Linear Algebra

Pierre Glaser, William Dorrell, Dimitri Meunier

September 27, 2025

Contents

Chapter 1

Preliminaries

1.1 Statements, Variables and Vocabulary

Statements, Variables

- A *statement* in mathematics is a declarative sentence, for instance, “the sky is red”.
- A *variable* is a symbol that can represent different objects. For instance, n often stands for an arbitrary integer. In the following, P will denote an arbitrary statement. If the statement P depends on a variable n, we note $P(n)$. Example: “n is even”.
- once fully specified, a statement is either true or false.

Logical Operations

- If P is a statement, the negation of P is the assertion “P is false” (denoted as: $\neg P$).
- If P and Q are two statements, the statement
  - (P and Q) is the assertion “P is true and Q is true” (denoted as $(P \wedge Q)$).
  - (P or Q) is the assertion “at least one of P, Q is true” (denoted as $(P \vee Q)$).

Implications

- For two statements P, Q, we define the statement $(P \implies Q)$ as meaning “if P is true, then Q is true”. Q can be true even when P is not, but it can’t be false when P holds.
- Example: It rains $\implies$ There are clouds in the sky.
- Important: $(P \implies Q) \implies (\neg Q \implies \neg P)$, and $(\neg Q \implies \neg P) \implies (P \implies Q)$.

Equivalencies

For two statements P and Q, we say that P and Q are equivalent, and note $P \iff Q$ when $P \implies Q$ and $Q \implies P$.

1.2 Set Theory Concepts

1.2.1 Sets

Set

- A set is a collection of objects. Examples: $\mathbb{N}$: set of natural numbers. $\mathbb{R}$: set of real numbers. $\{1, 2, \pi\}$, the set of 2-d rotations, the empty set $\emptyset$.
- If E is a set and x is an object, the statement "x is an element of E" is written as "$x \in E$". The statement "x is not an element of E" is written as "$x \notin E$".
- Equality of two sets. Two sets A and B are equal when they have the same elements.

1.2.2 Inclusion, subsets

Subsets

Definition 1 (Inclusion between sets, subsets of a set). *Let A and E be two sets. We say that A is a subset of E, and note $A \subset E$, if all elements of A belong to E, e.g. $A \subset E : \forall x \in A,\ x \in E$.*

Cartesian Product of a Finite Family of Sets

Definition 2 (Cartesian Product of a Finite Family of Sets). *Let $A_1, A_2, \dots, A_n$ be n sets for some $n \in \mathbb{N}$. The Cartesian product $A_1 \times A_2 \times \cdots \times A_n$ is defined as*

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \dots, a_n) \mid a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n\}.$$

if $A_1 = \cdots = A_n = A$, we denote it as A^n.

1.3 Universal and Existential Quantifiers

Quantifiers: "for all" ($\forall$) and "there exists" ($\exists$)

Suppose given a set $\mathcal{X}$ and, for every element x of $\mathcal{X}$, an assertion $P(x)$.

- The statement " $\forall\ x \in \mathcal{X}, P(x)$" means: "for every x in $\mathcal{X}$, $P(x)$ holds." Example: $\forall n \in \mathbb{N},\ n \geq 0$.
- The statement " $\exists\ x \in \mathcal{X} : P(x)$" means: "there exists (at least) one element x of $\mathcal{X}$ such that $P(x)$ holds." Example: $\exists n \in \mathbb{N},\ n \leq 2$. If that element is unique, we write " $\exists!\ x \in \mathcal{X} : P(x)$".

- Such quantifiers can be assembled into more complex statements. Example a subset $A \subset \mathbb{N}$ is bounded if $\exists M \in \mathbb{N} : \forall \in A \; n \leq M$.
- Important exercise: define an unbounded (e.g. a set which is not bounded) set.

Method – Negation of a statement with nested quantifiers

To negate a statement containing multiple quantifiers,

- change all $\forall$ to $\exists$ and all $\exists$ to $\forall$, carefully maintaining the order,
- write the negation of the non-quantified part.

1.4 Reasoning: methods and examples

To prove statements of the type:

- "$\forall x \in E, P(x)$", start with an arbitrary element of E. When E is finite, one can also check $P(x)$ for every concrete value in $\mathcal{X}$ and show $P(x)$.
- $\exists x \in E : \; P(x)$, find a concrete example of such an element, when possible.
- $A \implies B$, one can either:
  - Start by assuming A, and show B.
  - Assume $\neg B$, and show $\neg A$.
- $A \iff B$ one often needs to show separately $A \implies B$ and $B \implies A$.
- $\forall n \in \mathbb{N}, \; P(n)$, show separately
  - assertion $P(0)$ is true,
  - $[P(n) \implies P(n+1)]$.

1.5 Sums and Products

Sum & Product Notation

Given some elements $a_1, a_2, \ldots, a_n$ in $\mathbb{R}$, we will note

- $a_1 + a_2 + \ldots + a_n$ as $\sum_{i=1}^{n} a_i$
- $a_1 \times a_2 \times \ldots \times a_n$ as $\prod_{i=1}^{n} a_i$

Product and sums can be nested. Example:

$$\sum_{i=1}^{n} \sum_{j=1}^{i} a_j = a_1 + (a_1 + a_2) + (a_1 + a_2 + a_3) + ... + (a_1 + a_2 + ... + a_n) \; .$$

$$\sum_{i=1}^{n} \sum_{j=1}^{i} a_j^i = a_1^1 + (a_1^2 + a_2^2) + (a_1^3 + a_2^3 + a_3^3) + ... + (a_1^n + a_2^n + ... + a_n^n) \; .$$

1.6 Exercise Sheet 1: Preliminaries

Exercise: Necessary and Sufficient Conditions

Exercise 1. *1. Complete each statement in ... with one of the symbols* $\Longrightarrow$ *,* $\Longleftarrow$ *,* $\Longleftrightarrow$ *:*

(a) For $n \in \mathbb{N}$*,* n *is multiple of 2 ...* n *is multiple of 4 or* n *is multiple of 6.*

(b) For $x \in \mathbb{R}$*,* $\sqrt{x^2+4x+5} = 1 \ldots x^2+4x+5 = 1$

(c) For $x \in \mathbb{R}$*,* $\sqrt{x^2+4x+5} = 0 \ldots x^2+4x+5 = 0$

(d) For $x \in \mathbb{R}$*,* $x - 3 = x^2 + 2x \ldots e^{x-3} = e^{x^2+2x}$

2. Complete each statement with "necessary", "sufficient", or "necessary and sufficient". In all assertions, x *and* y *are real numbers and* n *is a natural number.*

(a) $x > 2$ *is ... for* $x^2 > 4$*.*

(b) $x + y = 5$ *is ... to have* $x = 2$ *and* $y = 3$*.*

(c) for n *to be multiple of 4, it is ... that* n *be the square of an even integer.*

(d) for $x+y$ *to be equal to 5 and* xy *to be equal to 6, it is ... that* $x = 2$ *and* $y = 3$*.*

Solution. 1. First exercise:

(a) $\Longleftarrow$, since n is a multiple of 4 or and 6, and both 4 and 6 are multiples of 2, n is a multiple of 2. The converse is not true, since 2 is not a multiple of 4 or 6.

(b) $\Longrightarrow$, squaring $\sqrt{x^2+4x+5} = 1$ recovers $x^2+4x+5 = 1$, however the converse is not true, since knowing only that $x^2+4x+5 = 1$, we see that $\sqrt{x^2+4x+5} = \pm 1$.

(c) $\Longleftrightarrow$, This problem is the same as the above, except $\pm 0 = 0$, so the equivalence goes both ways.

(d) $\Longleftrightarrow$, since the exponential is a 1-to-1 map (it is bijective, see later classes, this question will be removed!), $a = b$ is equivalent to $e^a = e^b$.

2. Second exercise:

(a) Sufficient. $x > 2$ implies, by squaring, that $x^2 > 4$. However, if $x < -2$ it is also true that $x^2 > 4$, hence it is not necessary.

(b) Necessary. $x + y = 5$ has to be true for $x = 2$ and $y = 3$, but there are many other choices of x and y such that $x + y = 5$, hence it is not sufficient.

(c) Sufficient. If n is the square of an even number, $n = (2p)^2 = 4p$ for some integer p, hence it is divisible by 4. However, there are n divisible by 4 such that it is not the square of an even integer, e.g. $8 = (2\sqrt{2})^2$.

(d) Sufficient. It is clear that if $x = 2$ and $y = 3$, $x + y = 5$ and $xy = 6$. However an alternative is that $x = 3$ and $y = 2$, hence it is not necessary that $x = 2$ and $y = 3$.

■

Exercise: Reading Statements

Exercise 2. *Consider the sets* $A = \{1, 2, 3\}$*,* $B = \{1, 2, 3, 4, 5\}$*,* $C = \{2, 4\}$*, and* $D = \mathbb{N}$*. Determine whether the following statements are true or false.*

(a) $\forall x \in A, x \in B$

(b) $\exists x \in B, x \in A$

(c) $\exists x \in A, x \notin B$

(d) $\exists x \in B, x \notin A$

(e) $\forall x \in C, \exists y \in B, x \leq y$

(f) $\forall x \in A, \exists y \in B, x \leq y$

(g) $\exists x \in C, \forall y \in A, y \leq x$

(h) $\exists x \in B, \forall y \in A, \forall z \in C, y + z \leq 2x$

(i) $\exists x \in B, \forall y \in B, x \leq y.$

(j) $\exists x \in D, \forall y \in D, x \leq y.$

(k) $\exists x \in D, \forall y \in D, x \geq y.$

(l) $\forall x \in D, \exists y \in A, x = y.$

(m) $\exists x \in D, \exists y \in A, x = y.$

(n) $\exists x \in D, \forall y \in A, x = y.$

Solution. (a) True, by exhaustion.

(b) True, 2.

(c) False, all of A is in B.

(d) True, 4.

(e) True, $5 \in B$ is bigger than all elements of C.

(f) True, $5 \in B$ is bigger than all elements of A.

(g) False, $1 \in A$ is smaller than all elements of C, hence no $y \in C$ is smaller than all $x \in A$.

(h) True, $10 = 2 \times 5, 5 \in B$ is greater than the sum of the greatest elements of A and C:7, therefore it is greater than $y + z \forall y \in A, z \in C$.

(i) True, $1 \in B$ is the minimal element of B, so $1 \leq y \forall y \in B$.

(j) True, 0 is less than or equal to all nonnegative integers.

(k) False, for any specific $x \in D$ we can find $y \in D$ that is larger, e.g. $x + 1$.

(l) False, take 4, that is in $\mathbb{N}$ but not A, so there is no $x \in A$ such that $x = 4$.

(m) True, 1 is in both D and A.

(n) False, there is no one number such that 1,2 and 3 all equal that number.

■

Nested Quantifiers and Negation

Exercise 3. *1. Write the negation of each of the following statements:*

(a) $\exists x \in \mathbb{R} : \forall n \in \mathbb{N}, x > 2n$

(b) $\forall n \in \mathbb{N} : \exists x \in \mathbb{R}, x > 2n$

(c) For all real numbers x and y, if $x \geq y$ then $x^2 \geq y^2$.

2. For each of the statements above, determine if it is true or false, and justify carefully. You may use the following fact without justification: if x is a real number, then there exists an integer $n \in \mathbb{N}$ such that $n > x$.

3. Consider the statement:

$$\exists x \in \mathbb{R} : \forall y \in \mathbb{R}, x^2 + 2xy = 1. \quad (P)$$

(a) Write the negation of (P).

(b) Is statement (P) true?

4. Is the following statement true or false?

$$\forall x \in \mathbb{N}, \exists y \in \mathbb{N} : \exists z \in \mathbb{N} : (x \leq y \text{ and } y = 2z)$$

Solution. 1. (a) $\forall x \in \mathbb{R} : \exists n \in \mathbb{N}, x \leq 2n$

(b) $\exists n \in \mathbb{N} : \forall x \in \mathbb{R}, x \leq 2n$

(c) $\exists x, y \in \mathbb{R} : x \geq y$ and $x^2 < y^2$

2. (a) False. For any $x \in \mathbb{R}$ choose an $n \in \mathbb{N} : n > x$. Clearly $x < 2n$.

(b) True. Given an $n \in \mathbb{N}$, choose $x = 2n + 0.1$. Clearly $x > 2n$.

(c) True. If $x \geq y$ we can square both sides and derive that $x^2 \geq y^2$.

3. (a) $\forall x \in \mathbb{R} : \exists y \in \mathbb{R}, x^2 + 2xy \neq 1$

(b) True, solving the quadratic equaiton: $x = -y \pm \sqrt{y^2 + 1}$ satisifes the equality and can be solved for all $y \in \mathbb{R}$.

4. True, choose $y = 2x$, then $x \leq y$ and $y = 2z$ for $z \in \mathbb{N}$ since $z = x \in \mathbb{N}$

■

Exercise 4. *Consider $A = \{n \in \mathbb{N} \mid \exists k \in \mathbb{N} : n = k(k+1)\}$ and let B be the set of even integers. Show the inclusion $A \subset B$.*

Solution. Begin with an element $a \in A$. Since it is an element of A, $a = p(p+1)$ for some $p \in \mathbb{N}$. Let's consider two cases, where p is even or odd. If p is even then $p = 2q$ for some $q \in \mathbb{N}$. Then $a = 2q(2q+1)$, $q(2q+1)$ is just some integer, hence a is 2 times some integer, so is divisible by 2. Conversely, if p is odd, $p = 2q+1$ for some $q \in \mathbb{N}$. Then $a = (2q+1)(2q+2) = 2(2q+1)(q+1)$, where $(2q+1)(q+1)$ is some integer, so similarly, a is divisible by 2. Finally, to show that A is a strict subset of B, we need an element of B that is not in A. An example of this is 4. ■

Exercise 5.

- *What is* $\prod_{i=1}^n ca_i$ *in terms of* $\prod_{i=1}^n a_i$*?*
- *What is* $\sum_{i=1}^m a_i - \sum_{i=1}^n a_i$ *for* $n < m$*?*
- *Write* $e^{\sum_{i=1}^n a_i}$ *in terms of each term* e^{a_i}*.*
- *What is* $\sum_{i=1}^n c$*?*
- $S_n = \sum_{i=1}^n a^i$ *is the sum of a geometric progression. By examining an expression for* $aS_n - S_n$ *show that* $S_n = \frac{a(a^n-1)}{a-1}$*. If* $|a| < 1$ *what is the limit as* $n \implies \infty$*? (you could also try and show this by induction)*

Solution. 1. $c^n \prod_{i=1}^n a_i$

2. $\sum_{i=n+1}^m a_i$

3. $\prod_{i=1}^n e^{a_i}$

4. nc

5. $aS_n - S_n = \sum_{i=2}^{n+1} a^i - \sum_{i=1}^n a_i = a^{n+1} - a = (a-1)S_n$, hence: $S_n = \frac{a(a^n-1)}{a-1}$. If $|a| < 1$ then a^n tends to zero, hence S_n tends to $\frac{a}{1-a}$.

■

Exercise 6.

1. *Show (by induction, although other proofs exist) that* $\sum\limits_{i=1}^n i = \frac{n(n+1)}{2}$ *for all* $n \in \mathbb{N}^\star$*.*

2. *Show that* $\sum\limits_{i=1}^n i + c = \frac{n(n+1)}{2} + cn$*, for all* $c \in \mathbb{R}, n \in \mathbb{N}^\star$*.*

3. *(Nested Sums) Provide expressions for*
   - $\sum\limits_{i=1}^n \sum\limits_{k=i}^n c$*, for all* $c \in \mathbb{R}, n \in \mathbb{N}^\star$*.*
   - $\sum\limits_{i=1}^n \sum\limits_{j=1}^n ij$ *(hint: develop* $(\sum\limits_{i=1}^n a_i) \times (\sum\limits_{j=1}^n b_i)$*, where for* $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$*.*

4. *(Reversing orders of summation in nested sums) Let* $(a_{i,j})_{1\leq i,j\leq n} \in R^{n^2}$*. Complete the bullets*

$$\sum_{i=1}^n \sum_{j=1}^n a_{i,j} = \sum_{j=\bullet}^{\bullet} \sum_{i=\bullet}^{\bullet} a_{i,j}$$

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{i,j} = \sum_{j=\bullet}^{\bullet} \sum_{i=\bullet}^{\bullet} a_{i,j}$$

$$\sum_{i=1}^{n-1} \sum_{j=1}^i a_{i,j} = \sum_{j=\bullet}^{\bullet} \sum_{i=\bullet}^{\bullet} a_{i,j}$$

5. *(Harder) Show by induction that* $(a+b)^n \leq 2^{n-1}(a^n + b^n)$ *for all* $a, b \in \mathbb{R}$ *and* $n \in \mathbb{N}^\star$ *for* $a, b \geq 0$.

Solution. 1. First, for n = 1 we see that both sides equal 1. Then let's assume the induction hypothesis, that $\sum_{i=1}^n = \frac{n(n+1)}{2}$ for some n, and show that implies the same hypothesis holds for $n+1$. We can see this as follows:

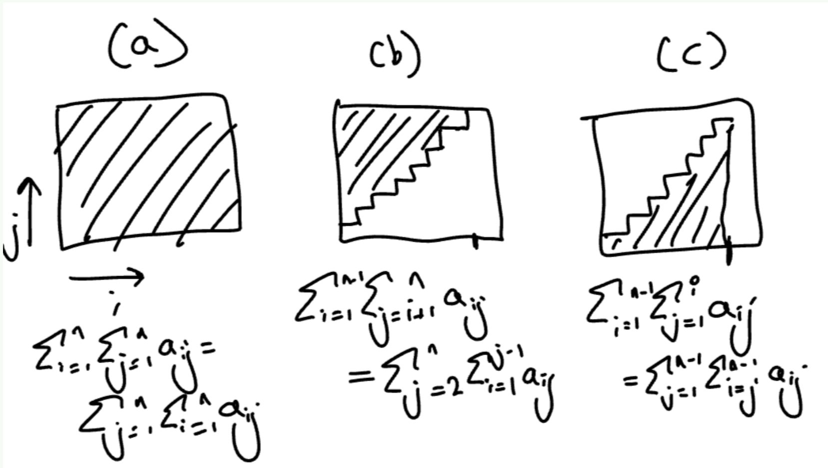
$$\sum_{i=1}^{n+1} i = \sum_{i=1}^{n} i + (n+1) = \frac{n(n+1)}{2} + (n+1) = \frac{(n+1)(n+2)}{2}$$

where the second equality used the induction hypothesis. We're done.

2. $\sum_{i=1}^n (i+c) = \sum_{i=1}^n i + \sum_{i=1}^n c$ which by previous questions is $\frac{n(n+1)}{2} + nc$.

