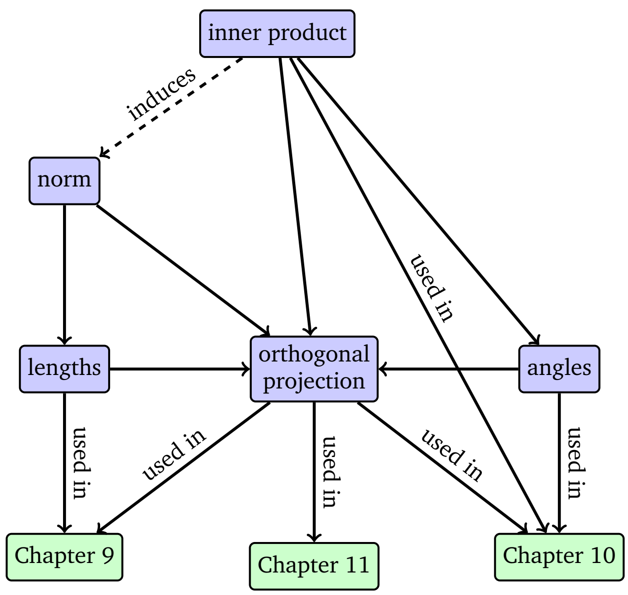
3

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 10. 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 11 and regression via maximum likelihood estimation in Chapter 9.

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.



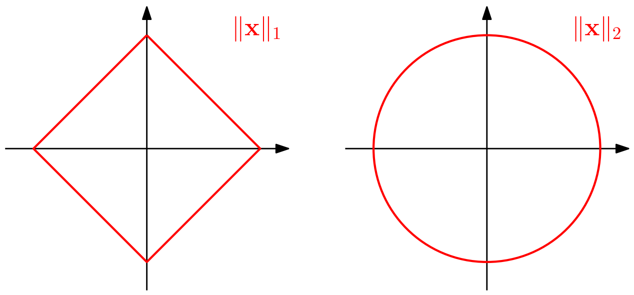


Figure 3.3 For different norms, the red lines indicate the set of vectors with norm 1. Left: Manhattan distance; Right: Euclidean distance.

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.

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

norm

$$\|\cdot\| : V \to \mathbb{R}, \tag{3.1}$$
$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|, \tag{3.2}$$

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:

length

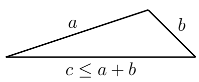
- Absolutely homogeneous: $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
- *Triangle inequality*: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- *Positive definite*: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$.

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.

Figure 3.2 Triangle inequality.



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 $\boldsymbol{x}$.

Example (Manhattan Distance)

The *Manhattan distance* (or ℓ_1 norm) on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}$ as

Manhattan distance

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

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

Example (Euclidean Norm)

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

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

Euclidean distance
Euclidean norm

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

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.

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

$$\boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i\,. \tag{3.5}$$

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.

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

$$\beta(\lambda\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}) = \lambda\beta(\boldsymbol{x}, \boldsymbol{z}) + \beta(\boldsymbol{y}, \boldsymbol{z}) \tag{3.6}$$

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

Equation (3.6) asserts that β is linear in the first argument, and equation (3.7) asserts that β is linear in the second argument.

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

symmetric

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

positive definite

- β is called *positive definite* if

$$\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \beta(\boldsymbol{x}, \boldsymbol{x}) > 0\,, \quad \beta(\boldsymbol{0}, \boldsymbol{0}) = 0 \tag{3.8}$$

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

- A positive definite, symmetric bilinear mapping $\beta : V \times V \to \mathbb{R}$ is called an *inner product* on V. We typically write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ instead of $\beta(\boldsymbol{x}, \boldsymbol{y})$.
- The pair $(V, \langle \cdot, \cdot \rangle)$ is called an *inner product space*, *normed vector space* or (real) *vector space with inner product*. If we use the dot product defined in (3.5), we call $(V, \langle \cdot, \cdot \rangle)$ a *Euclidean vector space*.

inner product
inner product space
normed vector space
vector space with inner product
Euclidean vector space

