Linear Algebra L1 - Vectors

Alexander Binder

February 17, 2023

1 Learning Goals

- vectors of real numbers
- norms of vectors and their properties
- inner products, their interpretation and properties
- representing a vector as a linear combination
- vector spaces
- independent sets of vectors
- orthogonal sets of vectors
- projecting onto a vector, removing the direction of a vector
- creating an orthogonal set of vectors

2 Motivation for Linear Algebra

- Examples for linear relationships in physics:
  - Newtons first axiom
    $$F = ma$$
    is linear in the mass m and in the acceleration a.
  - Einsteins energy-mass relationship for low speeds
    $$E = mc^2$$
    is linear in the mass m. This is the mass energy equivalence in special relativity theory for particles with a mass.
  - For mass-less photons it is instead:
    $$E = hf$$
    where $h = 6.626 * 10^{-34}\ J(Hz)^{-1}$ is Plancks constant and f the frequency (inverse of the wavelength) of the photon.
    ... Linear in the frequency f (visible light, UV, roentgen, gamma rays).
- Linear relationships often hold as an approximation within in a certain range. Linear mapping are like the first view on this world.
  - As an example the force which a spring exerts when it is stretched by length x:
    $$F = kx$$
    where k is a constant which depends on the material of the spring. Obviously this does not hold when one overstretches the string until it becomes something else.

  - The same holds for Einsteins energy-mass formula. For high speeds one has to use:

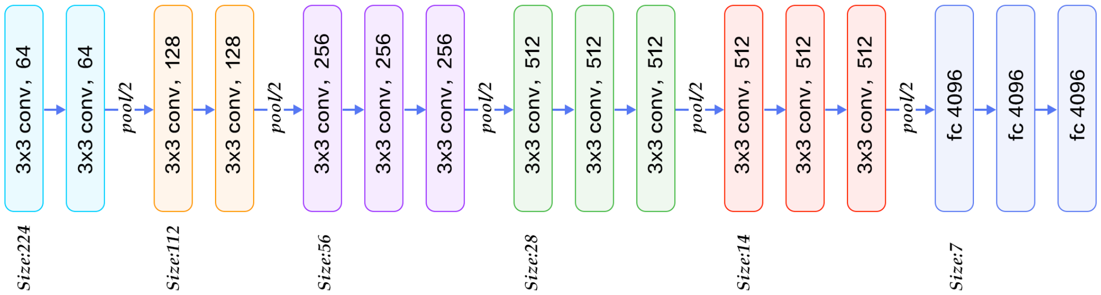
$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$$

- The linear model as simplest dependency on an input:

  A score s as a weighted sum of causes x_0, x_1, x_2 – provided the causes are on the same physical scale and can be summed together. One can sum up three real numbers or vectors with weights.

$$s = w_0 x_0 + w_1 x_1 + w_2 x_2$$

- Machine learning heavily relies on linear algebra: Neural nets are made of layers of linear operations, with non-linear activations on top.



Each block is a linear operation, implemented using matrix multiplication
source: `https://www.kaggle.com/code/blurredmachine/vggnet-16-architecture-a-complete-guide/notebook`

> Linear algebra is not essential for everything in life. For a study with aspirations of doing any deeper work in Machine Learning or engineering it is needed.

3 Vectors

vector of real numbers

Definition: A vector of real numbers is a sequence $v = (v_1, \dots, v_d)$ such that each component v_i is a real number. d is the dimensionality of the vector.

Examples:

$$\begin{aligned} v =& (-1.1, 2.7, 3.5) \\ v =& (0, 0, 1, 0, 0) \\ v =& (2.4, 9.3) \\ v =& (\cos(\alpha), \sin(\alpha)) \end{aligned}$$

The following just describes the set of all d-dim vectors with real values

The space of d-dimensional vectors with real values

Definition: The space of d-dimensional vectors with real values $\mathbb{R}^d$ is the set of all sequences $(v_1, \dots, v_d)$, of length d such that each component v_i is a real number.

In Literature it is commonly written as $\mathbb{R}^d$ as $\mathbb{R}$ denote the real numbers.

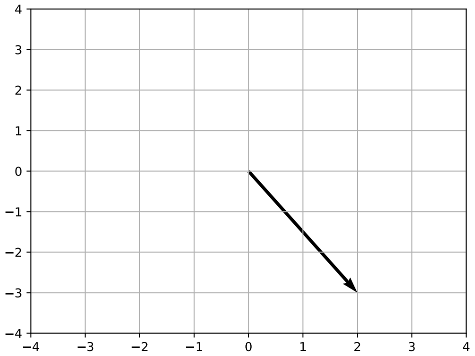
3.1 Vectors in Physics

A vector in a physical context can sometimes be interpreted as a measurement which has a magnitude and a direction.
Examples:

- Wind Speed with velocity of $3m/s$ and a direction of 227 degrees.
- A pushing force in 2D of $(2, -3)$ Newton.

  That is: 2 Newton along the first axis, and -3 Newton along the second axis – that is 3 newton against the direction of the second axis.

  Its direction can be plotted:



Questions:

- What is the magnitude and the angle in the second example ?

3.2 Vectors in machine learning

Take any feature map, for example those obtained from a layer of a neural network. It is usually called a tensor. It has multiple indices. Flattening / Rolling out a tensor yields a vector :) .
Feature spaces do not have directions with geographical meanings (cf. previous section), but they still are composed of sequences of numerical values.

4 Properties of Vectors

4.1 Can be written as column or row vector:

$$v = \begin{bmatrix} 5 & 7 & 9 \end{bmatrix} \qquad \text{in pytorch: } v.shape = (1,3) \text{ or } v.shape = (3)$$

$$v = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix} = (5,7,9)^\top \text{ in pytorch: } v.shape = (3,1) \text{ or } v.shape = (3)$$

4.2 Vector Addition: component-wise addition

$$v = (v_0, v_1, v_2),\ w = (w_0, w_1, w_2)$$
$$v + w = (v_0 + w_0, v_1 + w_1, v_2 + w_2)$$

Example:

$$(1,2,3) + (2,-1,3) = (3,1,6)$$

4.3 Vector Scaling: also component-wise

$$\begin{aligned} v &= (v_0, v_1, v_2) \\ av &= (av_0, av_1, av_2) \\ 5(2, -1, 3) &= (10, -5, 15) \end{aligned}$$

4.4 Euclidean vector norm:

This is also known as ℓ_2-norm

$$\|v\|_2 = \left(\sum_{k=1}^{d} v_k^2 \right)^{1/2}$$

This corresponds to the usual idea of classic Euclidean lengths. In physics the usual magnitude of a vector. Example:

$$\|(-1, 3, -2)\|_2 = \sqrt{(-1)^2 + 3^2 + (-2)^2} = \sqrt{1+9+4} = \sqrt{14} \approx ?$$

Note: There are many ways to define a norm. The euclidean norm is just one of many.