3. (a) $\sum_{i=1}^n \sum_{k=i}^n c = c\sum_{i=1}^n (n-i) = c(n^2 - \frac{n(n+1)}{2}) = \frac{cn(n-1)}{2}$
   (b) We can follow the hint and find that $(\sum_{i=1}^n a_i)(\sum_{j=1}^n b_i) = \sum_{i=1}^n \sum_{j=1}^n a_i b_i$. Then $\sum_{i=1}^n \sum_{j=1}^n ij = (\sum_{i=1}^n i)(\sum_{j=1}^n j) = \frac{n^2(n-1)^2}{4}$.
   Alternatively: $\sum_{i=1}^n \sum_{j=1}^n ij = \sum_{i=1}^n i \sum_{j=1}^n j = \sum_{i=1}^n i\frac{n(n-1)}{2} = \frac{n(n-1)}{2}\sum_{i=1}^n i = \frac{n^2(n-1)^2}{4}$

4. (a) $\sum_{i=1}^n \sum_{j=1}^n a_{ij} = \sum_{j=1}^n \sum_{i=1}^n a_{ij}$. These are just two ways of saying sum up all the elements of a_{ij}.
   (b) $\sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{ij} = \sum_{j=2}^n \sum_{i=1}^{j-1} a_{ij}$, by the drawing the square method.
   (c) $\sum_{i=1}^{n-1} \sum_{j=1}^i a_{ij} = \sum_{j=1}^{n-1} \sum_{i=j}^{n-1} a_{ij}$



5. To show by induction we first show that the statement holds for $n = 1$. This is clear, since $a + b \leq a + b$. Now we assume the statement holds for $n \in \mathbb{N}^*$, the induction hypothesis, and try to prove it for $n+1$. To do this let's examine the right hand side - $2^n(a^{n+1} + b^{n+1})$. Let's assume that $a > b$ for now, then:

$$2^n(a^{n+1} + b^{n+1}) = 2^n a(a^n + b^n \frac{b}{a}) \geq 2^n a(a^n + b^n) = 2a(2^{n-1}(a^n + b^n))$$

We can see the RHS (right hand side) of the induction hypothesis coming out, so we can use that to say:

$$2a(2^{n-1}(a^n + b^n)) \geq 2a(a+b)^n$$

Now we can make our final link. Since $a > b$, $2a > (a+b)$, therefore:

$$2a(a+b)^n \geq (a+b)(a+b)^n = (a+b)^{n+1}$$

So we have show that, only assuming the statement holds for a given n, it also holds for $n+1$, hence, the statement is true.

$\blacksquare$

Chapter 2

Vector and Matrices, basic operations and arithmetics

2.1 The Vector Space $\mathbb{R}^n$

Informal Motivation

A lot (most?) of science (and maths) is about decomposing things into their parts, then reasoning about the parts and their interactions. A lot of objects can be described by a set of numbers (position in 3D space, position in abstract space, vector of neural firing rates, weight vector connecting to one neuron, vector of fMRI voxel responses, etc.). This chapter is about decomposing and working with these objects.
We will begin by defining the objects, and how to combine them. Then we will discuss 'good' sets of atomic parts that we can use to decompose the whole space.

For $n \geq 1$, the space $\mathbb{R}^n$ is defined as the Cartesian product of $\mathbb{R}$ with itself n-times (recall Definition 2).

The n-dimensional real space $\mathbb{R}^n$

Definition 3 (n-dimensional real space)**.** *Let $n \in \mathbb{N}^\star$. For some elements $x, y \in \mathbb{R}^n$, and $\lambda \in \mathbb{R}$, we define the following two operations:*

- *vector addition:* $x + y \coloneqq (x_1 + y_1, \ldots, x_n + y_n)$
- *scalar multiplication:* $\lambda \times x \coloneqq (\lambda x_1, \ldots, \lambda x_n)$

From now on, we call elements of $\mathbb{R}^n$ vectors*.*

Properties of the Internal and External Binary Operations

Proposition 1. *The following properties hold:*

1. $\forall \lambda, \mu \in \mathbb{R}, \forall x \in \mathbb{R}^n, (\lambda + \mu) \times x = \lambda \times x + \mu \times x$ *(distributivity of + w.r.t ×)*

2. $\forall \lambda, \mu \in \mathbb{R}, \forall x \in \mathbb{R}^n, \lambda \times (x + y) = \lambda \times x + \lambda \times y$ *(distributivity of* $\times$ *w.r.t* $+$*)*

2.2 Matrices

Definition 4 (Matrices)**.** *A real matrix* A *with* n *rows and* p *columns is the indexed set* $(A_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$*, where* $A_{ij} \in R$*.* A *will also be denoted by:*

$$A := \begin{bmatrix} A_{11} & \ldots & A_{1n} \\ A_{21} & \ldots & A_{2n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \ldots & A_{np} \end{bmatrix}$$

A_{ij} *is the element "in* i*-th row and the* j*-the column". The set of all real-valued matrices with* n *rows and* p *columns is denoted by* $\mathcal{M}_{n,p}(\mathbb{R})$*. If* $n = p$*,* A *is called a square matrix, and we denote the set of all square matrices of size* n *by* $\mathcal{M}_n(\mathbb{R})$*.*

An important special case is the identity matrix, denoted $I_n \in \mathcal{M}_{n,n}(\mathbb{R})$, which is defined as:

$$I_n = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Definition 5 (Matrix Addition and Scalar Multiplication)**.** *For* A, B *be two matrices in* $\mathcal{M}_{n,p}(\mathbb{R})$*, and* $\lambda \in \mathbb{R}$*, we define the following two operations:*

- *(Addition)* $A + B \in \mathcal{M}_{n,p}(\mathbb{R})$*, defined as:*

$$(A + B)_{ij} = A_{ij} + B_{ij}$$

- *(Scalar Multiplication)* $\lambda A \in \mathcal{M}_{n,p}(\mathbb{R})$*, defined as:*

$$(\lambda A)_{ij} = \lambda A_{ij}$$

Definition 6 (Matrix-Matrix Multiplication)**.** *Let* $A \in \mathcal{M}_{n,p}(\mathbb{R})$*, and* $B \in \mathcal{M}_{p,q}(\mathbb{R})$*. Then the product* $AB \in \mathcal{A}_{m,p}(\mathbb{R})$ *is defined as:*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Definition 7 (Matrix-Vector Multiplication)**.** *Let* $M \in \mathcal{M}_{m,n}(\mathbb{R})$*, and* $x \in \mathbb{R}^n$*. Then the* *product $Mx \in \mathbb{R}^m$ is defined as:*

$$(Mx)_i = \sum_{k=1}^{n} M_{ik}x_k$$

Note that this recovers the matrix-matrix multiplication of M by x when viewing x as a $(n \times 1)$ matrix (by identifying vectors in $\mathbb{R}^n$ with $\mathcal{M}_{n,1}(\mathbb{R})$).

Note that $I_n x = x$ (show that yourself).

Definition 8 (Matrix and Vector Transpose)**.** *Let $M \in \mathcal{M}_{m,n}(\mathbb{R})$. The transpose of M, denoted $M^\top \in \mathcal{M}_{n,m}(\mathbb{R})$, is defined as:*

$$(M^\top)_{ij} = M_{ji}$$

If x is a vector in $\mathbb{R}^n$, the transpose of x, denoted $x^\top \in \mathcal{M}_{1,n}(\mathbb{R})$, is defined as the matrix-transpose of x when viewed as a $(n \times 1)$ matrix.

Matrix Inverse

Definition 9 (Inverse of a Matrix)**.** *Let $M \in R^{n\times n}$ (the number of columns and rows are equal!). Then M is said to be invertible if there exists a matrix $M^{-1} \in \mathbb{R}^{n\times n}$ such that:*

$$MM^{-1} = M^{-1}M = I$$

Inverse of product (Important!)

Proposition 2. *Let A, $B \in \mathbb{R}^{n\times n}$ be two invertible matrices. Then AB is invertible, and*

$$(AB)^{-1} = B^{-1}A^{-1}$$

Exercise

Proof. Exercise. □

Inverse of Matrix transpose

Proposition 3 (Inverse of Matrix Transpose)**.** *Let $M \in \mathbb{R}^{n\times n}$. Then if $M^\top$ is invertible:*

$$(M^\top)^{-1} = (M^{-1})^\top$$

Exercise

Proof. Exercise. □

2.3 Exercise Sheet 2: Vectors and Matrices

Some Concrete Examples

Exercise 7. *Let $x = (1, 2)$, $y = (-1, 3)$, $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Write out:*

1. $2x$
2. $3y$
3. $x + y$
4. $-2A$
5. Ax
6. $A(x + y)$

Show that A is invertible and compute its inverse. Show that B is not invertible.

Solution.

1. $(2, 4)$
2. $(-3, 9)$
3. $(0, 5)$
4. $\begin{pmatrix} -2 & -2 \\ -2 & 0 \end{pmatrix}$
5. $(3, 1)$
6. $(5, 0)$

To show A is invertible let's show we can find a matrix C such that $AC = I$, $CA = I$. If C satisfies this it must be the inverse of A. Let's set this up:

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

From this we can solve the equations and find $a = 0$, $b = 1$, $c = 1$, $d = -1$. Now we have a candidate, let's check that $CA = I$. Indeed we can verify this:

$$\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

But be careful! Not all C such that $CA = I$ satisfy $AC = I$. Think about the case where A and C are rectangular, AC and CA are not both defined! This leads to the concept of a right inverse and a left inverse, the right inverse of A is the matrix that when you multiply on the right produces identity $AC = I$, without guaranteeing that $CA = I$. If a matrix is both the left and right inverse of A, it is the inverse of A.
Since we have found the inverse of A, it is invertible.
Doing this for B:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

we find the condition $0 = 1$ for the bottom right element, which cannot be satisfied. Hence there is no inverse of B, it is not invertible. ■

Associativity of Matrix-Matrix Product

Exercise 8. *Let $A \in \mathcal{M}_{n,p}(\mathbb{R})$, $B \in \mathcal{M}_{p,q}(\mathbb{R})$, and $C \in \mathcal{M}_{q,r}(\mathbb{R})$. Show that $(AB)C = A(BC)$. (Hint: To do so, you must show that $((AB)C)_{i,j} = (A(BC))_{ij}$, for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,r\}$.)*

Solution. $((AB)C)_{ij} = \sum_{k=1}^{p}(AB)_{ik}C_{kj} = \sum_{k=1}^{q}(\sum_{l=1}^{p} A_{il}B_{lk})C_{kj} = \sum_{k=1}^{q}\sum_{l=1}^{p} A_{il}B_{lk}C_{kj} = \sum_{l=1}^{p} A_{il}(\sum_{k=1}^{q} B_{lk}C_{kj}) = \sum_{l=1}^{p} A_{il}(BC)_{lj} = (A(BC))_{ij}$ ■

Distributivity of Matrix and Vector Multiplications

Exercise 9.

- $A \in \mathcal{M}_{n,p}(\mathbb{R})$,
- $B_1,\dots,B_r \in \mathcal{M}_{p,q}(\mathbb{R})$,
- $x_1,\dots,x_r \in \mathbb{R}^p$
- $\lambda_1,\dots,\lambda_r \in \mathbb{R}$.

Show that:

1. $A\left(\sum_{i=1}^{r} \lambda_i B_i\right) = \sum_{i=1}^{r} \lambda_i AB_i$.
2. $A\left(\sum_{i=1}^{r} \lambda_i x_i\right) = \sum_{i=1}^{r} \lambda_i Ax_i$.

Solution.

1. $(A\left(\sum_{i=1}^{r} \lambda_i B_i\right))_{jk} = \sum_{l=1}^{p} A_{jl}\sum_{i=1}^{r} \lambda_i B_{lk} = \sum_{i=1}^{r} \lambda_i \sum_{l=1}^{p} A_{jl}B_{lk} = \sum_{i=1}^{r} \lambda_i (A_{jl}B_{lk})_{jk} = \sum_{i=1}^{r} \lambda_i (AB)_{jk}$
2. $(A(\sum_{i=1}^{r} \lambda_i x_i))_j = \sum_{l=1}^{p}\sum_{i=1}^{r} A_{jl}\lambda_i(x_i)_l = \sum_{i=1}^{r} \lambda_i(\sum_{l=1}^{p} A_{jl}(x_i)_l) = \sum_{i=1}^{r} \lambda_i (Ax_i)_j = (\sum_{i-1}^{r} \lambda_i Ax_i)_j$

■

Bilinear forms

Exercise 10. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a square matrix. Let us denote $B_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ the function defined by*

$$B_A(x,y) = x^T Ay,$$

Show that:

1. $B_A(x, y+z) = B_A(x,y) + B_A(x,z)$ *for all* $x,y,z \in \mathbb{R}^n$.
2. $B_A(x, \lambda y) = \lambda B_A(x,y)$ *for all* $x,y \in \mathbb{R}^n$ *and* $\lambda \in \mathbb{R}$.

With that in mind provide an expression for $B_A(x+\lambda y, z+\mu t)$, for all $x,y,z,t \in \mathbb{R}^n$ and $\lambda,\mu \in \mathbb{R}$.

Solution. **1. Linearity in the second argument**

- For all $x,y,z \in \mathbb{R}^n$,

$$\begin{aligned} B_A(x,y+z) &= x^T A(y+z) \\ &= x^T Ay + x^T Az \\ &= B_A(x,y) + B_A(x,z). \end{aligned}$$

- For all $x, y \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$,
$$\begin{aligned} B_A(x, \lambda y) &= x^T A(\lambda y) \\ &= \lambda x^T A y \\ &= \lambda B_A(x, y). \end{aligned}$$

Hence, $B_A(x, \cdot)$ is linear in the second argument. Similarly we can show that $B_A(\cdot, y)$ is linear in the first argument.

2. Expression for $B_A(x + \lambda y, z + \mu t)$

$$\begin{aligned} B_A(x + \lambda y, z + \mu t) &= B_A(x, z + \mu t) + \lambda B_A(y, z + \mu t) \\ &= B_A(x, z) + \mu B_A(x, t) + \lambda \left(B_A(y, z) + \mu B_A(y, t)\right) \\ &= B_A(x, z) + \mu B_A(x, t) + \lambda B_A(y, z) + \lambda\mu B_A(y, t). \end{aligned}$$

$$\boxed{B_A(x + \lambda y, z + \mu t) = B_A(x, z) + \mu B_A(x, t) + \lambda B_A(y, z) + \lambda\mu B_A(y, t)}$$

■

Triangular Matrices

Exercise 11. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a square matrix. We say that A is an upper triangular matrix if $A_{ij} = 0$ for all $i > j$, and a lower triangular matrix if $A_{ij} = 0$ for all $i < j$. Show that, if A, B are upper triangular matrices, and $\lambda \in \mathbb{R}$*

1. *$A + B$ is an upper triangular matrix, and λA.*
2. *AB is an upper triangular matrix.*

Solution. **1. The sum $A + B$ and the scalar multiple λA are upper triangular.**

- For the sum $C = A + B$, we compute each entry:
$$C_{ij} = A_{ij} + B_{ij}.$$
If $i > j$, then $A_{ij} = 0$ and $B_{ij} = 0$, so $C_{ij} = 0$. Hence, C is upper triangular.
- For the scalar multiple $D = \lambda A$, each entry is:
$$D_{ij} = \lambda A_{ij}.$$
If $i > j$, then $A_{ij} = 0$, so $D_{ij} = \lambda \cdot 0 = 0$. Hence, D is upper triangular.

2. The product AB is upper triangular.
Let $C = AB$. We compute each entry:
$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Suppose $i > j$. To show that $C_{ij} = 0$, observe the following:

- Since A is upper triangular, $A_{ik} = 0$ whenever $i > k$.

- Since B is upper triangular, $B_{kj} = 0$ whenever $k > j$.

So, for each k, either $A_{ik} = 0$ (if $i > k$) or $B_{kj} = 0$ (if $k > j$). But since $i > j$, there is no k for which both $A_{ik} \neq 0$ and $B_{kj} \neq 0$. Thus, every term in the sum is zero, and so:

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} = 0.$$

Hence, AB is upper triangular.

■

Nilpotent Matrix

Exercise 12. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a square matrix. We say that A is a nilpotent matrix if there exists an integer $n_0 \in \mathbb{N}$ such that $A^{n_0} := \underbrace{A \cdots A}_{n_0 \ times} = 0$. Show that if A is nilpotent,*

- *$A^k = 0$ for all $k \geq n_0$.*
- *$(I - A)$ is invertible with inverse $\sum_{k=0}^{n_0 - 1} A^k$.*

Solution. **1. $A^k = 0$ for all $k \geq n_0$**
We proceed by induction on $k \geq n_0$.

- Base case: $A^{n_0} = 0$ by definition of nilpotency.
- Inductive step: Suppose $A^k = 0$ for some $k \geq n_0$. Then

$$A^{k+1} = A \cdot A^k = A \cdot 0 = 0.$$

By induction, $A^k = 0$ for all $k \geq n_0$.

2. $I - A$ is invertible
We claim that

$$(I - A)^{-1} = \sum_{k=0}^{n_0 - 1} A^k.$$

To verify this, we compute:

$$(I - A) \left(\sum_{k=0}^{n_0 - 1} A^k \right) = \sum_{k=0}^{n_0 - 1} A^k - \sum_{k=0}^{n_0 - 1} A^{k+1}.$$

Note that:

$$\sum_{k=0}^{n_0 - 1} A^{k+1} = \sum_{k=1}^{n_0} A^k.$$

Therefore:

$$\sum_{k=0}^{n_0 - 1} A^k - \sum_{k=1}^{n_0} A^k = A^0 - A^{n_0} = I - 0 = I.$$

Thus, $(I - A)\left(\sum_{k=0}^{n_0-1} A^k\right) = I$, and similarly one can show

$$\left(\sum_{k=0}^{n_0-1} A^k\right)(I - A) = I.$$

Hence, $I - A$ is invertible with inverse $\sum_{k=0}^{n_0-1} A^k$.

■

Trace of a Square Matrix

Exercise 13. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a square matrix. We define the trace of A, denoted by $tr(A)$, as the sum of the diagonal entries of A, i.e.,*

$$tr(A) = \sum_{i=1}^{n} A_{ii}.$$

Show that:

1. $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$ *for all* $A, B \in \mathcal{M}_n(\mathbb{R})$.
2. $\mathrm{Tr}(\lambda A) = \lambda \mathrm{Tr}(A)$ *for all* $A \in \mathcal{M}_n(\mathbb{R})$ *and* $\lambda \in \mathbb{R}$.
3. *For* $A \in \mathcal{M}_{n,p}(\mathbb{R})$ *and* $B \in \mathcal{M}_{p,n}(\mathbb{R})$*, show that* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. *More generally, for* $A_1, \dots A_s$*, s matrices such that the number of columns of A_i equals the number of rows of A_{i+1}, show that* $\mathrm{Tr}\left(\prod_{i=1}^{s} A_i\right) = \mathrm{Tr}\left(\prod_{i=k}^{s} A_i \prod_{i=1}^{k} A_i\right)$. *We say that the trace is cyclic.*

Solution. **1. Linearity of the trace:**

- We compute:

$$\mathrm{Tr}(A + B) = \sum_{i=1}^{n}(A + B)_{ii} = \sum_{i=1}^{n}(A_{ii} + B_{ii}) = \sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n} B_{ii} = \mathrm{Tr}(A) + \mathrm{Tr}(B).$$

- Similarly,

$$\mathrm{Tr}(\lambda A) = \sum_{i=1}^{n}(\lambda A)_{ii} = \sum_{i=1}^{n} \lambda A_{ii} = \lambda \sum_{i=1}^{n} A_{ii} = \lambda\, \mathrm{Tr}(A).$$

2. Cyclic property of the trace:
Let $A \in \mathcal{M}_{n,p}(\mathbb{R})$ and $B \in \mathcal{M}_{p,n}(\mathbb{R})$. Then:

$$\mathrm{Tr}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{p} A_{ik}B_{ki},$$

$$\mathrm{Tr}(BA) = \sum_{k=1}^{p}(BA)_{kk} = \sum_{k=1}^{p}\sum_{i=1}^{n} B_{ki}A_{ik}.$$

Since both expressions are equal (sum over same terms, just reordered), we conclude:

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

3. General cyclicity:
Let $A_1, A_2, \ldots, A_s$ be matrices such that the product $A_1 A_2 \cdots A_s$ is well-defined and square. Let $1 \leq k \leq s$.
We define:

$$A := A_k A_{k+1} \cdots A_s, \quad B := A_1 A_2 \cdots A_{k-1}.$$

Then,

$$\mathrm{Tr}(A_1 A_2 \cdots A_s) = \mathrm{Tr}(AB).$$

Since AB and BA are both square matrices and the product is defined, we can apply the cyclic property of the trace:

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Thus,

$$\mathrm{Tr}(A_1 A_2 \cdots A_s) = \mathrm{Tr}(A_k A_{k+1} \cdots A_s A_1 A_2 \cdots A_{k-1}).$$

This shows that the trace is invariant under cyclic permutations — i.e., the trace is *cyclic*.