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

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := x_1 y_1 - (x_1 y_2 + x_2 y_1) + 2 x_2 y_2 \tag{3.9}$$

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

Remark For an inner product vector space $(V, \langle \cdot, \cdot \rangle)$, the induced norm $\| \cdot \|$ satisfies the *Cauchy-Schwarz inequality*

Cauchy-Schwarz inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\| \,. \tag{3.10}$$

3.2.1 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, an ordered basis $B = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ of V and an inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ (see Definition 3.3). Recall the definition of a basis (Section 2.6.1), which says that any vector can be written as a linear combination of basis functions. 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 is uniquely determined through $\boldsymbol{A}$. The symmetry of the inner product also implies 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

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\top \boldsymbol{A} \hat{\boldsymbol{y}} \tag{3.13}$$

defines an inner product with respect to a fixed 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*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\top \boldsymbol{A} \hat{\boldsymbol{y}} \,. \tag{3.14}$$

In Section 4.3, we will return to symmetric, positive definite matrices, and discuss some of their important properties.

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.

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \tag{3.15}$$

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 (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 (3.15). Let us take $\boldsymbol{x} = [1, 1]^\top \in \mathbb{R}^2$. If we use the dot product as the inner product, with (3.15) we obtain

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{1^2 + 1^2} = \sqrt{2} \tag{3.16}$$

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \boldsymbol{y} = x_1 y_1 - \frac{1}{2}(x_1 y_2 + x_2 y_1) + x_2 y_2 \,. \tag{3.17}$$

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

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = x_1^2 - x_1 x_2 + x_2^2 = 1 - 1 + 1 = 1 \implies \|\boldsymbol{x}\| = \sqrt{1} = 1 \,, \tag{3.18}$$

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

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle} \tag{3.19}$$

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

$$d : V \times V \to \mathbb{R} \tag{3.20}$$
$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto d(\boldsymbol{x}, \boldsymbol{y}) \tag{3.21}$$

is called *metric*.

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.

distance
Euclidean distance
metric
positive definite
symmetric
Triangular inequality

A metric d satisfies:

1. d is *positive definite*, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ 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*: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$.

3.4 Angles and Orthogonality

The Cauchy-Schwarz inequality (3.10) 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

$$-1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \leq 1\,. \tag{3.22}$$

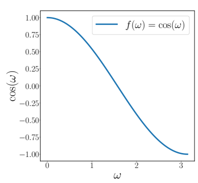
Therefore, there exists a unique $\omega \in [0, \pi]$ with

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|}\,, \tag{3.23}$$

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

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

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle\langle \boldsymbol{y}, \boldsymbol{y}\rangle}} = \frac{\boldsymbol{x}^\top\boldsymbol{y}}{\sqrt{\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{y}}} = \frac{3}{\sqrt{10}}\,, \tag{3.24}$$

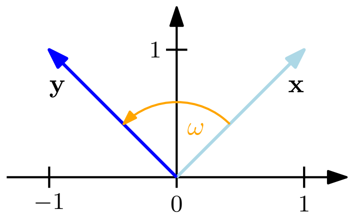
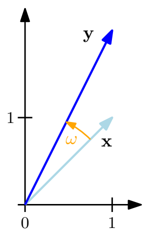


Figure 3.6 The angle ω between two vectors $\boldsymbol{x}, \boldsymbol{y}$ can change depending on the inner product.

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

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}\,, \tag{3.25}$$

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

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = -\frac{1}{3} \implies \omega \approx 1.91\,\mathrm{rad} \approx 109.5°\,, \tag{3.26}$$

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.

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 (3.5) for example, turns into an integral.

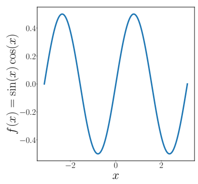
The inner product of two functions $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ is defined as the definite integral

$$\langle u, v \rangle := \int_a^b u(x)v(x)dx \tag{3.27}$$

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 (3.27) 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 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 (Inner Product of Functions)
If we choose $u = \sin(x)$ and $v = \cos(x)$, the integrand $f(x) = u(x)v(x)$ of (3.27), 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 $f(x) = \sin(x)\cos(x)$.



