its **transpose**, $\mathbf{A}^T \in \mathbb{R}^{m \times n}$ is defined by:

$$A^T{}_{ij} = A_{ji} \qquad \forall\, i, j$$

- Therefore:
  1. $(\mathbf{A}^T)^T = \mathbf{A}$
  2. $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
  3. $(\alpha \mathbf{A})^T = \alpha \mathbf{A}^T$
  4. $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

Matrix Identity

- The **identity matrix**, $\mathbf{I}$, is a **square** matrix with 1's on the diagonal and zeroes elsewhere

- $\mathbf{I}_n$ is the $n \times n$ identity matrix:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

- For $\mathbf{A} \in \mathbb{R}^{n \times m}$ then:

$$\mathbf{I}_n\mathbf{A} = \mathbf{A}\mathbf{I}_m = \mathbf{A}$$

Matrix Inverse

- For a square matrix, $\mathbf{A}$, its **inverse** (if it exists) satisfies:

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

- If $\mathbf{A}^{-1}$ does not exist then we say that $\mathbf{A}$ is **singular**
- For square matrices $\mathbf{A}$, $\mathbf{B}$ with inverses $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$ respectively, then:

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

- In particular, for $\mathbf{A} \in \mathbb{R}^{2\times 2}$:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \Longrightarrow \quad \mathbf{A}^{-1} = \frac{1}{(ad - bc)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

  Where $a, b, c, d \in \mathbb{R}$

Symmetric Matrices

- A square matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, is **symmetric** if $\mathbf{A}^T = \mathbf{A}$

Nullspace & Range

- If $\mathbf{A}$ is a linear map, $\mathbf{A} : V \to W$ then the **nullspace**, or **kernel**, of $\mathbf{A}$ is:

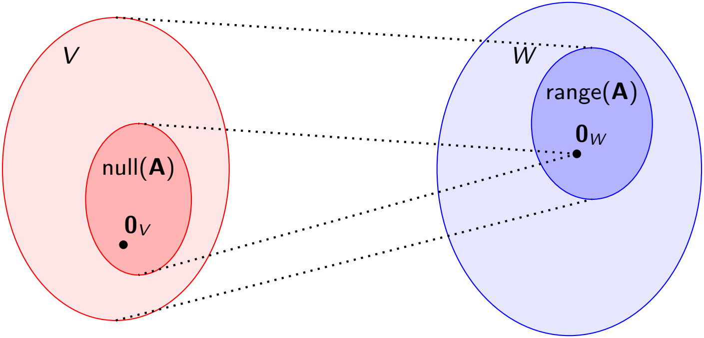
$$\text{null}(\mathbf{A}) = \{\mathbf{v} \in V | \mathbf{A}\mathbf{v} = \mathbf{0}\}$$

- and the **range**, or **image**, of $\mathbf{A}$ is:

$$\text{range}(\mathbf{A}) = \{\mathbf{w} \in W | \exists \mathbf{v} \in V \text{ such that } \mathbf{A}\mathbf{v} = \mathbf{w}\}$$

Nullspace & Range

$\mathbf{A} : V \to W$



Columnspace & Rowspace

- The **columnspace** of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the span of its m columns (which are all vectors in $\mathbb{R}^n$)
- The **rowspace** of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the span of its n rows (which are all row vectors in $\mathbb{R}^m$)
- So:

$$\text{columnspace}(\mathbf{A}) = \text{range}(\mathbf{A})$$
$$\text{rowspace}(\mathbf{A}) = \text{range}\left(\mathbf{A}^T\right)$$

- Recall from our definitions of **Span**, **Basis** and **Dimension** that the dimension of the columnspace (rowspace) is equal to the number of **linearly independent** vectors amongst the columns (rows) of $\mathbf{A}$

Rank

- It can be shown (See Appendix for proof) that dimension of the columnspace is equal to the dimension of the rowspace, and this dimension is known as the **rank** of $\mathbf{A}$:

$$\text{rank}(\mathbf{A}) = \dim\left(\text{range}(\mathbf{A})\right)$$

- So the **rank** of $\mathbf{A}$ is equal to the number of linearly independent vectors amongst the columns (rows) of $\mathbf{A}$

- Of course, $\dim\left(\text{range}(\mathbf{A})\right) \leq m$ and $\dim\left(\text{range}(\mathbf{A}^T)\right) \leq n$, so:

$$\text{rank}(\mathbf{A}) \leq \min(n, m)$$

Rank & Nullity

- For a matrix $\mathbf{A} \in \mathbb{R}^{n \times l}$ and a matrix $\mathbf{B} \in \mathbb{R}^{l \times m}$:

$$\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$$

$$\text{rank}(\mathbf{AB}) \leq \min\left(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\right)$$

- The dimension of the nullspace of $\mathbf{A} \in \mathbb{R}^{n \times m}$ is known as the **nullity** of $\mathbf{A}$:

$$\text{nullity}(\mathbf{A}) = \dim\left(\text{null}(\mathbf{A})\right)$$

- **Rank-Nullity Theorem**:

$$\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = m$$

Lecture Overview

Norms

- A **norm** equips a vector space with a notion of **length**
- More formally a norm on a real vector space, V, is a function, $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following, $\forall\, \mathbf{x}, \mathbf{y} \in V$ and $\forall\, \alpha \in \mathbb{R}$:

