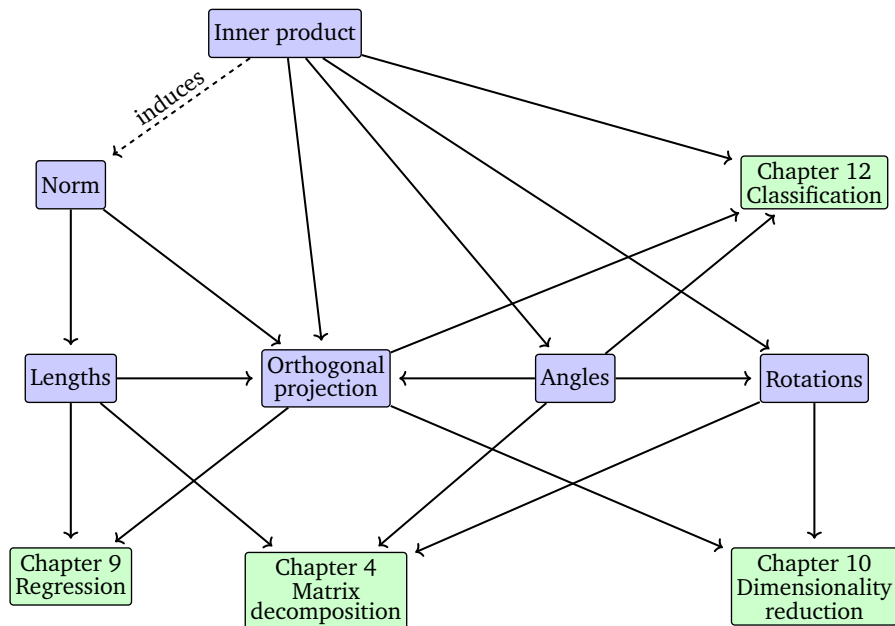


Analytic Geometry

In Chapter 2, we studied vectors, vector spaces and linear mappings at a general but abstract level. In this chapter, we will add some geometric interpretation and intuition to all of these concepts. In particular, we will look at geometric vectors, compute their lengths and distances or angles between two vectors. To be able to do this, we equip the vector space with an inner product that induces the geometry of the vector space. Inner products and their corresponding norms and metrics capture the intuitive notions of similarity and distances, which we use to develop the Support Vector Machine in Chapter 12. We will then use the concepts of lengths and angles between vectors to discuss orthogonal projections, which will play a central role when we discuss principal component analysis in Chapter 10 and regression via maximum likelihood estimation in Chapter 9. Figure 3.1 gives an overview of how concepts in this chapter are related and how they are connected to other chapters of the book.

Figure 3.1 A mind map of the concepts introduced in this chapter, along with when they are used in other parts of the book.



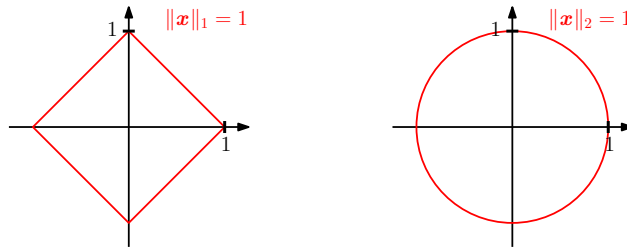


Figure 3.3 For different norms, the red lines indicate the set of vectors with norm 1. Left: Manhattan norm; Right: Euclidean distance.

3.1 Norms

When we think of geometric vectors, i.e., directed line segments that start at the origin, then intuitively the length of a vector is the distance of the “end” of this directed line segment from the origin. In the following, we will discuss the notion of the length of vectors using the concept of a norm.

Definition 3.1 (Norm). A norm on a vector space V is a function

norm

$$\|\cdot\| : V \rightarrow \mathbb{R}, \quad (3.1)$$

$$x \mapsto \|x\|, \quad (3.2)$$

which assigns each vector x its *length* $\|x\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $x, y \in V$ the following hold:

- Absolutely homogeneous: $\|\lambda x\| = |\lambda| \|x\|$
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$
- Positive definite: $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$.

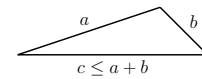
Triangle inequality

Positive definite

In geometric terms, the triangle inequality states that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side; see Figure 3.2 for an illustration.

Recall that for a vector $x \in \mathbb{R}^n$ we denote the elements of the vector using a subscript, that is x_i is the i^{th} element of the vector x .

Figure 3.2 Triangle inequality.



Example 3.1 (Manhattan Norm)

The *Manhattan norm* on \mathbb{R}^n is defined for $x \in \mathbb{R}^n$ as

Manhattan norm

$$\|x\|_1 := \sum_{i=1}^n |x_i|, \quad (3.3)$$

where $|\cdot|$ is the absolute value. The left panel of Figure 3.3 indicates all vectors $x \in \mathbb{R}^2$ with $\|x\|_1 = 1$. The Manhattan norm is also called ℓ_1 norm.

ℓ_1 norm

Example 3.2 (Euclidean Norm)

The length of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}, \quad (3.4)$$

Euclidean distance
Euclidean norm
 ℓ_2 norm

which computes the *Euclidean distance* of \mathbf{x} from the origin. This norm is called the *Euclidean norm*. The right panel of Figure 3.3 shows all vectors $\mathbf{x} \in \mathbb{R}^2$ with $\|\mathbf{x}\|_2 = 1$. The Euclidean norm is also called ℓ_2 norm.

1569 *Remark.* Throughout this book, we will use the Euclidean norm (3.4) by
1570 default if not stated otherwise. \diamond

Cauchy-Schwarz
inequality

Remark (Inner Products and Norms). Every inner product induces a norm, but there are norms (like the ℓ_1 norm) without a corresponding inner product. For an inner product vector space $(V, \langle \cdot, \cdot \rangle)$ the induced norm $\|\cdot\|$ satisfies the *Cauchy-Schwarz inequality*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (3.5)$$

1571

\diamond

1572 3.2 Inner Products

1573 Inner products allow for the introduction of intuitive geometrical con-
1574 cepts, such as the length of a vector and the angle or distance between
1575 two vectors. A major purpose of inner products is to determine whether
1576 vectors are orthogonal to each other.

1577 3.2.1 Dot Product

scalar product
dot product

We may already be familiar with a particular type of inner product, the *scalar product/dot product* in \mathbb{R}^n , which is given by

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i. \quad (3.6)$$

1578 We will refer to the particular inner product above as the dot product
1579 in this book. However, inner products are more general concepts with
1580 specific properties, which we will now introduce.

1581 3.2.2 General Inner Products

bilinear mapping

Recall the linear mapping from Section 2.7, where we can rearrange the mapping with respect to addition and multiplication with a scalar. A *bilinear*

mapping Ω is a mapping with two arguments, and it is linear in each argument, i.e., when we look at a vector space V then it holds that for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V, \lambda \in \mathbb{R}$

$$\Omega(\lambda \mathbf{x} + \mathbf{y}, \mathbf{z}) = \lambda \Omega(\mathbf{x}, \mathbf{z}) + \Omega(\mathbf{y}, \mathbf{z}) \quad (3.7)$$

$$\Omega(\mathbf{x}, \lambda \mathbf{y} + \mathbf{z}) = \lambda \Omega(\mathbf{x}, \mathbf{y}) + \Omega(\mathbf{x}, \mathbf{z}). \quad (3.8)$$

Equation (3.7) asserts that Ω is linear in the first argument, and equation (3.8) asserts that Ω is linear in the second argument.