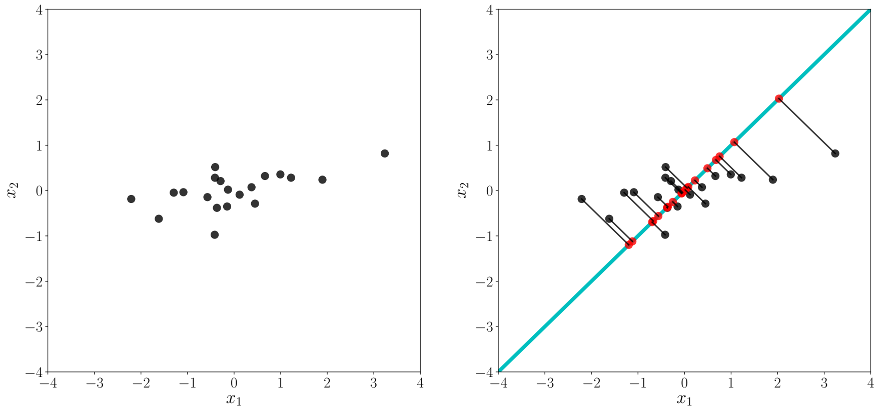
Remark It also holds that the collection of functions

$$\{1, \cos(x), \cos(2x), \cos(3x), \dots\} \tag{3.28}$$

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.

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



(a) Original data set.

(b) Original data points (black) and their corresponding orthogonal projections (red) onto the subspace (straight line).

3.5 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, see Figure 3.8 for an illustration. To minimize this compression loss, we ideally find the most informative dimensions in the data. Then, we can project the original very high-dimensional data into 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 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 11 for linear dimensionality reduction and in Chapter 10 for classification.

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

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

projection matrices

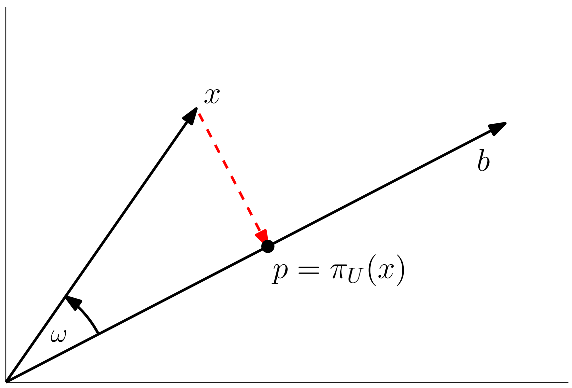


Figure 3.9 Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis $\boldsymbol{b}$.

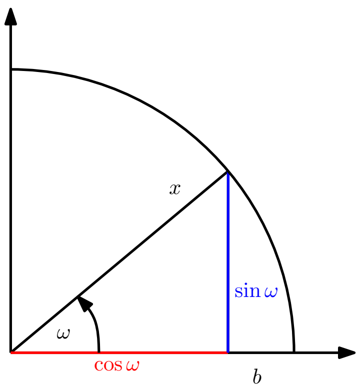


Figure 3.10 Projection of a two-dimensional vector $\boldsymbol{x}$ onto a one-dimensional subspace with $\|\boldsymbol{x}\| = 1$.

the property that $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$.

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

3.5.1 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 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

$$\langle \boldsymbol{x} - \pi_U(\boldsymbol{x}), \boldsymbol{b}\rangle = 0 \tag{3.29}$$

$$\overset{\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b}}{\iff} \langle \boldsymbol{x} - \lambda\boldsymbol{b}, \boldsymbol{b}\rangle = 0\,. \tag{3.30}$$

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

$$\langle \boldsymbol{x}, \boldsymbol{b}\rangle - \lambda\langle \boldsymbol{b}, \boldsymbol{b}\rangle = 0 \tag{3.31}$$

$$\iff \lambda = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\langle \boldsymbol{b}, \boldsymbol{b}\rangle} = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\|\boldsymbol{b}\|^2} \tag{3.32}$$

