EECS 553: Machine Learning (ECE) **University of Michigan**

Linear Algebra Background

Winter 2023 **Clayton Scott**

1 Overview

These notes review concepts that are critical for machine learning. Many of the concepts discussed below can be defined in more general settings, but for concreteness I focus on *Euclidean* vectors. I assume the reader is familiar with basic matrix and vector operations, as well as basic set notation.

2 Dot Product, Norm, and Outer Product

Let

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix}, \qquad \boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}$$

be two vectors. The *dot product* of $\boldsymbol{u}$ and $\boldsymbol{v}$ is

$$\begin{aligned} \langle \boldsymbol{u}, \boldsymbol{v} \rangle &= u_1 v_1 + \cdots + u_d v_d \\ &= \sum_{i=1}^{d} u_i v_i. \end{aligned}$$

The dot product is an example of an inner product. We'll talk about the more general definition of an inner product later in the course.

The *(Euclidean) norm* of a vector $\boldsymbol{u} \in \mathbb{R}^n$ is defined by

$$\|\boldsymbol{u}\| := \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle} = \left(\sum_{j=1}^{n} u_j^2 \right)^{\frac{1}{2}}$$

Example 1. If $\boldsymbol{u} = \begin{bmatrix} 4 \\ 2 \\ -7 \end{bmatrix}$, then $\|\boldsymbol{u}\| = \sqrt{16 + 4 + 49} = \sqrt{69} \approx 8.3$.

The Cauchy-Schwartz inequality is a key property of inner products:

Theorem 1. *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$, with equality iff $\boldsymbol{x}$ and $\boldsymbol{y}$ are scalar multiples of one another (i.e., they are linearly dependent).*

We say $\boldsymbol{u}$ and $\boldsymbol{v}$ are *orthogonal* if $\boldsymbol{u} \neq \boldsymbol{0} \neq \boldsymbol{v}$ (where $\boldsymbol{0}$ denotes the zero vector), and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$. See Figure 1.

The dot product of $\boldsymbol{u}$ and $\boldsymbol{v}$ can be expressed using matrix multiplication as

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^T \boldsymbol{v}.$$

The product $\boldsymbol{u}\boldsymbol{v}^T$ is a $d \times d$ matrix called the *outer product* of $\boldsymbol{u}$ and $\boldsymbol{v}$.

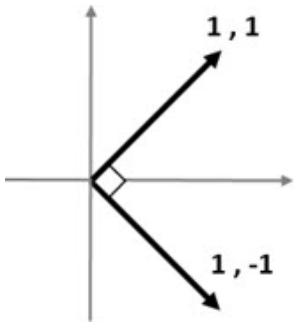


Figure 1: Two orthogonal vectors.

3 Matrix-Vector Multiplication

Let

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^n.$$

By the definition of matrix multiplication,

$$\boldsymbol{Ab} = \begin{bmatrix} \langle \boldsymbol{A}_{1,:}, \boldsymbol{b} \rangle \\ \vdots \\ \langle \boldsymbol{A}_{m,:}, \boldsymbol{b} \rangle \end{bmatrix} \in \mathbb{R}^m,$$

where $\boldsymbol{A}_{i,:}$ denotes the i^{th} row of $\boldsymbol{A}$.

However, there is another (very useful) way to represent $\boldsymbol{Ab}$. Define

$$\boldsymbol{a}_j = \boldsymbol{A}_{:,j},$$

the j^{th} column of $\boldsymbol{A}$. Then

$$\boldsymbol{Ab} = \sum_{j=1}^{n} \underbrace{b_j}_{\text{scalar}} \underbrace{\boldsymbol{a}_j}_{\text{vector}}.$$

To see this, it helps to write

$$\begin{aligned} \sum_j b_j \boldsymbol{a}_j &= b_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + b_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} \\ &= \begin{bmatrix} \sum b_j a_{1j} \\ \vdots \\ \sum b_j a_{mj} \end{bmatrix} \\ &= \boldsymbol{Ab}. \end{aligned}$$

4 Span

Now consider an arbitrary collection of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$. A *linear combination* of these vectors is any vector of the form

$$\sum_{j=1}^{n} x_j \boldsymbol{a}_j,$$

where $x_1, \ldots, x_n \in \mathbb{R}$. The *span* of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ is the set of all linear combinations of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. If we denote by $\boldsymbol{A}$ the matrix whose j^{th} column is $\boldsymbol{a}_j$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} | & & | \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \\ | & & | \end{bmatrix},$$

and we use the insight of the previous section, the *(linear) span* of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ is

$$\text{span}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\} = \{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\}.$$

This collection of vectors is also referred to as the *column span* or *image* of $\boldsymbol{A}$, and denoted colspan($\boldsymbol{A}$). See Figure 2.