Definition 3.2. Let V be a vector space and $\Omega : V \times V \rightarrow \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- Ω is called *symmetric* if $\Omega(\mathbf{x}, \mathbf{y}) = \Omega(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in V$, i.e., the order of the arguments does not matter. symmetric
- Ω is called *positive definite* if positive definite

$$\forall \mathbf{x} \in V \setminus \{\mathbf{0}\} : \Omega(\mathbf{x}, \mathbf{x}) > 0, \quad \Omega(\mathbf{0}, \mathbf{0}) = 0 \quad (3.9)$$

Definition 3.3. Let V be a vector space and $\Omega : V \times V \rightarrow \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- A positive definite, symmetric bilinear mapping $\Omega : V \times V \rightarrow \mathbb{R}$ is called an *inner product* on V . We typically write $\langle \mathbf{x}, \mathbf{y} \rangle$ instead of $\Omega(\mathbf{x}, \mathbf{y})$. inner product
 - The pair $(V, \langle \cdot, \cdot \rangle)$ is called an *inner product space* or (real) *vector space with inner product*. If we use the dot product defined in (3.6), we call $(V, \langle \cdot, \cdot \rangle)$ a *Euclidean vector space*. inner product space
vector space with inner product
Euclidean vector space
- We will refer to the spaces above as inner product spaces in this book. space

Example 3.3 (Inner Product that is not the Dot Product)

Consider $V = \mathbb{R}^2$. If we define

$$\langle \mathbf{x}, \mathbf{y} \rangle := x_1 y_1 - (x_1 y_2 + x_2 y_1) + 2x_2 y_2 \quad (3.10)$$

then $\langle \cdot, \cdot \rangle$ is an inner product but different from the dot product. The proof will be an exercise.

3.2.3 Symmetric, Positive Definite Matrices

Symmetric, positive definite matrices play an important role in machine learning, and they are defined via the inner product.

Consider an n -dimensional vector space V with an inner product $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ (see Definition 3.3) and an ordered basis $B = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ of V . Recall the definition of a basis (Section 2.6.1), which says that any vector can be written as a linear combination of basis vectors. That is, for any vectors $\mathbf{x} \in V$ and $\mathbf{y} \in V$ we can write them as $\mathbf{x} = \sum_{i=1}^n \psi_i \mathbf{b}_i \in V$

and $\mathbf{y} = \sum_{j=1}^n \lambda_j \mathbf{b}_j \in V$, for $\psi_i, \lambda_j \in \mathbb{R}$. Due to the bilinearity of the inner product it holds that for all \mathbf{x} and \mathbf{y} that

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{i=1}^n \psi_i \mathbf{b}_i, \sum_{j=1}^n \lambda_j \mathbf{b}_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \psi_i \langle \mathbf{b}_i, \mathbf{b}_j \rangle \lambda_j = \hat{\mathbf{x}}^\top \mathbf{A} \hat{\mathbf{y}}, \quad (3.11)$$

where $A_{ij} := \langle \mathbf{b}_i, \mathbf{b}_j \rangle$ and $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ are the coordinates of \mathbf{x} and \mathbf{y} with respect to the basis B . This implies that the inner product $\langle \cdot, \cdot \rangle$ is uniquely determined through \mathbf{A} . The symmetry of the inner product also means that \mathbf{A} is symmetric. Furthermore, the positive definiteness of the inner product implies that

$$\forall \mathbf{x} \in V \setminus \{\mathbf{0}\} : \mathbf{x}^\top \mathbf{A} \mathbf{x} > 0. \quad (3.12)$$

1599 **Definition 3.4** (Symmetric, positive definite matrix). A symmetric matrix
 symmetric, positive definite 1600 $\mathbf{A} \in \mathbb{R}^{n \times n}$ that satisfies (3.12) is called *symmetric, positive definite* or
 positive definite 1601 just *positive definite*. If only \geq holds in (3.12) then \mathbf{A} is called *symmetric,*
 symmetric, positive 1602 *positive semi-definite*.

Example 3.4 (Symmetric, Positive Definite Matrices)

Consider the following matrices:

$$\mathbf{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix} \quad (3.13)$$

Then, \mathbf{A}_1 is positive definite because it is symmetric and

$$\mathbf{x}^\top \mathbf{A}_1 \mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (3.14)$$

$$= 9x_1^2 + 12x_1x_2 + 5x_2^2 = (3x_1 + 2x_2)^2 + x_2^2 > 0 \quad (3.15)$$

for all $\mathbf{x} \in V \setminus \{\mathbf{0}\}$. However, \mathbf{A}_2 is symmetric but not positive definite because $\mathbf{x}^\top \mathbf{A}_2 \mathbf{x} = 9x_1^2 + 12x_1x_2 + 3x_2^2 = (3x_1 + 2x_2)^2 - x_2^2$ can be smaller than 0, e.g., for $\mathbf{x} = [2, -3]^\top$.

If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, positive definite then

$$\langle \mathbf{x}, \mathbf{y} \rangle = \hat{\mathbf{x}}^\top \mathbf{A} \hat{\mathbf{y}} \quad (3.16)$$

1603 defines an inner product with respect to an ordered basis B where $\hat{\mathbf{x}}$ and
 1604 $\hat{\mathbf{y}}$ are the coordinate representations of $\mathbf{x}, \mathbf{y} \in V$ with respect to B .

Theorem 3.5. For a real-valued, finite-dimensional vector space V and an ordered basis B of V it holds that $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with

$$\langle \mathbf{x}, \mathbf{y} \rangle = \hat{\mathbf{x}}^\top \mathbf{A} \hat{\mathbf{y}}. \quad (3.17)$$

1605 The following properties hold if $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric and positive
 1606 definite of order n :

- The null space (kernel) of \mathbf{A} consists only of $\mathbf{0}$ because $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. This implies that $\mathbf{A} \mathbf{x} \neq \mathbf{0}$ if $\mathbf{x} \neq \mathbf{0}$.
- The diagonal elements of \mathbf{A} are positive because $a_{ii} = \mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i > 0$, where \mathbf{e}_i is the i th vector of the standard basis in \mathbb{R}^n .

In Section 4.3, we will return to symmetric, positive definite matrices in the context of matrix decompositions.

3.3 Lengths and Distances

In Section 3.1, we already discussed norms that we can use to compute the length of a vector. Inner products and norms are closely related in the sense that any inner product induces a norm

Inner products induce norms.

$$\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \quad (3.18)$$

in a natural way, such that we can compute lengths of vectors using the inner product. However, not every norm is induced by an inner product. The Manhattan norm (3.3) is an example of a norm that is not induced by an inner product. In the following, we will focus on norms that are induced by inner products and introduce geometric concepts, such as lengths, distances and angles.

Example 3.5 (Length of Vectors using Inner Products)

In geometry, we are often interested in lengths of vectors. We can now use an inner product to compute them using (3.18). Let us take $\mathbf{x} = [1, 1]^\top \in \mathbb{R}^2$. If we use the dot product as the inner product, with (3.18) we obtain

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{1^2 + 1^2} = \sqrt{2} \quad (3.19)$$

as the length of \mathbf{x} . Let us now choose a different inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \mathbf{y} = x_1 y_1 - \frac{1}{2}(x_1 y_2 + x_2 y_1) + x_2 y_2. \quad (3.20)$$

If we compute the norm of a vector, then this inner product returns smaller values than the dot product if x_1 and x_2 have the same sign (and $x_1 x_2 > 0$), otherwise it returns greater values than the dot product. With this inner product we obtain

$$\langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 - x_1 x_2 + x_2^2 = 1 - 1 + 1 = 1 \implies \|\mathbf{x}\| = \sqrt{1} = 1, \quad (3.21)$$

such that \mathbf{x} is “shorter” with this inner product than with the dot product.

Definition 3.6 (Distance and Metric). Consider an inner product space $(V, \langle \cdot, \cdot \rangle)$. Then

$$d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\| = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle} \quad (3.22)$$

distance

Euclidean distance

is called *distance* of $\mathbf{x}, \mathbf{y} \in V$. If we use the dot product as the inner product, then the distance is called *Euclidean distance*. The mapping

$$d : V \times V \rightarrow \mathbb{R} \quad (3.23)$$

$$(\mathbf{x}, \mathbf{y}) \mapsto d(\mathbf{x}, \mathbf{y}) \quad (3.24)$$

metric

1620 is called *metric*.

1621 *Remark.* Similar to the length of a vector, the distance between vectors
1622 does not require an inner product: a norm is sufficient. If we have a norm
1623 induced by an inner product, the distance may vary depending on the
1624 choice of the inner product. \diamond

1625 A metric d satisfies:

positive definite

1626 1. d is *positive definite*, i.e., $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in V$ and $d(\mathbf{x}, \mathbf{y}) =$
1627 $0 \iff \mathbf{x} = \mathbf{y}$

symmetric

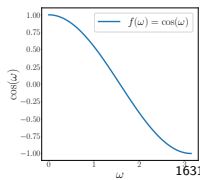
1628 2. d is *symmetric*, i.e., $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in V$.