  1. $\|\mathbf{x}\| \geq 0$ with equality iff $\mathbf{x} = \mathbf{0}$
  2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
  3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

Norms

- Examples of norms include:

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

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2}$$

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \text{ where: } p \geq 1$$

Inner Product

- An **inner product** equips a vector space with a notion of **similarity**
- More formally an inner product on a real vector space, V, is a function, $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following, $\forall\, \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\forall\, \alpha \in \mathbb{R}$:
    1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ with equality iff $\mathbf{x} = \mathbf{0}$
    2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$; and $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$
    3. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
- An inner product on V induces a norm on V:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

Orthonormality

- Two vectors $\mathbf{x}, \mathbf{y}$ are said to be **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$
- If two orthogonal vectors $\mathbf{x}, \mathbf{y}$ have unit length, i.e. $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, then they are described as **orthonormal**

Scalar Product

- For Euclidean space (i.e. $V = \mathbb{R}^n$) then the standard inner product is the **dot product** or **scalar product**:

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} &= \sum_{i=1}^{n} x_i y_i \\ &= \mathbf{x}^T \mathbf{y}\end{aligned}$$

- And the norm induced by this inner product is the **two-norm**, $\| \cdot \|_2$

- In addition for Euclidean space we can define the notion of **angle**, θ, between two vectors $\mathbf{x}, \mathbf{y}$ via the dot product:

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos \theta$$

Orthonormality in $\mathbb{R}^n$

- Members of the standard basis in $\mathbb{R}^n$, $\{\mathbf{e}_i\}_{i=1}^n$ are **orthonormal**
- For example, (for $n = 2$):

$$\mathbf{x} = \begin{bmatrix} 8 \\ 4 \end{bmatrix} = 8\mathbf{e}_1 + 4\mathbf{e}_2$$

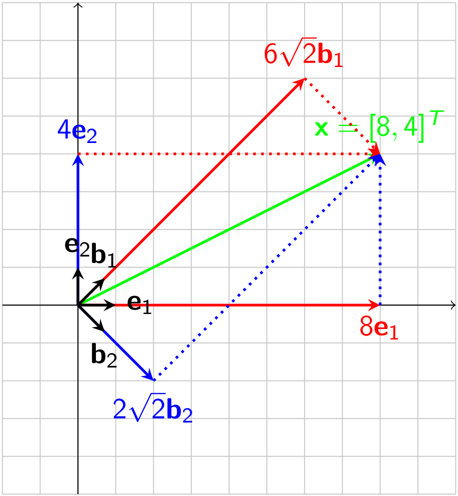
- But other orthonormal bases exist, for example, (for $n = 2$):

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

And this basis yields the following expression:

$$\mathbf{x} = \begin{bmatrix} 8 \\ 4 \end{bmatrix} = 6\sqrt{2}\mathbf{b}_1 + 2\sqrt{2}\mathbf{b}_2$$

Orthonormality in $\mathbb{R}^2$



Orthogonal Matrices

- A square matrix, $\mathbf{Q} \in \mathbb{R}^{n \times n}$, is **orthogonal** if its columns are orthonormal:
$$\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$$
- Therefore:
$$\mathbf{Q}^T = \mathbf{Q}^{-1}$$
- Orthogonal matrices preserve inner products:
$$(\mathbf{Qx}) \cdot (\mathbf{Qy}) = (\mathbf{Qx})^T(\mathbf{Qy}) = \mathbf{x}^T\mathbf{Q}^T\mathbf{Q}\mathbf{y} = \mathbf{x}^T\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$$
- ...consequently they preserve 2-norms
- So multiplication by an orthogonal matrix can be considered to be a mapping that preserves vector length, but **rotates** or **reflects** the vector about the origin

Projections

- In general a **projection**, $\mathbf{P}_U : V \to U$, is a linear mapping from a vector space, V, to a subspace of V, $U \subseteq V$, such that:

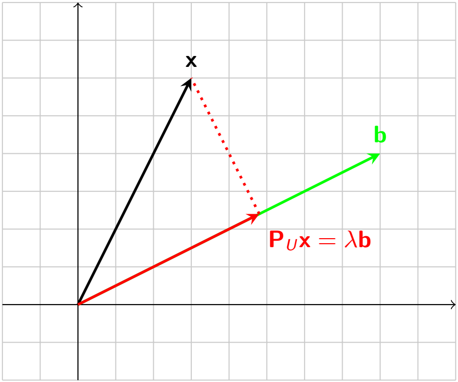
$$\mathbf{P}_U \mathbf{P}_U = \mathbf{P}_U$$

- We will have particular interest in **orthogonal projections** in Euclidean space, which map a vector $\mathbf{x} \in \mathbb{R}^n$ to a vector $\mathbf{P}_U \mathbf{x} \in U \subseteq \mathbb{R}^n$, that is 'closest' to $\mathbf{x}$, such that $\|\mathbf{x} - \mathbf{P}_U \mathbf{x}\|_2$ is minimal.