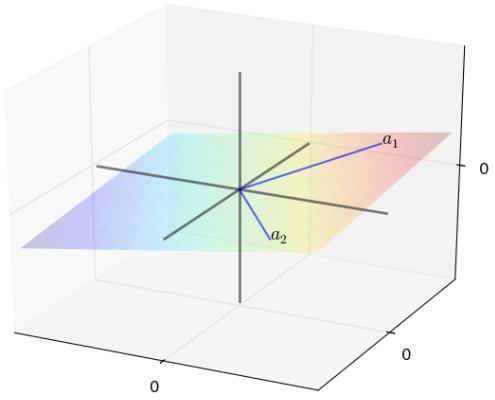


Figure 2: Linear span of two vectors.

5 Linear Independence

Vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$ are *linearly independent* iff the following implication holds:

$$\sum_{j=1}^{n} x_j \boldsymbol{a}_j = \boldsymbol{0} \implies x_j = 0 \; \forall j.$$

We can also think of linear independence in terms of the matrix $\boldsymbol{A}$ as in the previous section. In particular, for any matrix $\boldsymbol{A}$, the *nullspace* of $\boldsymbol{A}$ is the set

$$N(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^m | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

If

$$\boldsymbol{A} = \begin{bmatrix} | & & | \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \\ | & & | \end{bmatrix},$$

then $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are linearly independent iff $N(\boldsymbol{A}) = \{\boldsymbol{0}\}$. This is simply a restatement of the definition.

If a set of vectors is not linearly independent, the vectors are said to be *linearly dependent*.

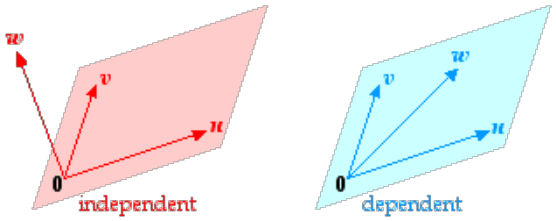


Figure 3: Left: Linearly independent vectors. Right: Linearly dependent vectors.

6 Subspaces

A *subspace* is a set of vectors that is closed under scalar multiplication and vector addition. In particular, a set $S \subseteq \mathbb{R}^m$ is a subspace provided the following two properties hold:

1. For all $\alpha \in \mathbb{R}$ and $\boldsymbol{u} \in S$, $\alpha \boldsymbol{u} \in S$.
2. For all $\boldsymbol{u}, \boldsymbol{v} \in S$, $\boldsymbol{u} + \boldsymbol{v} \in S$.

Example 2. Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Then colspan($\boldsymbol{A}$) is a subspace of $\mathbb{R}^m$, and $N(\boldsymbol{A})$ is a subspace of $\mathbb{R}^n$.

A *basis* for a subspace $S \subseteq \mathbb{R}^d$ is a finite collection of vectors $B \subseteq S$ satisfying any one of the following equivalent properties:

1. B is a minimal spanning set (i.e., a spanning set that is no longer a spanning set if some element is removed)
2. B is a maximal linearly independent set (i.e., a linearly independent set that is no longer a linearly independent set if any vector is added)
3. span(B) $= S$ and B is a linearly independent set
4. Every element of S can be written as a unique linear combination of elements of B.

Example 3. Consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The vectors $\boldsymbol{a}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\boldsymbol{a}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ are a basis of the image of $\boldsymbol{A}$. To see this, we can use the 3^{rd} characterization of bases: The vectors are clearly linearly independent. Furthermore, they span colspan($\boldsymbol{A}$). To see this, observe that $\boldsymbol{a}_3 = \boldsymbol{a}_1 + \boldsymbol{a}_2$. Then for any $\boldsymbol{x}$,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + x_3\boldsymbol{a}_3 \\ &= (x_1 + x_3)\boldsymbol{a}_1 + (x_2 + x_3)\boldsymbol{a}_2. \\ &\in \text{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2\}. \end{aligned}$$

By the third characterization of bases, these two vectors constitute a basis.