Triangular
inequality

1629 3. *Triangular inequality*: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

3.4 Angles and Orthogonality

1630 **Figure 3.4** When
restricted to $[0, \pi]$
then $f(\omega) = \cos(\omega)$
returns a unique
number in the
interval $[-1, 1]$.



angle

The Cauchy-Schwarz inequality (3.5) allows us to define angles ω in inner product spaces between two vectors \mathbf{x}, \mathbf{y} . Assume that $\mathbf{x} \neq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$. Then

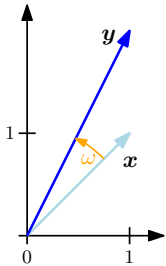
$$-1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \leq 1. \quad (3.25)$$

Therefore, there exists a unique $\omega \in [0, \pi]$ with

$$\cos \omega = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}, \quad (3.26)$$

see Figure 3.4 for an illustration. The number ω is the *angle* between the vectors \mathbf{x} and \mathbf{y} . Intuitively, the angle between two vectors tells us how similar their orientations are. For example, using the dot product, the angle between \mathbf{x} and $\mathbf{y} = 4\mathbf{x}$, i.e., \mathbf{y} is a scaled version of \mathbf{x} , is 0: Their orientation is the same.

1635 **Figure 3.5** The
angle ω between
two vectors \mathbf{x}, \mathbf{y} is
computed using the
inner product.



Example 3.6 (Angle between Vectors)

Let us compute the angle between $\mathbf{x} = [1, 1]^\top \in \mathbb{R}^2$ and $\mathbf{y} = [1, 2]^\top \in \mathbb{R}^2$, see Figure 3.5, where we use the dot product as the inner product. Then we get

$$\cos \omega = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle}} = \frac{\mathbf{x}^\top \mathbf{y}}{\sqrt{\mathbf{x}^\top \mathbf{x} \mathbf{y}^\top \mathbf{y}}} = \frac{3}{\sqrt{10}}, \quad (3.27)$$

and the angle between the two vectors is $\arccos(\frac{3}{\sqrt{10}}) \approx 0.32$ rad, which corresponds to about 18° .

The inner product also allows us to characterize vectors that are orthogonal.

Definition 3.7 (Orthogonality). Two vectors \mathbf{x} and \mathbf{y} are *orthogonal* if and only if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and we write $\mathbf{x} \perp \mathbf{y}$. If additionally $\|\mathbf{x}\| = 1 = \|\mathbf{y}\|$, i.e., the vectors are unit vectors, then \mathbf{x} and \mathbf{y} are *orthonormal*.

An implication of this definition is that the $\mathbf{0}$ -vector is orthogonal to every vector in the vector space.

Remark. Orthogonality is the generalization of the concept of perpendicularity to bilinear forms that do not have to be the dot product. In our context, geometrically, we can think of orthogonal vectors as having a right angle with respect to a specific inner product. \diamond

Example 3.7 (Orthogonal Vectors)

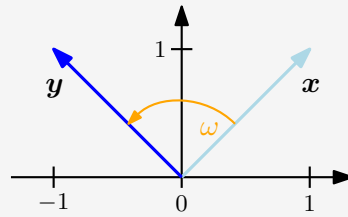


Figure 3.6 The angle ω between two vectors \mathbf{x}, \mathbf{y} can change depending on the inner product.

Consider two vectors $\mathbf{x} = [1, 1]^\top, \mathbf{y} = [-1, 1]^\top \in \mathbb{R}^2$, see Figure 3.6. We are interested in determining the angle ω between them using two different inner products. Using the dot product as inner product yields an angle ω between \mathbf{x} and \mathbf{y} of 90° , such that $\mathbf{x} \perp \mathbf{y}$. However, if we choose the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{y}, \quad (3.28)$$

we get that the angle ω between \mathbf{x} and \mathbf{y} is given by

$$\cos \omega = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} = -\frac{1}{3} \implies \omega \approx 1.91 \text{ rad} \approx 109.5^\circ, \quad (3.29)$$

and \mathbf{x} and \mathbf{y} are not orthogonal. Therefore, vectors that are orthogonal with respect to one inner product do not have to be orthogonal with respect to a different inner product.

Definition 3.8 (Orthogonal Matrix). A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is an *orthogonal matrix* if and only if its columns are orthonormal so that

$$\mathbf{A} \mathbf{A}^\top = \mathbf{I} = \mathbf{A}^\top \mathbf{A}, \quad (3.30)$$

which implies that

$$\mathbf{A}^{-1} = \mathbf{A}^\top, \quad (3.31)$$

It is convention to call these matrices “orthogonal” but a more precise description would be “orthonormal”.

Transformations with orthogonal matrices preserve distances and angles.

i.e., the inverse is obtained by simply transposing the matrix.

Remark. Transformations by orthogonal matrices are special because the length of a vector \mathbf{x} is not changed when transforming it using an orthogonal matrix \mathbf{A} . For the dot product we obtain

$$\|\mathbf{Ax}\|^2 = (\mathbf{Ax})^\top (\mathbf{Ax}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} = \mathbf{x}^\top \mathbf{Ix} = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|^2. \quad (3.32)$$

Moreover, the angle between any two vectors \mathbf{x}, \mathbf{y} , as measured by their inner product, is also unchanged when transforming both of them using an orthogonal matrix \mathbf{A} . Assuming the dot product as the inner product, the angle of the images \mathbf{Ax} and \mathbf{Ay} is given as

$$\cos \omega = \frac{(\mathbf{Ax})^\top (\mathbf{Ay})}{\|\mathbf{Ax}\| \|\mathbf{Ay}\|} = \frac{\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ay}}{\sqrt{\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} \mathbf{y}^\top \mathbf{A}^\top \mathbf{Ay}}} = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}, \quad (3.33)$$

which gives exactly the angle between \mathbf{x} and \mathbf{y} . This means that orthogonal matrices \mathbf{A} with $\mathbf{A}^\top = \mathbf{A}^{-1}$ preserve both angles and distances. \diamond

3.5 Orthonormal Basis

In Section 2.6.1, we characterized properties of basis vectors and found that in an n -dimensional vector space, we need n basis vectors, i.e., n vectors that are linearly independent. In Sections 3.3 and 3.4, we used inner products to compute the length of vectors and the angle between vectors. In the following, we will discuss the special case where the basis vectors are orthogonal to each other and where the length of each basis vector is 1. We will call this basis then an orthonormal basis.

Let us introduce this more formally.

Definition 3.9 (Orthonormal basis). Consider an n -dimensional vector space V and a basis $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ of V . If

$$\langle \mathbf{b}_i, \mathbf{b}_j \rangle = 0 \quad \text{for } i \neq j \quad (3.34)$$

$$\langle \mathbf{b}_i, \mathbf{b}_i \rangle = 1 \quad (3.35)$$

orthonormal basis
ONB
orthogonal basis

for all $i, j = 1, \dots, n$ then the basis is called an *orthonormal basis* (ONB). If only (3.34) is satisfied then the basis is called an *orthogonal basis*.

Note that (3.35) implies that every basis vector has length/norm 1. The Gram-Schmidt process (Strang, 2003) is a constructive way to iteratively build an orthonormal basis $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ given a set $\{\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_n\}$ of non-orthogonal and unnormalized basis vectors.

Example 3.8 (Orthonormal basis)

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

In \mathbb{R}^2 , the vectors

$$\mathbf{b}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{b}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (3.36)$$

form an orthonormal basis since $\mathbf{b}_1^\top \mathbf{b}_2 = 0$ and $\|\mathbf{b}_1\| = 1 = \|\mathbf{b}_2\|$.

We will exploit the concept of an orthonormal basis in Chapter 12 and Chapter 10 when we discuss Support Vector Machines and Principal Component Analysis.

3.6 Inner Product of Functions

Thus far, we looked at properties of inner products to compute lengths, angles and distances. We focused on inner products of finite-dimensional vectors.