4.5 Other norms

Other examples: the 1-norm

$$\|v\|_1 = \sum_{k=1}^{d} |v_k|$$

p-norms in General:

$$\|v\|_p = (\sum_{k=1}^{d} |v_k|^p)^{1/p}$$

Note that the Euclidean and the 1-norm are special cases of this for $p = 2$ and $p = 1$
The maximum norm, which can be seen as a limit of $\|v\|_p$ as $p \to \infty$:

$$\|v\|_\infty = \max_{k=1}^{d} |v_k|$$

Examples: Where other norms are used? In this part you do not need to remember the math.

- in LASSO: linear regression.

  One wants to obtain a linear prediction of y based on x

  $$y = f_w(x) = \sum_{k=1,\ldots,d} w_k x_k$$

  Inputs: pairs of a vector and its corresponding desired output $(x^{\{i\}}, y^{\{i\}}) = ((x_1^{\{i\}}, \ldots, x_d^{\{i\}}), y^{\{i\}})$.

  Standard, unregularized, linear regression would try to find the best weights $w = (w_1, \ldots, w_d)$ based on a set of training samples minimizing the squared difference between prediction $f(x^{\{i\}})$ and desired value $y^{\{i\}}$:

  $$\begin{aligned} &\text{find } w \text{ such that } \frac{1}{n} \sum_{i=1}^{n} (y^{\{i\}} - \sum_{k} w_k x_k^{\{i\}})^2 \\ &= \frac{1}{n} \|Y - Xw\|_2^2 \text{ is minimal} \end{aligned}$$

LASSO adds a 1-norm penalty on the weights.

$$\text{find } w \text{ such that } \frac{1}{n}\sum_{i=1}^{n}(y^{\{i\}} - \sum_k w_k x_k^{\{i\}})^2 + \lambda \|w\|_1$$
$$= \frac{1}{n}\|Y - Xw\|_2^2 + \lambda\|w\|_1 \text{ is minimal}$$

Effect: one can ensure that the weights are sparse, that is that we have $w_k = 0$ for many weight vector components.

Math to be remembered starts here again.

4.6 Properties of any norm:

Properties of any norm

- $\|0\| = 0$
- $\|\lambda v\| = |\lambda|\|v\|$
- $\|v + u\| \le \|v\| + \|u\|$
- a norm induces a distance measure $d(\cdot, \cdot)$ via $d(u, v) = \|u - v\|$

General: Norms encode a notion of length of a vector.

4.7 Unit length vectors:

Unit length vector (euclidean norm)

Definition: v is unit length vector with respect to the euclidean norm if

$$\|v\|_2 = 1$$

Making a vector to have unit length:

$$v \neq 0 \Rightarrow \frac{v}{\|v\|_2}$$

has unit length.

4.8 Inner product:

Inner product

Let be $u \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, Definition: Then the inner product between u and v is defined as:

$$u \cdot v = \sum_{k=1}^{d} u_k v_k$$

Examples:

$$(3, -1, -2) \cdot (1, -2, 1.5) = 3 + 2 - 3 = 2$$

Interpretation of the inner product

It holds for the inner product defined above that:

$$u \cdot v = \|u\|_2 \|v\|_2 \cos(\angle(u,v))$$

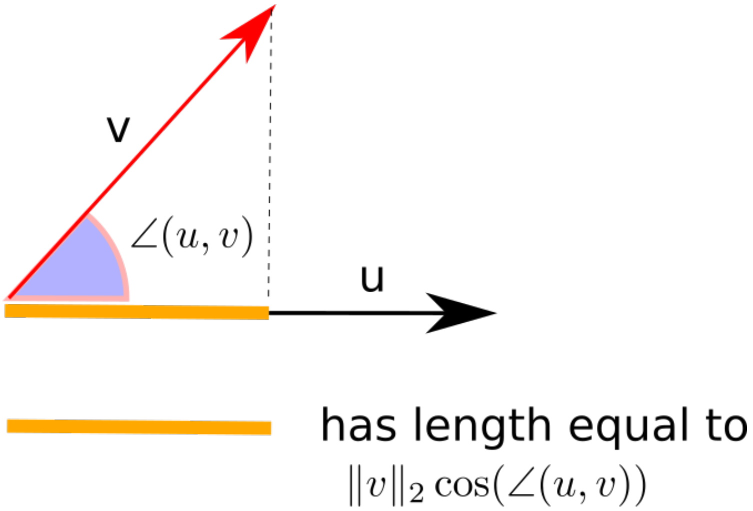
it is the product of the euclidean length of u, of v and the cosine of the angle between these two vectors.

See the cosine here:
https://en.wikipedia.org/wiki/Sine_and_cosine#/media/File:Sine_cosine_one_period.svg

- $\frac{u}{\|u\|_2} \cdot \frac{v}{\|v\|_2} = \cos(\angle(u,v))$ is a similarity measure between two vectors.

  The cosine angle gets larger if the angle between the two vector decreases:

  - it is -1 if $u = -v$
  - it is 0 for orthogonal vectors
  - it is 1 is $u = v$
  - for angles between 90 and 270 degrees (in radians the interval $(\pi/2, 3/2\pi)$), the cosine and thus the inner product is negative
  - for angles between 270 and 90 degrees (in radians the intervals $(0, \pi/2) \cup (3/2\pi, 2\pi)$), the cosine and thus the inner product is positive



- matrix-vector and matrix-matrix multiplications compute a set of inner products
- the idea of inner products to define a similarity is widely used in machine learning:

  See for Example (you do not need to understand or memorize anything from this paper!) https://arxiv.org/pdf/2003.04297.pdf Self-supervised pretraining. See the loss in equation (1) there. See also Table 3 for realistic training times)-: :-)

4.9 Properties of the inner product

- it is symmetric as a function of its two input vectors u, v:

$$u \cdot v = v \cdot u$$

- it is bi-linear, that is it is linear in each of its two input arguments: Let a_1, a_2 be real numbers, $u^{\{1\}}, u^{\{2\}}, v$ be vectors. Then:

$$\text{Linearity in the first argument: } (a_1 u^{\{1\}} + a_2 u^{\{2\}}) \cdot v = a_1(u^{\{1\}} \cdot v) + a_2(u^{\{2\}} \cdot v)$$
$$\text{Linearity in the 2nd argument: } v \cdot (a_1 u^{\{1\}} + a_2 u^{\{2\}}) = a_1(v \cdot u^{\{1\}}) + a_2(v \cdot u^{\{2\}})$$

What does the above say ? If we input a linear combination, then we can put the linear combination on the outside, with the inner product being inside.