The *dimension* of a subspace is the cardinality of a basis. The set $\{\boldsymbol{0}\}$ is a subspace, and its dimension is taken to be zero.

Example 4. $\dim(\mathbb{R}^n) = n$, with the *standard basis*

$$\boldsymbol{e}_i, = [0, \ldots, 0, \underbrace{1}_{i^{th} \text{ position}}, 0, \ldots, 0]^\intercal, \qquad i = 1, \ldots, n$$

as one possible basis.

7 Rank and Nullity

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, define the *rank* of $\boldsymbol{A}$

$$\text{rank}(\boldsymbol{A}) = \dim(\text{colspan}(\boldsymbol{A})),$$

and the *nullity* of $\boldsymbol{A}$

$$\text{nullity}(\boldsymbol{A}) = \dim(N(\boldsymbol{A})).$$

If $\text{rank}(\boldsymbol{A}) = \min(m, n)$, $\boldsymbol{A}$ is said to have *full rank*. If $n < m$, then $\boldsymbol{A}$ has full rank iff the columns of $\boldsymbol{A}$ are linearly independent. In this case, we say $\boldsymbol{A}$ has *full column rank*.

Rank and nullity are related by the following result:

Theorem 2 (Rank-Nullity Theorem)**.** *If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, then rank$(\boldsymbol{A})$ + nullity$(\boldsymbol{A}) = n$.*

Example 5. In the example of the previous section, $\text{rank}(\boldsymbol{A}) = 2$ and therefore $\text{nullity}(\boldsymbol{A}) = 1$.

8 Inverses

A $d \times d$ square matrix $\boldsymbol{A}$ is invertible when there exists another square matrix $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$, the identity matrix. In this case the matrix $\boldsymbol{B}$ is denoted $\boldsymbol{A}^{-1}$. The matrix $\boldsymbol{A}$ is invertible iff any one of the following equivalent conditions hold:

1. $\text{rank}(\boldsymbol{A}) = d$
2. The columns of $\boldsymbol{A}$ are linearly independent
3. $\text{nullity}(\boldsymbol{A}) = 0$
4. If $\boldsymbol{Ax} = \boldsymbol{0}$, then $\boldsymbol{x} = \boldsymbol{0}$.
5. $\det(\boldsymbol{A}) \neq 0$
6. For all $\boldsymbol{b} \in \mathbb{R}^d$, the equation $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution.

9 Orthogonal and Orthonormal Sets

The nonzero vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ comprise an *orthogonal set* if $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0$ for all $i \neq j$. If in addition $\|\boldsymbol{u}_i\| = 1$ for all i, we say $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ is an *orthonormal set*.

Example 6.

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

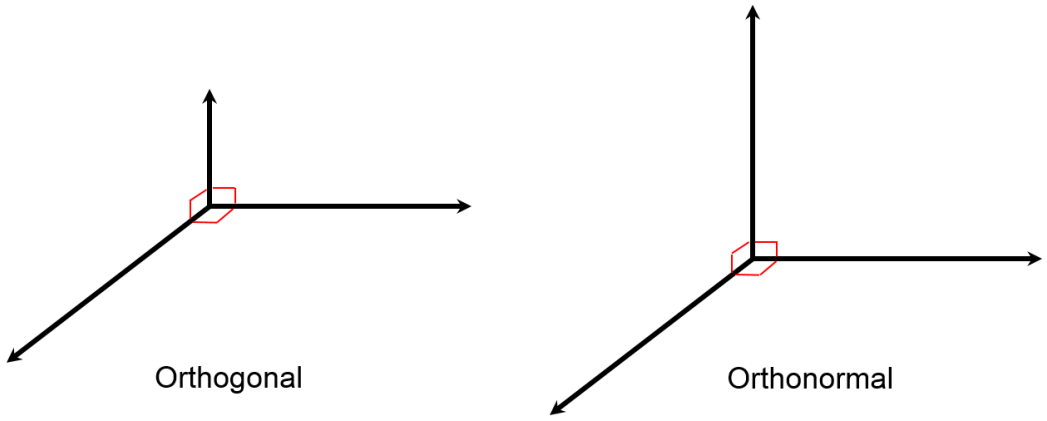
are orthonormal.