In the following, we will look at an example of inner products of a different type of vectors: inner products of functions.

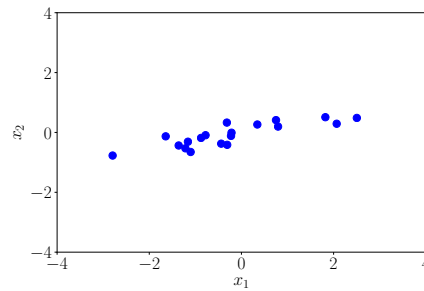
The inner products we discussed so far were defined for vectors with a finite number of entries. We can think of these vectors as discrete functions with a finite number of function values. The concept of an inner product can be generalized to vectors with an infinite number of entries (countably infinite) and also continuous-valued functions (uncountably infinite). Then, the sum over individual components of vectors, see (3.6) for example, turns into an integral.

An inner product of two functions $u : \mathbb{R} \rightarrow \mathbb{R}$ and $v : \mathbb{R} \rightarrow \mathbb{R}$ can be defined as the definite integral

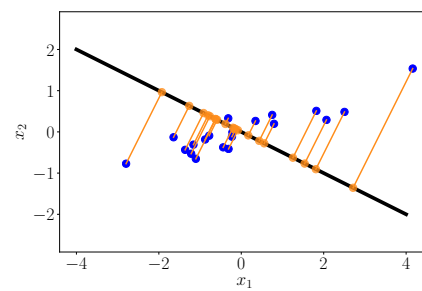
$$\langle u, v \rangle := \int_a^b u(x)v(x)dx \quad (3.37)$$

for lower and upper limits $a, b < \infty$, respectively. As with our usual inner product, we can define norms and orthogonality by looking at the inner product. If (3.37) evaluates to 0, the functions u and v are orthogonal. To make the above inner product mathematically precise, we need to take care of measures, and the definition of integrals. Furthermore, unlike inner product on finite-dimensional vectors, inner products on functions may diverge (have infinite value). Some careful definitions need to be observed, which requires a foray into real and functional analysis which we do not cover in this book.

Figure 3.8
Orthogonal
projection of a
two-dimensional
data set onto a
one-dimensional
subspace.



(a) Original dataset.

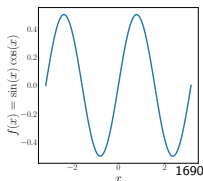


(b) Original data (blue) and their corresponding orthogonal projections (orange) onto a lower-dimensional subspace (straight line).

Example 3.9 (Inner Product of Functions)

If we choose $u = \sin(x)$ and $v = \cos(x)$, the integrand $f(x) = u(x)v(x)$ of (3.37), is shown in Figure 3.7. We see that this function is odd, i.e., $f(-x) = -f(x)$. Therefore, the integral with limits $a = -\pi, b = \pi$ of this product evaluates to 0. Therefore, \sin and \cos are orthogonal functions.

Figure 3.7 $f(x) = \sin(x) \cos(x)$.



Remark. It also holds that the collection of functions

$$\{1, \cos(x), \cos(2x), \cos(3x), \dots\} \quad (3.38)$$

is orthogonal if we integrate from $-\pi$ to π , i.e., any pair of functions are orthogonal to each other. \diamond

In Chapter 6, we will have a look at a second type of unconventional inner products: the inner product of random variables.

3.7 Orthogonal Projections

Projections are an important class of linear transformations (besides rotations and reflections). Projections play an important role in graphics, coding theory, statistics and machine learning. In machine learning, we often deal with data that is high-dimensional. High-dimensional data is often hard to analyze or visualize. However, high-dimensional data quite often possesses the property that only a few dimensions contain most information, and most other dimensions are not essential to describe key properties of the data. When we compress or visualize high-dimensional data we will lose information. To minimize this compression loss, we ideally find the most informative dimensions in the data. Then, we can project the original high-dimensional data onto a lower-dimensional feature space and work in this lower-dimensional space to learn more about the dataset and extract patterns. For example, machine learning algorithms, such as Principal Component Analysis (PCA) by Pearson (1901b);

“Feature” is a commonly used word for “data representation”.

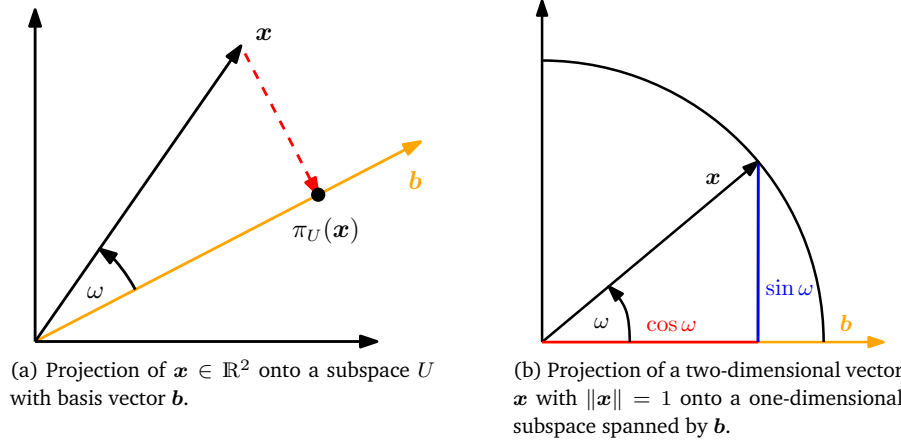


Figure 3.9
Examples of
projections onto
one-dimensional
subspaces.

Hotelling (1933) and Deep Neural Networks (e.g., deep auto-encoders Deng et al. (2010)), heavily exploit the idea of dimensionality reduction. In the following, we will focus on orthogonal projections, which we will use in Chapter 10 for linear dimensionality reduction and in Chapter 12 for classification. Even linear regression, which we discuss in Chapter 9, can be interpreted using orthogonal projections. For a given lower-dimensional subspace, orthogonal projections of high-dimensional data retain as much information as possible and minimize the difference/error between the original data and the corresponding projection. An illustration of such an orthogonal projection is given in Figure 3.8.

Before we detail how to obtain these projections, let us define what a projection actually is.

Definition 3.10 (Projection). Let V be a vector space and $W \subseteq V$ a subspace of V . A linear mapping $\pi : V \rightarrow W$ is called a *projection* if $\pi^2 = \pi \circ \pi = \pi$.

Remark (Projection matrix). Since linear mappings can be expressed by transformation matrices, the definition above applies equally to a special kind of transformation matrices, the *projection matrices* \mathbf{P}_π , which exhibit the property that $\mathbf{P}_\pi^2 = \mathbf{P}_\pi$.

In the following, we will derive orthogonal projections of vectors in the inner product space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ onto subspaces. We will start with one-dimensional subspaces, which are also called *lines*. If not mentioned otherwise, we assume the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ as the inner product.

3.7.1 Projection onto 1-Dimensional Subspaces (Lines)

Assume we are given a line (1-dimensional subspace) through the origin with basis vector $\mathbf{b} \in \mathbb{R}^n$. The line is a one-dimensional subspace $U \subseteq \mathbb{R}^n$

spanned by \mathbf{b} . When we project $\mathbf{x} \in \mathbb{R}^n$ onto U , we want to find the point $\pi_U(\mathbf{x}) \in U$ that is closest to \mathbf{x} . Using geometric arguments, let us characterize some properties of the projection $\pi_U(\mathbf{x})$ (Fig. 3.9 serves as an illustration):

- The projection $\pi_U(\mathbf{x})$ is closest to \mathbf{x} , where “closest” implies that the distance $\|\mathbf{x} - \pi_U(\mathbf{x})\|$ is minimal. It follows that the segment $\pi_U(\mathbf{x}) - \mathbf{x}$ from $\pi_U(\mathbf{x})$ to \mathbf{x} is orthogonal to U and, therefore, the basis \mathbf{b} of U . The orthogonality condition yields $\langle \pi_U(\mathbf{x}) - \mathbf{x}, \mathbf{b} \rangle = 0$ since angles between vectors are defined by means of the inner product.
- The projection $\pi_U(\mathbf{x})$ of \mathbf{x} onto U must be an element of U and, therefore, a multiple of the basis vector \mathbf{b} that spans U . Hence, $\pi_U(\mathbf{x}) = \lambda \mathbf{b}$, for some $\lambda \in \mathbb{R}$.