It is not linear as a function of the concatenated vector (u, v):

$$\begin{aligned} L((u,v)) &= u \cdot v \\ L(a(u,v)) = L((au, av)) &= (au) \cdot (av) && = a^2 u \cdot v \neq aL((u,v)) \end{aligned}$$

A Linear function L would behave like that: $L(a(u,v)) = aL((u,v))$, while above causes a to get squared.

- Every inner product $\cdot$ defines a norm:

$$\|u\|_{(\cdot)} = \sqrt{u \cdot u}$$

The inner product as above defines the Euclidean norm. There are other ways to define inner products not shown (here). The converse is not true: not every norm is created by an inner product.

- something simple: $u \cdot 0 = 0$

Inner product

Remember these properties.

4.10 Motivation: Inner products in statistics and ML

Linear regression: one builds a prediction model which takes as an input a vector $x = (x_0, \ldots, x_{d-1})$

$$f_w(x) = \sum_{k=0}^{d-1} w_k x_k = w \cdot x$$

Also models with a bias term can be represented by an inner product if we extend the vector x by a last component, which is just 1:

$$\begin{aligned} \tilde{w} &= (w_0, \ldots, w_{d-1}, b) = \mathit{concat}(w, b) \\ \tilde{x} &= (x_0, \ldots, x_{d-1}, 1) = \mathit{concat}(x, 1) \\ f_w(x) &= \sum_{k=0}^{d-1} w_k x_k + b = w \cdot x + b1 = \tilde{w} \cdot \tilde{x} \end{aligned}$$

fully connected layers in Neural nets:
https://pytorch.org/docs/stable/generated/torch.nn.Linear.html (see also torch.functional)

$$y = xA^\top + b$$

is a vector times matrix multiplication. You will see that this is effectively a set of inner products.

Convolution Layers: https://github.com/vdumoulin/conv_arithmetic
They compute inner products between the convolution kernel and a (sliding) window in the input features. You do not need to understand the mechanism of convolution in detail. It is not part of graded knowledge.

Conclusion: all those compute inner products, and thus similarities (weighted by the norms of both vectors) between the input features x and the trainable weights.

5 Representability as a linear combination, Vector Spaces, Linear Independence

Motivation: This will be useful to measure quantities along coordinate axes, to express vectors y as a function of a few selected vectors.

Representability as a linear combination

A vector y can be represented as a linear combination over a set of vectors $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$, if there exist real numbers $a_0, \dots, a_{d-1}$ such that

$$y = \sum_{k=0}^{d-1} a_k v^{\{k\}}.$$

Examples:

- $(1,0)$ and $(0,1)$ are able to represent all vectors (a_0, a_1)
- $(1,1)$ and $(0,1)$ are able to represent all vectors (a_0, a_1), too
- Each element tv of the line $\{tv, t \in \mathbb{R}\}$ is a linear combination over the vector v.
- $(1,0,0)$ and $(0,1,0)$ cannot represent any vector $(0,0,a_2)$, if $a_2 \neq 0$. Also they cannot represent any vector (a_0, a_1, a_2) if $a_2 \neq 0$.
- $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ can represent any vector (a_0, a_1, a_2)

Vector space

A set of objects V is a vector space if three properties are met:

- we know how to multiply an object $v \in V$ with a real number, that is we can define a multiplication operation $a * v$, where $v \in V$ is an object from the set, and a is a real number
- we know how to add two objects $u \in V$, $v \in V$, that is we can define an addition operation $u + v$
- the set V is closed under the above two operations, that is if $u \in V$, $v \in V$, then their multiplications with real numbers a_1, a_2 and sums of vectors are also contained within the set V:

$$u \in V, v \in V \Rightarrow a_1 u + a_2 v \in V$$

- Example: this obviously holds for vectors of real numbers. Summing two vectors or multiplying a vector with a real number results again in a vector.
- Counterexample: the set of all vectors with the constraint of having norms below 1 do not form a vector space. Proof: multiply a vector which is not the zero vector with a sufficiently large real number. Its norm will then be above 1.

Definition of a vector space spanned by a set of vectors

The vector space spanned by a set of vectors $w^{\{0\}}, \dots, w^{\{n-1\}}$ is the set of all their linear combinations using all possible real numbers $a_0, \dots, a_{n-1} \in \mathbb{R}$

$$V = \{w \quad \text{such that}$$
$$w = \sum_{r=0}^{n-1} a_r w^{\{r\}}, a_0, \dots, a_{n-1} \in \mathbb{R}\}$$

The dimension of this set is the largest number of independent vectors which we can obtain from $w^{\{0\}}, \dots, w^{\{n-1\}}$.

Essentially: this is the set of all vectors which can be represented using these n vectors $v^{\{0\}}$, $w^{\{1\}}$, $w^{\{2\}}$, …, $w^{\{n-1\}}$.

(Linearly) independent set of vectors

A set of vectors $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$ is independent (in the sense of linear algebra), if no vector $v^{\{i\}}$ from this set can be represented by the vectors from the set $\{v^{\{0\}}, \dots, v^{\{d-1\}}\} \setminus \{v^{\{i\}}\}$ (that is from the set without using $v^{\{i\}}$ itself).

That is for any $v^{\{i\}}$ from the set $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$, there exist **no solution** of shape

$$v^{\{i\}} = \sum_{k=0, k\neq i}^{d-1} a_k v^{\{k\}}.$$

such that $v^{\{i\}}$ and all $v^{\{k\}}$ are taken from the set $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$.

Note: in the above we try to represent $v^{\{i\}}$ by all other vectors from the set except $v^{\{i\}}$. Thats why the sum has the qualifier $k \neq i$ for index k.

Next step: show why independent sets are useful.

Uniqueness of linear representation w.r.t. an independent set

If $(v^{\{0\}}, \dots, v^{\{d-1\}})$ is an independent set with a fixed ordering of vectors (sets are unordered!!), and we have a decomposition of a vector $y = \sum_{k=0}^{d-1} a_k v^{\{k\}}$, then the coefficients $\{a_0, \dots, a_{d-1}\}$ are unique.
That is, if there exists another decomposition $y = \sum_{k=0}^{d-1} b_k v^{\{k\}}$ with respect to the same set with the same ordering of vectors, then it must hold $a_k = b_k$ for all $k = 0, \dots, d-1$.