Orthogonal Projections onto Lines

- Consider a 1-dimensional subspace passing through the origin (i.e. a line), described by the basis vector, **b**
- An orthogonal projection of a vector $\mathbf{x} \in \mathbb{R}^n$ onto this line must satisfy:
  1. $\mathbf{P}_U\mathbf{x} = \lambda\mathbf{b} \qquad \text{for some:} \quad \lambda \in \mathbb{R}$
  2. $\|\mathbf{x} - \mathbf{P}_U\mathbf{x}\|_2^2$ must be minimal

Orthogonal Projections onto Lines



Orthogonal Projections onto Lines

- Thus we seek λ that satisfies:

$$\operatorname*{argmin}_{\lambda} \|\mathbf{x} - \lambda \mathbf{b}\|_2^2$$

- Differentiating with respect to λ and seeking stationary points implies:

$$\begin{aligned} & 2(\mathbf{x} - \lambda \mathbf{b}) \cdot \mathbf{b} = 0 \\ \implies \qquad & \lambda = \frac{\mathbf{x} \cdot \mathbf{b}}{\|\mathbf{b}\|_2^2} \end{aligned}$$

Orthogonal Projections onto Lines

- This tells us that $(\mathbf{x} - \mathbf{P}_U\mathbf{x})$ is orthogonal to $\mathbf{b}$ since:

$$
\begin{aligned}
(\mathbf{x} - \mathbf{P}_U\mathbf{x}) \cdot \mathbf{b} &= \mathbf{x} \cdot \mathbf{b} - \lambda \|\mathbf{b}\|_2^2 \\
&= \mathbf{x} \cdot \mathbf{b} - \frac{\mathbf{x} \cdot \mathbf{b}}{\|\mathbf{b}\|_2^2} \|\mathbf{b}\|_2^2 = 0
\end{aligned}
$$

- Furthermore, we can derive an explicit form for the matrix $\mathbf{P}_U$:

$$
\begin{aligned}
\mathbf{P}_U\mathbf{x} &= \lambda\mathbf{b} = \mathbf{b}\lambda \\
&= \mathbf{b}\frac{\mathbf{x} \cdot \mathbf{b}}{\|\mathbf{b}\|_2^2} = \mathbf{b}\frac{\mathbf{b}^T\mathbf{x}}{\|\mathbf{b}\|_2^2} = \frac{\mathbf{b}\mathbf{b}^T}{\|\mathbf{b}\|_2^2}\mathbf{x}
\end{aligned}
$$

- Thus:

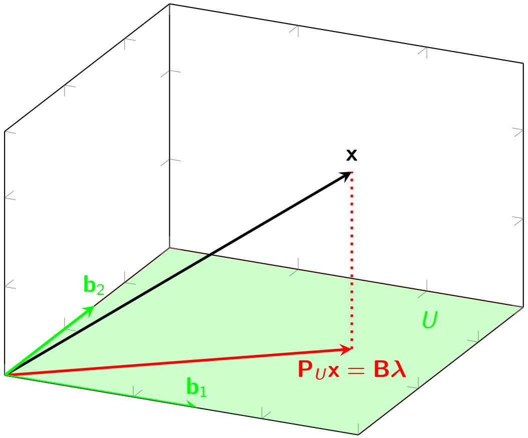
$$
\mathbf{P}_U = \frac{\mathbf{b}\mathbf{b}^T}{\|\mathbf{b}\|_2^2}
$$

Orthogonal Projections onto Subspaces

- Consider an m-dimensional subspace, $U \subseteq \mathbb{R}^n$, passing through the origin, described by the basis vectors, $\{\mathbf{b}_i\}_{i=1}^m$
- An orthogonal projection of a vector $\mathbf{x} \in \mathbb{R}^n$ onto this subspace must satisfy:

  1. $\mathbf{P}_U\mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{b}_i = \mathbf{B}\boldsymbol{\lambda}$
     Where:
     $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{R}^{n\times m}$, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^T \in \mathbb{R}^m$
  2. $\|\mathbf{x} - \mathbf{P}_U\mathbf{x}\|_2^2$ must be minimal

Orthogonal Projections onto Subspaces: 2-D Example



Orthogonal Projections onto Subspaces

- Thus we seek λ that satisfies:

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}\|_2^2$$

- Differentiating with respect to $\boldsymbol{\lambda}$ and seeking stationary points implies:

$$\begin{aligned} 2\mathbf{B}^T(\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) &= 0 \\ \implies \qquad \boldsymbol{\lambda} &= (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x} \end{aligned}$$

(These are known as the **normal equations**, they will be discussed in more detail in the context of **Linear Regression**)

Orthogonal Projections onto Subspaces

- This tells us that $(\mathbf{x} - \mathbf{P}_U\mathbf{x})$ is orthogonal to $\mathbf{b}_i$ for all i, since:

$$\begin{aligned}(\mathbf{x} - \mathbf{P}_U\mathbf{x}) \cdot \mathbf{b}_i &= \mathbf{x} \cdot \mathbf{b}_i - \mathbf{B}\boldsymbol{\lambda} \cdot \mathbf{b}_i \\ &= \mathbf{x} \cdot \mathbf{b}_i - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x} \cdot \mathbf{b}_i = 0\end{aligned}$$