$\blacksquare$

2.4 Neuro Q1: Ant Phasors

Some insects, like many other animals, are capable of dead-reckoning: they walk on a convoluted path away from their nest in search of food, then to return home they take the direct route. They do this by keeping track of a homing vector, i.e. the direction and displacement of the nest from their position. We know they are not using a more complicated 'cognitive map' strategy because if you pick them up and move them some vector away, they will return to a point that is displaced from the nest by that same vector![1] How might this be neurally implemented?

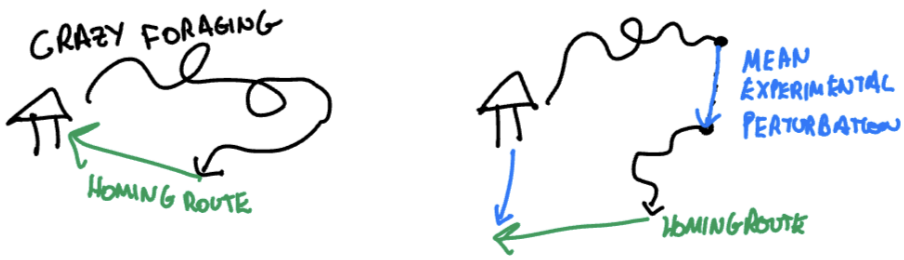


Figure 2.1: Left: Ants can take convoluted routes while foraging, but they return on the shortest path. Rigth: We known they are using vector navigation, or dead reckoning, because when shifted by a vector during their exploration they will home back to the nest, plus that vector!

One way to understand the circuitry is via phasors. Phasors are a way of representing sinusoids with a given frequency ω as points in 2D space: $f(t) = A\sin(\omega t + \phi)$ is assigned the point in 2D space with co-ordinantes $(A\cos(\phi), A\sin(\phi))$. Phasors are convenient because they make adding sinusoids easy.

1. To see how phasors make life easy, let's first see how it is hard.

   (a) First, use the double angle formula ($\sin(x+y) = \sin(x)\cos(y)+\cos(x)\sin(y)$) to show that any sum of a sine and cosine can be rewritten using just one sine: $a\sin(t) + b\cos(t) = A\sin(t+\phi)$, and find expressions relating each of A and ϕ to a and b. Hint: try to find expressions for A^2 and $\tan(\phi)$ in terms of a and b.

   (b) Now consider two sinusoids: $f_1(t) = A_1\sin(\omega t + \phi_1)$ and $f_2(t) = A_2\sin(\omega t + \phi_2)$. By using the double angle formula and the result above, or otherwise, show that $f(t) = f_1(t) + f_2(t) = A\sin(\omega t + \phi)$ for some (A, ϕ) and find similar expressions relating A and ϕ to A_1, A_2, ϕ_1, and ϕ_2.

2. So we painfully added the sinusoids. Now let's see how much easier life is in phasor land.

   Show that adding together the phasor representations of $f_1(t)$ and $f_2(t)$ using vector addition produces the phasor associated with $f_1(t) + f_2(t)$. Wasn't that simpler?

We turned adding sinusoids into adding 2D vectors, because we find adding 2D vectors easier. This is how phasors get used in physics and electrical engineering. Neural circuits, however, find adding sinusoids easier than adding 2D vectors, so they use the link in the other direction!!

[1] Similar, beautiful, work has shown they keep track of distance by integrating their steps: if you put an ant on stilts it will overshoot the nest, while if you, less whimsically, cut their legs to shorten them they will undershoot.

The ant has to keep track of its homing vector to the nest. Denote the current vector to nest as $u = (u_x, u_y) = (A_1 \cos(\phi_1), A_1 \sin(\phi_1))$, which the ant is storing somewhere internally. Then the ant takes a step $v = (v_x, v_y) = (A_2 \cos(\phi_2), A_2 \sin(\phi_2))$, and it needs to update its internal encoding to $u + v$. It can instead do this by adding sinusoids in one of two ways. Meditate and justify to yourself that both work:

- Many brain areas oscillate at some frequency. If you had one neuron that oscillated with phase and amplitude corresponding to f_1, another to f_2, and summed them this would perform vector arithmetic.
- Rather than distributing the sinusoid across time, you can distribute it across neurons! Create a neural population, $g, h \in \mathbb{R}^n$ where n is the number of neurons, with firing rates:

$$\text{Homing Encoding} \qquad g_i(t) = A_1(t)\sin(\omega i + \phi_1(t)) \tag{2.1}$$
$$\text{Heading Direction} \qquad h_i(t) = A_2(t)\sin(\omega i + \phi_2(t)) \tag{2.2}$$

  Then you can recurrently connect the populations so that they add the sinusoids!

$$g_i(t+1) = g_i(t) + h_i(t) \tag{2.3}$$

Now onto some more biological topics, consider the second way of implementing the computation:

3. This is not the most biological system, for one, real neurons only have positive firing rates but the neurons above are negative. Propose a new representation and corresponding algorithm (Eq. (2.1) and Eq. (2.3)) that avoids this problem.
4. Neurons also have a maximum firing rate. The combination of these two mean we are bounded from both ends. How does that restrict the computational power of the system?
5. We're bumping into the constraints of a limited range of firing. In some sense our counting system suffers from a similar problem. We want to express all the numbers between 1 and 1000 (or a million, or a billion, ...), but we don't want to use 1000 different symbols. Instead we use the same 10 symbols (0-9) in conjunction to code for many more numbers. Based on this motivation can you propose a way to use multiple copies of a homing tracking system to keep track of a larger range of homing vectors? How would you perform this computation and does it strike you as very biological?
6. Grid cells represent an alternative circuitry for encoding and tracking 2D vectors. Rather than encoding the amplitude in firing rate amplitude, and the phase in which neuron is most active, grid cells simply store the $u_x = A_1 \cos(\phi)$ and $u_y = A_2 \sin(\phi)$ components of the vector in the same way.

   Let's say we had two grid cells, defined as:

$$g_s(u) = \sin(\omega_x u_x + \omega_y u_y)$$
$$g_c(u) = \cos(\omega_x u_x + \omega_y u_y)$$

   Now we need to wire up a circuit that implements tracking of the homing vector, like Eq. (2.3). Show that updating the activity using the following matrix that depends on v in the following way accurately integrates (use the following identities $\sin(A+B) = \sin(A)\cos(B) + \cos(A)\sin(B)$, $\cos(A+B) = \cos(A)\cos(B) - \sin(A)\sin(B)$:

$$W(v) = \begin{bmatrix} \cos(\omega_x v_x + \omega_y v_y) & -\sin(\omega_x v_x + \omega_y v_y) \\ \sin(\omega_x v_x + \omega_y v_y) & \cos(\omega_x v_x + \omega_y v_y) \end{bmatrix} \qquad W(v)\begin{bmatrix} g_s(u) \\ g_c(u) \end{bmatrix} = \begin{bmatrix} g_s(u+v) \\ g_c(u+v) \end{bmatrix}$$

(Matrices of this type, $\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$, are called rotation matrices, for obvious reasons)

Explain how this avoids the unbounded firing rate problem, and how the system could use different grid cells to code for loads of positions using a small number of neurons.

This is also unbiological, because now $W(v)$ are effectively synaptic weights and we're asking them to be modulated by v: how do you change a synapse on the basis of an input? But it turns out there are good workarounds.

Solution. 1. (a) Let's use the suggested double angle formula:

$$A\sin(t+\phi) = A(\sin(t)\cos(\phi) + \cos(t)\sin(\phi)) = A\cos(\phi)\sin(t) + A\sin(\phi)\cos(t)$$

Hence $a = A\cos(\phi)$ and $b = A\sin(\phi)$. Squaring and adding these two, and using the fact that $\cos^2(\phi) + \sin^2(\phi) = 1$, we get:

$$a^2 + b^2 = A^2(\cos^2(\phi) + \sin^2(\phi)) = A^2, \quad A^2 = a^2 + b^2$$

By dividing a and b we find that:

$$\tan(\phi) = \frac{a}{b}$$

Actually solving for A and ϕ is a bit annoying. There's an annoying subtlety here relating when A is positive vs. negative. This problem is solved by the atan2 function that you can look up, first implemented in fortran, such that: $A = \sqrt{a^2 + b^2}$ and $\phi = \text{atan2}(b, a)$. We don't need this for the question, but it is useful if you ever want to convert from polar to cartesian coordinates.

(b) Now using double angle formula twice:

$$\begin{aligned} f_1(t) + f_2(t) &= A_1\cos(\phi_1)\sin(\omega t) + A_1\sin(\phi_1)\cos(\omega t) \\ &\quad + A_2\cos(\phi_2)\sin(\omega t) + A_2\sin(\phi_2)\cos(\omega t) \\ &= \Big(A_1\cos(\phi_1) + A_2\cos(\phi_2)\Big)\sin(\omega t) \\ &\quad + \Big(A_1\sin(\phi_1) + A_2\sin(\phi_2)\Big)\cos(\omega t) \end{aligned}$$

Hence:

$$A^2 = (A_1\cos(\phi_1) + A_2\cos(\phi_2))^2 + (A_1\sin(\phi_1) + A_2\sin(\phi_2))^2 \tag{2.4}$$

While:

$$\tan(\phi) = \frac{A_1\cos(\phi_1) + A_2\cos(\phi_2)}{A_1\sin(\phi_1) + A_2\sin(\phi_2)} \tag{2.5}$$

2. A phasor representation of a wave is a point (x, y) such that the amplitude of the wave is its length $A^2 = x^2 + y^2$, while its phase is the angle to the x axis $\tan(\phi) = \frac{y}{a}$. Representing each wave as a phasor we have the two points $(A_1\cos(\phi_1), A_1\sin(\phi_1))$ and $(A_2\cos(\phi_2), A_2\sin(\phi_2))$. Adding these is very simple:

$$(A_1\cos(\phi_1) + A_2\cos(\phi_2), A_1\sin(\phi_1) + A_2\sin(\phi_2))$$

And we can see that, our two rules match our results from the previous question. If we square and add the two elements using our rule to get the amplitude of the wave we do indeed find Eq. (2.4), while if we divide them we do indeed get tan of the phase as in Eq. (2.5).

3. You could add a constant offset to make things positive for some range of amplitude values:

$$\begin{aligned} g_i(t) &= A_1(t)\sin(\omega i + \phi_1(t)) + b_g \qquad && b_g > 0 \\ h_i(t) &= A_2(t)\sin(\omega i + \phi_2(t)) + b_h \qquad && b_h > 0 \end{aligned}$$

   Then if we choose our update equation such that:

$$g_i(t+1) = g_i(t) + h_i(t) - b_h$$

   We're good, everything is self-consistent.

   Having found this expression, how should b_h be related to the max running speed per discrete time interval?

4. Because there is a limited firing range, this coding scheme can only encode a limited range of amplitudes. If the amplitude goes too high the firing rates will max out and your encoding will break. Similarly, since $\sin(x) \in [-1, 1]$, we're also constrained from below by our nonnegativity constraint. If we don't shift the firing rates up by b_g enough then there will be a limit from below on the amplitude. The best choice for maximal range will be for b_g to be in the middle of the range of values. Then the range of firing rates is twice the maximal amplitude encodable using this scheme.

5. The question is trying to get you to say that you could have many circuits. Once the circuit reaches it's maximum amplitude you could treat it just like when a number is 9 and goes one bigger and is set to 0 again. Set that amplitude to 0, and set another neighbouring set of neurons to 1. Even for one-dimensional numbers this is a bit unbiological because we need a mechanism in neural circuitry to carry the 1 over to the next population when one reaches the maximum, and similarly when the units column goes below 0 amplitude it needs to reset to 9 and decrement the tens column. I can't think of good neural circuitry to do this, but it might exist.

   To do the two dimensions of the vector you could use two such one-dimensional schemes, but in keeping with the question you might want your two dimensions to be polar coordinates: the angle and length of the vector. That means that you need some quite subtle update mechanisms. For example, when the most detailed encoding reaches its max value you need to set it to 0, and shift the next level up by the right amount. There will be plausible neural circuitry to do this, but it will be complex...

6. We can simply check that the operation works as advertised. Let's denote $\theta_v = \omega_x v_x + \omega_y v_y$ and $\theta_u = \omega_x u_x + \omega_y u_y$:

$$\begin{aligned} &\begin{bmatrix} \cos(\theta_v) & -\sin(\theta_v) \\ \sin(\theta_v) & \cos(\theta_v) \end{bmatrix} \begin{bmatrix} \sin(\theta_u) \\ \cos(\theta_u) \end{bmatrix} \\ = &\begin{bmatrix} \cos(\theta_v)\sin(\theta_u) - \sin(\theta_v)\cos(\theta_u) \\ \sin(\theta_v)\sin(\theta_u) + \cos(\theta_v)\sin(\theta_u) \end{bmatrix} = \begin{bmatrix} \sin(\theta_u + \theta_v) \\ \cos(\theta_u + \theta_v) \end{bmatrix} \end{aligned}$$

This avoids the problem, because if we just add a constant this firing pattern will say between a fixed range for ever! At the moment many places are encoded identically, move by one lattice lengthscale and you get the same code. If we had loads of these 'grid modules' oscillating at different frequencies we could code for a much larger range of points.

■

Chapter 3

Vector Subspaces, Bases, Coordinates

3.1 Subspaces in $\mathbb{R}^n$

3.1.1 Definition and Basic Examples

We began with $\mathbb{R}^n$. Now we describe the objects which are the 'meaningful subparts' of $\mathbb{R}^n$, in the sense that, given our rules of combining vectors, if we begin in this subpart we'll never get out!

Definition 10 (Subspace). *Let $n \in \mathbb{N}^*$. A subset $F \subseteq \mathbb{R}^n$ such that $F \neq \emptyset$ is said to be a subspace of $\mathbb{R}^n$ if it is stable under the internal and external binary operations of $\mathbb{R}^n$, i.e.*

- $\forall x, y \in F,\ x + y \in F$;
- $\forall x \in F$ *and* $\lambda \in \mathbb{R}$, $\lambda \times x \in F$.

Proposition 4. *F is a subspace of $\mathbb{R}^n$ if and only if $\forall x, y \in F$ and $\lambda \in \mathbb{R}$, $\lambda \times x + y \in F$.*

Proof. Let us assume that $\forall x, y \in F$ and $\lambda \in \mathbb{R}$, $\lambda \times x + y \in F$. Then, on one hand, taking $\lambda = 1$, we obtain that $\forall x, y \in F$, $x + y \in F$. On the other hand, taking $y = 0$, we obtain that $\forall x \in F$ and $\lambda \in \mathbb{R}$, $\lambda \times x \in F$.
Let us now assume that

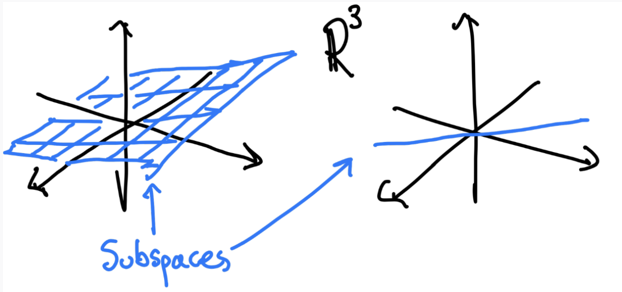
i) $\forall x, y \in F$, $x + y \in F$;

ii) $\forall x \in F$ and $\lambda \in \mathbb{R}$, $\lambda \times x \in F$.

Take $x \in F$ and $\lambda \in \mathbb{R}$, by ii), $z = \lambda \times x \in F$. Take $y \in F$, since $z \in F$, by i), $z + y \in F$. But since $z = \lambda \times x$, we conclude that $\forall x, y \in F$ and $\lambda \in \mathbb{R}$, $\lambda \times x + y \in F$. $\square$

Visualising Subspace

In 3D a subspace might be a plane or a line through the origin.



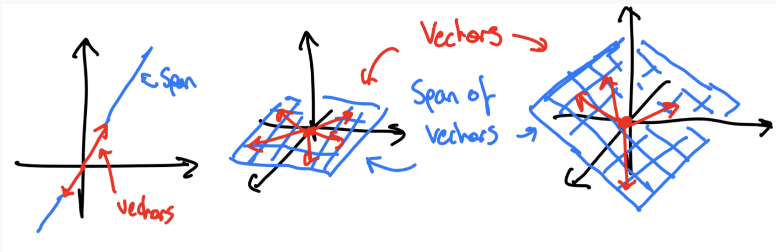
One very natural way to create a subspace is using a set of vectors. Find all the ways to combine them, then that will be a subspace, as you'll prove below.

Definition 11 (Span of a set of vectors)**.** *Let* $\{e_1, \ldots, e_m\}$ *be a set of vectors in* $\mathbb{R}^n$*. Then we call span* $\{e_1, \ldots, e_m\}$ *the set:*

$$\left\{ \sum_{i=1}^{m} \lambda_i e_i, \ \lambda_1, \ldots, \lambda_m \in \mathbb{R} \right\}$$

Visualising Span

The span of a set of vectors is intuitively the smallest subspace that contains those vectors.



Proposition 5. *Let* F*,* G *be two vector subspaces of* $\mathbb{R}^n$*. Then* $F \cap G$ *is a vector space, where for two sets* A, B*, we denoted* $A \cap B$ *the intersection of* A *and* B*, i.e., the set of elements that are in both* A *and* B*.*

Proof. Exercise. □

Usually, when F, G are vector subspaces of $\mathbb{R}^n$, $F \cup G$ is not a vector space (here, we denoted $A \cup B$ the union of two sets A and B, i.e., the set of elements that are in either A or B). Think about why, and try to characterize when the union of two vector subspaces is a vector space.

3.1.2 Sum of Vector Subspaces

Definition 12 (Sum of vector subspaces)**.** *Let* $n \in \mathbb{N}^*$*. Let* F *and* G *be two vector subspaces of* $\mathbb{R}^n$*. Then the set:*

$$\{x_F + x_G : x_F \in F, x_G \in G\}$$

is called the sum of F *and* G*, and is denoted* $F + G$*.*

Proposition 6. *Let* F*,* G *be two vector subspaces of* $\mathbb{R}^n$*. Then* $F + G$ *is a vector space.*

Proof. Exercise! □

Casually Explained

Sums of vector subspaces will show up a lot, often, we will decompose $\mathbb{R}^n$ into smaller chunks (see the rank-nullity theorem). A particular important case is when the sum of two vector subspaces is direct, e.g., when F and G only overlap at 0.