Proof by contradiction: Assumption: Suppose we have two decompositions with two different sets of coefficients $\{a_0, \dots, a_{d-1}\}$, $\{b_0, \dots, a_{b-1}\}$:

If the sets are different, then there must be one pair of differing coefficients: $a_r \neq b_r$. We use this as follows:

$$\begin{aligned}
y &= \sum_{k=0}^{d-1} a_k v^{\{k\}} \\
y &= \sum_{k=0}^{d-1} b_k v^{\{k\}} \\
\Rightarrow \sum_{k=0}^{d-1} a_k v^{\{k\}} &= \sum_{k=0}^{d-1} b_k v^{\{k\}} \\
a_r v^{\{r\}} + \sum_{k=0, k\neq r}^{d-1} a_k v^{\{k\}} &= b_r v^{\{r\}} + \sum_{k=0, k\neq r}^{d-1} b_k v^{\{k\}} \\
\underbrace{(a_r - b_r)}_{\neq 0} v^{\{r\}} &= \sum_{k=0, k\neq r}^{d-1} (b_k - a_k) v^{\{k\}} \\
v^{\{r\}} &= \frac{1}{a_r - b_r} \sum_{k=0, k\neq r}^{d-1} (b_k - a_k) v^{\{k\}}
\end{aligned}$$

This means that $v^{\{r\}}$ can be represented as a linear combination of the other $v^{\{k\}}, k \neq r$. This violates the assumptions of independence and therefore it is not possible. Therefore our assumption (two decompositions with two different sets of coefficients) can never become true.

Basis in a finite dimensional vector space and dimension of a vector space

A set of vectors $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$ is a basis of a vector space $\mathcal{V}$, if the set is independent, and any vector $y \in \mathcal{V}$ can be represented as a linear combination of the set. The number of vectors in this set, d, is the dimensionality of the vector space.

Note: There is no clash with the definition of independence.

$v^{\{i\}}$ from the set can be represented as a linear combination of the set, by using only itself. So there is no problem with respect to the definition of independence - where using $v^{\{i\}}$ to represent itself is excluded.

Compare to the definition above for the vector space $\mathbb{R}^d$

Basis of $\mathbb{R}^d$

The set

$$\begin{aligned}
&\{e^{\{0\}}, \dots, e^{\{d-1\}}\} \text{ with} \\
e^{\{0\}} &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}^\top \\
e^{\{1\}} &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^\top \\
e^{\{2\}} &= \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \end{bmatrix}^\top \\
e^{\{k\}} &= \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{at pos. } k \\
e^{\{d-1\}} &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \end{bmatrix}^\top
\end{aligned}$$

is one possible basis of $\mathbb{R}^d$. Its elements are called one-hot vectors.

In math this can be expressed as:

$$e_i^{\{k\}} = \begin{cases} 1 \text{ if } i = k \\ 0 \text{ if } i \neq k \end{cases}$$

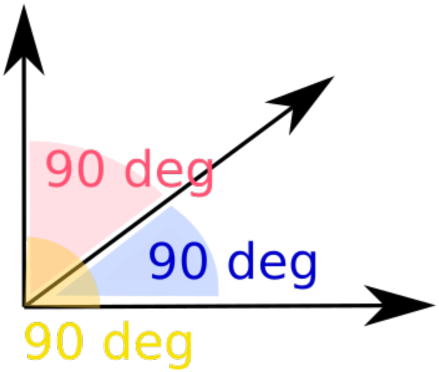
6 Orthogonal vectors

Orthogonal set of vectors

Definition: $\{v^{\{0\}}, \dots, v^{\{d-1\}}\}$ is an orthogonal set of vectors if all $v^{\{i\}} \neq 0$ and

$$i \neq k \Rightarrow v^{\{i\}} \cdot v^{\{k\}} = 0$$

Obviously such a set has angles of 90 degrees between each pair of different vectors. What is it good for ?



They are suitable as vectors defining the axes (plural of axis) in a coordinate system!

You can think of a set of orthogonal sets as vectors defining the axes of a coordinate system – no matter how high dimensional the space is. The point of having axes with 90 degrees between each other is to measure ...for every vector x in the space ... the component of x along one chosen axis independent of the other axes. Its about measurement and representation.

What is the value of having an orthogonal set ? For an orthogonal set, we can compute the coefficients of a linear combination very easily.
Suppose we want to represent y as a linear combination of the $v^{\{k\}}$ and find out the corresponding coefficients a_k:

$$y = \sum_{k=0}^{d-1} a_k v^{\{k\}}$$

compute inner product of both sides of the eq. with $v^{\{0\}}$

$$\begin{aligned} y \cdot v_0 &= \left(\sum_{k=0}^{d-1} a_k v^{\{k\}} \right) \cdot v^{\{0\}} \\ &= \sum_{k=0}^{d-1} a_k \; v^{\{k\}} \cdot v^{\{0\}} \\ &= a_0 \; v^{\{0\}} \cdot v^{\{0\}} + 0 \\ \frac{y \cdot v^{\{0\}}}{v^{\{0\}} \cdot v^{\{0\}}} &= a_0 \end{aligned}$$

The same holds if we take the inner product with any $v^{\{r\}}$:

$$y = \sum_{k=0}^{d-1} a_k v^{\{k\}}$$

$$\Rightarrow \frac{y \cdot v^{\{r\}}}{v^{\{r\}} \cdot v^{\{r\}}} = \frac{y \cdot v^{\{r\}}}{\|v^{\{r\}}\|_2^2} = a_r$$

Orthogonal sets and linear combination

If $\{v^{\{0\}},\dots,v^{\{d-1\}}\}$ is an Orthogonal set of vectors and y can be linearly represented with respect to this set, that is

$$y = \sum_{k=0}^{d-1} a_k v^{\{k\}},$$

then we have

$$\frac{y \cdot v^{\{r\}}}{v^{\{r\}} \cdot v^{\{r\}}} = \frac{y \cdot v^{\{r\}}}{\|v^{\{r\}}\|_2^2} = a_r$$

This has an interpretation: if $\{v^{\{0\}},\dots,v^{\{d-1\}}\}$ is an Orthogonal set, then $\frac{y \cdot v^{\{r\}}}{\|v^{\{r\}}\|_2^2}$ is the amount of $v^{\{r\}}$ which is present in y.

Orthogonal sets are independent sets!