- Furthermore, we can derive an explicit form for the matrix $\mathbf{P}_U$:

$$\begin{aligned}\mathbf{P}_U\mathbf{x} &= \mathbf{B}\boldsymbol{\lambda} \\ &= \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x}\end{aligned}$$

- Thus:

$$\mathbf{P}_U = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$$

Lecture Overview

Solving Linear Equations

- Let's say we want to solve the following system of linear equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{b} \in \mathbb{R}^{n}$ are both known, and $\mathbf{x} \in \mathbb{R}^{m}$ is unknown

- This is equivalent to a set of n equations:

$$\begin{aligned}
A_{11}x_1 + A_{12}x_2 + \dots + A_{1m}x_m &= b_1 \\
A_{21}x_1 + A_{22}x_2 + \dots + A_{2m}x_m &= b_2 \\
&\vdots \\
A_{n1}x_1 + A_{n2}x_2 + \dots + A_{nm}x_m &= b_n
\end{aligned}$$

Solving Linear Equations

- This system of equations can have:
  - A **unique solution**, **no solutions**, or **infinitely many solutions**
  - But it cannot have more than 1 and less than an infinite number of solutions. Why?
  - If $\mathbf{x}_A$ and $\mathbf{x}_B$ are solutions, then $\mathbf{x}_C = \alpha \mathbf{x}_A + (1 - \alpha)\mathbf{x}_B$ must also be a solution, for all α

- In order to solve this system we need to:
  - Check how many **distinct** simultaneous equations we have in our system
  - Check the **consistency** of these equations
  - Compare the **number** of these equations to the number of unknowns in our system

Solving Linear Equations

- To check how many distinct equations we have:
  - Form the **augmented matrix**, $\mathbf{A}|\mathbf{b} = [\mathbf{A}, \mathbf{b}] \in \mathbb{R}^{n \times (m+1)}$
  - Recall that the number of linearly independent rows in a matrix is equal to the matrix rank, so:

$$\begin{aligned} \#\ \text{distinct equations} &= \#\ \text{linearly independent rows in } \mathbf{A}|\mathbf{b} \\ &= \text{rank}(\mathbf{A}|\mathbf{b}) \leq \min(n, m+1) \end{aligned}$$

- To check the consistency of these equations:
  - Note that the number of distinct 'left-hand sides' of these equations should equal the number of these equations for consistency:

$$\begin{aligned} \#\ \text{distinct 'LHS' of equations} &= \#\ \text{linearly independent rows in } \mathbf{A} \\ &= \text{rank}(\mathbf{A}) \leq \min(n, m) \end{aligned}$$

Solving Linear Equations

- If $\text{rank}(\mathbf{A}) < \text{rank}(\mathbf{A}|\mathbf{b})$:
  - The equations are **inconsistent**
  - We have no solutions
- If $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}|\mathbf{b}) < m$:
  - There are too few equations to fully specify the number of unknowns
  - The system is **underdetermined**
  - We have infinitely many solutions
- If $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}|\mathbf{b}) = m \leq n$:
  - There are at least the same number of *consistent* equations as unknowns
  - We have a unique solution

Solving Linear Equations: An Algebraic Solution

- Let us try to solve our system directly using the machinery of linear algebra:

$$\begin{aligned} \mathbf{A}\mathbf{x} &= \mathbf{b} \\ \implies \qquad \mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

  ...if $\mathbf{A}^{-1}$ exists!

- It turns out that $\mathbf{A}^{-1}$ exists iff $\mathbf{A}$ is square and full rank [see **Invertible Matrix Theorem**]

- If $\mathbf{A}$ is square and of full rank then of course $\operatorname{rank}(\mathbf{A}) = m$

- So the algebraic solution, if it exists, yields a unique solution

Lecture Overview

Determinants

- The **determinant** is associated with square matrices only. For some square matrix $\mathbf{A}$ the determinant is denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$
- It is a function which maps a matrix to a real scalar
- Geometrically the determinant of $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be thought of as a signed volume of the n-dimensional paralleltope that results from the action of $\mathbf{A}$ on the unit cube

Determinants

- For $\mathbf{A} \in \mathbb{R}^{2\times 2}$:

$$\det(\mathbf{A}) = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = A_{11}A_{22} - A_{21}A_{12}$$

- For $\mathbf{A} \in \mathbb{R}^{3\times 3}$:

$$\det(\mathbf{A}) = \begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = A_{11}\begin{vmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{vmatrix} - A_{12}\begin{vmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{vmatrix} + A_{13}\begin{vmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{vmatrix}$$

- For $\mathbf{A} \in \mathbb{R}^{n\times n}$:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} |\mathbf{M}_{ij}|$$

where $|\mathbf{M}_{ij}|$ is the determinant of the $(n-1)\times(n-1)$ matrix that results from $\mathbf{A}$ by removing the i-th row and the j-th column

