3

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.

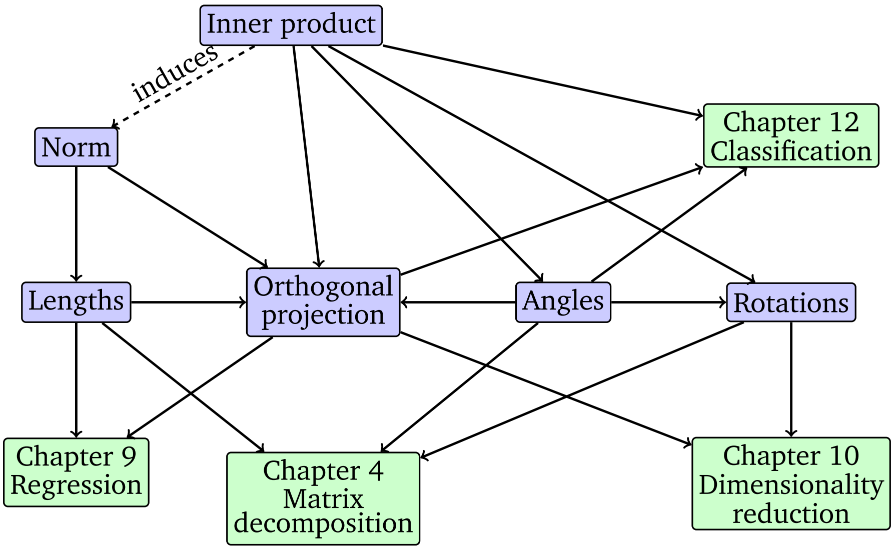


Figure 3.1 A mind map of the concepts introduced in this chapter, along with when they are used in other parts of the book.

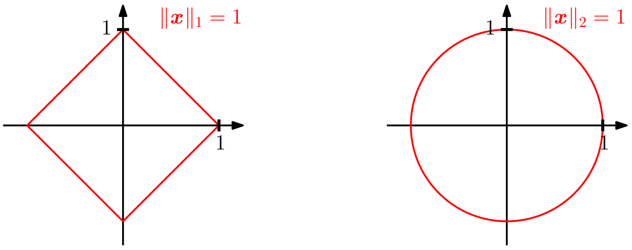


Figure 3.3 For different norms, the red lines indicate the set of vectors with norm 1. Left: Manhattan distance; Right: Euclidean distance.

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.

norm

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

$$\|\cdot\| : V \to \mathbb{R}\,, \tag{3.1}$$

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|\,, \tag{3.2}$$

length

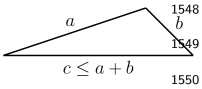
which assigns each vector $\boldsymbol{x}$ its *length* $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

Triangle inequality

Positive definite

- Absolutely homogeneous: $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$
- *Triangle inequality*: $\|\boldsymbol{x}+\boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- *Positive definite*: $\|\boldsymbol{x}\| \geqslant 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$.

Figure 3.2 Triangle inequality.



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

Manhattan distance

Example 3.1 (Manhattan Distance)
The *Manhattan distance* on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}$ as

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|\,, \tag{3.3}$$

ℓ_1 norm

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

Example 3.2 (Euclidean Norm)

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

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}\,, \tag{3.4}$$

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

Euclidean distance

Euclidean norm

ℓ_2 norm

Remark. Throughout this book, we will use the Euclidean norm (3.4) by default if not stated otherwise. ♢

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*

Cauchy-Schwarz inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leqslant \|\boldsymbol{x}\| \|\boldsymbol{y}\|\,. \tag{3.5}$$

♢

3.2 Inner Products

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

3.2.1 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

scalar product

dot product

$$\boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i\,. \tag{3.6}$$

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

3.2.2 General Inner Products

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*

bilinear mapping

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 $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V, \lambda \in \mathbb{R}$

$$\Omega(\lambda \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{z}) + \Omega(\boldsymbol{y}, \boldsymbol{z}) \tag{3.7}$$

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

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 $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

symmetric

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

positive definite

- Ω is called *positive definite* if

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

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

inner product

inner product space

vector space with inner product

Euclidean vector space

- A positive definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an *inner product* on V. We typically write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ instead of $\Omega(\boldsymbol{x}, \boldsymbol{y})$.
- 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*.

We will refer to the spaces above as inner product spaces in this book.

Example 3.3 (Inner Product that is not the Dot Product)
Consider $V = \mathbb{R}^2$. If we define

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

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 (SPD) 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 $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ (see Definition 3.3) and an ordered basis $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ 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 $\boldsymbol{x} \in V$ and $\boldsymbol{y} \in V$ we can write them as $\boldsymbol{x} = \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i \in V$

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

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

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

$$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} > 0 \,. \tag{3.12}$$

A symmetric matrix $\boldsymbol{A}$ that satisfies (3.12) is called *positive definite*.

positive definite

Example 3.4 (SPD Matrices)
Consider the following matrices:

$$\boldsymbol{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix} \tag{3.13}$$

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

$$\boldsymbol{x}^\top \boldsymbol{A}_1 \boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.14}$$

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

for all $\boldsymbol{x} \in V \setminus \{\boldsymbol{0}\}$. However, $\boldsymbol{A}_2$ is symmetric but not positive definite because $\boldsymbol{x}^\top \boldsymbol{A}_2 \boldsymbol{x} = 9x_1^2 + 12x_1x_2 + 4x_2^2 = (3x_1 + 2x_2)^2 - x_2^2$ can be smaller than 0, e.g., for $\boldsymbol{x} = [2, -3]^\top$.