λ is then the coordinate of $\pi_U(\mathbf{x})$ with respect to \mathbf{b} .

In the following three steps, we determine the coordinate λ , the projection $\pi_U(\mathbf{x}) \in U$ and the projection matrix \mathbf{P}_π that maps arbitrary $\mathbf{x} \in \mathbb{R}^n$ onto U .

1. Finding the coordinate λ . The orthogonality condition yields

$$\langle \mathbf{x} - \pi_U(\mathbf{x}), \mathbf{b} \rangle = 0 \quad (3.39)$$

$$\stackrel{\pi_U(\mathbf{x}) = \lambda \mathbf{b}}{\iff} \langle \mathbf{x} - \lambda \mathbf{b}, \mathbf{b} \rangle = 0. \quad (3.40)$$

With a general inner product, we get $\lambda = \langle \mathbf{x}, \mathbf{b} \rangle$ if $\|\mathbf{b}\| = 1$.

We can now exploit the bilinearity of the inner product and arrive at

$$\langle \mathbf{x}, \mathbf{b} \rangle - \lambda \langle \mathbf{b}, \mathbf{b} \rangle = 0 \quad (3.41)$$

$$\iff \lambda = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{b} \rangle} = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\|\mathbf{b}\|^2} \quad (3.42)$$

If we choose $\langle \cdot, \cdot \rangle$ to be the dot product, we obtain

$$\lambda = \frac{\mathbf{b}^\top \mathbf{x}}{\mathbf{b}^\top \mathbf{b}} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2} \quad (3.43)$$

If $\|\mathbf{b}\| = 1$, then the coordinate λ of the projection is given by $\mathbf{b}^\top \mathbf{x}$.

2. Finding the projection point $\pi_U(\mathbf{x}) \in U$. Since $\pi_U(\mathbf{x}) = \lambda \mathbf{b}$ we immediately obtain with (3.43) that

$$\pi_U(\mathbf{x}) = \lambda \mathbf{b} = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\|\mathbf{b}\|^2} \mathbf{b} = \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2} \mathbf{b}, \quad (3.44)$$

where the last equality holds for the dot product only. We can also compute the length of $\pi_U(\mathbf{x})$ by means of Definition 3.1 as

$$\|\pi_U(\mathbf{x})\| = \|\lambda \mathbf{b}\| = |\lambda| \|\mathbf{b}\|. \quad (3.45)$$

This means that our projection is of length $|\lambda|$ times the length of \mathbf{b} . This also adds the intuition that λ is the coordinate of $\pi_U(\mathbf{x})$ with respect to the basis vector \mathbf{b} that spans our one-dimensional subspace U .

If we use the dot product as an inner product we get

$$\|\pi_U(\mathbf{x})\| \stackrel{(3.44)}{=} \frac{|\mathbf{b}^\top \mathbf{x}|}{\|\mathbf{b}\|^2} \|\mathbf{b}\| \stackrel{(3.26)}{=} |\cos \omega| \|\mathbf{x}\| \|\mathbf{b}\| \frac{\|\mathbf{b}\|}{\|\mathbf{b}\|^2} = |\cos \omega| \|\mathbf{x}\|. \quad (3.46)$$

Here, ω is the angle between \mathbf{x} and \mathbf{b} . This equation should be familiar from trigonometry: If $\|\mathbf{x}\| = 1$ then \mathbf{x} lies on the unit circle. It follows that the projection onto the horizontal axis spanned by \mathbf{b} is exactly $\cos \omega$, and the length of the corresponding vector $\pi_U(\mathbf{x}) = |\cos \omega|$. An illustration is given in Figure 3.9.

The horizontal axis is a one-dimensional subspace.

3. Finding the projection matrix \mathbf{P}_π . We know that a projection is a linear mapping (see Definition 3.10). Therefore, there exists a projection matrix \mathbf{P}_π , such that $\pi_U(\mathbf{x}) = \mathbf{P}_\pi \mathbf{x}$. With the dot product as inner product and

$$\pi_U(\mathbf{x}) = \lambda \mathbf{b} = \mathbf{b} \lambda = \mathbf{b} \frac{\mathbf{b}^\top \mathbf{x}}{\|\mathbf{b}\|^2} = \frac{\mathbf{b} \mathbf{b}^\top}{\|\mathbf{b}\|^2} \mathbf{x} \quad (3.47)$$

we immediately see that

$$\mathbf{P}_\pi = \frac{\mathbf{b} \mathbf{b}^\top}{\|\mathbf{b}\|^2}. \quad (3.48)$$

Note that $\mathbf{b} \mathbf{b}^\top$ is a symmetric matrix (with rank 1) and $\|\mathbf{b}\|^2 = \langle \mathbf{b}, \mathbf{b} \rangle$ is a scalar.

Projection matrices are always symmetric.

The projection matrix \mathbf{P}_π projects any vector $\mathbf{x} \in \mathbb{R}^n$ onto the line through the origin with direction \mathbf{b} (equivalently, the subspace U spanned by \mathbf{b}).

Remark. The projection $\pi_U(\mathbf{x}) \in \mathbb{R}^n$ is still an n -dimensional vector and not a scalar. However, we no longer require n coordinates to represent the projection, but only a single one if we want to express it with respect to the basis vector \mathbf{b} that spans the subspace U : λ . \diamond

Example 3.10 (Projection onto a Line)

Find the projection matrix \mathbf{P}_π onto the line through the origin spanned by $\mathbf{b} = \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}^\top$. \mathbf{b} is a direction and a basis of the one-dimensional subspace (line through origin).

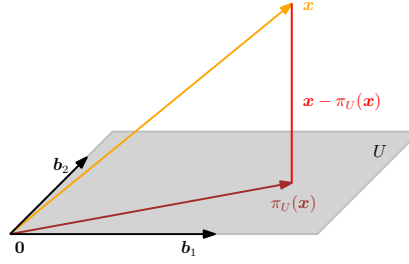
With (3.48), we obtain

$$\mathbf{P}_\pi = \frac{\mathbf{b} \mathbf{b}^\top}{\mathbf{b}^\top \mathbf{b}} = \frac{1}{9} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 4 \end{bmatrix}. \quad (3.49)$$

Let us now choose a particular \mathbf{x} and see whether it lies in the subspace

Figure 3.10

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



spanned by \mathbf{b} . For $\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$, the projection is

$$\pi_U(\mathbf{x}) = \mathbf{P}_\pi \mathbf{x} = \frac{1}{9} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 5 \\ 10 \\ 10 \end{bmatrix} \in \text{span} \left[\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \right]. \quad (3.50)$$

Note that the application of \mathbf{P}_π to $\pi_U(\mathbf{x})$ does not change anything, i.e., $\mathbf{P}_\pi \pi_U(\mathbf{x}) = \pi_U(\mathbf{x})$. This is expected because according to Definition 3.10 we know that a projection matrix \mathbf{P}_π satisfies $\mathbf{P}_\pi^2 \mathbf{x} = \mathbf{P}_\pi \mathbf{x}$ for all \mathbf{x} .

With the results from Chapter 4 we can show that $\pi_U(\mathbf{x})$ is also an eigenvector of \mathbf{P}_π , and the corresponding eigenvalue is 1.

If U is given by a set of spanning vectors which are not a basis, make sure you determine a basis $\mathbf{b}_1, \dots, \mathbf{b}_m$ before proceeding.

The basis vectors form the columns of $\mathbf{B} \in \mathbb{R}^{n \times m}$, where $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m]$.

3.7.2 Projection onto General Subspaces

In the following, we look at orthogonal projections of vectors $\mathbf{x} \in \mathbb{R}^n$ onto higher-dimensional subspaces $U \subseteq \mathbb{R}^n$ with $\dim(U) = m \geq 1$. An illustration is given in Figure 3.10.