Definition 13 (Direct sum of vector subspaces)**.** *Let* F*,* G *be two vector subspaces of* $\mathbb{R}^n$*. Then* F *and* G *are called "in direct sum" if* $F \cap G = \{0\}$*. Their sum is then denoted* $F \oplus G$*.*

Characterization of the direct sum of vector subspaces

Proposition 7. *Let* F, G *be two subspaces, then* F *and* G *are in direct sum if and only if*

$$\forall x \in F + G, \exists!(x_1, x_2) \in F \times G \text{ such that } x = x_1 + x_2.$$

Proof

Proof. Let us first assume that the decomposition in $F + G$ is unique, i.e., $\forall x \in F + G, \exists!(x_1, x_2) \in F \times G$ such that $x = x_1 + x_2$. Let us show that $F \cap G = \{0\}$. Since F and G are subspaces, we immediately have $\{0\} \subset F \cap G$ (do you see why?). We are left to show $F \cap G \subset \{0\}$. Let us take $x \in F \cap G$. x, then admits two decompositions:

$$x = \underbrace{x}_{\in F} + \underbrace{0}_{\in G} = \underbrace{0}_{\in F} + \underbrace{x}_{\in G}$$

By assumption, the decomposition is unique, therefore, $x = 0$. This proves $F \cap G \subset \{0\}$ and therefore $F \cap G = \{0\}$.

Let us now assume that $F \cap G = \{0\}$. Take $x \in F + G$ admitting two decompositions $x = x_1 + x_2 = x_1' + x_2'$ with $x_1, x_1' \in F$ and $x_2, x_2' \in G$. We therefore have $x_1 - x_1' = x_2' - x_2$. Since F and G are subspaces we have that $x_1 - x_1' \in F$ and $x_2' - x_2 \in G$. Furthermore, since $x_1 - x_1' = x_2' - x_2$, we have $x_1 - x_1' \in G$. Therefore, $x_1 - x_1' \in F \cap G$, implying

$x_1 = x_1'$ since by assumption $F \cap G = \{0\}$. We conclude similarly that $x_2 = x_2'$, therefore the decomposition is unique. □

One additional desirable property we could wish for vector spaces in direct sum is that they span the entire space. In that case, we say that F is the complementary of G (in the vector space sense).

Complementary of a vector subspace

Definition 14 (Complementary of a vector subspace)**.** *Let F, G be two vector subspaces of $\mathbb{R}^n$. Then F and G are called complementary if*

1. *F and G are in direct sum*
2. *$F \oplus G = \mathbb{R}^n$*

Complementary of a vector subspace

Proposition 8 (Complementary of a vector subspace)**.** *Any vector subspace F of $\mathbb{R}^n$ has a complementary subspace G in $\mathbb{R}^n$.*

The proof is outside of the scope of this class.

Characterization of the complementary of a vector subspace

Corollary 1. *Let F, G be two subspaces of $\mathbb{R}^n$. Then F, G are complementary subspaces if and only if*

$$\forall x \in \mathbb{R}^n, \exists!(x_1, x_2) \in F \times G \text{ such that } x = x_1 + x_2.$$

Proof. This is a direct corollary of Proposition 7 and the definition of the complementary of a vector subspace. □

3.2 Family of Vectors

We are going to define three sets of families of vectors that are of paramount importance in linear algebra. They will tell us when we can decompose vectors, and when we can do it in only one way.

- Spanning sets
- Linearly independent sets
- Bases

3.2.1 Spanning Sets

Definition 15 (Spanning Set). *Let $n \in \mathbb{N}^*$, F be a vector subspace of $\mathbb{R}^n$ and $S = \{e_1, \ldots, e_m\}$ be m vectors in F. Then S is said to be a spanning set of F if $\forall x \in F, \exists \lambda_1, \ldots, \lambda_m$ s.t.*

$$x = \sum_{i=1}^{m} \lambda_i e_i.$$

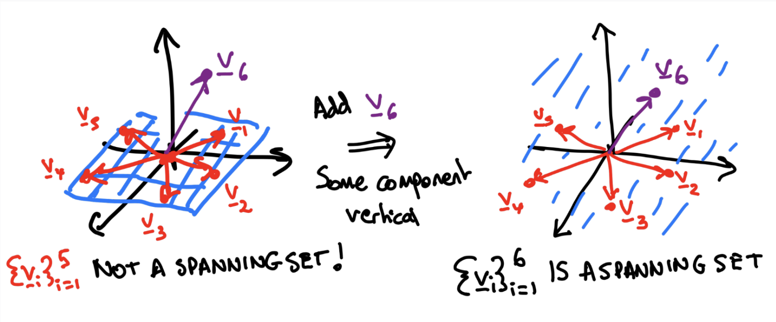
We can connect the span of a set of vectors to the notion of spanning set.

Proposition 9. *Let $S = \{e_1, \ldots, e_m\}$ be a set of vectors in $\mathbb{R}^n$. Then S is a spanning set of F if and only $F \subset span\{e_1, \ldots, e_m\}$.*

Proof. Exercise! □

Intuitively

Spanning sets provide a way to describe any vector $x \in \mathbb{R}^n$ (or a subspace F) using m scalars, $\lambda_1, \ldots, \lambda_m$. This means the vectors reach everywhere in the space.



3.2.2 Linearly (In)dependent Sets

Spanning sets let us travel everywhere in the space by combining members of the set. However, there may be more than one way to reach the same vector. it would be nice if we were able to enforce that for any x, that there exists a *unique* set of scalars that can be used to represent x. This is what the concept of linearly independent vectors is about:

Definition 16 (Linearly dependent set of vectors). *A set $S = \{e_1, \ldots, e_m\}$ of vectors is called linearly dependent if, there exists a $i_0 \in [m]$, and $\lambda_1, \ldots, \lambda_{i_0 - 1}, \lambda_{i_0 + 1}, \lambda_n$ s.t*

$$e_{i_0} = \sum_{i=1, i \neq i_0}^{m} \lambda_i e_i$$

Definition 17 (Linearly independent set of vectors)**.** *A set $S = \{e_1, \ldots, e_m\}$ of vectors is called linearly independent if it is not linearly dependent.*

Proposition 10.

- *S is linearly dependent i.i.f $\exists \lambda_1, \ldots, \lambda_m \neq (0, \ldots, 0)$ s.t. $\sum_{i=1}^{m} \lambda_i e_i = 0$.*
- *Conversely, S is linearly independent i.i.f $\sum_{i=1}^{m} \lambda_i e_i = 0 \implies \lambda_1 = \cdots = \lambda_m = 0$.*

Proof. Let's assume $\lambda_1 e_1 + \cdots + \lambda_m e_m = 0$ holds with some non-zero coefficient, λ_{k_0}. Dividing by λ_{k_0} and rearranging we find:

$$e_{k_0} = - \sum_{i=1, k \neq k_0}^{K} \frac{\lambda_i}{\lambda_{k_0}} e_m$$

There are two cases. Either all the other coefficients , $\{\lambda_i\}_{i=1, i \neq k_0}^{l}$, are zero, in which case $e_{k_0} = 0$, and S is linearly dependent. Alternatively, another λ_i is non-zero, and we have found that one vector in the set can be expressed as a linear combination of the others, so S is linearly dependent. Therefore, if S is linearly independent $\lambda_1 = ... = \lambda_m = 0$. □

3.2.3 Bases

Spanning sets let us get everywhere in the space, but did not guarantee unique decompositions. Linearly independent vectors uniquely decompose all vectors in their span, but don't necessarily reach everywhere. A basis combines these two nice properties:

Definition 18 (Basis)**.** *Let $n \in \mathbb{N}^*$, F be a vector subspace of $\mathbb{R}^n$ and $S = \{e_1, \ldots, e_m\}$ be a set of k vectors in F. Then S is said to be a basis for F if it is both a spanning set for F linearly independent.*

Proposition 11. *If S is a basis for F, then for any xF, $\exists! \{\lambda_1, \ldots, \lambda_m\}$ such that*

$$x = \sum_{i=1}^{k} \lambda_i e_i$$

Motivation: a basis provides a unique way to describe vectors in $\mathbb{R}^n$ (or a subspace F)

The definition of a basis above transcribes in formal terms that for any x in $\mathbb{R}^n$ (or a subspace F), there exists a unique set of k scalars that can be used to "describe" x. This description procedure is thus

- powerful enough to describe any vector in the vector space;

- tight enough to describe each vector in a unique way.

3.2.4 Dimension of a Vector Space

Theorem 1. *Let $n \in \mathbb{N}^*$. Let F be a vector subspace of $\mathbb{R}^n$.*

- *There exists a basis for F.*
- *All bases of F have the cardinality m, called the* dimension *of F (or* dim(F)*).*

Additionally, let $S = \{e_1, \ldots, e_p\}$ a set of p vectors in F.

- *If S is linearly independent, then:*
  - $p \leq \dim(F)$.
  - *if $p = \dim(F)$, then S is a basis for F.*
- *If S is a spanning set for F, then:*
  - $p \geq \dim(F)$.
  - *if $p = \dim(F)$, then S is a basis for F.*

Remark

After an elaborate journey we have arrived at a formalization of the very intuitive notion of the dimension of a space. You can check that it matches your intuition for one, two, or three dimensional space. But, by formalizing it, you can travel to some far off mathematical land, and if you still find something that obeys these rules you will know it matches your idea of a dimension.

Dimesion of subspaces of vector subspaces

Proposition 12. *Let E, F be a two finite dimensional subspace of $\mathbb{R}^n$ s.t. $F \subset E$. Then* $\dim(F) \leq \dim(E)$*, and* $\dim(F) = \dim(E) \iff E = F$.

Dimensionality of the direct sum of vector subspaces

Lemma 1. *Let F, G be two vector subspaces of $\mathbb{R}^n$ that are in direct sum. Then:*

$$dim(F \oplus G) = dim(F) + dim(G)$$

Proof. If B is a basis of F and C is a basis of G, then $B \cup C$ is a basis of $F \oplus G$. Since the basis of a vector space has the same cardinality as its dimension, we conclude the result. $\square$

Dimensionality of the sum of vector subspaces

Proposition 13. *Let F, G be two vector subspaces of $\mathbb{R}^n$. Then:*

$$dim(F + G) = dim(F) + dim(G) - dim(F \cap G)$$

Proof. Let F' be the complement of $F \cap G$ in F, e.g. $F' \oplus (F \cap G) = F$. One has $F' + G = F + G$. Note moreover that F' and G are in direct sum. Conclude the proof. $\square$

3.3 Exercise Sheet 3: Subspaces, Bases

Vector Subspaces

Exercise 14. *Let $\{e_1, \ldots, e_m\}$ be a set of vectors in $\mathbb{R}^n$. Then $F = span\{e_1, \ldots, e_m\}$ is a vector subspace of $\mathbb{R}^n$.*

Solution.

- Let us take two elements x_1, x_2 of F, by the definition of the span, $\exists \lambda_1, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_m$ such that $x_1 = \sum_{i=1}^m \lambda_i e_i$ and $x_2 = \sum_{i=1}^m \mu_i e_i$. Let us take $\theta \in \mathbb{R}$. Since

$$\theta \times x_1 + x_2 = \sum_{i=1}^{m} (\theta \lambda_i + \mu_i) e_i$$

belongs to F, we conclude that F is a vector subspace.

■

Exercise 15. *Show that the set $F = \{(x, y, 0) \mid x \in \mathbb{R}, y \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$.*

Solution. Let us take two elements z_1, z_2 of F, by the definition of F, $\exists x_1, y_1, x_2, y_2 \in \mathbb{R}$ such that $z_1 = (x_1, y_1, 0)$ and$z_2 = (x_2, y_2, 0)$. Let us take $\lambda \in \mathbb{R}$. Since

$$\lambda \times z_1 + z_2 = (\lambda x_1 + x_2, \lambda y_1 + y_2, 0)$$

belongs to F, we conclude that F is a vector subspace. ■

Exercise 16. *Write down a subset of $\mathbb{R}^n$ that is not a subspace*

Solution. Many options, for example $\{(100, 0), (50, 0)\}$. ■

Exercise 17. *Show that the origin is a member of every subspace.*

Solution. Take a subspace F. Take any element $f \in F$. $\lambda f \in F, \forall \lambda \in \mathbb{R}$. Therefore choose $\lambda = 0$, then $0f = 0$, and hence $0 \in F$. ■

Exercise 18. *Let E be a vector subspace of $\mathbb{R}^n$, F, G be two vector subspaces of E. Show that $F \cup G$ is a vector subspace of E if and only if $F \subseteq G$ or $G \subseteq F$.*

Solution. Let's first show that if $F \subseteq G$, $F \cup G$ is a subspace. This is clear because $F \cup G = G$ and we already know G is a vector subspace. The same argument holds if $G \subseteq F$.
Now we have to go the other way, showing that if $F \cup G$ is a subspace, it must be that $F \subseteq G$ or $G \subseteq F$. Let's say $F \cup G$ is a subspace but $F \nsubseteq G$ and $G \nsubseteq F$. Since F is not a subset of G, $\exists x \in F, x \notin G$, similarly $\exists y \in G, y \notin F$. Both $x \in F \cup G$ and $y \in F \cup G$, then, since it is a subspace $x + y \in F \cup G$. Therefore either $x + y \in F$, but then $x + y - x = y \in F$, since $x \in F$ and F is a subspace, which breaks our assumption. Or $x + y \in G$, but similarly then $x + y - y = x \in G$, which again breaks our assumption. Therefore one must be a subset of the other. ■

Exercise 19. *1. Which of these sets are vector subspaces of $\mathbb{R}^3$?*

(a) $\{(x,y,z) : (x-2y)z = 0\}$

(b) $\{(x,y,z) : x-2y = z\}$

(c) $\{(x,y,z) : x-2y = y+z = 1\}$

(d) $\{(x,y,z) : x-2y = x+y+z = 0\}$

2. *Let $m, n \in \mathbb{N}$.*

(a) *is $\mathbb{R}^m$ a vector subspace of $\mathbb{R}^n$?*

(b) *Let $E \subset \mathbb{R}^n$ the sets of vectors x in $\mathbb{R}^n$ such that $x_k = 0$ for $k > m$. Is E a vector subspace of $\mathbb{R}^n$?*

Solution. 1. (a) Nope, given (a,b,c) and (a',b',c') we can add them to get a vector that doesn't satisfy the property: $(a+a'-2(b+b'))(c+c') = (a-2b)c+(a'-2b')c'+(a-2b)c'+(a'-2b')c = (a-2b)c'+(a'-2b')c$.

(b) Yes, because if $a-2b = c$, then $\lambda a - 2\lambda b = \lambda c$. Similarly given (a,b,c) and (a',b',c') in the set we can see that $(a+a', b+b', c+c')$ is in the set, since: $a+a'-2(b+b') = a-2b+a'-2b' = c+c'$.

(c) Yes. Similar arguments to above.

(d) Yes. Similar arguments to above.

2. (a) Only if $m = n$. If $m \neq n$ then the elements of the two spaces are different, a 2D vector can't live within 3D space. ($(1,2,0)$ is a 3D vector that lives in a 2D plane, but that is different from the 2D vector $(1,2)$)

(b) Yes. Let's check $x + \lambda y$ for $x, y \in E, \lambda \in \mathbb{R}$. The first m dimensions are arbitrary, but the last $n-m$ dimensions will always be 0, since the sum and scaling of 0 vectors is zero.

■

Sums of Vector Subspaces

Exercise 20. *In each case, are the vector subspaces F and G in direct sum? Complementary?*

1. $E = \mathbb{R}^2$, $F = span\{(1,0)\}$, $G = \{0\}$

2. $E = \mathbb{R}^2$, $F = span\{(1,0)\}$, $G = span\{(0,1)\}$

3. $E = \mathbb{R}^2$, $F = span\{(1,0)\}$, $G = span\{(1,1)\}$

Solution. 1. Since $F \cap G = \{0\}$ they are in direct sum. However $(0,1) \notin F \bigoplus G$, therefore they are not complementary.

2. Again $F \cap G = \{0\}$ so they are in direct sum (To show this you can study a vector that is in $F \cap G$: $(a,b) = \lambda(1,0) = \mu(0,1)$, from which you conclude $a = b = 0$). But now $F \bigoplus G = \mathbb{R}^2$, since for any $x = (a,b) \in \mathbb{R}^2$, we can choose $x = a(1,0) + b(0,1)$.

3. Similar arguments show you that they are in direct sum and are complementary.

■

Bases, Linear Dependence, Dimension

Exercise 21. *Show that if any vector in the set $S = \{e_1, \dots, e_m\}$ is the zero-vector (the origin), the set is linearly dependent.*

Solution. $\{e_1, \dots, e_m\}$ is linearly dependent if $\sum_{i=1}^m \mu_i e_i = 0$ for some nonzero set of $\{\mu_i\}_{i=1}^m$. Assume there is a zero element in the set, e_{k_0}, then we can set all μ_i to zero except μ_{k_0} and this linear combination will equal zero for a non-zero set of μ. Hence S is linearly dependent. ■

Exercise 22. *Show that if S is linearly independent, for any $x \in \operatorname{span}(S)$, there exists a unique $(\lambda_1, \dots, \lambda_m)$ such that $x = \sum\limits_{i=1}^{k} \lambda_i e_i$.*

Solution. Let's show this by assuming both S is linearly independent and there are multiple ways of writing $x = \sum_{i=1}^k \lambda_i e_i$ we get a contradiction, hence the converse must be true, that S being linearly independent implies a unique decomposition. Then $x = \sum_{i=1}^k \lambda_i e_i = \sum_{i=1}^k \lambda_i' e_i$ where these are two different sets $\{\lambda_i\}$, and $\{\lambda_i'\}$. Then:

$$\sum_{i=1}^{k} (\lambda_i - \lambda_i') e_i = 0$$

Now one case is that $\{\lambda_i\}$ and $\{\lambda_i'\}$ differ in only one element, let's call it k_0, then only $(\lambda_{k_0} - \lambda_{k_0}')$ is non-zero, and $(\lambda_{k_0} - \lambda_{k_0}') e_{k_0} = 0 = e_{k_0}$, which by our previous exercise, breaks the assumption of linear independence.
Alternatively there are multiple non-zero $(\lambda_i - \lambda_i')$ values. Divide the whole expression by one of them, e.g. k_0, then:

$$e_{k_0} + \sum_{i=1, i \neq k_0}^{k} \frac{\lambda_i - \lambda_i'}{\lambda_{k_0} - \lambda_{k_0}'} e_i = 0$$

Hence we can write one vector as a linear combination of the others, completing the proof. ■

Exercise 23. *Determine the dimension of the span of the following vectors (the last two ones require you to know how to solve systems of equations)*

- $u = (1, 2, 0), v = (-1, 1, 1)$
- $u = (1, 0, 2), v = (0, 2, 1), w = (-1, 4, 2)$
- $u = (1, 1, -2), v = (1, 3, 2), w = (-2, 0, 1), z = (1, -1, 0)$

Solution. **First case:** $u = (1, 2, 0), \quad v = (-1, 1, 1)$
Suppose $\alpha u + \beta v = 0$. That gives:

$$\alpha(1, 2, 0) + \beta(-1, 1, 1) = (0, 0, 0).$$

Compute:

$$(\alpha - \beta, 2\alpha + \beta, \beta) = (0, 0, 0).$$

From the third coordinate: $\beta = 0$, then from the first: $\alpha = 0$. Hence, the only solution is the trivial one, and the vectors are linearly independent.

$$\Rightarrow \dim(\operatorname{span}\{u, v\}) = \boxed{2}.$$

Second case: $u = (1, 0, 2), \quad v = (0, 2, 1), \quad w = (-1, 4, 2)$
Step 1: Check if u and v are linearly independent.
Assume $\alpha u + \beta v = 0$. Then:

$$\alpha(1, 0, 2) + \beta(0, 2, 1) = (0, 0, 0) \Rightarrow (\alpha,\ 2\beta,\ 2\alpha + \beta) = (0, 0, 0).$$

From the first coordinate: $\alpha = 0$. From the second: $2\beta = 0 \Rightarrow \beta = 0$. Then the third equation is also satisfied: $2 \cdot 0 + 0 = 0$.
So the only solution is $\alpha = \beta = 0$, and u, v are linearly independent.