Determinants

- **Properties**:

$$\det(\mathbf{I}) = 1$$

$$\det\left(\mathbf{A}^T\right) = \det(\mathbf{A})$$

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$$

$$\det\left(\mathbf{A}^{-1}\right) = (\det(\mathbf{A}))^{-1}$$

$$\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$$

Traces

- The **trace** is associated with square matrices only. For some square matrix $\mathbf{A} \in \mathbb{R}^n$ the trace is denoted by $\text{tr}(\mathbf{A})$
- It is a function which maps a matrix to a real scalar, and is defined as the sum of the diagonal elements of $\mathbf{A}$:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

Traces

- **Properties**:

$$\mathrm{tr}\left(\mathbf{A}+\mathbf{B}\right)=\mathrm{tr}\left(\mathbf{A}\right)+\mathrm{tr}\left(\mathbf{B}\right) \qquad \text{where: } \mathbf{A},\mathbf{B}\in\mathbb{R}^{n\times n}$$

$$\mathrm{tr}\left(\alpha\mathbf{A}\right)=\alpha\mathrm{tr}\left(\mathbf{A}\right) \qquad \text{where: } \alpha\in\mathbb{R},\ \mathbf{A}\in\mathbb{R}^{n\times n}$$

$$\mathrm{tr}\left(\mathbf{I}_n\right)=n$$

$$\mathrm{tr}\left(\mathbf{A}\mathbf{B}\right)=\mathrm{tr}\left(\mathbf{B}\mathbf{A}\right) \qquad \text{where: } \mathbf{A}\in\mathbb{R}^{n\times k},\mathbf{B}\in\mathbb{R}^{k\times n}$$

Invertible Matrix Theorem

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, then the following statements are equivalent:
  - $\mathbf{A}$ is invertible
  - $\text{rank}(\mathbf{A}) = n$
  - $\det(\mathbf{A}) \neq 0$
  - The columns of $\mathbf{A}$ are linearly independent
  - The rows of $\mathbf{A}$ are linearly independent
  - $\nexists\ \mathbf{x}$ such that $\mathbf{Ax} = \mathbf{0}$
  - ...and many more...

Eigenvectors & Eigenvalues

- For a square matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, we say that $\mathbf{x}$ is an **eigenvector** of $\mathbf{A}$ corresponding to an **eigenvalue**, λ, if:

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

- If $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$, then $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$
- Therefore if $(\mathbf{A} - \lambda \mathbf{I})^{-1}$ exists, then the unique solution is $\mathbf{x} = \mathbf{0}$
- Therefore for non-trivial solutions we must demand that $\nexists (\mathbf{A} - \lambda \mathbf{I})^{-1}$
- Therefore for non-trivial solutions $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ [by the **Invertible Matrix Theorem**]
- $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ is known as the **characteristic polynomial** of $\mathbf{A}$
  - Its (possibly non-unique) roots determine the possible eigenvalues of our problem

Eigendecomposition

- The **eigen-** or **spectral decomposition** of a square matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, with n linearly independent eigenvectors, $\{\mathbf{q}_i\}_{i=1}^{n}$ states that:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

  Where:

  - $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n]$;
  - $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ where λ_i is the eigenvalue associated with the eigenvector $\mathbf{q}_i$

- From this it follows that:

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

Eigendecomposition: Symmetric Matrices

- If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is (real and) symmetric, then there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$, $\{\mathbf{q}_i | \mathbf{q}_i \cdot \mathbf{q}_j = \delta_{ij}, \quad \forall\, j\}_{i=1}^{n}$, and the associated eigenvalues are real

- From this it follows that the eigendecomposition of a real symmetric matrix, $\mathbf{A}$, is given by:

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$$

  Where:

  - $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n]$;
  - $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ where λ_i is the eigenvalue associated with the eigenvector $\mathbf{q}_i$

Eigendecomposition: Geometric Intuition

- Consider the linear map, $\mathbf{Q} \in \mathbb{R}^{n \times n}$, described above, and the set of standard basis vectors which span $\mathbb{R}^n$, $\{\mathbf{e}_i\}_{i=1}^n$
- Now consider this map acting on the i-th such vector:

$$\mathbf{Q}\mathbf{e}_i = \sum_{j=1}^{n} \mathbf{q}_j [\mathbf{e}_i]_j = \mathbf{q}_i$$

- Thus $\mathbf{Q}$ maps $\mathbf{e}_i$ to the corresponding eigenvector of $\mathbf{A}$.
- Similarly, the matrix $\mathbf{Q}^{-1}$ maps the i-th eigenvector of $\mathbf{A}$ to the corresponding standard basis vector:

$$\mathbf{Q}^{-1}\mathbf{q}_i = \mathbf{e}_i$$

Eigendecomposition: Geometric Intuition

- Now we consider the linear map, $\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$, as performing the following operations when it acts on an eigenvector of $\mathbf{A}$:

  1. A transformation from the basis spanned by the eigenvectors to that spanned by the standard basis vectors:
  $$\mathbf{Q}^{-1}\mathbf{q}_i = \mathbf{e}_i$$
  2. A scaling of the resulting standard basis vector by the corresponding eigenvalue:
  $$\mathbf{\Lambda Q}^{-1}\mathbf{q}_i = \mathbf{\Lambda e}_i = \lambda_i \mathbf{e}_i$$
  3. A basis change of the resulting scaled vector back to the basis spanned by the eigenvectors:
  $$\mathbf{Q\Lambda Q}^{-1}\mathbf{q}_i = \mathbf{Q}\lambda_i \mathbf{e}_i = \lambda_i \mathbf{Q e}_i = \lambda_i \mathbf{q}_i$$