Assume that $(\mathbf{b}_1, \dots, \mathbf{b}_m)$ is an ordered basis of U . Any projection $\pi_U(\mathbf{x})$ onto U is necessarily an element of U . Therefore, they can be represented as linear combinations of the basis vectors $\mathbf{b}_1, \dots, \mathbf{b}_m$ of U , such that $\pi_U(\mathbf{x}) = \sum_{i=1}^m \lambda_i \mathbf{b}_i$.

As in the 1D case, we follow a three-step procedure to find the projection $\pi_U(\mathbf{x})$ and the projection matrix \mathbf{P}_π :

1. Find the coordinates $\lambda_1, \dots, \lambda_m$ of the projection (with respect to the basis of U), such that the linear combination

$$\pi_U(\mathbf{x}) = \sum_{i=1}^m \lambda_i \mathbf{b}_i = \mathbf{B} \boldsymbol{\lambda}, \quad (3.51)$$

$$\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m] \in \mathbb{R}^{n \times m}, \quad \boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]^\top \in \mathbb{R}^m, \quad (3.52)$$

is closest to $\mathbf{x} \in \mathbb{R}^n$. As in the 1D case, “closest” means “minimum distance”, which implies that the vector connecting $\pi_U(\mathbf{x}) \in U$ and $\mathbf{x} \in \mathbb{R}^n$ must be orthogonal to all basis vectors of U . Therefore, we obtain m simultaneous conditions (assuming the dot product as the inner product)

$$\langle \mathbf{b}_1, \mathbf{x} - \pi_U(\mathbf{x}) \rangle = \mathbf{b}_1^\top (\mathbf{x} - \pi_U(\mathbf{x})) = 0 \quad (3.53)$$

$$\vdots \quad (3.54)$$

$$\langle \mathbf{b}_m, \mathbf{x} - \pi_U(\mathbf{x}) \rangle = \mathbf{b}_m^\top (\mathbf{x} - \pi_U(\mathbf{x})) = 0 \quad (3.55)$$

which, with $\pi_U(\mathbf{x}) = \mathbf{B}\boldsymbol{\lambda}$, can be written as

$$\mathbf{b}_1^\top (\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) = 0 \quad (3.56)$$

$$\vdots \quad (3.57)$$

$$\mathbf{b}_m^\top (\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) = 0 \quad (3.58)$$

such that we obtain a homogeneous linear equation system

$$\begin{bmatrix} \mathbf{b}_1^\top \\ \vdots \\ \mathbf{b}_m^\top \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{B}\boldsymbol{\lambda} \end{bmatrix} = \mathbf{0} \iff \mathbf{B}^\top (\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) = \mathbf{0} \quad (3.59)$$

$$\iff \mathbf{B}^\top \mathbf{B}\boldsymbol{\lambda} = \mathbf{B}^\top \mathbf{x}. \quad (3.60)$$

The last expression is called *normal equation*. Since $\mathbf{b}_1, \dots, \mathbf{b}_m$ are a basis of U and, therefore, linearly independent, $\mathbf{B}^\top \mathbf{B} \in \mathbb{R}^{m \times m}$ is regular and can be inverted. This allows us to solve for the coefficients/coordinates

normal equation

$$\boldsymbol{\lambda} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{x}. \quad (3.61)$$

The matrix $(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$ is also called the *pseudo-inverse* of \mathbf{B} , which can be computed for non-square matrices \mathbf{B} . It only requires that $\mathbf{B}^\top \mathbf{B}$ is positive definite, which is the case if \mathbf{B} is full rank.

pseudo-inverse

2. Find the projection $\pi_U(\mathbf{x}) \in U$. We already established that $\pi_U(\mathbf{x}) = \mathbf{B}\boldsymbol{\lambda}$. Therefore, with (3.61)

$$\pi_U(\mathbf{x}) = \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{x}. \quad (3.62)$$

3. Find the projection matrix \mathbf{P}_π . From (3.62) we can immediately see that the projection matrix that solves $\mathbf{P}_\pi \mathbf{x} = \pi_U(\mathbf{x})$ must be

$$\mathbf{P}_\pi = \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top. \quad (3.63)$$

Remark. Comparing the solutions for projecting onto a one-dimensional subspace and the general case, we see that the general case includes the 1D case as a special case: If $\dim(U) = 1$ then $\mathbf{B}^\top \mathbf{B} \in \mathbb{R}$ is a scalar and we can rewrite the projection matrix in (3.63) $\mathbf{P}_\pi = \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$ as $\mathbf{P}_\pi = \frac{\mathbf{B}\mathbf{B}^\top}{\mathbf{B}^\top \mathbf{B}}$, which is exactly the projection matrix in (3.48). \diamond

In practical applications (e.g., linear regression), we often add a “jitter term” $\epsilon \mathbf{I}$ to $\mathbf{B}^\top \mathbf{B}$ to guarantee increased numerical stability and positive definiteness. This “ridge” can be rigorously derived using Bayesian inference. See Chapter 9 for details.

Example 3.11 (Projection onto a Two-dimensional Subspace)

For a subspace $U = \text{span}\left[\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}\right] \subseteq \mathbb{R}^3$ and $\mathbf{x} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^3$ find the

coordinates λ of x in terms of the subspace U , the projection point $\pi_U(x)$ and the projection matrix P_π .

First, we see that the generating set of U is a basis (linear independence) and write the basis vectors of U into a matrix $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

Second, we compute the matrix $B^\top B$ and the vector $B^\top x$ as

$$B^\top B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}, \quad B^\top x = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}. \quad (3.64)$$

Third, we solve the normal equation $B^\top B\lambda = B^\top x$ to find λ :

$$\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix} \iff \lambda = \begin{bmatrix} 5 \\ -3 \end{bmatrix}. \quad (3.65)$$

Fourth, the projection $\pi_U(x)$ of x onto U , i.e., into the column space of B , can be directly computed via

$$\pi_U(x) = B\lambda = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix}. \quad (3.66)$$

projection error

The corresponding *projection error* is the norm of the difference vector between the original vector and its projection onto U , i.e.,

$$\|x - \pi_U(x)\| = \left\| \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}^\top \right\| = \sqrt{6}. \quad (3.67)$$

The projection error is also called the *reconstruction error*.

Fifth, the projection matrix (for any $x \in \mathbb{R}^3$) is given by

$$P_\pi = B(B^\top B)^{-1}B^\top = \frac{1}{6} \begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}. \quad (3.68)$$

To verify the results, we can (a) check whether the displacement vector $\pi_U(x) - x$ is orthogonal to all basis vectors of U , (b) verify that $P_\pi = P_\pi^2$ (see Definition 3.10).

1787 *Remark.* The projections $\pi_U(x)$ are still vectors in \mathbb{R}^n although they lie in
1788 an m -dimensional subspace $U \subseteq \mathbb{R}^n$. However, to represent a projected
1789 vector we only need the m coordinates $\lambda_1, \dots, \lambda_m$ with respect to the
1790 basis vectors b_1, \dots, b_m of U . \diamond

1791 *Remark.* In vector spaces with general inner products, we have to pay
1792 attention when computing angles and distances, which are defined by
1793 means of the inner product. \diamond

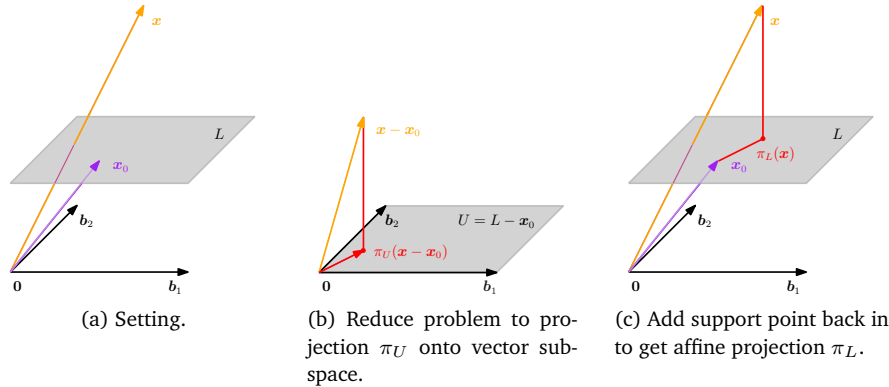


Figure 3.11 Projection onto an affine space. (a) The original setting; (b) The setting is shifted by $-x_0$, so that $x - x_0$ can be projected onto the direction space U ; (c) The projection is translated back to $x_0 + \pi_U(x - x_0)$, which gives the final orthogonal projection $\pi_L(x)$.