Every orthogonal set is linearly independent. To see this, suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are orthogonal and $c_1\boldsymbol{u}_1 + \cdots + c_n\boldsymbol{u}_n = \boldsymbol{0}$ for some scalars $c_1, \ldots, c_n \in \mathbb{R}$. Then for each i

$$\begin{aligned} 0 = \langle \boldsymbol{0}, \boldsymbol{u}_i \rangle &= \langle c_1\boldsymbol{u}_1 + \cdots + c_n\boldsymbol{u}_n, \boldsymbol{u}_i \rangle \\ &= c_i \|\boldsymbol{u}_i\|^2 \end{aligned}$$

which implies $c_i = 0$ since $\|\boldsymbol{u}_i\| > 0$.

If $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathbb{R}^n$ are an orthogonal set, they are a basis of $\mathbb{R}^n$, by the second characterization of basis: they are LI and clearly maximal since $\dim(\mathbb{R}^n) = n$.



If $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ is an orthonormal set and a basis of a subspace S, it is refered to as an *orthonormal basis* (ONB) of that subspace.

An *orthogonal matrix* is a square matrix $\boldsymbol{U}$ such that $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. That is, the transpose of $\boldsymbol{U}$ is its inverse. Note that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ implies the columns of $\boldsymbol{U}$ are an ONB, and $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ implies the rows of $\boldsymbol{U}$ are also an ONB.

Example 7. The rotation matrix $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ is an orthogonal matrix.

10 Eigenvalues and Eigenvectors

Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. If

$$\boldsymbol{A}\boldsymbol{u} = \lambda\boldsymbol{u}$$

for some $\lambda \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$, we say λ is an *eigenvalue* of $\boldsymbol{A}$ and $\boldsymbol{u}$ is a corresponding *eigenvector*. If $\boldsymbol{A}$ is *symmetric*, i.e., $\boldsymbol{A} = \boldsymbol{A}^T$, we can characterize $\boldsymbol{A}$ in terms of its eigenvalues and eigenvectors.

Theorem 3 (Spectral Theorem)**.** *If $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is symmetric, then*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

where $\boldsymbol{U}$ is an orthogonal matrix with real entries and $\boldsymbol{\Lambda}$ is a diagonal matrix with real entries.

What is the connection to eigenvalues and eigenvectors? Multiplying $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ on the right by $\boldsymbol{U}$, we have

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}.$$

Let

$$\boldsymbol{U} = \begin{bmatrix} | & & | \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_d \\ | & & | \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{bmatrix}.$$

If we look at the matrix equation $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}$ one column at a time, we have

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i, i = 1, \ldots, n.$$

Thus, λ_i are eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{u}_i$ are corresponding eigenvectors. Since $\boldsymbol{U}$ is an orthogonal matrix, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ are an orthonormal basis of $\mathbb{R}^d$. Thus, the spectral theorem implies the existence of an ONB consisting of eigenvectors of $\boldsymbol{A}$.

An important identity is

$$\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T.$$

This can be verified just by calculating entries on both sides and checking that they agree. For example, the $(1,1)$ entry is $\sum_{i=1}^{d} \lambda_i u_{i1}^2$. This identity expresses $\boldsymbol{A}$ as a sum of rank 1 outer products. $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ and $\sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ are referred to as the *spectral* or *eigenvalue decomposition* of $\boldsymbol{A}$. $\lambda_1, \ldots, \lambda_d$ constitute the *spectrum* of $\boldsymbol{A}$.

The determinant of a matrix is the product of its eigenvalues, while the trace of a matrix (the sum of the diagonal entries) is the sum of its eigenvalues.

11 Positive (Semi-)Definite Matrices

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a square matrix. We say

- $\boldsymbol{A}$ is *positive semi-definite* if $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$
- $\boldsymbol{A}$ is *positive definite* if $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$.

Clearly if $\boldsymbol{A}$ is PD, it is also PSD.