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

$$\lambda = \frac{\boldsymbol{b}^\top \boldsymbol{x}}{\boldsymbol{b}^\top \boldsymbol{b}} = \frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2} \tag{3.33}$$

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 (3.33) that

$$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b} = \frac{\langle \boldsymbol{x}, \boldsymbol{b}\rangle}{\|\boldsymbol{b}\|^2}\boldsymbol{b} = \frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2}\boldsymbol{b}\,, \tag{3.34}$$

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

$$\|\boldsymbol{p}\| = \|\lambda\boldsymbol{b}\| = |\lambda|\,\|\boldsymbol{b}\|\,. \tag{3.35}$$

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

$$\|\pi_U(\boldsymbol{x})\| \overset{(3.34)}{=} \frac{|\boldsymbol{b}^\top \boldsymbol{x}|}{\|\boldsymbol{b}\|^2}\,\|\boldsymbol{b}\| \overset{(3.23)}{=} |\cos\omega|\,\|\boldsymbol{x}\|\,\|\boldsymbol{b}\|\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{b}\|^2} = |\cos\omega|\,\|\boldsymbol{x}\|\,. \tag{3.36}$$

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[1] 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

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

$$\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b} = \boldsymbol{b}\lambda = \boldsymbol{b}\frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2} = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}\boldsymbol{x} \tag{3.37}$$

we immediately see that

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}. \tag{3.38}$$

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.

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.

Example (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 (3.38), we obtain

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\boldsymbol{b}^\top\boldsymbol{b}} = \frac{1}{9}\begin{bmatrix}1\\2\\2\end{bmatrix}\begin{bmatrix}1 & 2 & 2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}. \tag{3.39}$$

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

$$\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi\boldsymbol{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in [\begin{bmatrix}1\\2\\2\end{bmatrix}]. \tag{3.40}$$

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

[1] This is a one-dimensional subspace.

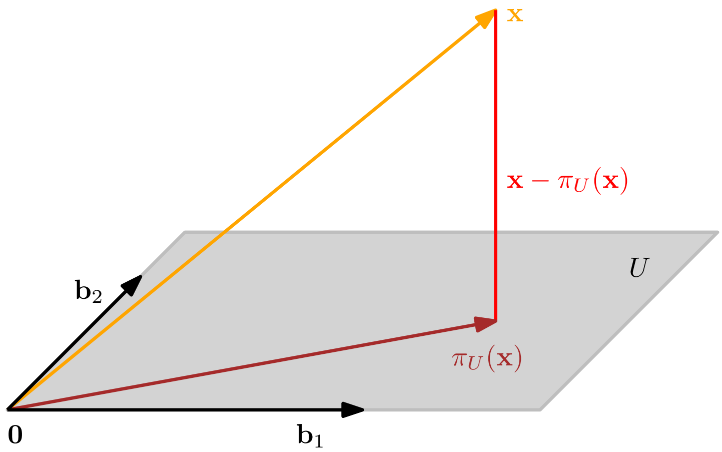


Figure 3.11 Projection onto a two-dimensional subspace U with basis $\boldsymbol{b}_1, \boldsymbol{b}_2$. The projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x} \in \mathbb{R}^3$ onto U can be expressed as a linear combination of $\boldsymbol{b}_1, \boldsymbol{b}_2$ and the displacement vector $\boldsymbol{x}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

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.

3.5.2 Projection onto General Subspaces

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 $\dim(U) = m \geq 1$. An illustration is given in Figure 3.11.

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.

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 $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ of U, such that $\pi_U(\boldsymbol{x}) = \sum_{i=1}^m \lambda_i \boldsymbol{b}_i$.

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

The basis vectors form the columns of $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, where $\boldsymbol{B} = (\boldsymbol{b}_1|\ldots|\boldsymbol{b}_m)$.