Step 2: Is $w \in span\{u, v\}$?
We ask whether there exist $\alpha, \beta \in \mathbb{R}$ such that:

$$\alpha(1, 0, 2) + \beta(0, 2, 1) = (-1, 4, 2).$$

Compute each component:
- First coordinate: $\alpha = -1$
- Second coordinate: $2\beta = 4 \Rightarrow \beta = 2$
- Third coordinate: $2\alpha + \beta = 2(-1) + 2 = -2 + 2 = 0$, but we want 2. Contradiction.
So no such α, β exist, and $w \notin \operatorname{span}\{u, v\}$.
Conclusion: u and v are linearly independent, and w is not in their span, so $\{u, v, w\}$ is linearly independent.

$$\dim(\operatorname{span}\{u, v, w\}) = \boxed{3}.$$

Third case: $u = (1, 1, -2), \quad v = (1, 3, 2), \quad w = (-2, 0, 1), \quad z = (1, -1, 0)$
We are given four vectors in $\mathbb{R}^3$, so they must be linearly dependent (since the dimension of $\mathbb{R}^3$ is 3). But to determine whether their span has dimension 3 or less, we now verify whether three of them are linearly independent.

Step 1: Check if u, v, w are linearly independent.
Assume:

$$\alpha u + \beta v + \gamma w = 0.$$

That is,

$$\alpha(1, 1, -2) + \beta(1, 3, 2) + \gamma(-2, 0, 1) = (0, 0, 0).$$

Compute each coordinate:

1. First: $\alpha + \beta - 2\gamma = 0$
2. Second: $\alpha + 3\beta = 0$
3. Third: $-2\alpha + 2\beta + \gamma = 0$

From (2): $\alpha = -3\beta$
Substitute into (1):

$$-3\beta + \beta - 2\gamma = 0 \Rightarrow -2\beta - 2\gamma = 0 \Rightarrow \gamma = -\beta$$

Now substitute into (3):

$$-2(-3\beta) + 2\beta + (-\beta) = 6\beta + 2\beta - \beta = 7\beta = 0 \Rightarrow \beta = 0 \Rightarrow \alpha = 0, \gamma = 0.$$

Hence, the only solution is the trivial one, so u, v, w are linearly independent.

Step 2: The set $\{u, v, w, z\}$ is linearly dependent (since it has 4 vectors in $\mathbb{R}^3$), but since $\{u, v, w\}$ is already linearly independent, the dimension of the span is 3.

$$\dim(\text{span}\{u, v, w, z\}) = \boxed{3}.$$

■

Exercise 24. *Let $u, v, w \in \mathbb{R}^n$, and $\alpha, \beta, \lambda \in \mathbb{R}$, such that $\alpha, \beta \neq 0$, and*

$$\alpha u + \beta v + \lambda w = 0$$

Show that $\text{span}(u, w) = \text{span}(v, w)$.

Solution. Given $x \in span(u, w)$ we know $x = au + bw$. But using the constraint we can also write this as $x = bw + a(-\frac{\lambda}{\alpha}w - \frac{\beta}{\alpha}v) = (b - \frac{a\lambda}{\alpha})w - \frac{a\beta}{\alpha}v$, so is in $span(u, w)$. Similar arguments get you back the other way. ■

Bases

Exercise 25.

- *Show that $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ is a basis for $\mathbb{R}^3$. This basis is called the canonical basis for $\mathbb{R}^3$.*
- *In general, show that for $n \in \mathbb{N}^\star$, then the set $\{e_1, \ldots, e_n\}$, where $(e_i)_i = 1$ and $(e_j)_{j \neq i} = 0$, is a basis for $\mathbb{R}^n$. This basis is called the* canonical basis
- *Show that $\{(1, 0, 0), (1, 1, 0), (1, 1, 1)\}$ is a basis for $\mathbb{R}^3$.*
- *In general, show that for $n \in \mathbb{N}^\star$, then the set $\{e_1, \ldots, e_n\}$, where $(e_i)_j = 1$ if $j \leq i$ and $(e_j)_{j > i} = 0$, is a basis for $\mathbb{R}^n$.*

Solution. **1. Show that $\{(1, 0, 0),\ (0, 1, 0),\ (0, 0, 1)\}$ is a basis for $\mathbb{R}^3$.**
This set is linearly independent because any linear combination

$$\alpha(1, 0, 0) + \beta(0, 1, 0) + \gamma(0, 0, 1) = (0, 0, 0)$$

implies $\alpha = \beta = \gamma = 0$, and it spans $\mathbb{R}^3$ since any vector $(x, y, z) \in \mathbb{R}^3$ can be written as:

$$x(1, 0, 0) + y(0, 1, 0) + z(0, 0, 1).$$

So the canonical basis of $\mathbb{R}^3$ is:

$$\boxed{\{(1, 0, 0),\ (0, 1, 0),\ (0, 0, 1)\}}.$$

2. Show that for any $n \in \mathbb{N}^\star$, the set $\{e_1, \ldots, e_n\}$, where $(e_i)_j = \delta_{ij}$, is a basis for $\mathbb{R}^n$.
The vectors $e_1, \ldots, e_n$ are defined by:

$$(e_i)_j = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Any vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ can be written as:

$$x = x_1 e_1 + \cdots + x_n e_n,$$

so the set spans $\mathbb{R}^n$. The representation is unique, and any linear combination

$$\sum_{i=1}^{n} \lambda_i e_i = 0 \Rightarrow \lambda_i = 0 \text{ for all } i$$

implies linear independence.
Hence, $\{e_1, \ldots, e_n\}$ is a basis, called the *canonical basis* of $\mathbb{R}^n$.

3. Show that $\{(1, 0, 0),\ (1, 1, 0),\ (1, 1, 1)\}$ is a basis for $\mathbb{R}^3$.
Let us check linear independence.
Suppose

$$\alpha(1, 0, 0) + \beta(1, 1, 0) + \gamma(1, 1, 1) = (0, 0, 0).$$

Compute the sum:

$$(\alpha + \beta + \gamma,\ \beta + \gamma,\ \gamma) = (0, 0, 0).$$

From the third coordinate: $\gamma = 0$. Then second: $\beta + 0 = 0 \Rightarrow \beta = 0$. Then first: $\alpha + 0 + 0 = 0 \Rightarrow \alpha = 0$.
So the set is linearly independent. There are 3 vectors in $\mathbb{R}^3$, so they form a basis.

$$\dim(\text{span}) = 3, \quad \text{and the set is a basis.}$$

4. Show that for $n \in \mathbb{N}^\star$, the set $\{e_1, \ldots, e_n\}$, where $(e_i)_j = 1$ if $j \leq i$, and 0 otherwise, is a basis for $\mathbb{R}^n$.
We define $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (1, 1, 0, \ldots, 0)$, $e_3 = (1, 1, 1, 0, \ldots, 0)$, etc., up to $e_n = (1, 1, \ldots, 1)$.
We show that this set is linearly independent and spans $\mathbb{R}^n$.
We proceed by induction or triangular argument. Let $E = \{e_1, \ldots, e_n\}$, and suppose:

$$\sum_{i=1}^{n} \lambda_i e_i = 0.$$

Look at the last coordinate: only e_n contributes to it (equal to 1), so $\lambda_n = 0$. Then consider the $(n-1)$-th coordinate: only e_{n-1} and e_n contribute to it, and $\lambda_n = 0$, so $\lambda_{n-1} = 0$, etc.
This descending chain gives $\lambda_i = 0$ for all i, so the vectors are linearly independent.
Since we have n linearly independent vectors in $\mathbb{R}^n$, they form a basis.

$$\boxed{\{e_1, \ldots, e_n\} \text{ is a basis for } \mathbb{R}^n.}$$

■

Bases of Subspaces

Exercise 26. *Show that the subset of* $\mathbb{R}^3$ *given by:* $E = \{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}$ *is a subspace of* $\mathbb{R}^3$. *Find a basis for* E *and determine its dimension.*

Solution. **1. Show that $E \subseteq \mathbb{R}^3$ is a subspace.**
We are given:

$$E = \{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}.$$

To show that E is a subspace of $\mathbb{R}^3$, we verify the three subspace properties:

- **Closed under addition:** Let $u = (x_1, y_1, z_1) \in E$ and $v = (x_2, y_2, z_2) \in E$, so $x_1 + y_1 + z_1 = 0$ and $x_2 + y_2 + z_2 = 0$. Then

$$(x_1 + x_2) + (y_1 + y_2) + (z_1 + z_2) = (x_1 + y_1 + z_1) + (x_2 + y_2 + z_2) = 0 + 0 = 0.$$

  Hence, $u + v \in E$.

- **Closed under scalar multiplication:** Let $\lambda \in \mathbb{R}$ and $(x, y, z) \in E$, so $x+y+z = 0$. Then

$$\lambda x + \lambda y + \lambda z = \lambda(x + y + z) = \lambda \cdot 0 = 0.$$

  Thus, $\lambda(x, y, z) \in E$.

Therefore, E is a subspace of $\mathbb{R}^3$.

2. Find a basis for E and its dimension.
We solve the equation:

$$x + y + z = 0 \iff x = -y - z.$$

So every element $(x, y, z) \in E$ can be written as:

$$(x, y, z) = (-y - z, y, z) = y(-1, 1, 0) + z(-1, 0, 1).$$

This shows that:

$$E = \operatorname{span}\{(-1, 1, 0), (-1, 0, 1)\}.$$

We now prove that these two vectors are linearly independent. Suppose:

$$\alpha(-1, 1, 0) + \beta(-1, 0, 1) = (0, 0, 0).$$

Then:

$$-\alpha - \beta = 0, \quad \alpha + 0 \times \beta = 0, \quad 0 \times \alpha + \beta = 0.$$

From the second equation: $\alpha = 0$, and from the third: $\beta = 0$. Plugging into the first confirms the equation is satisfied. Hence, the only solution is $\alpha = \beta = 0$, so the vectors are linearly independent.

$$\text{Basis of } E : \{(-1, 1, 0), (-1, 0, 1)\}, \quad \text{Dimension: } \boxed{2}.$$

∎

3.4 Coordinates

3.4.1 Definition and Examples

Definition 19 (Coordinates). *Let F be a m-dimensional subspace of $\mathbb{R}^n$. Given a basis $B_F := \{e_1, \ldots, e_m\}$. Let $x \in F$, then since B_F is a basis, there exists unique scalars $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ such that $x = \sum_{i=1}^{k} \lambda_i e_i$. Then the object $[x]_{B_F} = (\lambda_i)_{1 \le i \le k} \in \mathbb{R}^k$ is called its vector of coordinates in the basis B_F.*

Why is the notion of coordinates useful?

Coordinates provide any vector x in some arbitrary m-dimensional subspace F with a description in terms of m numbers. These numbers depend on the basis B_F chosen; thus, to one vector x, we can associate as many set of coordinates as there are bases in F!

A visual explanation

Below is a drawing showing the impact of the basis choice on the coordinates of vectors in $\mathbb{R}^2$. Both figures choose a different basis B, and represent "isolines", e.g., lines of constant first or second coordinate value z_0, for some equally spaces values of z_0: (-1, -0.5, ..., 1). The two arrows are used to denote the two basis vectors of each basis.

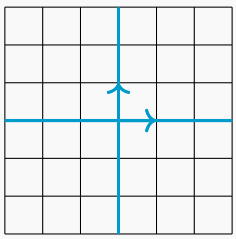


Figure 3.1: Isolines of $\mathbb{R}^2$ for the canonical basis $\{(1,0),(0,1)\}$

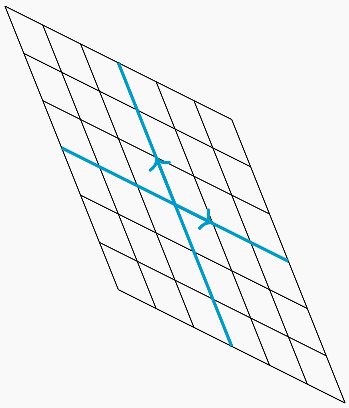


Figure 3.2: Isolines for the basis $\{(1, -\frac{1}{2}), (-\frac{1}{2}, 1)\}$

Canonical Bases do not always exist!

For the specific yet important case of $F = \mathbb{R}^n$, the numbers $(x_1, \ldots, x_n)$ forming the vector $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ are 1-1 correspondence with coordinates of x in the canonical basis, which is why this basis deserves a name (canonical has to be understood as "natural"). However, canonical bases do not always exist for all vector spaces, as the next exercise shows.

Exercise 27. *Let $F \subset \mathbb{R}^2 := \left\{x \in \mathbb{R}^2 : x_1 + x_2 = 0\right\}$. Give a basis for F. $F \subset \mathbb{R}^n :=$*

$\left\{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 0\right\}$. *Give a basis for F.*

Solution. **1. Case of $F \subset \mathbb{R}^2$ defined by $x_1 + x_2 = 0$:**
We want to describe the subspace:

$$F = \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 = 0\right\}.$$

We can express every vector in F as:

$$(x_1, x_2) = (x_1, -x_1) = x_1 \cdot (1, -1).$$

So every element of F is a scalar multiple of the vector $(1, -1)$, and this vector is nonzero.
Therefore,

$$\boxed{\{(1, -1)\}} \text{ is a basis for } F, \quad \text{and } \dim(F) = 1.$$

2. General case: $F \subset \mathbb{R}^n$ defined by $\sum_{i=1}^{n} x_i = 0$:
This is a linear subspace of $\mathbb{R}^n$ defined by one linear equation, so it has dimension $n - 1$. To build a basis, observe that we can solve the constraint by expressing the last coordinate in terms of the others:

$$x_n = -\sum_{i=1}^{n-1} x_i.$$

So any $x \in F$ has the form:

$$x = (x_1, \ldots, x_{n-1}, -x_1 - \cdots - x_{n-1}).$$

For each $i \in \{1, \ldots, n-1\}$, define:

$$v_i = (\underbrace{0, \ldots, 1, \ldots, 0}_{i\text{-th position}}, 0, \ldots, 0, \ -1),$$

i.e., $v_i \in \mathbb{R}^n$ has 1 at the i-th position, 0 elsewhere (except at the last coordinate, which is -1).
Conclusion:

$$\boxed{F = \operatorname{Span}\{v_1, \ldots, v_{n-1}\}}.$$

This set forms a basis for F, and $\dim(F) = n - 1$. You should check that $\{v_1, \ldots, v_{n-1}\}$ is indeed a basis. ■

In that case, the coordinates of some $x \in F$ with respect to B_F are different from the numbers in $x \in F$. Thus, in that case, there is no clear one-one correspondence between the traditional representation of vectors in F and its coordinate in any given basis!

3.5 Changing Basis with Matrices

Definition

Let $n \in \mathbb{N}^*$, and let $B \coloneqq \{e_1, \ldots, e_n\}, B' \coloneqq \{f_1, \ldots, f_n\}$ be two bases of $\mathbb{R}^n$. We define the change of basis matrix from B to B', denoted $P_B^{B'}$, as the matrix given by:

$$((P_B^{B'})_{ij})_{1 \le i,j \le n} = (([e_j]_{B'})_i)_{1 \le i,j \le n} = \begin{pmatrix} | & \ldots & | \\ [e_1]_{B'} & \ldots & [e_n]_{B'} \\ | & \ldots & | \end{pmatrix}$$

Given some $x \in \mathbb{E}$, and two basis B, B' of E, $P_{B,B'}$ allows to compute the coordinates of x in the basis B' from the coordinates of x in the basis B:

Proposition 14. *Let E be a n-dimensional vector space, and B, B' two bases of E. Then, for all $x \in E$, we have that:*

$$[x]_{B'} = P_B^{B'}[x]_B$$

Proof. Let $x \in E$ and let $(\lambda_1, \ldots, \lambda_n) \coloneqq [x]_B$ be the coordinates of x in B. Then

$$\begin{aligned} x &= \sum_{i=1}^n \lambda_i e_i \\ &= \sum_{i=1}^n \lambda_i \sum_{j=1}^n ([e_i]_{B'})_j f_j \\ &= \sum_{j=1}^n \left(\sum_{i=1}^n ([e_i]_{B'})_j \lambda_i \right) f_j \\ &= \sum_{i=1}^n \left(\sum_{j=1}^n ([e_j]_{B'})_i \lambda_j \right) f_i \\ &= \sum_{i=1}^n (P_B^{B'}[x]_B)_i f_i \end{aligned}$$

Thus, we have that $[x]_{B'} = (P_B^{B'}[x]_B)_{1 \le i \le n} = P_B^{B'}[x]_B$, which concludes the proof. $\square$

Informal discussion: why should we care about change of basis matrices

Change of basis matrices will appear when studying the matrix-representation of linear maps. As we will see, some linear maps are best represented (in matrix form) using a particular basis, which may not be the original basis the map was represented with in the first place. The link between its original representation and its "best representation" will involve a change of basis matrices.

Familiarizing Exercise

Exercise 28. *In the following settings, you will be given a vector space E, two basis B and B' and the coordinates $[x]_B$ of some vector x in the basis B. In each case, write the change of basis matrix $P_B^{B'}$, and compute $[x]_{B'}$*

- $n = 2$, $B = \{(1,0),(0,1)\}$ $B' = \{(2,0),(0,2)\}$, $[x]_B = (1,1)$.
- $n = 3$, $B := \{(a_1,a_2,a_3),(b_1,b_2,b_3),(c_1,c_2,c_3)\}$ *where the* a_i, b_i, c_i, $i \in \{1,2,3\}$ *are picked such that B actually forms a basis, and* $B' := \{(1,0,0),(0,1,0),(0,0,1)\}$. $[x]_B = (1,1,1)$.