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\top \boldsymbol{A} \hat{\boldsymbol{y}} \tag{3.16}$$

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

Theorem 3.4. *For a real-valued, finite-dimensional vector space V and an ordered basis B of V it holds that $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\top \boldsymbol{A} \hat{\boldsymbol{y}} \,. \tag{3.17}$$

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

- The nullspace (kernel) of $\boldsymbol{A}$ consists only of $\boldsymbol{0}$ because $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$. This implies that $\boldsymbol{A}\boldsymbol{x} \neq 0$ if $\boldsymbol{x} \neq 0$.

- The diagonal elements of $\boldsymbol{A}$ are positive because $a_{ii} = \boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{e}_i > 0$, where $\boldsymbol{e}_i$ is the ith 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.

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} \tag{3.18}$$

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 distance (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 $\boldsymbol{x} = [1, 1]^\top \in \mathbb{R}^2$. If we use the dot product as the inner product, with (3.18) we obtain

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{1^2 + 1^2} = \sqrt{2} \tag{3.19}$$

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

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \boldsymbol{x}^\top \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \boldsymbol{y} = x_1y_1 - \frac{1}{2}(x_1y_2 + x_2y_1) + x_2y_2 \,. \tag{3.20}$$

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 $x_1x_2 > 0$), otherwise it returns greater values than the dot product. With this inner product we obtain

$$\langle \boldsymbol{x}, \boldsymbol{x}\rangle = x_1^2 - x_1x_2 + x_2^2 = 1 - 1 + 1 = 1 \implies \|\boldsymbol{x}\| = \sqrt{1} = 1\,, \tag{3.21}$$

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

Definition 3.5 (Distance and Metric)**.** Consider an inner product space $(V, \langle\cdot,\cdot\rangle)$. Then

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

distance
Euclidean distance

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

$$d : V \times V \to \mathbb{R} \tag{3.23}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto d(\boldsymbol{x}, \boldsymbol{y}) \tag{3.24}$$

is called *metric*.

metric

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

A metric d satisfies:

1. d is *positive definite*, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) \geqslant 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$ and $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$
2. d is *symmetric*, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$.
3. *Triangular inequality*: $d(\boldsymbol{x}, \boldsymbol{z}) \leqslant d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$.

positive definite

symmetric

Triangular inequality

3.4 Angles and Orthogonality

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

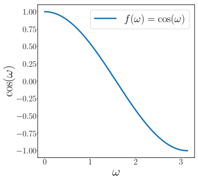
$$-1 \leqslant \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} \leqslant 1\,. \tag{3.25}$$

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

$$\cos \omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|}\,, \tag{3.26}$$

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

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



angle

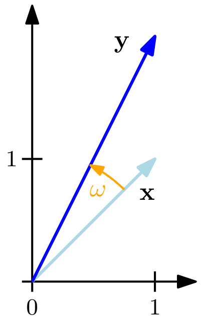
Example 3.6 (Angle between Vectors)

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

$$\cos \omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle}} = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{x} \boldsymbol{y}^\top \boldsymbol{y}}} = \frac{3}{\sqrt{10}}\,, \tag{3.27}$$

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

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



The inner product also allows us to characterize vectors that are most dissimilar, i.e., orthogonal.

orthogonal

Definition 3.6 (Orthogonality)**.** Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, and we write $\boldsymbol{x} \perp \boldsymbol{y}$.

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 to have a right angle with respect to a specific inner product. ♢

Example 3.7 (Orthogonal Vectors)

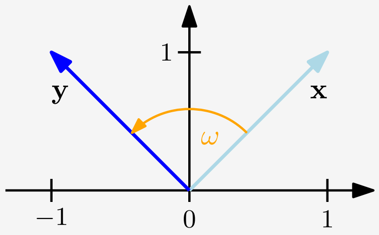


Figure 3.6 The angle ω between two vectors $\boldsymbol{x}, \boldsymbol{y}$ can change depending on the inner product.

Consider two vectors $\boldsymbol{x} = [1, 1]^\top, \boldsymbol{y} = [-1, 1]^\top \in \mathbb{R}^2$, 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ of $90°$, such that $\boldsymbol{x} \perp \boldsymbol{y}$. However, if we choose the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y} \,, \tag{3.28}$$

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

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} = -\frac{1}{3} \implies \omega \approx 1.91 \,\text{rad} \approx 109.5° \,, \tag{3.29}$$

and $\boldsymbol{x}$ and $\boldsymbol{y}$ 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.