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

$$\pi_U(\boldsymbol{x}) = \sum_{i=1}^m \lambda_i \boldsymbol{b}_i = \boldsymbol{B\lambda}\,, \tag{3.41}$$

$$\boldsymbol{B} = (\boldsymbol{b}_1|\ldots|\boldsymbol{b}_m) \in \mathbb{R}^{n\times m},\ \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\top \in \mathbb{R}^m\,, \tag{3.42}$$

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)

$$\langle \boldsymbol{b}_1, \boldsymbol{x} - \pi_U(\boldsymbol{x})\rangle = \boldsymbol{b}_1^\top(\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.43}$$

$$\vdots \tag{3.44}$$

$$\langle \boldsymbol{b}_m, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle = \boldsymbol{b}_m^\top (\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.45}$$

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

$$\boldsymbol{b}_1^\top (\boldsymbol{x} - \boldsymbol{B\lambda}) = 0 \tag{3.46}$$

$$\vdots \tag{3.47}$$

$$\boldsymbol{b}_m^\top (\boldsymbol{x} - \boldsymbol{B\lambda}) = 0 \tag{3.48}$$

such that we obtain a homogeneous linear equation system

$$\begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{x} - \boldsymbol{B\lambda} \end{bmatrix} = \boldsymbol{0} \iff \boldsymbol{B}^\top (\boldsymbol{x} - \boldsymbol{B\lambda}) = \boldsymbol{0} \tag{3.49}$$

$$\iff \boldsymbol{B}^\top \boldsymbol{B\lambda} = \boldsymbol{B}^\top \boldsymbol{x} \,. \tag{3.50}$$

The last expression is called *normal equation*. Since $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ 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 optimal coefficients/coordinates

normal equation

$$\boldsymbol{\lambda} = (\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{x} \,. \tag{3.51}$$

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.[2]

pseudo-inverse

2. Find the projection $\pi_U(\boldsymbol{x}) \in U$. We already established that $\pi_U(\boldsymbol{x}) = \boldsymbol{B\lambda}$. Therefore, with (3.51)

$$\pi_U(\boldsymbol{x}) = \boldsymbol{p} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{x} \,. \tag{3.52}$$

3. Find the projection matrix $\boldsymbol{P}_\pi$. From (3.52) we can immediately see that the projection matrix that solves $\boldsymbol{P}_\pi \boldsymbol{x} = \pi_U(\boldsymbol{x})$ must be

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \,. \tag{3.53}$$

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 (3.53) $\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top$ as $\boldsymbol{P}_\pi = \frac{\boldsymbol{B}\boldsymbol{B}^\top}{\boldsymbol{B}^\top \boldsymbol{B}}$, which is exactly the projection matrix in (3.38).

[2] In practical applications (e.g., linear regression), we often add a "jitter term" $\epsilon \boldsymbol{I}$ to $\boldsymbol{B}^\top \boldsymbol{B}$ to guarantee increase numerical stability and positive definiteness. This "ridge" can be rigorously derived using Bayesian inference.

Example (Projection onto a two-dimensional subspace)

For a subspace $U = [\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0\\1\\2\end{bmatrix}] \subseteq \mathbb{R}^3$ and $\boldsymbol{x} = \begin{bmatrix}6\\0\\0\end{bmatrix} \in \mathbb{R}^3$ find $\boldsymbol{\lambda}$, 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

$$\boldsymbol{B}^\top\boldsymbol{B} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\\1 & 2\end{bmatrix} = \begin{bmatrix}3 & 3\\3 & 5\end{bmatrix} \qquad \boldsymbol{B}^\top\boldsymbol{x} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}6\\0\\0\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix} \tag{3.54}$$

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

$$\begin{bmatrix}3 & 3\\3 & 5\end{bmatrix}\begin{bmatrix}\lambda_1\\\lambda_2\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix} \iff \boldsymbol{\lambda} = \begin{bmatrix}5\\-3\end{bmatrix}. \tag{3.55}$$