Projections allow us to look at situations where we have a linear system $Ax = b$ without a solution. Recall that this means that b does not lie in the span of A , i.e., the vector b does not lie in the subspace spanned by the columns of A . Given that the linear equation cannot be solved exactly, we can find an *approximate solution*. The idea is to find the vector in the subspace spanned by the columns of A that is closest to b , i.e., we compute the orthogonal projection of b onto the subspace spanned by the columns of A . This problem arises often in practice, and the solution is called the *least squares solution* (assuming the dot product as the inner product) of an overdetermined system. This is discussed further in Chapter 9.

Remark. We just looked at projections of vectors x onto a subspace U with basis vectors $\{b_1, \dots, b_k\}$. If this basis is an ONB, i.e., (3.34)–(3.35) are satisfied, the projection equation (3.62) simplifies greatly to

$$\pi_U(x) = BB^T x \quad (3.69)$$

since $B^T B = I$ with coordinates

$$\lambda = B^T x. \quad (3.70)$$

This means that we no longer have to compute the tedious inverse from (3.62), which saves us much computation time. \diamond

We can find approximate solutions to unsolvable linear equation systems using projections.

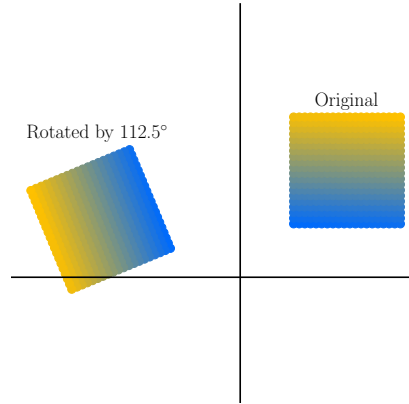
least squares solution

3.7.3 Projection onto Affine Subspaces

Thus far, we discussed how to project a vector onto a lower-dimensional subspace U . In the following, we provide a solution to projecting a vector onto an affine subspace.

Consider the setting in Figure 3.11(a). We are given an affine space $L = x_0 + U$ where b_1, b_2 are basis vectors of U . To determine the orthogonal projection $\pi_L(x)$ of x onto L , we transform the problem into a problem that we know how to solve: the projection onto a vector subspace. In

Figure 3.12 A rotation rotates objects in a plane about the origin. If the rotation angle is positive, we rotate counterclockwise.



order to get there, we subtract the support point \mathbf{x}_0 from \mathbf{x} and from L , so that $L - \mathbf{x}_0 = U$ is exactly the vector subspace U . We can now use the orthogonal projections onto a subspace we discussed in Section 3.7.2 and obtain the projection $\pi_U(\mathbf{x} - \mathbf{x}_0)$, which is illustrated in Figure 3.11(b). This projection can now be translated back into L by adding \mathbf{x}_0 , such that we obtain the orthogonal projection onto an affine space L as

$$\pi_L(\mathbf{x}) = \mathbf{x}_0 + \pi_U(\mathbf{x} - \mathbf{x}_0), \quad (3.71)$$

1810 where $\pi_U(\cdot)$ is the orthogonal projection onto the subspace U , i.e., the
1811 direction space of L , see Figure 3.11(c).

From Figure 3.11 it is also evident that the distance of \mathbf{x} from the affine space L is identical to the distance of $\mathbf{x} - \mathbf{x}_0$ from U , i.e.,

$$d(\mathbf{x}, L) = \|\mathbf{x} - \pi_L(\mathbf{x})\| = \|\mathbf{x} - (\mathbf{x}_0 + \pi_U(\mathbf{x} - \mathbf{x}_0))\| \quad (3.72)$$

$$= d(\mathbf{x} - \mathbf{x}_0, \pi_U(\mathbf{x} - \mathbf{x}_0)). \quad (3.73)$$

3.8 Rotations

1812

1813 Length and angle preservation, as discussed in Section 3.4, are the two
1814 characteristics of linear mappings with orthogonal transformation matri-
1815 ces. In the following, we will have a closer look at specific orthogonal
1816 transformation matrices, which describe rotations.

rotation

A *rotation* is a linear mapping (more specifically, an automorphism of a Euclidean vector space) that rotates a plane by an angle θ about the origin, i.e., the origin is a fixed point. For a positive angle $\theta > 0$, by common convention, we rotate in a counterclockwise direction. An example is shown in Figure 3.12, where the transformation matrix is

$$\mathbf{R} = \begin{bmatrix} -0.38 & -0.92 \\ 0.92 & -0.38 \end{bmatrix}. \quad (3.74)$$

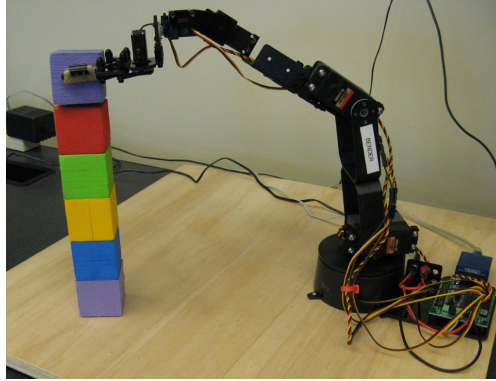


Figure 3.13 The robotic arm needs to rotate its joints in order to pick up objects or to place them correctly. Figure taken from (Deisenroth et al., 2015).

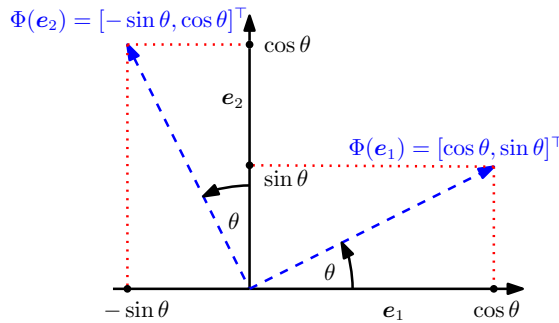


Figure 3.14 Rotation of the standard basis in \mathbb{R}^2 by an angle θ .

Important application areas of rotations include computer graphics and robotics. For example, in robotics, it is often important to know how to rotate the joints of a robotic arm in order to pick up or place an object, see Figure 3.13.

3.8.1 Rotations in \mathbb{R}^2

Consider the standard basis $\left\{ e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ of \mathbb{R}^2 , which defines the standard coordinate system in \mathbb{R}^2 . We aim to rotate this coordinate system by an angle θ as illustrated in Figure 3.14. Note that the rotated vectors are still linearly independent and, therefore, are a basis of \mathbb{R}^2 . This means that the rotation performs a basis change.

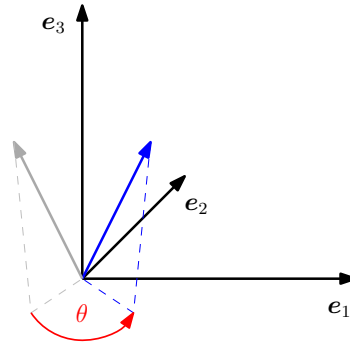
Rotations Φ are linear mappings so that we can express them by a *rotation matrix* $R(\theta)$. Trigonometry (see Figure 3.14) allows us to determine the coordinates of the rotated axes (the image of Φ) with respect to the standard basis in \mathbb{R}^2 . We obtain

$$\Phi(e_1) = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}, \quad \Phi(e_2) = \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix}. \quad (3.75)$$

Therefore, the rotation matrix that performs the basis change into the

rotation matrix

Figure 3.15
Rotation of a vector
(gray) in \mathbb{R}^3 by an
angle θ about the
 e_3 -axis. The rotated
vector is shown in
blue.



rotated coordinates $\mathbf{R}(\theta)$ is given as

$$\mathbf{R}(\theta) = [\Phi(e_1) \quad \Phi(e_2)] = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}. \quad (3.76)$$

3.8.2 Rotations in \mathbb{R}^3

1827

1828 In contrast to the \mathbb{R}^2 case, in \mathbb{R}^3 we can rotate any two-dimensional plane
1829 about a one-dimensional axis. The easiest way to specify the general rota-
1830 tion matrix is to specify how the images of the standard basis e_1, e_2, e_3 are
1831 supposed to be rotated, and making sure these images $\mathbf{R}e_1, \mathbf{R}e_2, \mathbf{R}e_3$ are
1832 orthonormal to each other. We can then obtain a general rotation matrix
1833 \mathbf{R} by combining the images of the standard basis.