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

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 = (\boldsymbol{B}\boldsymbol{x})^\top (\boldsymbol{B}\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{x} = \|\boldsymbol{x}\|^2 \,. \tag{3.30}$$

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

$$\cos\omega = \frac{(\boldsymbol{Bx})^\top(\boldsymbol{By})}{\|\boldsymbol{Bx}\|\,\|\boldsymbol{By}\|} = \frac{\boldsymbol{x}^\top\boldsymbol{B}^\top\boldsymbol{By}}{\sqrt{\boldsymbol{x}^\top\boldsymbol{B}^\top\boldsymbol{Bx}\boldsymbol{y}^\top\boldsymbol{B}^\top\boldsymbol{By}}} = \frac{\boldsymbol{x}^\top\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|}\,, \tag{3.31}$$

which is exactly the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. This means that orthogonal matrices $\boldsymbol{B}$ with $\boldsymbol{B}^\top = \boldsymbol{B}^{-1}$ preserve both angles and distances. ♢

3.5 Inner Products 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 $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ can be defined as the definite integral

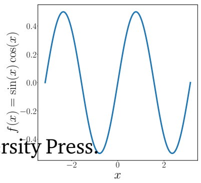
$$\langle u, v\rangle := \int_a^b u(x)v(x)dx \tag{3.32}$$

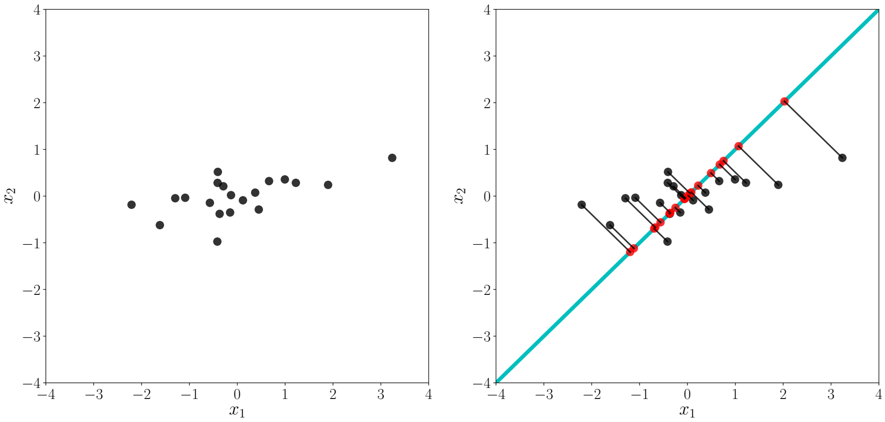
for lower and upper limits $a, b < \infty$, respectively. As with our usual inner product, we can define norms and orthogonality by looking at the inner product. If (3.32) evaluates to 0, the functions u and v are orthogonal. To make the above inner product mathematically precise, we need to take care 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.

Example 3.8 (Inner Product of Functions)

If we choose $u = \sin(x)$ and $v = \cos(x)$, the integrand $f(x) = u(x)v(x)$ of (3.32), 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 $f(x) = \sin(x)\cos(x)$.





(a) Original data set.

(b) Original data points (black) and their corresponding orthogonal projections (red) onto a lower-dimensional subspace (straight line).

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

Remark. It also holds that the collection of functions

$$\{1, \cos(x), \cos(2x), \cos(3x), \dots\} \tag{3.33}$$

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

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

3.6 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 data set and extract patterns. For example, machine learning algorithms, such as Principal Component Analysis (PCA) by Pearson (1901); 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

"Feature" is a commonly used word for "data representation".

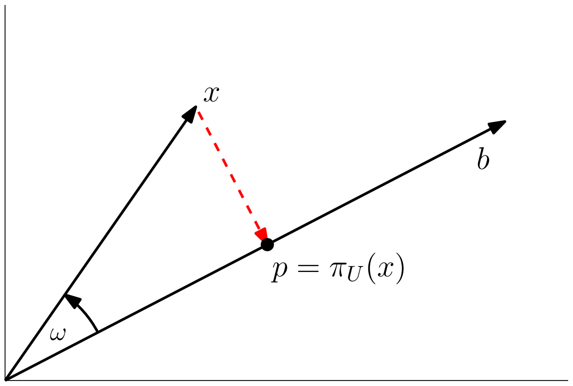


Figure 3.9 Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis $\boldsymbol{b}$.

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

projection

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* $\boldsymbol{P}_\pi$, which exhibit the property that $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$.

projection matrices

Since $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$ it follows that all eigenvalues of $\boldsymbol{P}_\pi$ are either 0 or 1. The corresponding eigenspaces are the kernel and image of the projection, respectively. More details about eigenvalues and eigenvectors are provided in Chapter 4. ♢

A good illustration is given here: `http://tinyurl.com/p5jn5ws`.

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 $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^\top \boldsymbol{y}$ as the inner product.

lines

3.6.1 Projection onto 1-Dimensional Subspaces (Lines)

Assume we are given a line (1-dimensional subspace) through the origin with basis vector $\boldsymbol{b} \in \mathbb{R}^n$. The line is a one-dimensional subspace $U \subseteq \mathbb{R}^n$ spanned by $\boldsymbol{b}$. When we project $\boldsymbol{x} \in \mathbb{R}^n$ onto U, we want to find the point $\pi_U(\boldsymbol{x}) \in U$ that is closest to $\boldsymbol{x}$. Using geometric arguments, let us

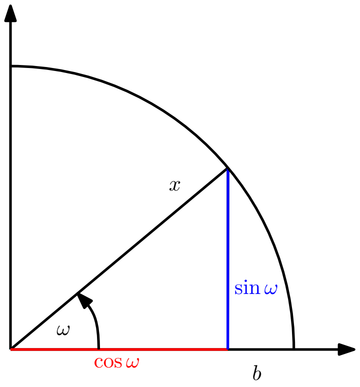


Figure 3.10 Projection of a two-dimensional vector $\boldsymbol{x}$ onto a one-dimensional subspace with $\|\boldsymbol{x}\| = 1$.