Solution. **First case:** We are told $[x]_B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, so we compute $[x]_{B'} = P_B^{B'} \cdot [x]_B$.
Each column of $P_B^{B'}$ is the coordinate of a basis vector of B in the basis B'. That is:
$(1,0) = \frac{1}{2}(2,0) \Rightarrow [e_1]_{B'} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$,
$(0,1) = \frac{1}{2}(0,2) \Rightarrow [e_2]_{B'} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$.
So:

$$P_B^{B'} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad [x]_{B'} = P_B^{B'} \cdot [x]_B = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Second case: Let $B = \{v_1, v_2, v_3\} \subset \mathbb{R}^3$, where:

$$v_1 = (a_1,a_2,a_3), \quad v_2 = (b_1,b_2,b_3), \quad v_3 = (c_1,c_2,c_3).$$

We are told that:

$$[x]_B = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

which means:

$$x = 1 \cdot v_1 + 1 \cdot v_2 + 1 \cdot v_3 = v_1 + v_2 + v_3.$$

Step 1: Change of basis matrix $P_B^{B'}$. The matrix $P_B^{B'}$ is the matrix whose j-th column is the coordinate vector of v_j in the basis B' (i.e., its coordinates in the canonical basis). So:

$$P_B^{B'} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}.$$

Step 2: Compute $[x]_{B'}$.
Using:

$$[x]_{B'} = P_B^{B'} \cdot [x]_B = P_B^{B'} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

we get:

$$[x]_{B'} = v_1 + v_2 + v_3 = \begin{pmatrix} a_1 + b_1 + c_1 \\ a_2 + b_2 + c_2 \\ a_3 + b_3 + c_3 \end{pmatrix}.$$

Conclusion: The vector x, originally expressed in basis B by coordinates $(1, 1, 1)$, has canonical coordinates:

$$[x]_{B'} = v_1 + v_2 + v_3 = \begin{pmatrix} a_1 + b_1 + c_1 \\ a_2 + b_2 + c_2 \\ a_3 + b_3 + c_3 \end{pmatrix}.$$

■

Properties of Change of Basis Matrices

The following convenient properties hold for change of basis matrices:

Chaining change of basis matrices

Proposition 15. *Let E be a n-dimensional vector space. Let B, B', B'' be three bases of E. Then we have:*

$$P_B^{B''} = P_{B'}^{B''} P_B^{B'}.$$

Proof. Let $B \coloneqq \{e_1, \ldots, e_n\}, B' \coloneqq \{f_1, \ldots, f_n\}, B'' \coloneqq \{g_1, \ldots, g_n\}$, for some $e_i, f_i, g_i \in \mathbb{R}^n$, $i = 1, \ldots, n$. We are going to express e_i in the basis B'', using the basis B' as a pivot to show the desired equality.

$$\begin{aligned} e_j &= \sum_{k=1}^{n} (P_B^{B'})_{kj} f_k \\ &= \sum_{k=1}^{n} (P_B^{B'})_{kj} \sum_{i=1}^{n} (P_{B'}^{B''})_{ik} g_i \\ &= \sum_{i=1}^{n} (\sum_{i=1}^{n} (P_{B'}^{B''})_{ik} (P_B^{B'})_{kj} g_i \\ &= \sum_{i=1}^{n} (P_{B'}^{B''} P_B^{B'})_{ij} g_i \end{aligned}$$

Because the last line is also, by definition of change of basis matrix, $P_B^{B''}$, we have that $P_{B'}^{B''} P_B^{B'} = P_B^{B''}$. □

Change of basis matrices are invertible

Proposition 16. *Let E be a n-dimensional vector space, and let B, B' be two bases of E. $P_B^{B'}$ is invertible, and $(P_B^{B'})^{-1} = P_{B'}^{B}$.*

Proof. We can use the chaining property of change of basis matrices: indeed, setting $B'' = B$, we have that:

$$P_B^B = I_n = P_{B'}^B P_B^{B'}$$

$$P_{B'}^{B'} = I_n = P_B^{B'} P_{B'}^B$$

From the two equalities above, $P_{B'}^B$ satisfies the definition of the inverse of $P_B^{B'}$. Thus, $P_B^{B'}$ is invertible, and $(P_B^{B'})^{-1} = P_{B'}^B$. $\square$

3.6 Exercise Sheet 4: Coordinates

Familiarization Exercise

Exercise 29. *Let us focus first on the case $\mathbb{R}^2$. Let us take one $z \in \mathbb{R}^2$. By definition, $z := [x, y]$ for some $x, y \in \mathbb{R}$. Then give the coordinates of z in:*

- *the base $B := \{[0, 1], [1, 0]\}$,*
- *the base $B' := \{[1, 1], [2, -1]\}$*

Solution.

- for $B := \{[0, 1], [1, 0]\}$, then since $[z] = x[1, 0] + y[0, 1]$, then $[z]_B = [y, x]$,
- for $B' := \{[1, 1], [1, -1]\}$ then since $[z] = \frac{x+y}{2}[1, 1] + \frac{x-y}{2}[1, -1]$, then $[z]_{B'} = [\frac{x+y}{2}, \frac{x-y}{2}]$.

■

Exercise 30. *Let $n > 0$ and let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Give the coordinates of x in*

1. *The canonical basis*
2. *The basis $B' = \{(1, 0, \ldots, 0), (-1, 1, 0, \ldots, 0), (0, -1, 1, 0, \ldots, 0) \ldots, (0, \ldots, -1, 1)\}$*

Solution.

1. We have that:
$$(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i e_i$$
thus, by definition, $[x]_B = (x_1, \ldots, x_n)$.

2. Let y_i be the i-th coordinate of x in the basis B'. Then by definition of coordinates, we have that
$$(x_1, \ldots, x_n) = \sum_{i=1}^{n} y_i e_i = (y_1 - y_2, y_2 - y_3, \ldots, y_{n-1} - y_n, y_n),$$
thus, solving for $y_1, \ldots, y_n$, we have $y_n = x_n$, $y_{n-1} = y_n + x_{n-1} = x_n + x_{n-1}$, etc. It is possible to show by induction that for a given $m \in \{1, \ldots, n\}$. $y_m = \sum_{i=m}^{n} x_i$. The proof is done by induction on m. Therefore,
$$[x]_{B'} = \left(\sum_{i=1}^{n} x_i, \sum_{i=2}^{n} x_i, \ldots, x_{n-1} + x_n, x_n \right).$$

■

Exercise 31. *Let $F \subseteq \mathbb{R}^4$ defined by the equation*

$$F = \{(x, y, z, t) : x + z = t + y\}$$

and G defined by :

$$G = \{(x, y, z, t) : y + t = x - z - y = 0\}$$

- *Find the dimension of F and a basis for F, and express the coordinates of the vector $(3, 1, 2, 4)$ in this basis.*
- *Find the dimension of G and a basis for G, and express the coordinates of the vector $(4, 1, 3, -1)$ in this basis.*
- *Find the dimension of $F \cap G$ and a basis for $F \cap G$.*

Solution. TO DO ■

Exercise 32. *Let $\{v_1, \ldots, v_p\}$ a family of vectors in $\mathbb{R}^n$. Let $\{u_1, \ldots, u_p\}$ defined by:*

$$u_i = \sum_{j=1}^{i} v_j, \quad i \in \{1, \ldots, p\}$$

Show that the family $\{u_1, \ldots, u_p\}$ is linearly independent if and only if the family $\{v_1, \ldots, v_p\}$ is linearly independent.

Solution. TO DO ■

Chapter 4

Inner Products and Orthogonality

4.1 Norm and Inner Product on $\mathbb{R}^n$

The inner product in $\mathbb{R}^n$

Definition 20. *Let $n \in \mathbb{N}^*$.*

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i, \quad x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$$

is called the inner product on $\mathbb{R}^n$.

Inner Product

Proposition 17. *The inner product $\langle \cdot, \cdot \rangle$ is a map from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$ that is*

- *bilinear (linear in each variable)*
- *symmetric: $\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in E$*
- *positive: $\langle x, x \rangle \geq 0 \quad \forall x \in E$*
- *definite: $\langle x, x \rangle = 0 \iff x = 0$*

Proof. The bilinearity and symmetry are immediate. For the positivity, note that $\langle x, x \rangle = \sum_{i=1}^{n} x_i^2 \geq 0$, and $\langle x, x \rangle = 0 \implies x_i^2 = 0$ for all i, implying that $x = 0_{\mathbb{R}^n}$. □

Proposition 18 (Cauchy–Schwarz inequality)**.** *For all $x, y \in \mathbb{R}^n$:*

$$|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle}\sqrt{\langle y, y \rangle}$$

Proof. Let $x, y \in \mathbb{R}^n$, and let $\lambda \in \mathbb{R}$. Then, by bilinearity and symmetry, we have that:

$$\langle x - \lambda y, x - \lambda y \rangle \iff \langle x, x \rangle - 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \geq 0$$

The l.h.s is a second-order polynomial in λ, and does not change sign. Thus, its discriminant is negative, e.g.:

$$\begin{aligned} &4 \langle x, y \rangle^2 - 4 \langle x, x \rangle \langle y, y \rangle \leq 0 \\ \implies &\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle \\ \implies &|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle} \end{aligned}$$

□

Definition 21. $\|\cdot\|_n := \left(x \longmapsto \sqrt{\langle x, x \rangle}\right)$ *is called the canonical norm for* $\mathbb{R}^n$.

Norm

Proposition 19. $\|\cdot\|$ *is such that it is*

- *homogeneous:* $\|x\| \geq 0 \quad \forall x \in E$
- *positive:* $\|x\| \geq 0 \quad \forall x \in E$
- *definite:* $\|x\| = 0 \iff x = 0.$
- *verifies the triangular inequality:* $\|x + y\| \leq \|x\| + \|y\|.$

Casually Explained

The norm should be thought of as the "length" of a vector. The norm arising from the inner product recovers the traditional properties that length have. For instance, for $n = 2$,

$$\|(x, y)\| = \sqrt{x^2 + y^2}$$

Which is precisely the length of straight line from the origin to the point (x, y). The triangular identity signals that it is always faster to go from 0 to $(x + y)$ in a straight line, than to go from 0 to x, and then from x to $x + y$.

Proof. The only nontrivial point is the triangular inequality, which can be proven using

Proposition 18. Indeed:

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle \\ &= \langle x,x\rangle + \langle y,y\rangle + 2\langle x,y\rangle \\ &= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x,y\rangle| \\ &\leq \|x\|^2 + \|y\|^2 + 2\sqrt{\langle x,x\rangle}\sqrt{\langle y,y\rangle} \\ &= \|x\|^2 + \|y\|^2 + 2\|x\|\,\|y\| \\ &= (\|x\| + \|y\|)^2\end{aligned}$$

From which it follows, by taking the square root of the last inequality, that $\|x+y\| \leq \|x\| + \|y\|$. □

The "inner product" norm is not the only norm!

Any function on $\mathbb{R}^n$ that satisfies the properties listed in Proposition 19 is called a norm. The canonical norm is not the only norm on $\mathbb{R}^n$. It is possible to show that the following functions define a norm on $\mathbb{R}^n$:

- $\|\cdot\|_1$, defined by $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- $\|\cdot\|_\infty$, defined by $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$
- More generally: $\|\cdot\|_p$, defined by $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$

Under that notation, the "inner product" norm is $\|\cdot\|_2$. Their unit spheres (e.g., the set $\{x \in \mathbb{R}^n : \|x\| = 1\}$ is drawn below, for the case $n = 2$.

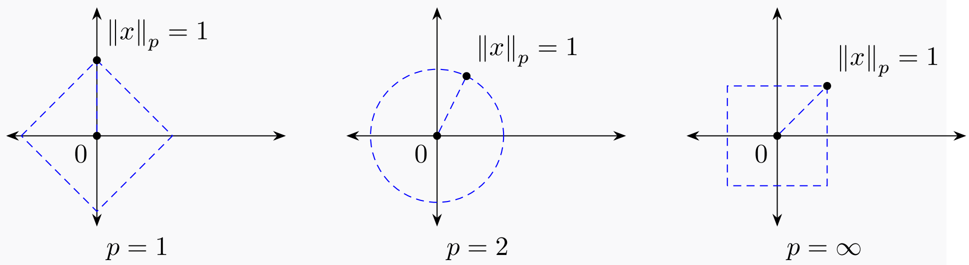


Figure 4.1: The unit balls for the norms $\|\cdot\|_p$, e.g the set $\left\{x : \|x\|_p = 1\right\}$ of the norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ in $\mathbb{R}^2$.

Exercise 33. *Let $x, y \in \mathbb{R}^2$, such that the angle between x and y is θ. By knowing that*

$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$, *show that;*

$$\langle x, y \rangle = \|x\|_2 \, \|y\|_2 \cos(\theta)$$

Solution. Let $x, y \in \mathbb{R}^2$, and suppose the angle between them is θ. Let us write the vectors in polar coordinates:

$$x = \|x\|_2(\cos(\alpha), \sin(\alpha)), \quad y = \|y\|_2(\cos(\beta), \sin(\beta))$$

Then the angle between x and y is given by $\theta = \beta - \alpha$. We compute their dot product:

$$\langle x, y \rangle = \|x\|_2\|y\|_2 \left(\cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)\right)$$

Now, recall the identity:

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$$

Thus,

$$\langle x, y \rangle = \|x\|_2\|y\|_2 \cos(\beta - \alpha)$$

Since $\theta = \beta - \alpha$, we obtain:

$$\langle x, y \rangle = \|x\|_2\|y\|_2 \cos(\theta)$$

as required. ■

Inner product as a "similarity metric"

Given two vectors x, y such that $\|x\| = \|y\| = 1$, $\langle x, y \rangle$ can be seen as a measure of how similar x and y are:

- $\langle x, y \rangle$ is maximized when $x = y$.
- $\langle x, y \rangle$ is minimized when $x = -y$.
- in $\mathbb{R}^2$ - we will see how that notion carries over to $\mathbb{R}^n$ - $\langle x, y \rangle$ is 0 when x and y are orthogonal.

The view of inner products as a measure of similarity motivates its use in neural networks, in which the firing rate of a neuron is often chosen to take the form:

$$\phi(w^\top x)$$

where x is the input, and w is the "receptive field" of the neuron. The nonlinearity indicates the saturation of the neuron, which cannot reach arbitrarily high (or negative) firing rates, and is often chosen to cap the firing rate from 0 to some positive value a.

4.2 Orthogonality, Orthogonal Basis, Orthogonal Subspaces

The concept of orthogonality will allow us to construct a new kind of basis whose elements are "independent" in a sense that we will make precise later.

Definition 22. *Let $n\mathbb{N}^*$,*

- *Let $x, y \in \mathbb{R}^n$. Then x, y are called **orthogonal** if $\langle x, y\rangle = 0$.*
- *A set of vectors $\{x_1, \ldots, x_n\}$ is called **orthogonal** if $\langle x_i, x_j\rangle = 0$ for all $i \neq j$.*
- *A set of vectors $\{x_1, \ldots, x_n\}$ is called **orthonormal** if it is orthogonal and $\|x_i\| = 1$ for all i.*
- *A basis B of E is called **orthogonal** if B is an orthogonal set, and **orthonormal** if B is an orthonormal set.*

Important Example

The canonical basis $\{e_1, \ldots, e_n\}$ is an orthonormal basis of $\mathbb{R}^n$.

4.3 Representation of vectors in Orthonormal Bases

The coefficients of a vector x in an orthonormal basis B are easy to compute:

Decomposition in Orthogonal Bases

Theorem 2. *Let $\mathbb{R}^n$ be an Euclidean space, and let $B = \{e_1, \ldots, e_n\}$ be an orthonormal basis of $\mathbb{R}^n$. Then*

$$x = \sum_{i=1}^{n} \langle x, e_i\rangle\, e_i$$

implying that

$$([x]_B)_i = \langle x, e_i\rangle$$

Proof

Proof. By definition, we have $x = \sum_{i=1}^{n} [x]_{B,i}\, e_i$. Taking the scalar product with e_j and leveraging the orthonormality of B gives the desired result. □

4.4 Exercise Sheet 5: Orthogonality

Familiarizing Exercise

Exercise 34.

- *Let* $x = (1, 2, 3), y = (-1, 1, 0)$. *Compute:* $\langle x, y\rangle$, $\|x\|$, $\|y\|$.
- *Construct an orthonormal basis of* $\mathbb{R}^2$ *that is not the canonical basis.*
- *Construct an orthonormal basis of* $\mathbb{R}^n$ *that is not the canonical basis.*
- *Let* $F = \{(x, y, z) : x + y + z = 0\}$. *Find an orthonormal basis for* F. *(Hint: find a standard basis first, and then try to "orthonormalize it"—ask teacher for help!)*
- *Construct a basis of* $\mathbb{R}^2$ *which is not orthonormal.*

Solution. TO DO ∎

Orthogonal Complements

Exercise 35. *Let* $F \subset \mathbb{R}^n$ *be a vector subspace. Let* $F^\perp$ *be the set defined as:*

$$F^\perp = \{x \in \mathbb{R}^n : \forall y \in F, \langle x, y\rangle = 0\}$$

Show that:

- $F^\perp$ *is a vector subspace of* $\mathbb{R}^n$.
- F *and* $F^\perp$ *are in direct sum*

In fact, it can be shown that $F \oplus F^\perp = \mathbb{R}^n$. $F^\perp$ *is called the* orthogonal complement *of* F.

Solution. TO DO ∎

Exercise 36. *Given two square matrices* $A, B \in \mathcal{M}_n(\mathbb{R})$, *we define the function* $f(A, B)$ *as:*

$$f(A, B) = \mathrm{Tr}(A^\top B)$$

where we recall that $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$. *Show that:*

- f *is symmetric, i.e.:* $f(A, B) = f(B, A)$ *for all* $A, B \in \mathcal{M}_n(\mathbb{R})$.
- f *is bilinear, i.e.:* $f(\lambda A + \mu B, C) = \lambda f(A, C) + \mu f(B, C)$ *for all* $\lambda, \mu \in \mathbb{R}$ *and* $A, B, C \in \mathcal{M}_n(\mathbb{R})$.
- f *is positive definite, i.e.:* $f(A, A) \geq 0$ *for all* $A \in \mathcal{M}_n(\mathbb{R})$, *and* $f(A, A) = 0$ *if and only if* $A = 0$.

f *can be shown to serve as the equivalent of an inner product, for the space of square matrices.*

Solution. TO DO ∎

Chapter 5

Preliminaries on Functions

5.1 Mappings

5.1.1 First Definitions

Mappings

Definition 23 (Mapping - slightly informal definition). *A mapping f consists of three elements:*

- *a domain or set of definition E;*
- *a co-domain F;*
- *a mechanism or relationship that associates to every element x of E a unique element $f(x)$ of F.*

We use the following notation to summarize this information:

$$\begin{aligned} f: \quad & E \to F \\ & x \mapsto f(x) \end{aligned}$$

We also say that f is defined on E to indicate that the set of definition is E and that f maps to values in F to indicate that the co-domain is F.

Casually Explained: The notation $f(x)$

For a given function f and some $x \in E$, the notation $f(x)$ often lays out a procedure to *compute* the $y = f(x)$. However, a procedure may not always stop, and thus may never produce an output (think for instance of procedures defined recursively $f(x) = x \times f(x-1)$). The definition of a mapping gets rid of having to deal with such concerns by abstracting away the procedure, and only characterizing a function through its input-output pairs. In practice of course, most procedures of interest are known to actually produce an output (for instance, $f(x) = x \times x$).

Identity Mapping

Definition 24 (Identity mapping)**.** *If E is any set, we can define an identity application id_E as follows:*

$$id_E : E \to E, \quad x \mapsto x$$

(for every $x \in E$, we have $id_E(x) = x$).

5.1.2 Compositions of maps

Compositions of two maps

Definition 25. *Let E, F, G be three sets, $f : E \longmapsto F$ and $g : F \longmapsto G$. The composition of f and g, denoted $g \circ f$, is the map*

$$\begin{aligned} g \circ f : E &\longrightarrow G \\ x &\longmapsto g \circ f(x) = g(f(x)) \end{aligned}$$

Iterating compositions of maps

Definition 26. *Let E be a set, $f : E \longmapsto E$ a map, and $n \in \mathbb{N}$. We define the n-th iterate of f, denoted f^n, as*

$$f^n = \underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}}$$

with the convention that $f^0 = Id_E$.

Examples

If

- $f : \mathbb{R} \to \mathbb{R}$ is the function defined by $\forall x \in \mathbb{R}$, $f(x) = x^2$,
- $g : \mathbb{R} \to \mathbb{R}$ is the function defined by $\forall x \in \mathbb{R}$, $g(x) = \sin(x)$,

then $g \circ f$ and $f \circ g$ are functions from $\mathbb{R}$ to $\mathbb{R}$, and for all $x \in \mathbb{R}$, we have

$$(f \circ g)(x) = (\sin(x))^2 \quad \text{and} \quad (g \circ f)(x) = \sin(x^2).$$

Note: The notation $g \circ f$ reads "from right to left": it denotes the function where f is applied first, followed by g.