Orthogonal sets are independent sets.

How to prove: by contradiction. We assume that an orthogonal set would be not independent, and then show a contradiction. The coefficients would be $\frac{v^{\{k\}} \cdot v^{\{r\}}}{\|v^{\{r\}}\|_2^2} = 0$.

In short, orthogonal sets are useful:

- they are an independent set, thus coefficients of a linear representation are unique
- when trying to represent a vector we can get the coefficients in a simple way
- as we can see below, there is an algorithm how to obtain an orthogonal set (with some possible zero vectors as undesired radioactive waste)

6.1 Projecting onto a vector and removing the direction of a vector

This is generally useful to know, independent from orthogonal sets – how to project onto a vector and how to remove its component from another vector, but we will also use it in the next subsection to construct an orthogonal set.

We will use this idea:

This has an interpretation: if $\{v^{\{0\}},\dots,v^{\{d-1\}}\}$ is an Orthogonal set, then $\frac{y \cdot v^{\{r\}}}{\|v^{\{r\}}\|_2^2}$ is the amount of $v^{\{r\}}$ which is present in y.

to remove the component of $v^{\{k\}}$ present in vector $v^{\{r\}}$ from the vector $v^{\{r\}}$. How to remove? Subtract it :)

1. Project a vector x onto a vector v:

$$x_{\|v} = \frac{x \cdot v}{v \cdot v} v = \left(x \cdot \frac{v}{\|v\|_2} \right) \frac{v}{\|v\|_2}$$

This means: $x_{\|v}$ is the component of x in direction of vector v.
Above also shows that the projection can be written in terms of a unit-length vector $\frac{v}{\|v\|_2}$.

Note: $\underbrace{\left(x \cdot \frac{v}{\|v\|_2} \right)}_{\text{real number}} \underbrace{\frac{v}{\|v\|_2}}_{\text{a vector}}$

2. Remove from a vector v a vector x:

$$x_{\perp v} = x - x_{\|v} = x - \frac{x \cdot v}{v \cdot v} v = x - \left(x \cdot \frac{v}{\|v\|_2} \right) \frac{v}{\|v\|_2}$$

This means: $x_{\perp v}$ has no component in direction of vector v, that is $x_{\perp v} \cdot v = 0$.

We can show that $x_{\perp v}$ it is indeed orthogonal to v

$$x_{\perp v} \cdot v = x \cdot v - x_{\|v} \cdot v = x \cdot v - \frac{x \cdot v}{v \cdot v}(v \cdot v) = x \cdot v - x \cdot v = 0$$

Source of mistake: This works if you remove the direction of a single vector from x. However (!!!), if you subtract two vectors $v^{\{0\}}$ and $v^{\{1\}}$ from x, and these two vectors share a component in common ($v^{\{0\}} \cdot v^{\{1\}} \neq 0$), then you can reintroduce a component back by mistake. Here an extreme example, where we use for convenience $v^{\{0\}} = v^{\{1\}} = v$:

$$\begin{aligned} x_{\text{bad } v^{\{0\}}, v^{\{1\}}} = x - x_{\|v^{\{0\}}} - x_{\|v^{\{1\}}} = x - x_{\|v} - x_{\|v} = x - 2x_{\|v} \\ x_{\text{bad } v^{\{0\}}, v^{\{1\}}} \cdot v = 0 - 1x_{\|v} \cdot v = -x \cdot v \end{aligned}$$

Good news: one can avoid this mistake by applying this iteratively, that is removing $v^{\{1\}}$ from $x_{\perp v^{\{0\}}}$ (and not from x itself as wrongly done above)!

What you got to remember:

Projecting onto a vector and removing the direction of a vector

Project a vector x onto a vector v:

$$x_{\|v} = \frac{x \cdot v}{v \cdot v} v = \left(x \cdot \frac{v}{\|v\|_2} \right) \frac{v}{\|v\|_2}$$

Remove from a vector x a vector v:

$$x_{\perp v} = x - x_{\|v}$$

6.2 Constructing an orthogonal set using inner products

This is also known as Gram-Schmidt orthogonalization.

It has two properties:

- The input is a set of vectors. The output is a set of orthogonal vectors. Some of the vector can be zero vectors[1].
- The vector space spanned by the first i output vectors is the same as the vector space spanned by the first i input vectors.

Algorithm Input: non-zero vectors $v^{\{0\}}, \dots, v^{\{d-1\}}$

- Run a for loop from Step 0 to step $d - 1$:
- Step 0: Initialize:

$$\tilde{v}^{\{0\}} = v^{\{0\}}$$

- Step 1: Remove from $v^{\{1\}}$ the direction of $\tilde{v}^{\{0\}}$, as much as there is present:

$$\tilde{v}^{\{1\}} = v^{\{1\}} - v^{\{1\}}{}_{\|\tilde{v}^{\{0\}}}$$

[1] when doing this on a computer, they usually will not become zero but have a very small norm

We just compute $(v^{\{1\}})_{\perp v^{\{0\}}}$ as in the previous section.

If $\widetilde{v}^{\{1\}} \neq 0$, they are now an orthogonal system of two vectors ($\widetilde{v}^{\{0\}}, \widetilde{v}^{\{1\}}$).

Next step is to add $v^{\{2\}}$ to it!

- Step 2: Remove from $v^{\{2\}}$ the direction of $\widetilde{v}^{\{0\}}$ and $\widetilde{v}^{\{1\}}$, as much as there is present:

$$\begin{aligned}\widetilde{v}^{\{2\}} &= v^{\{2\}} - v^{\{2\}}{}_{\|\widetilde{v}^{\{0\}}} - v^{\{2\}}{}_{\|\widetilde{v}^{\{1\}}} \\ &= v^{\{2\}} - \sum_{k=0}^{1} v^{\{2\}}{}_{\|\widetilde{v}^{\{k\}}}\end{aligned}$$

Are not we doing the mistake we just warned ourselves? The answer is no: $\widetilde{v}^{\{0\}}$ and $\widetilde{v}^{\{1\}}$ are orthogonal to each other: $\widetilde{v}^{\{0\}} \cdot \widetilde{v}^{\{1\}} = 0$, so one can subtract it directly from $\widetilde{v}^{\{2\}}$ without running into the mistake from above.

One can show: $\widetilde{v}^{\{2\}} \cdot \widetilde{v}^{\{1\}} = 0$, $\widetilde{v}^{\{2\}} \cdot \widetilde{v}^{\{0\}} = 0$

If $\widetilde{v}^{\{2\}} \neq 0$, them $\widetilde{v}^{\{0\}}, \widetilde{v}^{\{1\}}, \widetilde{v}^{\{2\}}$ are an orthogonal system.