characterize some properties of the projection $\pi_U(\boldsymbol{x})$ (Fig. 3.9 serves as an illustration):

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

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

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

1. Finding the coordinate λ. The orthogonality condition yields

$$\langle \boldsymbol{x} - \pi_U(\boldsymbol{x}), \boldsymbol{b}\rangle = 0 \tag{3.34}$$

$$\overset{\pi_U(\boldsymbol{x})=\lambda\boldsymbol{b}}{\iff} \langle \boldsymbol{x} - \lambda\boldsymbol{b}, \boldsymbol{b}\rangle = 0\,. \tag{3.35}$$

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

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

$$\langle \boldsymbol{x}, \boldsymbol{b}\rangle - \lambda\langle \boldsymbol{b}, \boldsymbol{b}\rangle = 0 \tag{3.36}$$

$$\iff \lambda = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\langle \boldsymbol{b}, \boldsymbol{b}\rangle} = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\|\boldsymbol{b}\|^2} \tag{3.37}$$

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

$$\lambda = \frac{\boldsymbol{b}^\top\boldsymbol{x}}{\boldsymbol{b}^\top\boldsymbol{b}} = \frac{\boldsymbol{b}^\top\boldsymbol{x}}{\|\boldsymbol{b}\|^2} \tag{3.38}$$

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

2. Finding the projection point $\pi_U(\boldsymbol{x}) \in U$. Since $\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b}$ we immediately obtain with (3.38) that

$$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b} = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\|\boldsymbol{b}\|^2}\boldsymbol{b} = \frac{\boldsymbol{b}^\top\boldsymbol{x}}{\|\boldsymbol{b}\|^2}\boldsymbol{b}\,, \tag{3.39}$$

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

$$\|\boldsymbol{p}\| = \|\lambda\boldsymbol{b}\| = |\lambda|\,\|\boldsymbol{b}\|\,. \tag{3.40}$$

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

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

$$\|\pi_U(\boldsymbol{x})\| \overset{(3.39)}{=} \frac{|\boldsymbol{b}^\top\boldsymbol{x}|}{\|\boldsymbol{b}\|^2}\,\|\boldsymbol{b}\| \overset{(3.26)}{=} |\cos\omega|\,\|\boldsymbol{x}\|\,\|\boldsymbol{b}\|\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{b}\|^2} = |\cos\omega|\,\|\boldsymbol{x}\|\,. \tag{3.41}$$

Here, ω is the angle between $\boldsymbol{x}$ and $\boldsymbol{b}$. This equation should be familiar from trigonometry: If $\|\boldsymbol{x}\| = 1$ then $\boldsymbol{x}$ lies on the unit circle. It follows that the projection onto the horizontal axis spanned by $\boldsymbol{b}$ is exactly $\cos\omega$, and the length of the corresponding vector $\pi_U(\boldsymbol{x}) = |\cos\omega|$. An illustration is given in Figure 3.10.

The horizontal axis is a one-dimensional subspace.

3. Finding the projection matrix $\boldsymbol{P}_\pi$. We know that a projection is a linear mapping (see Definition 3.7). Therefore, there exists a projection matrix $\boldsymbol{P}_\pi$, such that $\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi\boldsymbol{x}$. With the dot product as inner product and

$$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b} = \boldsymbol{b}\lambda = \boldsymbol{b}\frac{\boldsymbol{b}^\top\boldsymbol{x}}{\|\boldsymbol{b}\|^2} = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}\boldsymbol{x} \tag{3.42}$$

we immediately see that

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}. \tag{3.43}$$

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

Projection matrices are always symmetric.

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

Remark. The projection $\pi_U(\boldsymbol{x}) \in \mathbb{R}^n$ 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 $\boldsymbol{b}$ that spans the subspace U: λ. ♢

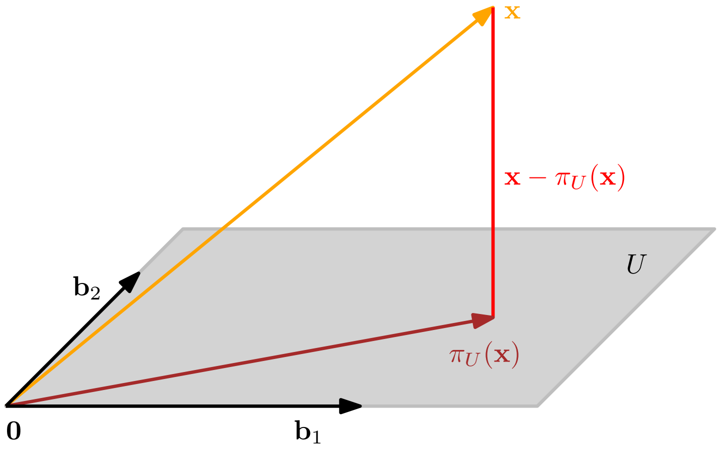


Figure 3.11 Projection onto a two-dimensional subspace U with basis $\boldsymbol{b}_1, \boldsymbol{b}_2$. The projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x} \in \mathbb{R}^3$ onto U can be expressed as a linear combination of $\boldsymbol{b}_1, \boldsymbol{b}_2$ and the displacement vector $\boldsymbol{x} - \pi_U(\boldsymbol{x})$ is orthogonal to both $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.