Lecture Overview

Quadratic Forms

- Assume a square matrix, $\mathbf{M} \in \mathbb{R}^{n \times n}$. Then the **quadratic form** of $\mathbf{M}$ is defined by:

$$\mathbf{x}^T \mathbf{M} \mathbf{x}$$

Quadratic Forms

- Note also that it is possible to write any matrix, $\mathbf{M}$, as follows:

$$\mathbf{M} = \frac{1}{2}\left(\mathbf{M} + \mathbf{M}^T\right) + \frac{1}{2}\left(\mathbf{M} - \mathbf{M}^T\right) = \mathbf{A} + \mathbf{B}$$

  Where:
  $\mathbf{A} = \frac{1}{2}\left(\mathbf{M} + \mathbf{M}^T\right)$ is a **symmetric** matrix
  $\mathbf{B} = \frac{1}{2}\left(\mathbf{M} - \mathbf{M}^T\right)$ is an **antisymmetric** matrix

- Note further that because $\mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}^T\mathbf{M}^T\mathbf{x}$, then:

$$\mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

- In other words it is always possible to express the quadratic form of a general square matrix, $\mathbf{M}$, as the quadratic form of an associated symmetric matrix, $\mathbf{A} = \frac{1}{2}\left(\mathbf{M} + \mathbf{M}^T\right)$

Definiteness of Matrices

- A real symmetric matrix, $\mathbf{A}$, is said to be:
  - **Positive Semidefinite** (psd), written as $\mathbf{A} \succeq 0$, iff $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0, \forall \mathbf{x} \neq \mathbf{0}$
  - **Positive Definite** (pd), written as $\mathbf{A} \succ 0$, iff $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \forall \mathbf{x} \neq \mathbf{0}$
  - **Negative Semidefinite** (nsd), written as $\mathbf{A} \preceq 0$, iff $\mathbf{x}^T\mathbf{A}\mathbf{x} \leq 0, \forall \mathbf{x} \neq \mathbf{0}$
  - **Negative Definite** (nd), written as $\mathbf{A} \prec 0$, iff $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0, \forall \mathbf{x} \neq \mathbf{0}$
  - **Indefinite** otherwise

PSD & PD Matrices: Properties

- For a real symmetric matrix, **psd** matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, with (real) eigenvalues, $\{\lambda_i\}_{i=1}^{n}$:

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0, \quad \forall \mathbf{x} \neq \mathbf{0} \quad &\iff \quad \{\lambda_i \geq 0\}_{i=1}^{n} \\ &\implies \quad \det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i \geq 0 \end{aligned}$$

- For a real symmetric matrix, **pd** matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, with (real) eigenvalues, $\{\lambda_i\}_{i=1}^{n}$:

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \quad &\iff \quad \{\lambda_i > 0\}_{i=1}^{n} \\ &\implies \quad \det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i > 0 \\ &\implies \quad \textbf{A is invertible} \end{aligned}$$

PSD Matrices: Further Properties

- For a real, symmetric, **psd** matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, we can write:

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{D} \end{bmatrix}$$

where: $\mathbf{B} \in \mathbb{R}^{u \times u}$, $\mathbf{D} \in \mathbb{R}^{v \times v}$, $\mathbf{C} \in \mathbb{R}^{u \times v}$, and $u + v = n$.

- Consider the quadratic form induced by some non-trivial vector, $\mathbf{x} = [\mathbf{a}^T, \mathbf{b}^T]^T$, where $\mathbf{a} \in \mathbb{R}^u$ and $\mathbf{b} \in \mathbb{R}^v$:

$$\begin{aligned} & \mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0 \\ \implies \quad & \mathbf{a}^T\mathbf{B}\mathbf{a} + \mathbf{b}^T\mathbf{C}^T\mathbf{a} + \mathbf{a}^T\mathbf{C}\mathbf{b} + \mathbf{b}^T\mathbf{D}\mathbf{b} \geq 0 \\ \implies \quad & \mathbf{a}^T\mathbf{B}\mathbf{a} + 2\mathbf{a}^T\mathbf{C}\mathbf{b} + \mathbf{b}^T\mathbf{D}\mathbf{b} \geq 0 \end{aligned}$$

PSD Matrices: Further Properties (Cont.)