- Step 3: Remove from $v^{\{3\}}$ the direction of $\widetilde{v}^{\{0\}}$, $\widetilde{v}^{\{1\}}$ and $\widetilde{v}^{\{2\}}$, as much as there is present:

$$\widetilde{v}^{\{3\}} = v^{\{3\}} - \sum_{k=0}^{2} v^{\{3\}}{}_{\|\widetilde{v}^{\{k\}}}$$

One can show: $\widetilde{v}^{\{3\}} \cdot \widetilde{v}^{\{2\}} = 0$, $\widetilde{v}^{\{3\}} \cdot \widetilde{v}^{\{1\}} = 0$, $\widetilde{v}^{\{3\}} \cdot \widetilde{v}^{\{0\}} = 0$

Why is that the case ? Well, we remove the amount of vector $\widetilde{v}^{\{k\}}$ from $v^{\{i\}}$ when we do:

$$v^{\{i\}} - v^{\{i\}}_{\|\widetilde{v}^{\{k\}}}$$

The second important insight is that we have $\widetilde{v}^{\{r\}} \cdot \widetilde{v}^{\{k\}} = 0$ if $k \neq i$, so that we do not reintroduce a component parallel to $\widetilde{v}^{\{k\}}$ in later steps.

Therefore this must be orthogonal to $\widetilde{v}^{\{k\}}$!

next: take it to step r and generalize it.

- Step r: Remove from $v^{\{r\}}$ the direction of $\widetilde{v}^{\{0\}}, \widetilde{v}^{\{1\}}, \ldots, \widetilde{v}^{\{r-1\}}$, as much as there is present:

$$\widetilde{v}^{\{r\}} = v^{\{r\}} - \sum_{k=0}^{r-1} v^{\{r\}}{}_{\|\widetilde{v}^{\{k\}}}$$

Again: We can directly subtract , because $\widetilde{v}^{\{k\}} \cdot \widetilde{v}^{\{l\}} = 0$ if $k \neq l$.

Orthogonalization (numerically not so stable!)

- Algorithm Input: non-zero vectors $v^{\{0\}}, \dots, v^{\{d-1\}}$
- Step 0: Initialize:
$$\widetilde{v}^{\{0\}} = v^{\{0\}}$$
- run a for loop over $0, \dots, d-1$
- step r (in a `for`-loop from $r = 1$ to $r = d-1$): have got so far: $\widetilde{v}^{\{0\}}, \dots, \widetilde{v}^{\{r-1\}}$
$$\begin{aligned}\widetilde{v}^{\{r\}} &= v^{\{r\}} - \sum_{k=0}^{r-1} \frac{v^{\{r\}} \cdot \widetilde{v}^{\{k\}}}{\widetilde{v}^{\{k\}} \cdot \widetilde{v}^{\{k\}}} \widetilde{v}^{\{k\}} \\ &= v^{\{r\}} - \sum_{k=0}^{r-1} v^{\{r\}}{}_{\|\widetilde{v}^{\{k\}}}\end{aligned}$$
at each step: Remove zero vectors $\widetilde{v}^{\{k\}} = 0$ (you cannot divide by $\|\widetilde{v}^{\{k\}}\|_2^2$ for a zero vector).
- Return the non-zero vectors from $\widetilde{v}^{\{0\}}, \dots, \widetilde{v}^{\{d-1\}}$ as orthogonal set

Note:

- If the result of this algorithm has some vectors being zero, then the set was not independent.
- If the result of this algorithm results in d non-zero vectors, and d is the dimensionality of the vector space, then we have obtained a basis of the vector space.
- when one obtains a zero vector along the way, one drops it from the final result.
- you can write the removal formula also in terms of unit length vectors $z^{\{k\}} = \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2}$:
$$\begin{aligned}\widetilde{v}^{\{r\}} &= v^{\{r\}} - \sum_{k=0}^{r-1} \left(v^{\{r\}} \cdot \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2} \right) \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2} \\ &= v^{\{r\}} - \sum_{k=0}^{r-1} \left(v^{\{r\}} \cdot z^{\{k\}} \right) z^{\{k\}}\end{aligned}$$
Here the removal of the k-th unit length vector $z^{\{k\}}$ from the input $v^{\{r\}}$ becomes clearer. $\left(v^{\{r\}} \cdot z^{\{k\}}\right)$ is the amount of the unit vector $z^{\{k\}}$ in the input $v^{\{r\}}$.
Note: this formula requires that this holds: $\|z^{\{k\}}\|_2 = 1$ (unit length property)
- disadvantage: Gram-Schmidt can give numerical problems when implemented on a computer due to subtractive cancellation. One may get vectors which are not zero but only almost zero when they should be. They will have non-zero inner products to the actual result vectors.

  It is suitable, however, to understand how to construct orthogonal bases for subspaces.

A numerically more stable version is:

Modified Gram Schmidt (better numerical stability)

- for k in range(0,d): $\widetilde{v}^{\{k\}} = v^{\{k\}}$ init
- for k in range(0,d):
  - $\widetilde{z}^{\{k\}} = \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2}$
  - for m in range(k+1,d): $\widetilde{v}^{\{m\}} = \widetilde{v}^{\{m\}} - (\widetilde{v}^{\{m\}} \cdot \widetilde{z}^{\{k\}})\widetilde{z}^{\{k\}}$

The difference is: when one obtained a normalized $\widetilde{z}^{\{k\}}$, then one removes its component from all “future” $\widetilde{v}^{\{l\}}$ with indices $l > k$. Then one moves forward by incrementing $k = k + 1$.

6.3 One last result:

Orthogonalization preserves vector subspaces

Any vector u which can represented as a linear combination of the vectors $v^{\{0\}}, \ldots, v^{\{i\}}$ up to index i, can also be represented by the output $\widetilde{v}^{\{0\}}, \ldots, \widetilde{v}^{\{i\}}$ of the Gram-Schmidt process up to the same index i. This holds for every index i.

What is the meaning of this ? Lets recap: “Definition of a vector space spanned by a set of vectors”.

Insight

Gram-Schmidt not only allows to create an orthogonal set, but the output has the same ability for representing vectors (as linear combination) as the input to the algorithm – at each step of the algorithm.