Example 3.9 (Projection onto a Line)

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

With (3.43), we obtain

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\boldsymbol{b}^\top\boldsymbol{b}} = \frac{1}{9}\begin{bmatrix}1\\2\\2\end{bmatrix}\begin{bmatrix}1 & 2 & 2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}. \tag{3.44}$$

Let us now choose a particular $\boldsymbol{x}$ and see whether it lies in the subspace spanned by $\boldsymbol{b}$. For $\boldsymbol{x} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^\top$, the projected point is

$$\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi\boldsymbol{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in \operatorname{span}[\begin{bmatrix}1\\2\\2\end{bmatrix}]. \tag{3.45}$$

Note that the application of $\boldsymbol{P}_\pi$ to $\pi_U(\boldsymbol{x})$ does not change anything, i.e., $\boldsymbol{P}_\pi\pi_U(\boldsymbol{x}) = \pi_U(\boldsymbol{x})$. This is expected because according to Definition 3.7 we know that a projection matrix $\boldsymbol{P}_\pi$ satisfies $\boldsymbol{P}_\pi^2\boldsymbol{x} = \boldsymbol{P}_\pi\boldsymbol{x}$. Therefore, $\pi_U(\boldsymbol{x})$ is also an eigenvector of $\boldsymbol{P}_\pi$, and the corresponding eigenvalue is 1.

3.6.2 Projection onto General Subspaces

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

In the following, we look at orthogonal projections of vectors $\boldsymbol{x} \in \mathbb{R}^n$ onto higher-dimensional subspaces $U \subseteq \mathbb{R}^n$ with $\dim(U) = m \geqslant 1$. An illustration is given in Figure 3.11.

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

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

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

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

$$\pi_U(\boldsymbol{x}) = \sum_{i=1}^m \lambda_i \boldsymbol{b}_i = \boldsymbol{B\lambda}\,, \tag{3.46}$$

$$\boldsymbol{B} = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_m] \in \mathbb{R}^{n\times m},\ \boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]^\top \in \mathbb{R}^m\,, \tag{3.47}$$

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

$$\langle \boldsymbol{b}_1, \boldsymbol{x} - \pi_U(\boldsymbol{x})\rangle = \boldsymbol{b}_1^\top(\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.48}$$

$$\vdots \tag{3.49}$$

$$\langle \boldsymbol{b}_m, \boldsymbol{x} - \pi_U(\boldsymbol{x})\rangle = \boldsymbol{b}_m^\top(\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.50}$$

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

$$\boldsymbol{b}_1^\top(\boldsymbol{x} - \boldsymbol{B\lambda}) = 0 \tag{3.51}$$

$$\vdots \tag{3.52}$$

$$\boldsymbol{b}_m^\top(\boldsymbol{x} - \boldsymbol{B\lambda}) = 0 \tag{3.53}$$

such that we obtain a homogeneous linear equation system

$$\begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} \begin{bmatrix} \\ \boldsymbol{x} - \boldsymbol{B\lambda} \\ \\ \end{bmatrix} = \boldsymbol{0} \iff \boldsymbol{B}^\top(\boldsymbol{x} - \boldsymbol{B\lambda}) = \boldsymbol{0} \tag{3.54}$$

$$\iff \boldsymbol{B}^\top\boldsymbol{B\lambda} = \boldsymbol{B}^\top\boldsymbol{x}\,. \tag{3.55}$$

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

normal equation

$$\boldsymbol{\lambda} = (\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top\boldsymbol{x}\,. \tag{3.56}$$

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

pseudo-inverse

[1] In practical applications (e.g., linear regression), we often add a “jitter term” $\epsilon\boldsymbol{I}$ to $\boldsymbol{B}^\top\boldsymbol{B}$

2. Find the projection $\pi_U(\boldsymbol{x}) \in U$. We already established that $\pi_U(\boldsymbol{x}) = \boldsymbol{B\lambda}$. Therefore, with (3.56)

$$\pi_U(\boldsymbol{x}) = \boldsymbol{p} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top \boldsymbol{x}\,. \tag{3.57}$$

3. Find the projection matrix $\boldsymbol{P}_\pi$. From (3.57) we can immediately see that the projection matrix that solves $\boldsymbol{P}_\pi \boldsymbol{x} = \pi_U(\boldsymbol{x})$ must be

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top\,. \tag{3.58}$$

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 $\boldsymbol{B}^\top \boldsymbol{B} \in \mathbb{R}$ is a scalar and we can rewrite the projection matrix in (3.58) $\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top$ as $\boldsymbol{P}_\pi = \frac{\boldsymbol{B}\boldsymbol{B}^\top}{\boldsymbol{B}^\top \boldsymbol{B}}$, which is exactly the projection matrix in (3.43). ♢

Example 3.10 (Projection onto a Two-dimensional Subspace)

For a subspace $U = \text{span}[\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0\\1\\2\end{bmatrix}] \subseteq \mathbb{R}^3$ and $\boldsymbol{x} = \begin{bmatrix}6\\0\\0\end{bmatrix} \in \mathbb{R}^3$ find the coordinates $\boldsymbol{\lambda}$ of $\boldsymbol{x}$ in terms of the subspace U, the projection point $\pi_U(\boldsymbol{x})$ and the projection matrix $\boldsymbol{P}_\pi$.