Composition with identity. Here is another more theoretical example. If $f : E \to F$ is a function, then we have

$$\mathrm{id}_F \circ f = f \quad \text{and} \quad f \circ \mathrm{id}_E = f.$$

5.1.3 Image and pre-Image

Image and pre-Image

Definition 27 (Image and pre-Image)**.** *For every element $x \in E$, $f(x)$ is called the image of x by f. The image of f is the set of images of the elements of E by f, that is*

$$\mathrm{Im}(f) = \{y \in F \text{ such that } \exists x \in E \text{ verifying } f(x) = y\}.$$

For every $y \in \mathrm{Im}(f)$, x is called a pre-image of y if $f(x) = y$. Note that, by the very definition of a function, an element of E has a unique image. However, an element of $\mathrm{Im}(f)$ *can have multiple pre-images.*

Example

Consider the function

$$f : \mathbb{R} \to \mathbb{R} \quad x \mapsto x^2.$$

- The element $y = -1$ of the co-domain $\mathbb{R}$ is not reached by f.
- The element $y = 0$ of $\mathbb{R}$ is reached by f: the number $x = 0$ is a pre-image of it, and it is the only pre-image of $y = 0$ by f.
- The element $y = 3$ is also reached by f: it has exactly two pre-images by f, the numbers $x = \sqrt{3}$ and $x = -\sqrt{3}$.

Casually Explained

Casually Explained: a function is not one-to-many This definition is the formalization of the intuitive definition of a mapping, which is some *relation* which, for each element in x in E, associates a single element in F, denoted $f(x)$. In other terms, we say that f cannot be *one-to-many*. The fact that a function is not *one-to-many* makes it differ other mapping-like objects like conditional distributions $p(y|x)$: for a given x, each time one samples from $p(\cdot|x)$, one might end up with a different y.
Casually Explained: a function can be many-to-one! Even though f cannot be one-to-many, nothing in its definition prevents it from being *many-to-one*: for a given $y \in F$, there might be several $x \in E$ such that $f(x) = y$.

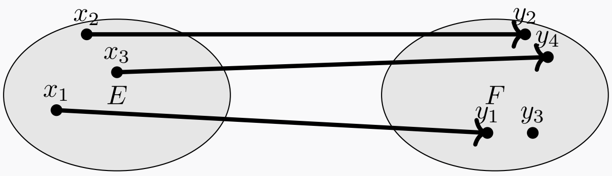
5.1.4 Injectivity, Surjectivity, Bijectivity and Inversion of Maps

Injective Map

Let E and F be *two sets*, and let f be a map from E to F. We say that f is injective if:

$$\forall x, y \in E, \quad f(x) = f(y) \implies x = y$$

A Visual Explanation



Here, f is given by the triplet

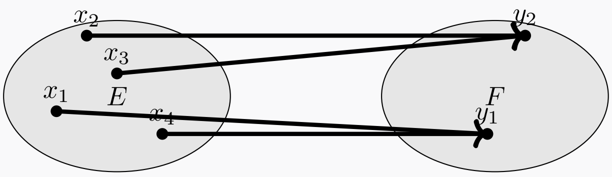
- $E = \{x_1, x_2, x_3\}$
- $F = \{y_1, y_2, y_3, y_4\}$
- $G \subset E \times F = \{(x_1, y_1), (x_2, y_2), (x_3, y_4)\}$

Surjective Map

Let E and F be *two sets*, and let f be a map from E to F. We say that f is surjective if:

$$\forall y \in F, \quad \exists x \in E, \quad f(x) = y$$

A Visual Explanation



Here, f is given by the triplet

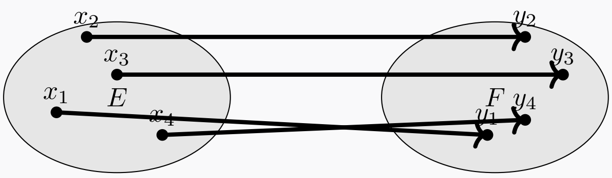
- $E = \{x_1, x_2, x_3, x_4\}$
- $F = \{y_1, y_2\}$
- $G \subset E \times F = \{(x_1, y_1), (x_2, y_2), (x_3, y_2), (x_4, y_1)\}$

Bijective Map

Let E and F be two sets, and let f be a map from E to F. We say that f is bijective if:

$$\forall y \in F, \quad \exists! x \in E, \quad f(x) = y$$

A Visual Explanation



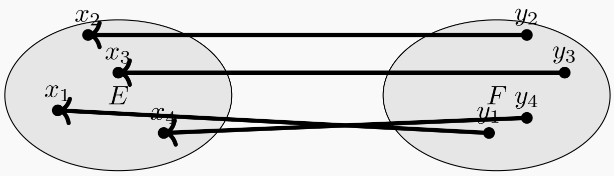
Proposition 20. *Let E, F be two sets, and f be a map from E to F. Then f is bijective if and only if f is injective and surjective.*

Proof. Exercise. □

Definition 28 (Inverse)**.** *Let E, F be two sets, and f be a bijection map from E to F. Then we can construct a map, called the* inverse *of f which to each $y \in F$ associates the unique $x \in E$ such that $f(x) = y$. By definition, we have that:*

$$\forall x \in E, \quad f^{-1}(f(x)) = x$$

A Visual Explanation



Composing Bijections

Proposition 21. *Let E, F and G be three sets, f be a bijection from E to F, and g be a bijection from F to G. Then the composition $g \circ f$ is a bijection from E to G, and we have:*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}$$

Composing Bijections

Proof. Exercise. □

5.2 Exercise Sheet 6: Functions

Familiarizing Exercises

Exercise 37. *In each case, determine whether the map is injective, surjective, bijective.*

1. $E = \mathbb{R}^2$, $F = \mathbb{R}^2$, $f(x, y) = (x^3, y^3)$
2. $E = \mathbb{R}^2$, $F = \mathbb{R}^2$, $f(x, y) = (x + y, x - y)$
3. $E = \mathbb{R}^2$, $F = \mathbb{R}^3$, $f(x, y) = (x + y, x - y, 0)$
4. $E = \mathbb{R}^3$, $F = \mathbb{R}^2$, $f(x, y, z) = (x + y + z, x - y + z)$
5. $E = \mathbb{R}^2$, $F = \mathbb{R}^2$, $f(x, y) = (x + y, x + y)$

Solution.

1. f is bijective.
2. f is bijective.
3. f is injective, but not surjective, and thus not bijective.
4. f is surjective, but not injective, and thus not bijective.
5. f is neither injective, nor surjective, and thus not bijective.

■

Direct Image

Exercise 38. 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by: $\forall x \in \mathbb{R}, f(x) = x^2$.*

   (a) *If $A = [0, 2]$ and $B = [1, 4]$, what are $f(A)$, $f(B)$, $f(A \cap B)$, $f(A \cup B)$, $f(A) \cap f(B)$, and $f(A) \cup f(B)$?*

   (b) *Find two sets A and B for which $f(A \cap B) \neq f(A) \cap f(B)$.*

2. *Let $g : \mathbb{R} \to \mathbb{R}$ be an arbitrary function.*

   (a) *Let A and B be two subsets of $\mathbb{R}$. Show that $g(A \cap B) \subseteq g(A) \cap g(B)$.*

   (b) *What can be generally said about the relationship between $g(A \cup B)$ and $g(A) \cup g(B)$?*

3. *Consider the function*

$$f : [0, 1] \times [0, 1] \to \mathbb{R}, \quad (x, y) \mapsto x + y.$$

   Determine the set $f([0, 1] \times [0, 1])$.

Solution. 1. (a) $f(A) = [0, 4]$, $f(B) = [1, 16]$, $f(A \cap B) = f([1, 2]) = [1, 4]$, $f(A \cup B) = f([0, 4]) = [0, 16]$, $f(A) \cap f(B) = [0, 4] \cap [1, 16] = [1, 4]$, $f(A) \cup f(B) = [0, 4] \cup [1, 16] = [0, 16]$

   (b) $A = [0, 1]$, $B = [-1, 0]$

2. (a) Take $x \in g(A \cap B)$. That means there is an element $y \in A \cap B$ such that $g(y) = x$. But then $y \in A$, so $g(y) \in g(A)$, and $y \in B$ so $g(y) \in g(B)$, hence $x = g(y) \in g(A) \cap g(B)$. Hence $g(A \cap B) \subseteq g(A) \cap g(B)$.
   (b) $g(A \cup B) = g(A) \cup g(B)$, since the places you can get to from either A or B are exactly either $g(A)$ or $g(B)$.
3. $[0, 2]$

■

Inverse Image

Exercise 39. *1. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by: $\forall x \in \mathbb{R}, f(x) = x^2$. If $A = [0, 4]$ and $B = [-1, 1]$, what are $f^{-1}(A)$, $f^{-1}(B)$, $f^{-1}(A \cap B)$, and $f^{-1}(A \cup B)$?*

2. Let $g : \mathbb{R} \to \mathbb{R}$ be an arbitrary function. Let A and B be two subsets of $\mathbb{R}$. Show that $g^{-1}(A \cap B) = g^{-1}(A) \cap g^{-1}(B)$ and $g^{-1}(A \cup B) = g^{-1}(A) \cup g^{-1}(B)$.

Solution. 1. $f^{-1}(A) = [-2, 2]$, $f^{-1}(B) = [-1, 1]$, $f^{-1}(A \cap B) = [-1, 1]$, $f^{-1}(A \cup B) = [-2, 2]$.

$g^{-1}(A \cap B) = g^{-1}(A) \cap g^{-1}(B)$. Let $x \in g^{-1}(A \cap B)$. Then $g(x) \in A \cap B$, so $g(x) \in A$ and $g(x) \in B$. Therefore, $x \in g^{-1}(A) \cap g^{-1}(B)$.
Conversely, if $x \in g^{-1}(A) \cap g^{-1}(B)$, then $g(x) \in A$ and $g(x) \in B$, so $g(x) \in A \cap B$, hence $x \in g^{-1}(A \cap B)$.

$$\boxed{g^{-1}(A \cap B) = g^{-1}(A) \cap g^{-1}(B)}.$$

$g^{-1}(A \cup B) = g^{-1}(A) \cup g^{-1}(B)$. Let $x \in g^{-1}(A \cup B)$. Then $g(x) \in A \cup B$, so either $g(x) \in A$ or $g(x) \in B$. Thus, $x \in g^{-1}(A) \cup g^{-1}(B)$.
Conversely, if $x \in g^{-1}(A) \cup g^{-1}(B)$, then $g(x) \in A \cup B$, so $x \in g^{-1}(A \cup B)$.

$$\boxed{g^{-1}(A \cup B) = g^{-1}(A) \cup g^{-1}(B)}.$$

■

Abstract Manipulations

Exercise 40. *1. Let E and F be two sets, and let $f : E \to F$ be a function. Let A be a subset of E and B a subset of F. Show the following equality:*

$$f(A \cap f^{-1}(B)) = f(A) \cap B.$$

2. Let E and F be two sets and let $f : E \to F$. Show the following equality:

$$E = \bigcup_{y \in F} f^{-1}(y).$$

Solution. **1.** We prove the two inclusions separately. Inclusion $\subseteq$: Let $y \in f(A \cap f^{-1}(B))$. Then there exists $x \in A \cap f^{-1}(B)$ such that $f(x) = y$. Since $x \in A$, we have $y = f(x) \in$

$f(A)$, and since $x \in f^{-1}(B)$, we also have $f(x) \in B$. Hence:

$$y \in f(A) \cap B.$$

This inclusion holds for any function f, without needing injectivity or surjectivity. Convince yourself that we could have $f^{-1}(B) = \emptyset$ is f is not surjective but that the inclusion would still be true.

Inclusion $\supseteq$: Let $y \in f(A) \cap B$. Then:
- $y \in f(A) \Rightarrow \exists x \in A$ such that $f(x) = y$,
- $y \in B \Rightarrow x \in f^{-1}(B)$, so $x \in A \cap f^{-1}(B)$, and hence:

$$y = f(x) \in f(A \cap f^{-1}(B)).$$

2. Let $x \in E$. Then, $y = f(x) \in F$, and by definition: $x \in f^{-1}(y)$. Therefore, $x \in \bigcup_{y \in F} f^{-1}(y)$, so:

$$E \subseteq \bigcup_{y \in F} f^{-1}(y).$$

Conversely, if $x \in \bigcup_{y \in F} f^{-1}(y)$, then there exists $y \in F$ such that $x \in f^{-1}(y)$. This means $f(x) = y$, so $x \in E$. ■

Characteristic Function of a Subset

Exercise 41. *Let E be a set. For each subset A of E, define a function $1_A : E \to \mathbb{R}$ by the formula*

$$\forall x \in E, \quad 1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

1. *Let A be a subset of E. Show the following equality of functions: $1_{A^c} = 1 - 1_A$. Where the set A^c is all the elements of E that are not in A.*
2. *Let A and B be two sets.*
   (a) *Show the following equality of functions: $1_{A \cap B} = 1_A 1_B$*
   (b) *Show the equality $1_{A \cup B} = 1_A + 1_B - 1_A 1_B$.*
3. *Let A and B be two subsets of E. Find and prove a formula for $1_{A \setminus B}$ where the set $A \setminus B$ is A with those elements that are also in B removed.*

Solution. **1. Show that $1_{A^c} = 1 - 1_A$.**
Let $x \in E$. We consider two cases:

- If $x \in A$, then $1_A(x) = 1$, and $x \notin A^c$, so $1_{A^c}(x) = 0$. Hence:

$$1 - 1_A(x) = 1 - 1 = 0 = 1_{A^c}(x).$$

- If $x \notin A$, then $1_A(x) = 0$, and $x \in A^c$, so $1_{A^c}(x) = 1$. Hence:

$$1 - 1_A(x) = 1 - 0 = 1 = 1_{A^c}(x).$$

Therefore, the identity holds for all $x \in E$, so $\boxed{1_{A^c} = 1 - 1_A}$.

2(a). Show that $1_{A\cap B} = 1_A \cdot 1_B$.
Let $x \in E$. We consider the value of both sides:
- If $x \in A \cap B$, then $x \in A$ and $x \in B$, so $1_A(x) = 1$, $1_B(x) = 1$, and:

$$1_A(x) \cdot 1_B(x) = 1 \cdot 1 = 1 = 1_{A\cap B}(x).$$

- If $x \notin A \cap B$, then either $x \notin A$ or $x \notin B$, so at least one of $1_A(x)$, $1_B(x)$ is 0, hence:

$$1_A(x) \cdot 1_B(x) = 0 = 1_{A\cap B}(x).$$

Thus, $\boxed{1_{A\cap B} = 1_A \cdot 1_B}$.

2(b). Show that $1_{A\cup B} = 1_A + 1_B - 1_A \cdot 1_B$.
Let $x \in E$. There are four cases to consider:
- If $x \in A \cap B$: then $1_A(x) = 1$, $1_B(x) = 1$, so:

$$1_A(x) + 1_B(x) - 1_A(x) \cdot 1_B(x) = 1 + 1 - 1 = 1 = 1_{A\cup B}(x).$$

- If $x \in A \setminus B$: then $1_A(x) = 1$, $1_B(x) = 0$, so:

$$1 + 0 - 0 = 1 = 1_{A\cup B}(x).$$

- If $x \in B \setminus A$: then $1_A(x) = 0$, $1_B(x) = 1$, so:

$$0 + 1 - 0 = 1 = 1_{A\cup B}(x).$$

- If $x \notin A \cup B$: then $1_A(x) = 0$, $1_B(x) = 0$, so:

$$0 + 0 - 0 = 0 = 1_{A\cup B}(x).$$

Therefore, $\boxed{1_{A\cup B} = 1_A + 1_B - 1_A \cdot 1_B}$.

3. Find and prove a formula for $1_{A\setminus B}$.
Note that:

$$A \setminus B = A \cap B^c,$$

so by part (2a):

$$1_{A\setminus B} = 1_A \cdot 1_{B^c}.$$

Now applying part (1), we have $1_{B^c} = 1 - 1_B$, so:

$$1_{A\setminus B} = 1_A(1 - 1_B) = 1_A - 1_A \cdot 1_B.$$

Hence, the final identity is:

$$\boxed{1_{A\setminus B} = 1_A - 1_A \cdot 1_B}.$$

■

Exercise 42. *Show that if $f : E \to F$ is a bijection, then we have:*

$$f(f^{-1}(y)) = y, \quad \forall y \in F$$

Solution. Since f is a bijection, it is in particular:

- **Surjective**: for every $y \in F$, there exists $x \in E$ such that $f(x) = y$.
- **Injective**: for every $x_1, x_2 \in E$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.

Let $y \in F$. Because f is surjective, there exists at least one $x \in E$ such that $f(x) = y$. Define $f^{-1}(y) := x$. This makes sense because f is a bijection, so f^{-1} is well-defined and assigns a unique pre-image to each $y \in F$.
Then,

$$f(f^{-1}(y)) = f(x) = y,$$

as required.

Conclusion: For all $y \in F$, $f(f^{-1}(y)) = y$, so:

$$\boxed{f \circ f^{-1} = \mathrm{id}_F}.$$

■

Exercise 43. *Prove the propositions from the lecture, that a function is bijective if and only if it is injective and surjective, and that the composition of two bijections is a bijection.*

Solution. TO DO ■

5.3 Neuro Q2: Representational Geometry & Dot Products

You live in a bizarre dystopian future where aliens are studying your nervous system. In pursuit of understanding they are only allowing you to see a sequence of four badly drawn smiley faces (the aliens haven't mastered drawing human forms).

Previous evidence (from fMRI studies with ENORMOUS magnets) has led the aliens to think that two of your neurons encode the smiley faces, i.e. there is a function, $g : F \to \mathbb{R}^2$, that maps the set of four smiley faces, F, to the two neurons' activities. They begin the process of preparing you for surgery to examine how your two neurons do this. In your desperation you quickly shout out "Don't hurt me! I'm a theoretical neuroscientist! I already know what you'll find in these neurons - there's no need to do the experiments!"
They pause, bemused, as their babel fish translators lather complex chemical patterns into their taste based communication systems. "(Burnt chicken with grape overtones) (Metallic tang of gory fangs) (Bamboo after rainfall)" they reply, which can be poorly translated as "Though we doubt what you are excreting, tasting you does sound more pallatable than getting your disgusting brain juices all over us again. Go on."
You begin tentatively...

A reasonable model of these higher level visual neurons is that they are encoding information about these faces in a way that allows other neurons to decode the information they need to. For example, your system for emotional control must be able to extract whether the faces are happy or sad, and your finely tuned nasal analysing cortex has to know if the faces have noses. Let's imagine that you model your emotional extraction process as a neuron somewhere in the amygdala that swallows your visual neuron activity pattern, $g(f)$ for a face f, making it a function, $h : \mathbb{R}^2 \to \mathbb{R}$, defined by:

$$h(g(f)) = \phi(w^T g(f)))$$

For some weight vector, $w \in \mathbb{R}^2$, and a nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$. This nonlinearity is very simple:

$$\phi(y) = \begin{cases} 1 & \text{if} \quad y \geq 0 \\ 0 & \text{if} \quad y < 0 \end{cases}$$

For the neuron to readout emotion from the population it must activate ($h = 1$) for the happy faces and deactivate for the unhappy.