The proof of the above claim:

Lets assume that $u = \sum_{r=0}^{i} a_r v^{\{r\}}$ is a linear combination of $v^{\{0\}}, \ldots, v^{\{i\}}$ up to index i. Then:

$$\begin{aligned}
u &= \sum_{r=0}^{i} a_r v^{\{r\}} \\
\widetilde{v}^{\{r\}} &= v^{\{r\}} - \sum_{k=0}^{r-1} \left(v^{\{r\}} \cdot \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2} \right) \frac{\widetilde{v}^{\{k\}}}{\|\widetilde{v}^{\{k\}}\|_2} \\
&= v^{\{r\}} - \sum_{k=0}^{r-1} c_k \widetilde{v}^{\{k\}} \text{ for some real number } c_k \\
\text{solve for } \Rightarrow v^{\{r\}} &= \widetilde{v}^{\{r\}} + \sum_{k=0}^{r-1} c_k \widetilde{v}^{\{k\}} \\
\text{plug it in } \Rightarrow u &= \sum_{r=0}^{i} a_r \widetilde{v}^{\{r\}} + \sum_{r=0}^{i} a_r \sum_{k=0}^{r-1} c_k \widetilde{v}^{\{k\}}
\end{aligned}$$

In the last double sum: r runs up to i and k up to $r-1$, so k runs up to at most $i-1$. The first sum runs up to i, therefore we have a linear combination which uses $\widetilde{v}^{\{k\}}$ until $k = i$ and does not use higher indices. Done!

7 Matrix-vector and Matrix-Matrix multiplication

7.1 Matrix-vector

Given a matrix of shape (n, d) and a vector of dimensionality d, the multiplication Ax of them (with x being on the right hand side) is defined as a vector of length n such that

$$\begin{aligned}
Ax &= \begin{bmatrix} a_{0,0} & \cdots & a_{0,d-1} \\ a_{1,0} & \cdots & a_{1,d-1} \\ \vdots & \ddots & \vdots \\ a_{n-1,0} & \cdots & a_{n-1,d-1} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_{d-1} \end{bmatrix} \\
&= \begin{bmatrix} \begin{bmatrix} a_{0,0} & \cdots & a_{0,d-1} \end{bmatrix} \cdot x \\ \begin{bmatrix} a_{1,0} & \cdots & a_{1,d-1} \end{bmatrix} \cdot x \\ \vdots \\ \begin{bmatrix} a_{n-1,0} & \cdots & a_{n-1,d-1} \end{bmatrix} \cdot x \end{bmatrix}
\end{aligned}$$

Thus, Ax is a vector and the k-th component of vector Ax is given as an inner product

$$\begin{aligned}(Ax)_k &= A[k,:] \cdot x \\ &= (a_{k,0} \dots a_{k,d-1}) \cdot x \\ &= \sum_{r=0}^{d-1} a_{k,r} x_r\end{aligned}$$

between the k-th row of A and vector x. The matrix multiplication with a vector from the right is only defined, if the number of columns in A is equal to the dimensionality of the vector x.

Given a matrix of shape (n, d) and a vector of dimensionality n, the multiplication $x^\top A$ of them (with $x^\top$ being on the left hand side) is defined as a vector of length d such that

$$\begin{aligned}x^\top A &= \begin{bmatrix} x_0, x_1, \dots, x_{n-1} \end{bmatrix} \begin{bmatrix} a_{0,0} & \dots & a_{0,d-1} \\ a_{1,0} & \dots & a_{1,d-1} \\ \vdots & \ddots & \vdots \\ a_{n-1,0} & \dots & a_{n-1,d-1} \end{bmatrix} \\ &= \begin{bmatrix} x \cdot \begin{bmatrix} a_{0,0} \\ \vdots \\ a_{n-1,0} \end{bmatrix} & x \cdot \begin{bmatrix} a_{0,1} \\ \vdots \\ a_{n-1,1} \end{bmatrix} & \dots & x \cdot \begin{bmatrix} a_{0,d-1} \\ \vdots \\ a_{n-1,d-1} \end{bmatrix} \end{bmatrix}\end{aligned}$$

Thus, $x^\top A$ is a vector and the k-th component of vector $x^\top A$ is given as an inner product

$$\begin{aligned}(x^\top A)_k &= x \cdot A[:,k] \\ &= (a_{0,k} \dots a_{n-1,k}) \cdot x \\ &= \sum_{r=0}^{n-1} x_r a_{r,k}\end{aligned}$$

Consequence:

inner products as matrix vector multiplication

The inner product $x \cdot y$ of two column-shaped vectors can be written in matrix-vector multiplication notation as

$$x \cdot y = \begin{bmatrix} x_0, x_1, \dots, x_{d-1} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{d-1} \end{bmatrix} = x^\top y$$

where $x^\top$ is the transpose of a vector or matrix x.

The transpose was used here to convert the column-shaped vector into a row-shaped vector.

7.2 Matrix-Matrix

Given a matrix A of shape (n, d) and a matrix B of shape (d, f), their multiplication AB is defined as a matrix of shape (n, f) and its component $(AB)_{i,k}$ at row i and colunmn k is given as

$$(AB)_{i,k} = A_{(i,:)} \cdot B_{(:,k)} = \sum_{r=1}^{d} A_{i,r} B_{rk}$$

as an inner product between the i-th row of the left matrix and the $k-$th column of the right matrix.

Important: the number or dimensions in the second axis of A must be equal to the number or dimensions in the first axis of B. Otherwise matrix multiplication is not possible.

Therefore a matrix- matrix multiplication is a matrix consisting of inner products:

$$AB = \begin{bmatrix} A_{(0,:)} \cdot B_{(:,0)} & A_{(0,:)} \cdot B_{(:,1)} & A_{(0,:)} \cdot B_{(:,2)} & \dots & A_{(0,:)} \cdot B_{(:,f-1)} \\ A_{(1,:)} \cdot B_{(:,0)} & A_{(1,:)} \cdot B_{(:,1)} & A_{(1,:)} \cdot B_{(:,2)} & \dots & A_{(1,:)} \cdot B_{(:,f-1)} \\ A_{(2,:)} \cdot B_{(:,0)} & A_{(2,:)} \cdot B_{(:,1)} & A_{(2,:)} \cdot B_{(:,2)} & \dots & A_{(2,:)} \cdot B_{(:,f-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{(n-1,:)} \cdot B_{(:,0)} & A_{(n-1,:)} \cdot B_{(:,1)} & A_{(n-1,:)} \cdot B_{(:,2)} & \dots & A_{(n-1,:)} \cdot B_{(:,f-1)} \end{bmatrix}$$

Memorizing

- Matrix-Matrix multiplication results in a matrix, if shapes are permissible:
  $(n, d)(d, f) \to (n, f)$.
- The component $(AB)_{i,k}$ at row i and column k is given as the inner product between row i of the left matrix and column k of the right matrix.