PD and PSD matrices arise frequently in machine learning. For example,

- Gram matrices, which have the form

$$\begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle & \cdots & \langle \boldsymbol{x}_1, \boldsymbol{x}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}_n, \boldsymbol{x}_1 \rangle & \cdots & \langle \boldsymbol{x}_n, \boldsymbol{x}_n \rangle \end{bmatrix}$$

  for some vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \mathbb{R}^d$
- Covariance matrices
- Kernel matrices
- Hessian matrices of convex functions.

All of the above example are symmetric, and all of the PD/PSD matrices we will encounter will be symmetric. By the spectral theorem, such matrices have an eigenvalue decomposition with real eigenvalues. The following properties characterize PSD and PD matrices in terms of their spectrum.

- $\boldsymbol{A}$ is PSD iff $\lambda_i \geq 0 \ \forall i$.
- $\boldsymbol{A}$ is PD iff $\lambda_i > 0 \ \forall i$.

12 Orthogonal Complements

Let $S \subseteq \mathbb{R}^d$ be a subspace. The set

$$S^\perp := \{\boldsymbol{v} \mid \langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0 \text{ for all } \boldsymbol{u} \in S\}$$

is called the orthogonal complement of S. It can easily be shown to be a subspace itself. See Figure 4.

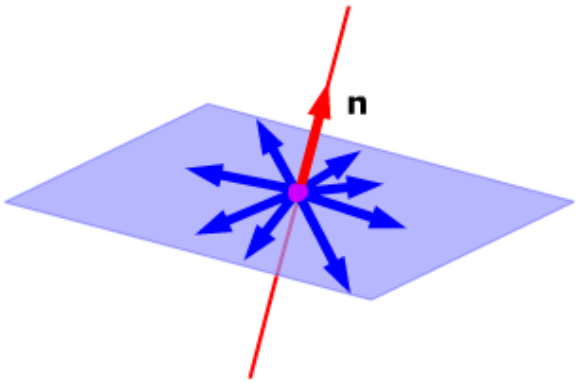


Figure 4: Illustration of orthogonal complement

13 Projections

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k \in \mathbb{R}^d$ be linearly independent column vectors. Denote

$$\boldsymbol{A} = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_k] \qquad (d \times k).$$

The linear span of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ is the column span of $\boldsymbol{A}$, written $\langle \boldsymbol{A} \rangle$ for brevity.

The *projection* onto the subspace $S = \langle \boldsymbol{A} \rangle$ is the mapping $\boldsymbol{\Pi}_A$ that sends $\boldsymbol{x} \in \mathbb{R}^d$ to the closest point in $\langle \boldsymbol{A} \rangle$. Every point in $\langle \boldsymbol{A} \rangle$ equals $\boldsymbol{A\theta}$ for some $\boldsymbol{\theta} \in \mathbb{R}^k$. Therefore, $\boldsymbol{\Pi}_A \boldsymbol{x} = \boldsymbol{A}\widehat{\boldsymbol{\theta}}$, where

$$\widehat{\boldsymbol{\theta}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^k} \|\boldsymbol{x} - \boldsymbol{A\theta}\|^2 .$$

The solution is

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{x}$$

where $(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T$ is called the pseudoinverse of $\boldsymbol{A}$. Since $\boldsymbol{A}$ has full rank, $\boldsymbol{A}^T \boldsymbol{A}$ is invertible (exercise). Therefore,

$$\boldsymbol{\Pi}_A \boldsymbol{x} = \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{x}.$$

Since $\boldsymbol{x}$ was arbitrary, we conclude that $\boldsymbol{\Pi}_A$ is the $d \times d$ matrix

$$\boldsymbol{\Pi}_A := \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T,$$

which is called the *projection matrix* for $\boldsymbol{A}$. Projection matrices are symmetric and positive-semidefinite.

If $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ are orthonormal, i.e.,

$$\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases},$$

then

$$\boldsymbol{\Pi}_A = \boldsymbol{A}\boldsymbol{A}^T \quad (d \times d).$$

Since the projection depends only on the subspace S and not the matrix used to represent it, we can assume the columns of $\boldsymbol{A}$ are orthonormal without loss of generality.