1834

To have a meaningful rotation angle we have to define what “counterclockwise” means when we operate in more than two dimensions. We
1835 use the convention that a “counterclockwise” (planar) rotation about an
1836 axis refers to a rotation about an axis when we look at the axis “head on,
1837 from the end toward the origin”. In \mathbb{R}^3 , there are therefore three (planar)
1838 rotations about the three standard basis vectors (see Figure 3.15):
1839

- Rotation about the e_1 -axis

$$\mathbf{R}_1(\theta) = [\Phi(e_1) \quad \Phi(e_2) \quad \Phi(e_3)] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}. \quad (3.77)$$

1840

Here, the e_1 coordinate is fixed, and the counterclockwise rotation is
1841 performed in the e_2e_3 plane.

- Rotation about the e_2 -axis

$$\mathbf{R}_2(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix}. \quad (3.78)$$

1842

If we rotate the e_1e_3 plane about the e_2 axis, we need to look at the e_2
1843 axis from its “tip” toward the origin.

- Rotation about the e_3 -axis

$$\mathbf{R}_3(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (3.79)$$

Figure 3.15 illustrates this.

3.8.3 Rotations in n Dimensions

The generalization of rotations from 2D and 3D to n -dimensional Euclidean vector spaces can be intuitively described as fixing $n - 2$ dimensions and restrict the rotation to a two-dimensional plane in the n -dimensional space. As in the three-dimensional case we can rotate any plane (two-dimensional subspace of \mathbb{R}^n).

Definition 3.11 (Givens Rotation). Let V be an n -dimensional Euclidean vector space and $\Phi : V \rightarrow V$ an automorphism with transformation matrix

$$\mathbf{R}_{ij}(\theta) := \begin{bmatrix} \mathbf{I}_{i-1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cos \theta & \mathbf{0} & -\sin \theta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{j-i-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sin \theta & \mathbf{0} & \cos \theta & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{I}_{n-j} \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad (3.80)$$

for $1 \leq i < j \leq n$ and $\theta \in \mathbb{R}$. Then $\mathbf{R}_{ij}(\theta)$ is called a *Givens rotation*. Essentially, $\mathbf{R}_{ij}(\theta)$ is the identity matrix \mathbf{I}_n with

$$r_{ii} = \cos \theta, \quad r_{ij} = -\sin \theta, \quad r_{ji} = \sin \theta, \quad r_{jj} = \cos \theta. \quad (3.81)$$

In two dimensions (i.e., $n = 2$), we obtain (3.75) as a special case.

3.8.4 Properties of Rotations

Rotations exhibit a number useful properties:

- Rotations preserve distances, i.e., $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{R}_\theta(\mathbf{x}) - \mathbf{R}_\theta(\mathbf{y})\|$. In other words, rotations leave the distance between any two points unchanged after the transformation.
- Rotations preserve angles, i.e., the angle between $\mathbf{R}_\theta \mathbf{x}$ and $\mathbf{R}_\theta \mathbf{y}$ equals the angle between \mathbf{x} and \mathbf{y} .
- Rotations in three (or more) dimensions are generally not commutative. Therefore, the order in which rotations are applied is important, even if they rotate about the same point. Only in two dimensions vector rotations are commutative, such that $\mathbf{R}(\phi)\mathbf{R}(\theta) = \mathbf{R}(\theta)\mathbf{R}(\phi)$ for all $\phi, \theta \in [0, 2\pi)$, and form an Abelian group (with multiplication) only if they rotate about the same point (e.g., the origin).

3.9 Further Reading

In this chapter, we gave a brief overview of some of the important concepts of analytic geometry, which we will use in later chapters of the book. For a broader and more in-depth overview of some the concepts we presented we refer to the following excellent books: Axler (2015) and Boyd and Vandenberghe (2018).

Inner products allow us to determine specific bases of vector (sub)spaces, where each vector is orthogonal to all others (orthogonal bases) using the Gram-Schmidt method. These bases are important in optimization and numerical algorithms for solving linear equation systems. For instance, Krylov subspace methods, such as Conjugate Gradients or GMRES, minimize residual errors that are orthogonal to each other (Stoer and Burlirsch, 2002).

In machine learning, inner products are important in the context of kernel methods (Schölkopf and Smola, 2002). Kernel methods exploit the fact that many linear algorithms can be expressed purely by inner product computations. Then, the “kernel trick” allows us to compute these inner products implicitly in a (potentially infinite-dimensional) feature space, without even knowing this feature space explicitly. This allowed the “non-linearization” of many algorithms used in machine learning, such as kernel-PCA (Schölkopf et al., 1997) for dimensionality reduction. Gaussian processes (Rasmussen and Williams, 2006) also fall into the category of kernel methods and are the current state-of-the-art in probabilistic regression (fitting curves to data points). The idea of kernels is explored further in Chapter 12.

Projections are often used in computer graphics, e.g., to generate shadows. In optimization, orthogonal projections are often used to (iteratively) minimize residual errors. This also has applications in machine learning, e.g., in linear regression where we want to find a (linear) function that minimizes the residual errors, i.e., the lengths of the orthogonal projections of the data onto the linear function (Bishop, 2006). We will investigate this further in Chapter 9. PCA (Hotelling, 1933; Pearson, 1901b) also uses projections to reduce the dimensionality of high-dimensional data. We will discuss this in more detail in Chapter 10.

Exercises

3.1 Show that $\langle \cdot, \cdot \rangle$ defined for all $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ in \mathbb{R}^2 by:

$$\langle \mathbf{x}, \mathbf{y} \rangle := x_1 y_1 - (x_1 y_2 + x_2 y_1) + 2(x_2 y_2)$$

is an inner product.

3.2 Consider \mathbb{R}^2 with $\langle \cdot, \cdot \rangle$ defined for all \mathbf{x} and \mathbf{y} in \mathbb{R}^2 as:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \underbrace{\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}}_{=: \mathbf{A}} \mathbf{y}$$

1901 Is $\langle \cdot, \cdot \rangle$ an inner product?

3.3 Consider the Euclidean vector space \mathbb{R}^5 with the dot product. A subspace $U \subseteq \mathbb{R}^5$ and $\mathbf{x} \in \mathbb{R}^5$ are given by

$$U = \text{span} \left[\begin{bmatrix} 0 \\ -1 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \\ 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} -3 \\ 4 \\ 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -3 \\ 5 \\ 0 \\ 7 \end{bmatrix} \right], \quad \mathbf{x} = \begin{bmatrix} -1 \\ -9 \\ -1 \\ 4 \\ 1 \end{bmatrix}$$

1902 1. Determine the orthogonal projection $\pi_U(\mathbf{x})$ of \mathbf{x} onto U

1903 2. Determine the distance $d(\mathbf{x}, U)$

3.4 Consider \mathbb{R}^3 with the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \mathbf{y}.$$

1904 Furthermore, we define $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ as the standard/canonical basis in \mathbb{R}^3 .

1. Determine the orthogonal projection $\pi_U(\mathbf{e}_2)$ of \mathbf{e}_2 onto

$$U = \text{span}[\mathbf{e}_1, \mathbf{e}_3].$$

1905 Hint: Orthogonality is defined through the inner product.

1906 2. Compute the distance $d(\mathbf{e}_2, U)$.

1907 3. Draw the scenario: standard basis vectors and $\pi_U(\mathbf{e}_2)$

1908 3.5 Prove the Cauchy-Schwarz inequality $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$ for $\mathbf{x}, \mathbf{y} \in V$, where
1909 V is a vector space.