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

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

$$\boldsymbol{B}^\top \boldsymbol{B} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\\1 & 2\end{bmatrix} = \begin{bmatrix}3 & 3\\3 & 5\end{bmatrix}, \quad \boldsymbol{B}^\top \boldsymbol{x} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}6\\0\\0\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix}. \tag{3.59}$$

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

$$\begin{bmatrix}3 & 3\\3 & 5\end{bmatrix}\begin{bmatrix}\lambda_1\\\lambda_2\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix} \iff \boldsymbol{\lambda} = \begin{bmatrix}5\\-3\end{bmatrix}. \tag{3.60}$$

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

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B\lambda} = \begin{bmatrix}5\\2\\-1\end{bmatrix}. \tag{3.61}$$

to guarantee increase numerical stability and positive definiteness. This “ridge” can be rigorously derived using Bayesian inference. See Chapter 9 for details.

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

projection error

$$\|\boldsymbol{x} - \pi_U(\boldsymbol{x})\| = \left\| \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}^\top \right\| = \sqrt{6}\,. \tag{3.62}$$

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

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

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}. \tag{3.63}$$

To verify the results, we can (a) check whether the displacement vector $\pi_U(\boldsymbol{x}) - \boldsymbol{x}$ is orthogonal to all basis vectors of U, (b) verify that $\boldsymbol{P}_\pi = \boldsymbol{P}_\pi^2$ (see Definition 3.7).

Remark. The projections $\pi_U(\boldsymbol{x})$ are still vectors in $\mathbb{R}^n$ although they lie in an m-dimensional subspace $U \subseteq \mathbb{R}^n$. However, to represent a projected vector we only need the m coordinates $\lambda_1, \ldots, \lambda_m$ with respect to the basis vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ of U. ♢

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

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

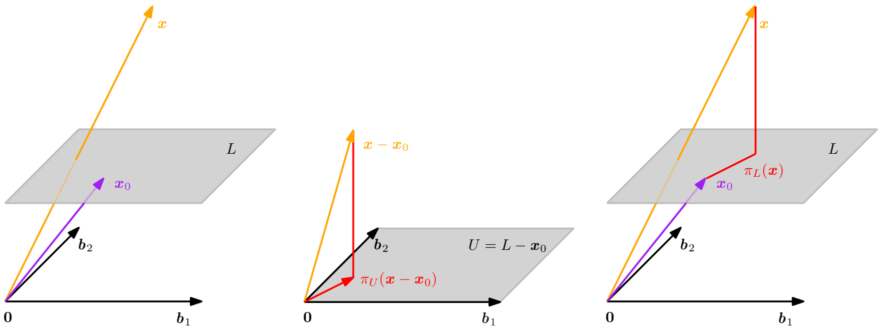
Projections allow us to look at situations where we have a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ without a solution. Recall that this means that $\boldsymbol{b}$ does not lie in the span of $\boldsymbol{A}$, i.e., the vector $\boldsymbol{b}$ does not lie in the subspace spanned by the columns of $\boldsymbol{A}$. 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 $\boldsymbol{A}$ that is closest to $\boldsymbol{b}$, i.e., we compute the orthogonal projection of $\boldsymbol{b}$ onto the subspace spanned by the columns of $\boldsymbol{A}$. 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.

least squares solution

3.6.3 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 3.12(a). We are given an affine space $L = \boldsymbol{x}_0 + U$ where $\boldsymbol{b}_1, \boldsymbol{b}_2$ are basis vectors of U. To determine the orthogonal projection $\pi_L(\boldsymbol{x})$ of $\boldsymbol{x}$ onto L, we transform the problem into a problem that we know how to solve: the projection onto a vector subspace. In order to get there, we subtract the support point $\boldsymbol{x}_0$ from $\boldsymbol{x}$ and from L, so that $L - \boldsymbol{x}_0 = U$ is exactly the vector subspace U. We can now use the



(a) Setting.

(b) Reduce problem to projection π_U onto vector subspace.

(c) Add support point back in to get affine projection π_L.

Figure 3.12 Projection onto an affine space. (a) The original setting; (b) The setting is shifted by $-\boldsymbol{x}_0$, so that $\boldsymbol{x} - \boldsymbol{x}_0$ can be projected onto the direction space U; (c) The projection is translated back to $\boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$, which gives the final orthogonal projection $\pi_L(\boldsymbol{x})$.

orthogonal projections onto a subspace we discussed in Section 3.6.2 and obtain the projection $\pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$, which is illustrated in Figure 3.12(b). This projection can now be translated back into L by adding $\boldsymbol{x}_0$, such that we obtain the orthogonal projection onto an affine space L as

$$\pi_L(\boldsymbol{x}) = \boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)\,, \tag{3.64}$$

where $\pi_U(\cdot)$ is the orthogonal projection onto the subspace U, i.e., the direction space of L, see Figure 3.12(c).