  That also tells you which axes must match in dimensionality: left matrix - the number of columns $=$ the dimensionality of the second axis. right matrix - the number of rows $=$ the dimensionality of the first axis.

8 Out-of-Exam Bonus: Examples of non-finite dimensional vector spaces with norms and inner products

For those who are bored with basic linear algebra. This is a part of the field of functional analysis. There are more vector spaces than $a = (a_0, a_1, a_2)$.

8.1 Vector spaces of sequences

Consider the set of all sequences:

$$(a_0, a_1, a_2, \dots), a_i \in \mathbb{R}$$

Obviously this is a vector space too. You can element-wise add and multiply them.

How to provide norms for such examples?

$$a = (a_0, a_1, a_2, \dots) \mapsto \|a\|_p = (\sum_{i=0}^{\infty} |a_i|^p)^{1/p}$$

if this would be a finite sum. This is the direct extension of the p-norm for finite vectors to sequences.

Examples can be taken from convergent sequences, and taking their p-th squareroot: We know $\sum_{i=0}^{\infty} 2^{-i} = 2$, $\sum_{i=0}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$. Therefore

$$a_i = 2^{-i/p}$$
$$a_i = \frac{1}{i^{2/p}}$$

are valid sequences.

You can define normed vector spaces from all those elements whose norm is finite:

$$S_p = \{(a_0, a_1, a_2, \dots), a_i \in \mathbb{R} : \|a\|_p < \infty\}$$

Are they closed under vector space operations? Lets consider multiplication with a scalar:

$$a = (a_0, a_1, a_2, \dots),\ (\sum_{i=0}^{\infty} |a_i|^p)^{1/p} < \infty ca = (ca_0, ca_1, ca_2, \dots) \Rightarrow (\sum_{i=0}^{\infty} |ca_i|^p)^{1/p} = c(\sum_{i=0}^{\infty} |a_i|^p)^{1/p} < \infty$$

Lets consider addition of two sequences. This is more tricky to prove.

$$a = (a_0, a_1, a_2, \ldots), \|a\|_p < \infty, b = (b_0, b_1, b_2, \ldots), \|b\|_p < \infty, \|a+b\|_p^p = \sum_{i=0}^{\infty} |a_i + b_i|^p$$

The trick is to use here convexity. $f(x) = |x|^p$ is a convex function if $x \geq 0$. Therefore: $f(1/2a + 1/2b) \leq 1/2f(a) + 1/2f(b)$ and in our case:

$$|1/2a_i + 1/2b_i|^p \leq |1/2|a_i| + 1/2|b_i||^p \leq 1/2|a_i|^p + 1/2|b_i|^p$$

Plug this back in:

$$\begin{aligned} \|a+b\|_p^p &= \sum_{i=0}^{\infty} 2^p |1/2a_i + 1/2b_i|^p \leq \sum_{i=0}^{\infty} 2^p (1/2|a_i|^p + 1/2|b_i|^p) \\ &= 2^{p-1} \sum_{i=0}^{\infty} |a_i|^p + 2^{p-1} \sum_{i=0}^{\infty} |b_i|^p < \infty \end{aligned}$$

We have shown:

$$S_p = \{(a_0, a_1, a_2, \ldots), a_i \in \mathbb{R} : \|a\|_p < \infty\}$$

is closed under multiplication and addition of elements, and therefore a valid normed vector space.

General Vector spaces with a norm are called normed vector spaces. The normed vector space above has an additional property of completeness under the norm (every sequence of elements such that $\sum_{i=0}^{\infty} \|v_i\| < \infty$ has the property that its limit $\sum_{i=0}^{\infty} v_i$ exists as a vector and lies within the space) which makes it a Banach-Space after Stefan Banach https://en.wikipedia.org/wiki/Stefan_Banach.
more eg on https://www.math.ucdavis.edu/~hunter/book/ch5.pdf.

8.2 Vector spaces of functions

consider the set of functions over an interval $[a, b]$ or (a, b).
Then you can define a vector space with norm as:

$$V = \{f : \|f\|_p = (\int_a^b |f(x)|^p dx)^{1/p} < \infty\}$$

Functions can be added and multiplied. The finiteness or the norm is preserved under multiplication and addition. The proof for addition is very similar to the one above for sequences.

One can also define more exotic norms on a closed interval like $[a, b]$:

$$V = \{f : f \text{ is a continuous function on } [a, b]\}, \; \|f\| = \max_{x \in [a,b]} |f(x)|$$

Yes, this defines a norm on continuous functions.

8.3 Vector spaces with inner product

You can define inner products for sequences and functions:
On sequences:

$$a = (a_0, a_1, a_2, \ldots), \; b = (b_0, b_1, b_2, \ldots)$$

$$a \cdot b = \sum_{i=0}^{\infty} a_i b_i$$

This is a pretty straightforward extension of the ideas for vectors. The corresponding vector space would be:

$$V = \{(a_0, a_1, a_2, \ldots) : a \cdot a = \|a\|_2^2 < \infty\}$$

On functions you can define inner products via the integral:

$$f \cdot g = \int_a^b f(x)g(x)dx$$

The corresponding vector space would be:

$$V = \{f : f \cdot f = \|f\|_2^2 = \int_a^b |f(x)|^2 dx < \infty\}$$

This contains all continuous functions, but it contains also bounded functions with discontinuities / jumps.

Btw, not every function satisfies $\int_a^b |f(x)|^2 dx < \infty$. A counterexample for the interval $(0, 1]$ would be $f(x) = 1/x$. Then

$$\int_0^1 \frac{1}{x^2} dx = -\frac{1}{x}|_0^1 = \infty$$

The pattern for defining inner products follows this idea:

- Aggregation via summing for finite elements $\{u_i, i = 0, \dots, d-1\}$

  $$u \cdot v = \sum_{i=0}^{d-1} u_i v_i$$

- summing for countably infinite elements $\{u_i, i = 0, \dots\}$

  $$u \cdot v = \sum_{i=0}^{\infty} u_i v_i$$

  subject to its existence

- integrating for a non-countable set $\{u(x), x \in \mathcal{X}\}$

  $$u \cdot v = \int_{x \in \mathcal{X}} u(x)v(x)dx$$

  subject to its existence

Banach spaces where the norm is defined by an inner product are called Hilbert-spaces after `https://en.wikipedia.org/wiki/David_Hilbert`.