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

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \begin{bmatrix}5\\2\\-1\end{bmatrix}. \tag{3.56}$$

projection error
reconstruction error

The corresponding *projection error/reconstruction error* is

$$\|\boldsymbol{x} - \pi_U(\boldsymbol{x})\| = \left\|\begin{bmatrix}1 & -2 & 1\end{bmatrix}^\top\right\| = \sqrt{6}. \tag{3.57}$$

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

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top = \frac{1}{6}\begin{bmatrix}5 & 2 & -1\\2 & 2 & 2\\-1 & 2 & 5\end{bmatrix}. \tag{3.58}$$

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 $\lambda_1, \dots, \lambda_m$ with respect to the basis vectors $\boldsymbol{b}_1, \dots, \boldsymbol{b}_m$ of U.

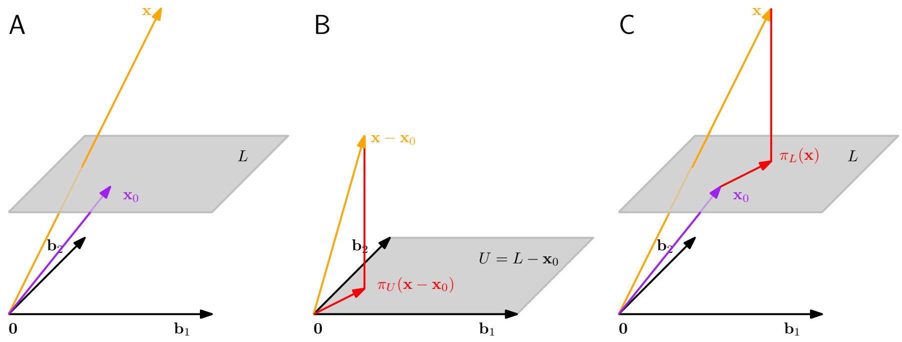


Figure 3.12 Projection onto an affine space. A: The original setting; B: The setting is shifted by $-\boldsymbol{x}_0$, so that $\boldsymbol{x} - \boldsymbol{x}_0$ can be projected onto the direction space U; C: The projection is translated back to $\boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$,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

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.

Projections allow us to look at situations where we have a linear system $\boldsymbol{Ax} = \boldsymbol{b}$ 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.

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

least squares solution

3.5.3 Projections 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 Panel A of Figure 3.12. 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 orthogonal projections onto a subspace we discussed in Section 3.5.2 and obtain the projection $\pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$, see Panel B of Figure 3.12. 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

$$\pi_L(\boldsymbol{x}) = \boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)\,, \tag{3.59}$$

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

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.,

$$d(\boldsymbol{x}, L) = \|\boldsymbol{x} - \pi_L(\boldsymbol{x})\| = \|\boldsymbol{x} - (\boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0))\| \tag{3.60}$$

$$= d(\boldsymbol{x} - \boldsymbol{x}_0, \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)) . \tag{3.61}$$

3.6 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

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0 \quad \text{for } i \neq j \tag{3.62}$$

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1 \tag{3.63}$$

orthonormal basis
ONB
orthogonal basis

for all $i, j = 1, \ldots, n$ then the basis is called an *orthonormal basis* (*ONB*). If only (3.62) is satisfied and then the basis is called an *orthogonal basis*.

Note that (3.63) 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 $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ given a set $\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n$ of non-orthogonal and unnormalized basis vectors.

Example (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

$$\boldsymbol{b}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{b}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{3.64}$$

form an orthonormal basis.

In Section 3.5, we derived projections of vectors $\boldsymbol{x}$ onto a subspace U with basis vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$. If this basis is an ONB, i.e., (3.62)–(3.63) are satisfied, the projection equation (3.52) simplifies greatly to

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x} \tag{3.65}$$

since $B^\top B = I$. This means that we no longer have to compute the tedious inverse from (3.52), which saves us much computation time.

We will exploit the concept of an orthonormal basis in Chapter 10 and Chapter 11 when we discuss Support Vector Machines and Principal Component Analysis.

3.7 Further Reading

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., 1998) 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 10.

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) also uses projections to reduce the dimensionality of high-dimensional data. We will discuss this in more detail in Chapter 11.