The *orthogonality principle* states that $\forall \boldsymbol{x}, \boldsymbol{x} - \boldsymbol{\Pi}_A \boldsymbol{x}$ is orthogonal to every element of $\langle \boldsymbol{A} \rangle$. That is, $\boldsymbol{x} - \boldsymbol{\Pi}_A \boldsymbol{x} \in \langle \boldsymbol{A} \rangle^\perp$ (see Figure 5 (b)). To see this, let $\boldsymbol{A\theta}$ denote an arbitary element of $\langle \boldsymbol{A} \rangle$. Then

$$\begin{aligned} \langle \boldsymbol{A\theta}, \boldsymbol{x} - \boldsymbol{\Pi}_A \boldsymbol{x} \rangle &= \boldsymbol{\theta}^T \boldsymbol{A}^T (\boldsymbol{x} - \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{x}) \\ &= \boldsymbol{\theta}^T (\boldsymbol{A}^T \boldsymbol{x} - \boldsymbol{A}^T \boldsymbol{x}) \\ &= \boldsymbol{0}. \end{aligned}$$

The *projection theorem* states that for any subspace $S \subseteq \mathbb{R}^d$, $\mathbb{R}^d = S \bigoplus S^\perp$. In other words, every $\boldsymbol{x} \in \mathbb{R}^d$ can be written uniquely as $\boldsymbol{u} + \boldsymbol{v}$ where $\boldsymbol{u} \in S$ and $\boldsymbol{v} \in S^\perp$. Existence of $\boldsymbol{u}$ and $\boldsymbol{v}$ follows from the orthogonality principle by taking $\boldsymbol{u} = \boldsymbol{\Pi}_A \boldsymbol{x}$ (where $\boldsymbol{A}$ is such that $S = \langle \boldsymbol{A} \rangle$) and $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{u}$. To see uniqueness, suppose $\boldsymbol{x} = \boldsymbol{u}' + \boldsymbol{v}'$. Then $\boldsymbol{u} - \boldsymbol{u}' = \boldsymbol{v}' - \boldsymbol{v}$. But the only common element of S and $S^\perp$ is the zero vector, so $\boldsymbol{u} = \boldsymbol{u}'$ and $\boldsymbol{v} = \boldsymbol{v}'$.

Projection matrices are *idempotent*, which means $\boldsymbol{\Pi}_A^2 = \boldsymbol{\Pi}_A$. This follows from

$$\begin{aligned}
\boldsymbol{\Pi}_A^2 &= \boldsymbol{\Pi}_A \cdot \boldsymbol{\Pi}_A \\
&= \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \cdot \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \\
&= \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \\
&= \boldsymbol{\Pi}_A.
\end{aligned}$$

Intuitively, the second projection has no effect because $\boldsymbol{\Pi}_A \boldsymbol{x} \in \langle \boldsymbol{A} \rangle$ already.

Remark 1. In the above treatment of projections, we defined projections in terms of the closest point property, from which we derived the orthogonality condition and the projection theorem. Projections can be defined in a much more general setting, namely projections onto closed subspaces of Hilbert spaces. In this more general setting, one first proves the projection theorem and uses it to define projections, and then deduces the orthogonality principle and closest point property as corollaries. Projection in Hilbert space is a very important concept in applied mathematics and data analysis.

Exercises

1. (★) Prove the Pythagorean theorem: If $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$, then $\|\boldsymbol{u} + \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$.
2. (★) Verify that the column span and nullspace of a matrix are subspaces.
3. (★) For the matrix in Example 3, find a basis for the nullspace.
4. (★) Show that if $\boldsymbol{u}$ and $\boldsymbol{v}$ are two vectors, then the outer product $\boldsymbol{u}\boldsymbol{v}^T$ has rank one.
5. (★★) Show that the sum of k rank-one matrices has rank at most k.
6. (★★) Let $k < d$. Show that if $\boldsymbol{A} \in \mathbb{R}^{d \times k}$ is full rank, then $\boldsymbol{A}^T\boldsymbol{A}$ is invertible.
7. (★★) Use the projection theorem to prove the rank plus nullity theorem. *Hint:* The row-rank of a matrix is equal to the column rank, which you may assume.
8. (★) Show that if λ is an eigenvalue of a matrix, then the set of all associated eigenvectors constitute a subspace. This subspace is called the *eigenspace* of λ and its dimension is called the (geometric) multiplicity of λ.
9. (a) (★) Show that if $\boldsymbol{U}$ is an orthogonal matrix, then for all $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\| = \|\boldsymbol{U}\boldsymbol{x}\|$, where the norm is the Euclidean norm.