- This must hold for all $\mathbf{x}$, in particular:

$$\begin{aligned} \mathbf{a} \neq \mathbf{0}, \mathbf{b} = \mathbf{0} \quad &\implies \mathbf{a}^T \mathbf{B} \mathbf{a} \geq 0 \\ \mathbf{a} = \mathbf{0}, \mathbf{b} \neq \mathbf{0} \quad &\implies \mathbf{b}^T \mathbf{D} \mathbf{b} \geq 0 \end{aligned}$$

- Thus, if $\mathbf{A}$ is positive semidefinite, then all its block diagonal matrices are all also psd

- (Similar results hold if $\mathbf{A}$ is positive definite, negative semidefinite, or negative definite)

Lecture Overview

Some Useful Results

1. For any matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, the matrix $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{m \times m}$ is symmetric

2. $\mathbf{X}^T\mathbf{X}$ is always positive semidefinite for all $\mathbf{X}$
   - **Proof:**
   $$\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = \|\mathbf{X}\mathbf{a}\|_2^2 \geq 0, \qquad \forall\ \mathbf{a} \neq \mathbf{0}$$

Some Useful Results

3. $\text{rank}(\mathbf{X}^T\mathbf{X}) = m$ if and only if $\mathbf{X}^T\mathbf{X} \succ 0$
   - **Proof:** (*only if*)

$$
\begin{aligned}
& \text{rank}(\mathbf{X}^T\mathbf{X}) = m \\
\implies \quad & \nexists\, \mathbf{a} \text{ such that } \mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0} && \text{[by \textbf{Linear Independence}]} \\
\implies \quad & \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = \|\mathbf{X}\mathbf{a}\|_2^2 > 0, \qquad \forall\, \mathbf{a} \neq \mathbf{0} \\
\implies \quad & \mathbf{X}^T\mathbf{X} \succ 0
\end{aligned}
$$

   - **Proof:** (*if*)

$$
\begin{aligned}
& \mathbf{X}^T\mathbf{X} \succ 0 \\
\implies \quad & \|\mathbf{X}\mathbf{a}\|_2^2 > 0, \qquad \forall\, \mathbf{a} \neq \mathbf{0} \\
\implies \quad & \nexists\, \mathbf{a} \text{ such that } \mathbf{X}\mathbf{a} = \mathbf{0} \\
\implies \quad & \nexists\, \mathbf{a} \text{ such that } \mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0} \\
\implies \quad & \text{rank}(\mathbf{X}^T\mathbf{X}) = m && \text{[by \textbf{Invertible Matrix Theorem}]}
\end{aligned}
$$

Some Useful Results

4. $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X})$
   - **Proof:**
     For any vector $\mathbf{a} \in \text{null}(\mathbf{X}^T\mathbf{X})$:

$$\begin{aligned}\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0} &\implies \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = 0\\ &\implies \|\mathbf{X}\mathbf{a}\|_2^2 = 0\\ &\implies \mathbf{X}\mathbf{a} = \mathbf{0}\end{aligned}$$

     For any vector $\mathbf{b} \in \text{null}(\mathbf{X})$:

$$\mathbf{X}\mathbf{b} = \mathbf{0} \implies \mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{0}$$

     So, any $\mathbf{a}$ which belongs to the nullspace of $\mathbf{X}^T\mathbf{X}$ also belongs to the nullspace of $\mathbf{X}$ and vice-versa, so:

$$\text{nullity}(\mathbf{X}^T\mathbf{X}) = \text{nullity}(\mathbf{X})$$

     By the **Rank-Nullity Theorem**:

$$\begin{aligned}\text{rank}(\mathbf{X}) + \text{nullity}(\mathbf{X}) &= m\\ \text{rank}(\mathbf{X}^T\mathbf{X}) + \text{nullity}(\mathbf{X}^T\mathbf{X}) &= m\end{aligned}$$

     So: $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X})$

Some Useful Results

5. $\text{rank}(\mathbf{X}^T\mathbf{X}) \leq \min(n, m)$
   - **Proof:**
     Recall $\text{rank}(\mathbf{X}) \leq \min(n, m)$

6. $(\mathbf{X}^T\mathbf{X})^{-1}$ exists if and only if $\text{rank}(\mathbf{X}^T\mathbf{X}) = m$
   - **Proof:**
     By **Invertible Matrix Theorem**

Some Useful Results

7. $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) \quad \forall\ \mathbf{y}$
   - **Proof:**
     Recall:
     $$\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^T)$$
     Because $(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y})$ has more columns than $\mathbf{X}^T\mathbf{X}$ then:
     $$\text{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) \geq \text{rank}(\mathbf{X}^T\mathbf{X}) \tag{1}$$
     Observe that: $(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) = \mathbf{X}^T[\mathbf{X},\mathbf{y}]$:
     $$\begin{aligned} \implies \quad \text{rank}(\mathbf{X}^T[\mathbf{X},\mathbf{y}]) &\leq \min(\text{rank}(\mathbf{X}^T), \text{rank}([\mathbf{X},\mathbf{y}])) \\ &= \text{rank}(\mathbf{X}^T) \\ &= \text{rank}(\mathbf{X}^T\mathbf{X}) \end{aligned} \tag{2}$$
     Combining (1) and (2):
     $$\begin{aligned} &\text{rank}(\mathbf{X}^T\mathbf{X}) \leq \text{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) \leq \text{rank}(\mathbf{X}^T\mathbf{X}) \\ \implies \quad &\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) \end{aligned}$$

Some Useful Results

- To sum up:
  - $\mathbf{X}^T\mathbf{X}$ is symmetric
  - $\mathbf{X}^T\mathbf{X} \succeq 0$
  - $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}) \leq \min{(n, m)}$
  - $\exists(\mathbf{X}^T\mathbf{X})^{-1} \iff (\text{rank}(\mathbf{X}^T\mathbf{X}) = m) \iff \mathbf{X}^T\mathbf{X} \succ 0$
  - $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) \quad \forall\, \mathbf{y}$

Lecture Overview

Summary

- **Linear Algebra** is an essential tool that helps us deal with systems of linear equations
- We have introduced the setting and some structure for the operation of linear algbera
- Building on this we have introduced some key results which will be of direct use in machine learning

Lecture Overview

Columspace & Rowspace: Revisited

- Recall that the dimension of the columnspace is equal to the dimension of the rowspace.