From Figure 3.12 it is also evident that the distance of $\boldsymbol{x}$ from the affine space L is identical to the distance of $\boldsymbol{x} - \boldsymbol{x}_0$ from U, i.e.,

$$\begin{aligned} d(\boldsymbol{x}, L) &= \|\boldsymbol{x} - \pi_L(\boldsymbol{x})\| = \|\boldsymbol{x} - (\boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0))\| && (3.65)\\ &= d(\boldsymbol{x} - \boldsymbol{x}_0, \pi_U(\boldsymbol{x} - \boldsymbol{x}_0))\,. && (3.66)\end{aligned}$$

3.7 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 3.8 (Orthonormal basis)**.** Consider an n-dimensional vector space V and a basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$ of V. If

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0 \quad \text{for } i \neq j \tag{3.67}$$

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1 \tag{3.68}$$

orthonormal basis
ONB
orthogonal basis

for all $i, j = 1, \ldots, n$ then the basis is called an *orthonormal basis* (*ONB*). If only (3.67) is satisfied and then the basis is called an *orthogonal basis*.

Note that (3.68) 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 $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_n\}$ given a set $\{\tilde{\boldsymbol{b}}_1, \dots, \tilde{\boldsymbol{b}}_n\}$ of non-orthogonal and unnormalized basis vectors.

Example 3.11 (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

$$\boldsymbol{b}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \quad \boldsymbol{b}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix} \tag{3.69}$$

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

In Section 3.6, we derived projections of vectors $\boldsymbol{x}$ onto a subspace U with basis vectors $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_k\}$. If this basis is an ONB, i.e., (3.67)–(3.68) are satisfied, the projection equation (3.57) simplifies greatly to

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x} \tag{3.70}$$

since $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}$. This means that we no longer have to compute the tedious inverse from (3.57), which saves us much computation time.

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.

3.8 Rotations

Length and angle preservation, as discussed in Section 3.4, are the two characteristics why linear mappings with orthogonal transformation matrices are the general case of two specific geometric operations: *rotations* and *reflections*. We focus here on rotations only since they are key to machine learning intuitions and machine learning applications, such as robotics.

rotations
reflections

A rotation rotates an object by an angle θ about the origin. For $\theta > 0$, by common convention, we rotate in a counterclockwise direction, and for $\theta < 0$ in the opposite direction. 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.

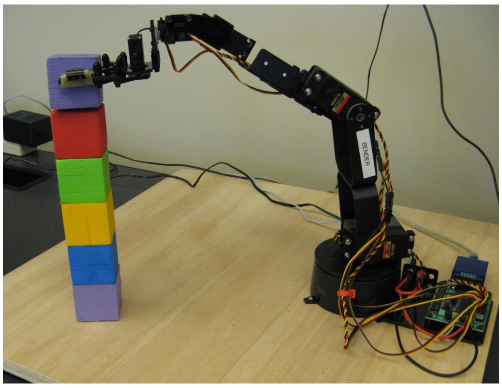


Figure 3.13 The robotic arm needs to rotate its joints in order to pick up objects or to place them correctly. Figure taken from (Deisenroth et al., 2015).

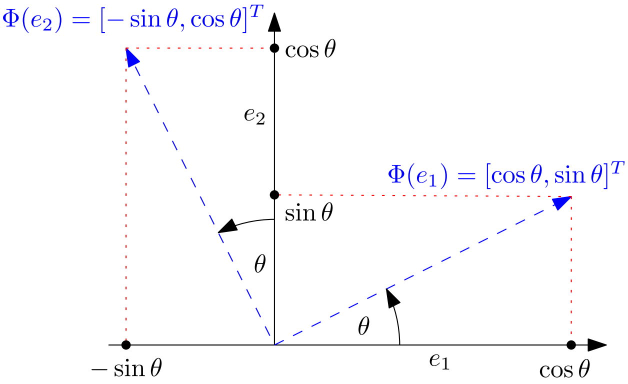

$\Phi(e_2) = [-\sin\theta, \cos\theta]^T$
$\cos\theta$
e_2
$\Phi(e_1) = [\cos\theta, \sin\theta]^T$
$\sin\theta$
θ
θ
$-\sin\theta$
e_1
$\cos\theta$

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

1805 3.8.1 Rotations in $\mathbb{R}^2$

1806 Consider the standard basis in $\mathbb{R}^2$ $\left\{\boldsymbol{e}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{e}_2 = \begin{bmatrix}0\\1\end{bmatrix}\right\}$, which defines
1807 the standard coordinate system in $\mathbb{R}^2$. Assume we want to rotate this co-
1808 ordinate system by an angle θ as illustrated in Figure 3.14. Note that the
1809 rotated vectors are still linearly independent and, therefore, are a basis of
1810 $\mathbb{R}^2$. This means that the rotation performs a basis change.

rotation matrix

Since rotations Φ are linear mappings, we can express them by a *rotation matrix* $\boldsymbol{R}(\theta)$. Trigonometry allows us to determine the coordinates of the rotated axes with respect to the standard basis in $\mathbb{R}^2$. We obtain

$$\Phi(\boldsymbol{e}_1) = \begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}, \quad \Phi(\boldsymbol{e}_2) = \begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}. \tag{3.71}$$

Therefore, the rotation matrix that performs the basis change into the rotated coordinates $\boldsymbol{R}(\theta)$ is given as

$$\boldsymbol{R}(\theta) = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}. \tag{3.72}$$

3.8.2 Rotations in $\mathbb{R}^3$

In contrast to the $\mathbb{R}^2$ case, in $\mathbb{R}^3$ we can rotate any two-dimensional plane about a one-dimensional axis. The easiest way to specify the general rotation matrix is to specify how the images of the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are supposed to be rotated, and making sure these images $\boldsymbol{R}\boldsymbol{e}_1, \boldsymbol{R}\boldsymbol{e}_2, \boldsymbol{R}\boldsymbol{e}_3$ are orthonormal to each other. We can then obtain a general rotation matrix $\boldsymbol{R}$ by combining the images of the standard basis.