   (b) (★★) Show that all 2×2 orthogonal matrices have the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

   for some θ. Give a geometric interpretation of the effect of these two transformations.

(c) (★★) An ellipse in $\mathbb{R}^2$ can be expressed in the form

$$\left\{ \boldsymbol{x} \,\middle|\, (\boldsymbol{x}-\boldsymbol{c})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{c}) = r^2 \right\},$$

where $\boldsymbol{c} \in \mathbb{R}^2$, $r > 0$, and $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Lambda}$ is diagonal and positive definite. Choose $\boldsymbol{c}, r, \boldsymbol{U}$, and $\boldsymbol{\Lambda}$ such that the major axis has length 3, the minor axis has length 1, and the center of the ellipse is at the point $[3 \; -1]^T$, and the major axis makes an angle of $+\pi/6$ radians with the positive x-axis.

10. (★★) Let $\boldsymbol{A}$ be a symmetric matrix. Show that $\boldsymbol{A}$ is invertible iff all of its eigenvalues are nonzero, and express the spectral decomposition of $\boldsymbol{A}^{-1}$ in terms of the spectral decomposition of $\boldsymbol{A}$. Conclude that if $\boldsymbol{A}$ is PD, then so is $\boldsymbol{A}^{-1}$.

11. Show that the following types of matrices are PSD:

    (a) (★) Any matrix of the form $\boldsymbol{A}^T\boldsymbol{A}$, where $\boldsymbol{A}$ is an arbitrary matrix.

    (b) (★) Covariance matrices, i.e., matrices of the form $\mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T]$, where $\boldsymbol{X}$ is a random column vector.

    (c) (★★) Gram matrices.

12. (☆☆) Show that a Gram matrix is PD iff the associated vectors are linearly independent.

13. (☆☆) Prove the Cauchy-Schwartz inequality by considering the determinant of the 2×2 Gram matrix associated to $\boldsymbol{x}$ and $\boldsymbol{y}$.

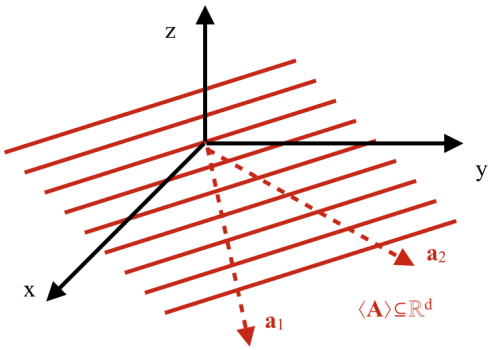
14. (☆) Show that the matrix

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

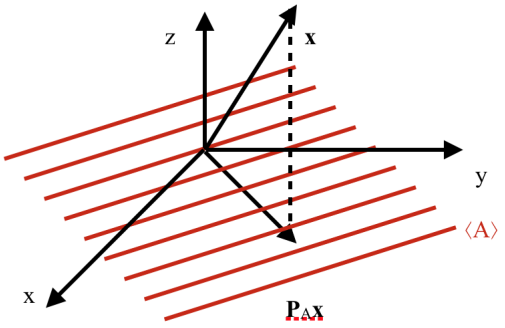
is positive definite but not symmetric.

15. (★★) Describe the eigenvalue decomposition of the projection matrix for a subspace S. What are the eigenvalues, what are their multiplicities, and what are the associated eigenspaces? (See an earlier problem for definitions of multiplicity and eigenspace).

16. (★★) Show that $(S^\perp)^\perp = S$, and use this fact together with the projection theorem to show that the projection onto $S^\perp$ is given by $\boldsymbol{I} - \boldsymbol{\Pi}_A$, where $S = \langle \boldsymbol{A} \rangle$.



Column span of $\boldsymbol{A}$ where $d = 3, k = 2$.



Projection of $\boldsymbol{x}$ onto $\langle \boldsymbol{A} \rangle$.

Figure 5: Linear span and projection