  For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ this can be expressed as:

  $$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$$

Columspace & Rowspace: Revisited

- **Proof:**

  Let the row rank of $\mathbf{A}$ be r, thus $\text{rank}(\mathbf{A}^T) = r$.
  And let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ be a basis of the rowspace of $\mathbf{A}$.

  Now, consider a set of scalar coefficients, $c_1, c_2, \ldots, c_r$, such that:

$$\begin{aligned} \mathbf{0} &= c_1\mathbf{A}\mathbf{x}_1 + c_2\mathbf{A}\mathbf{x}_2 + \ldots + c_r\mathbf{A}\mathbf{x}_r \\ &= \mathbf{A}(c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_r\mathbf{x}_r) = \mathbf{A}\mathbf{v} \end{aligned}$$

  Where: $\mathbf{v} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_r\mathbf{x}_r$

  Now, $\mathbf{v}$ is a linear combination of vectors in the row space of $\mathbf{A}$.
  Thus, $\mathbf{v} \in \text{rowspace}(\mathbf{A})$ (**Fact 1**).

  Furthermore, since $\mathbf{A}\mathbf{v} = \mathbf{0}$, then $\mathbf{v}$ is orthogonal to every row vector of $\mathbf{A}$.
  Thus, $\mathbf{v}$ is orthogonal to every vector in $\text{rowspace}(\mathbf{A})$ (**Fact 2**).

Columspace & Rowspace: Revisited

- **Proof (Cont.):**

  Facts 1 & 2 can only hold simultaneously if $\mathbf{v}$ is orthogonal to itself, and this occurs only if $\mathbf{v} = \mathbf{0}$.

  Therefore: $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_r\mathbf{x}_r = \mathbf{0}$

  But $\{\mathbf{x}_i\}_{i=1}^r$ are linearly independent, so $c_1 = c_2 = \ldots = c_r = 0$.

  Thus we have shown that:

  $$c_1\mathbf{A}\mathbf{x}_1 + c_2\mathbf{A}\mathbf{x}_2 + \ldots + c_r\mathbf{A}\mathbf{x}_r = \mathbf{0} \quad \Longrightarrow \quad c_1 = c_2 = \ldots = c_r = 0$$

  Thus, by the definitive property of linear independence, $\{\mathbf{A}\mathbf{x}_i\}_{i=1}^r$ are linearly independent.

Columspace & Rowspace: Revisited

- **Proof (Cont.):**

  Now, each $\mathbf{A}\mathbf{x}_i$ is a vector in the columnspace of $\mathbf{A}$, so $\{\mathbf{A}\mathbf{x}_i\}_{i=1}^r$ is a set of r linearly independent vectors in the columnspace of $\mathbf{A}$. Thus:

$$\begin{aligned} & \dim(\text{columnspace}(\mathbf{A})) \geq r \\ \implies \quad & \text{rank}(\mathbf{A}) \geq r \\ \implies \quad & \text{rank}(\mathbf{A}) \geq \text{rank}(\mathbf{A}^T) \end{aligned}$$

  Analagously, we can prove:

$$\text{rank}(\mathbf{A}^T) \geq \text{rank}(\mathbf{A})$$

  Since these inequalities hold simultaneously, this implies:

$$\begin{aligned} & \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T) \\ \implies \quad & \dim(\text{columnspace}(\mathbf{A})) = \dim(\text{rowspace}(\mathbf{A})) \end{aligned}$$

Rank-Nullity Theorem: Revisited

- Recall the Rank-Nullity Theorem for a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$:

$$\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = m$$

Rank-Nullity Theorem: Revisited

- **Proof:**

  Consider the matrix equation $\mathbf{Ax} = \mathbf{0}$, where $\mathbf{x} \in \mathbb{R}^m$.

  From our discussion of Gauss-Jordan elimination, recall that:

  $$\text{rank}(\text{rref}(\mathbf{A})) = \text{rank}(\mathbf{A}) = r$$

  Thus $\text{rref}(\mathbf{A})$ has only r non-zero rows, and hence $(m - r)$ of the variables in the solution $\mathbf{x}$ are free.

  But the number of free variables is the number of free parameters in a general parametric solution of $\mathbf{Ax} = \mathbf{0}$.

  And the number of free parameters defines the dimension of the space spanned by the solutions.

  Recall that the dimension of the space spanned by the solutions is, definitively, the dimension of the null space of $\mathbf{A}$.

  In other words the nullity of $\mathbf{A}$ is equal to $(m - r)$:

  $$\text{nullity}(\mathbf{A}) = m - r$$