To have a meaningful rotation angle we have to define what "counterclockwise" means, and we use the convention that a "counterclockwise" (planar) rotation about an axis refers to a rotation about an axis when we look at the axis "head on, from the end toward the origin". In $\mathbb{R}^3$, there are therefore three (planar) rotations about the three standard basis vectors (see Figure 3.15):

- Counterclockwise rotation about the $\boldsymbol{e}_1$-axis

$$\boldsymbol{R}_1(\theta) = \begin{bmatrix}\Phi(\boldsymbol{e}_1) & \Phi(\boldsymbol{e}_2) & \Phi(\boldsymbol{e}_3)\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & \cos\theta & -\sin\theta\\ 0 & \sin\theta & \cos\theta\end{bmatrix} \tag{3.73}$$

  Here, the $\boldsymbol{e}_1$ coordinate is fixed, and the counterclockwise rotation is performed in the $\boldsymbol{e}_2\boldsymbol{e}_3$ plane.

- Counterclockwise rotation about the $\boldsymbol{e}_2$-axis

$$\boldsymbol{R}_2(\theta) = \begin{bmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{bmatrix} \tag{3.74}$$

  If we rotate the $\boldsymbol{e}_1\boldsymbol{e}_3$ plane about the $\boldsymbol{e}_2$ axis, we need to look at the $\boldsymbol{e}_2$ axis from its "tip" toward the origin.

- Counterclockwise rotation about the $\boldsymbol{e}_3$-axis

$$\boldsymbol{R}_3(\theta) = \begin{bmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{3.75}$$

  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 keeping $n-2$ dimensions fix and restrict the rotation to a two-dimensional plane in the n-dimensional space. As in the 3D we can either rotate any plane.

Definition 3.9 (Givens Rotation)**.** Let V be an n-dimensional Euclidean

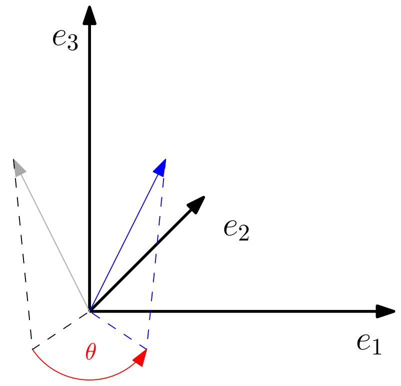


Figure 3.15 Rotation of a vector (gray) in $\mathbb{R}^3$ by an angle θ about the e_3-axis. The rotated vector is shown in blue.

vector space and $\Phi : V \to V$ an automorphism with transformation matrix

$$\boldsymbol{R}_{ij}(\theta) := \begin{bmatrix} \boldsymbol{I}_{i-1} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cos\theta & \boldsymbol{0} & -\sin\theta & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{j-i-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sin\theta & \boldsymbol{0} & \cos\theta & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} & \boldsymbol{I}_{n-j} \end{bmatrix} \in \mathbb{R}^{n\times n}\,, \qquad \theta \in \mathbb{R}\,, \tag{3.76}$$

Givens rotation

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

$$r_{ii} = \cos\theta\,, \quad r_{ij} = -\sin\theta\,, \quad r_{ji} = \sin\theta\,, \quad r_{jj} = \cos\theta\,. \tag{3.77}$$

In two dimensions (i.e., $n = 2$), we obtain (3.72) as a special case.

3.8.4 Properties of Rotations

Rotations exhibit a number useful properties:

- Rotations preserve distances, i.e., $\|\boldsymbol{x}-\boldsymbol{y}\| = \|\boldsymbol{R}_\theta(\boldsymbol{x})-\boldsymbol{R}_\theta(\boldsymbol{y})\|$. In other words, rotations leave the distance between any two points unchanged after the transformation.
- Rotations preserve angles, i.e., the angle between $\boldsymbol{R}_\theta\boldsymbol{x}$ and $\boldsymbol{R}_\theta\boldsymbol{y}$ equals the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.
- 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 $\boldsymbol{R}(\phi)\boldsymbol{R}(\theta) = \boldsymbol{R}(\theta)\boldsymbol{R}(\phi)$ 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).
- Rotations have no real eigenvalues, except we rotate about $n\pi$ where $n \in \mathbb{Z}$.

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 excellent book by Boyd and Vandenberghe (2018), for example.

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, 1901) 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 $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$ in $\mathbb{R}^2$ by:

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

is an inner product.

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

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

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 $\boldsymbol{x} \in \mathbb{R}^5$ are given by

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

1. Determine the orthogonal projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x}$ onto U
2. Determine the distance $d(\boldsymbol{x}, U)$

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

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

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

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

$$U = \operatorname{span}[\boldsymbol{e}_1, \boldsymbol{e}_3]\,.$$

Hint: Orthogonality is defined through the inner product.

2. Compute the distance $d(\boldsymbol{e}_2, U)$.
3. Draw the scenario: standard basis vectors and $\pi_U(\boldsymbol{e}_2)$

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