1. The encodings of the faces are points in 2D space, $\{g(f)\}_{f \in F}$. A readout neuron splits these inputs into two groups, those that it activates for, and those that it doesn't. Choose some w and draw and describe these two spaces, i.e. an $h(g(f))$.

2. How must the representation be structured (i.e. what must $g(f)$ be) to allow emotion to be extracted using this neuron? Draw two representations, one that would work and one that wouldn't.

3. Now you have another readout neuron that is supposed to decode whether the face has a nose or not. Draw a representation that permits both features to be readout at once.

The aliens are not satisfied, you have given them many options, but none of them match what they know about neurons: neurons never have negative firing rates! They sharpen their knives. "Wait!" You shout.

4. You quickly think about what would happen if the representations are only allowed to be positive, (i.e. $g : F \to \mathbb{R}^2_+$, where $\mathbb{R}^2_+$ is the positive quadrant of the plane) show that in the simple model described above you can't read out both features from a positive-only representation.

5. You propose a quick change:

   $h = \phi(w^T g + b)$

   With a new bias parameter $b \in \mathbb{R}$. Describe how this changes the classifications the readout neuron is able to learn.

6. Show how this allows you, for some choices of representation, to readout both emotion and nosed-ness from four positive representations at the same time.

7. Show that there are still representations from which emotion and nosedness cannot be decoded.

8. Explain how the same effect as a bias could have been achieved if you were allowed a third neuron to create your representation. How must that third neuron behave? (i.e. what is its response pattern as a function of faces)

The aliens are still unimpressed with your wishey-washey-ness. When your model is questioned you just add a new parameter, not the most enlightening behaviour. You need to impress them - you start thinking harder.
You begin to wonder about two perenniel neural concerns, energy costs and noisiness.

9. Let's say your neurons are noisy. Rather than always firing with rates $g(f)$ they tend to fire with a rate $\text{ReLU}(g(f)+\eta)$ where $\eta_i \sim \text{UNIFORM}[-1, 1]\sigma$ ($\text{ReLU}(x) = \max(x, 0)$, just to keep the neural activities positive). How should the structure of the representation be changed so that the readout neurons will still be able to extract emotion and nosedness? What is a sufficient condition on the representation such that noise will not effect the emotional and nosedness outputs?

10. But, we would also like to save energy, which in this case we'll take to mean the amount of neuron firing, i.e. the cost of the representation is $\mathcal{C}(g) = \sum_{f \in F} ||g(f)||_2$. We could push all the firing rates to zero which would then cost zero to represent, but then we wouldn't be able to readout what we need to, and they can't be negative as we discussed earlier. Given that we want to perfectly be able to readout emotion and nosedness despite noise, what is the structure of the lowest energy representation?

You should have ended with a very surprising thing: each of the neurons in the representation should encode either emotion or nosedness! This is known as a disentangled representation, different pieces of information are stored in different neurons, and in fact the above derivation can be formalised and generalised to many neurons and many variables under suitable assumptions. The aliens are impressed and slightly angry, why are they looking for readout neurons elsewhere in the brain when the information is already perfectly represented here in these two neurons?! They release you, asking for your help in the future, and go look for the authors of the original paper that misled them, knives still drawn... These aliens take science quite seriously...

> *Solution.* Hello
>
> ■

Chapter 6

Linear Maps & Matrices

6.1 Linear Maps

Motivation

Now the full game begins! We turn to functions of vectors, but we focus on a specific type of function that is of paramount importance: linear functions.
Let's follow the natural idea that we can understand something if we can break it into its parts, analyze the behavior of the parts, then recombine them to understand the whole. For linear functions this procedure works nicely: the function applied to a sum of two inputs is the sum of the function applied to each of the inputs. We can therefore study how the function behaves on some set of atoms, and from this reconstruct the whole behavior!

Definition 29 (Linear maps)**.** *Let $n \in \mathbb{N}^*$. A mapping $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ is called linear if the following holds:*

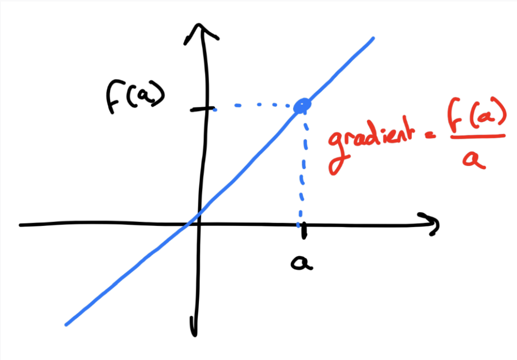
- $f(x + y) = f(x) + f(y)$, *for all* $x, y \in \mathbb{R}^n$;
- $f(\lambda x) = \lambda f(x)$, *for all* $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$.

1D Intuition

Let's think about the simplest linear function $f : \mathbb{R} \to \mathbb{R}$. In order to be linear:

$$f(\lambda a) = \lambda f(a) \quad \forall \lambda, a \in \mathbb{R}$$

Using $\lambda = 0$ we can see that $f(0) = 0$. Further, as soon as you know $f(a)$ for one number a, the equation above lets you find all others by multiplying by the right λ. Plotting $(x, f(x))$ we see that this is a straight line through the origin with gradient $\frac{f(a)}{a}$:



Definition 30 (Endomorphisms). *Let $n \in \mathbb{N}^*$. A linear map T from $\mathbb{R}^n$ to itself is called an endomorphism. The space of all endomorphisms is denoted $\mathcal{L}(\mathbb{R}^n)$.*

Definition 31 (Isomorphisms). *Let $f \in \mathcal{L}(\mathbb{R}^n)$. If f is a bijection, then f is called an isomorphism.*

Proposition 22. *If f is an isomorphism, then f^{-1} is an isomorphism.*

Proof. Let $a, b \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, then if there exist $y, z \in \mathbb{R}^n$ such that $f(x) = a$ and $f(y) = b$. We then have $f(x) + \lambda f(y) = f(x + \lambda y)$, that is to say, applying f^{-1} on both sides, $f^{-1}(a + \lambda b) = x + \lambda y = f^{-1}(a) + \lambda f^{-1}(b)$. □

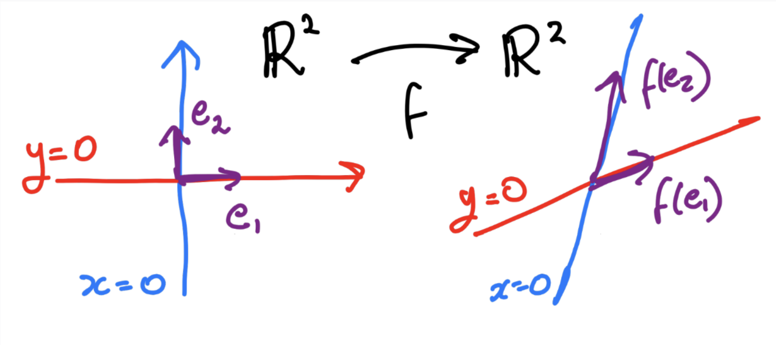
So, when it exists, the inverse of a linear function is also a linear function!

2D Intuition

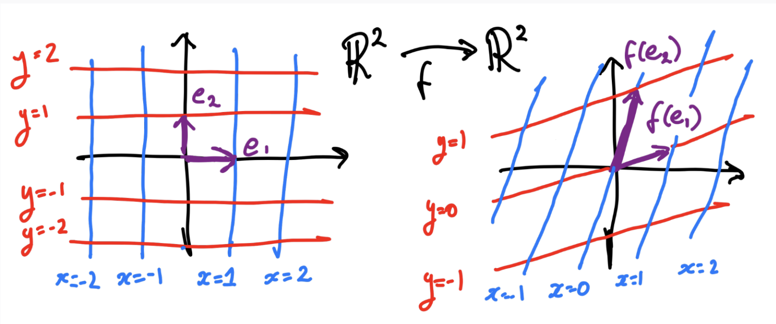
Let's think about a more complex linear function, $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x)$. Motivated by linearity we might try decomposing the function's behavior; a natural choice would be to decompose the vector in a basis (for example the canonical basis): $x = x_1 e_1 + x_2 e_2$, $x_1, x_2 \in \mathbb{R}$, $e_1, e_2 \in \mathbb{R}^2$. Then:

$$f(x) = f(x_1 e_1 + x_2 e_2) = x_1 f(e_1) + x_2 f(e_2)$$

Here $f(e_1)$ and $f(e_2)$ are two vectors in the output space. To understand this, consider input points with $x_2 = 0$: the x-axis in the input space. These are mapped to $x_1 f(e_1)$, i.e. another line in the output space, but pointing in the $f(e_1)$ direction! Similarly, the y-axis is mapped to the line $f(e_2)$.



Now what if both x_1 and x_2 are nonzero? Consider all the points with $x_1 = 1$, a vertical line in the input space. This gets mapped to $f(e_1) + x_2 f(e_2)$, i.e. another line in the output space, offset from the origin by $f(e_1)$, pointing along $f(e_2)$. Now we can generalize this. Any gridline we draw in the input space (i.e. a set of points with either constant x_1 or x_2) get's mapped to a new line in the output space, shifted by some amount of either $f(e_1)$ or $f(e_2)$, and pointing along the other vector.



This logic generalises beyond 2D; indeed, you can always think of a linear map as mapping grid lines in the input space to grid lines in the output space, from the decomposition:

$$f : \mathbb{R}^n \to \mathbb{R}^m \qquad f(x) = f(\sum_{i=1}^{n} x_i e_i) = \sum_{i=1}^{n} x_i f(e_i)$$

6.2 Matrices & Linear Maps

We now build a link between linear functions and matrices.

Matrix-Vector product is a Linear Function

Proposition 23. *The function $f : \mathbb{R}^n \to \mathbb{R}^m$, $n, m \in \mathbb{N}^*$, $f(x) = Mx$ for matrix $M \in \mathbb{R}^{m \times n}$ is linear.*

Proof. Recall the definition of matrix-vector product:

$$(Mx)_i = \sum_j M_{ij} x_j \tag{6.1}$$

We can see that this is a linear function:

$$(M(x+\lambda y))_i = \sum_j M_{ij}(x_j + \lambda y_j) = \sum_j M_{ij} x_j + \lambda \sum_j M_{ij} y_j = (Mx)_i + \lambda (My)_i \tag{6.2}$$

Hence, $f(x + \lambda y) = f(x) + \lambda f(y)$. □

Column-View of Matrix-Vector Product

We saw that matrix-vector product is a linear function. We now link it to our 'grid-lines to grid-lines' view of linear functions. We do this using the column vectors of $M \in \mathbb{R}^{m \times n}$, $c_i \in \mathbb{R}^m$:

$$M = \begin{bmatrix} | & | & \dots & | \\ c_1 & c_2 & \dots & c_n \\ | & | & \dots & | \end{bmatrix} \qquad Mx = \begin{bmatrix} | & | & \dots & | \\ c_1 & c_2 & \dots & c_n \\ | & | & \dots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i c_i$$

In other words, matrix-vector product is just a linear combination of the columns of the matrix M, and the weightings are given by the elements of x!!!
Further, we can make one more link. We saw that a linear function can be decomposed according to the canonical basis:

$$f(x) = \sum_{i=1}^{n} x_i f(e_i)$$

Let's choose to express these vectors $f(e_i)$ in the canonical basis, $[f(e_i)]_B = c_i$. Then we can see that matrix-vector product performs exactly the same operation as f when we express the inputs and outputs in the canonical basis! The axes of the new grid lines align with the columns of the matrix, M. This means we can instantly draw the behavior of the matrix-vector product in the same way we drew the behavior of linear functions.

This view shows us the link going the other way:

Linear Map can be enacted by Matrix-vector Product

Proposition 24. *For every linear function $f : \mathbb{R}^n \to \mathbb{R}^m$, $n, m \in \mathbb{N}^*$, we can create the matrix $M \in \mathbb{R}^{m \times n}$ defined as:*

$$M = \begin{bmatrix} | & | & \dots & | \\ [f(e_1)]_B & [f(e_2)]_B & \dots & [f(e_n)]_B \\ | & | & \dots & | \end{bmatrix}$$

where $\{e_i\}_{i=1}^n$ is the canonical basis, B. Then $[f(x)]_B = M[x]_B$.

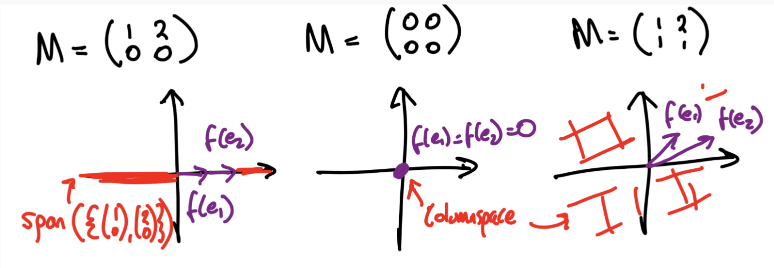
Given a linear map the associated matrix is Basis Dependent!!!

Remark 1. Important point: A linear map takes vectors in the input space and maps them to vectors in the output space. These quantities are independent of basis, and so a linear map is defined independently of a basis. However, as shown above, a matrix corresponds to a particular choice of basis. If we change the basis in either the input or output space we change the matrix. We will return to this point. For now we'll assume we are using the canonical basis.

Having established this link, life becomes simpler. Many things become easier to show using linear maps than using matrices. We shall see this repeatedly in the following sections. The elegant proofs presented are much more laborious in matrix-vector notation, though they express the same ideas.

6.3 Image, Kernel, & Rank

We saw that the output of Mx was a linear combination of the columns of M. Different choices of x lead to different linear combinations of the columns. It seems useful to think of the notion of column space, the span of the columns of the matrix. Mx is in the column space for all x, and at least one x reaches every point in the column space.



We will now formalize these notions in powerful ways.

Definition 32 (Image of a set)**.** *Let E and F be two sets and f a function between them, and let $H \subset E$. Then the set:*

$$\{f(x) : x \in H\}$$

is called the image of H by f, and is denoted by $f(H)$.

Definition 33 (Preimage of a set)**.** *Let E and F be two sets, and let $G \subset F$. Then the set:*

$$\{x : f(x) \in G\}$$

is called the pre-image of G by f, and is denoted by $f^{-1}(G)$.

Proposition 25 (The image of a subspace is a subspace)**.** *Let $n, m \in \mathbb{N}^*$, and H be a subspace of $\mathbb{R}^n$. Let f be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $f(H)$ is a subspace of $\mathbb{R}^m$.*

Proof. Let $u, v \in f(H)$, $\lambda \in \mathbb{R}$. By definition of $f(H)$, there exist $x, y \in H$ such that $f(x) = u$ and $f(y) = v$. Since H is a subspace of $\mathbb{R}^n$, $x + \lambda y \in H$. Using the linearity of f, we have $f(x + \lambda y) = f(x) + \lambda f(y) = u + \lambda v$. Therefore, $u + \lambda v \in f(H)$. Thus, $f(H)$ is a subspace of $\mathbb{R}^m$. $\square$

Proposition 26 (The preimage of a subspace is a subspace)**.** *Let $n, m \in \mathbb{N}^*$, and let G be a subspace of $\mathbb{R}^m$. Then $f^{-1}(G)$ is a subspace of $\mathbb{R}^n$.*

Proof. We have $f^{-1}(G) = \{x \in \mathbb{R}^n \mid f(x) \in G\}$. Let $x, y \in f^{-1}(G)$ then $f(x + \lambda y) = f(x) + \lambda f(y) \in G$ because G is a subspace, and thus $x + \lambda y \in f^{-1}(G)$. $\square$

Definition 34 (Image of a linear map)**.** *Let $n, m \in \mathbb{N}^*$, and let f be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then we call $f(\mathbb{R}^n)$ the image of f, and denote it $Im(f)$.*

Spanning set of Im(f)

Proposition 27. *Let $n, m \in \mathbb{N}^*$, and $B = \{e_1, \dots, e_n\}$ be a basis for $\mathbb{R}^n$. Then we have that:*

$$span\{f(e_1), \dots, f(e_n)\} = Im(f)$$

In other terms, $\{f(e_1), \dots, f(e_n)\}$ is a spanning set for $Im(f)$.

Rank of a linear map

Definition 35 (Rank)**.** *Let f be a linear function, $f : \mathbb{R}^n \to \mathbb{R}^m$. The **rank** of f is the integer defined as the dimension of its image:*

$$rank(f) := \dim(Im(f))$$

Definition 36 (Kernel of a linear map)**.** *Let $n, m \in \mathbb{N}^*$, and let f be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then we call $f^{-1}(\{0\})$ the kernel of f, and denote it $ker(f)$.*

Why do we care about characterizing kernels

A large amount of problems in mathematics can be framed as problems of the form:

$$\text{find } x : f(x) = y$$

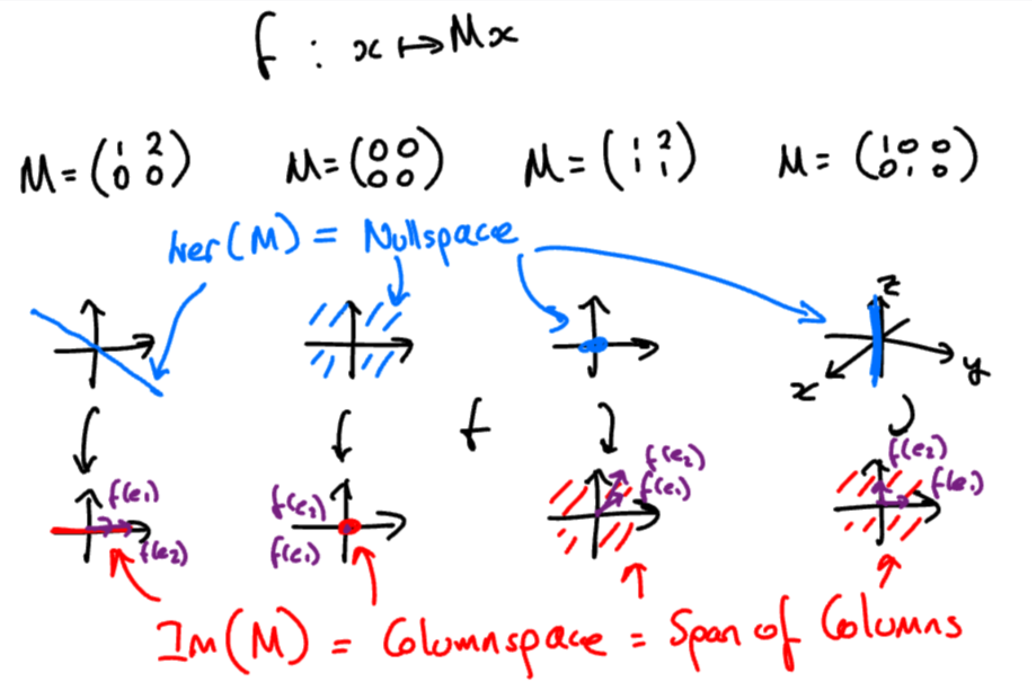
The elements of the kernel are those we can add to the input without changing the output:

$$f(x) = y \qquad x_k \in ker(f) \qquad f(x + x_k) = f(x) + f(x_k) = f(x) = y$$

So, given the constraint $f(x) = y$, we can only find x up to degeneracy given by the $ker(f)$.

Columnspace & Nullspace

Given our link between linear maps and the corresponding matrix applied to a vector in the canonical basis, there are more matrix-centric versions of the image and kernel. As discussed before, the image of $f(x) = Mx$ is the span of the columns of f, while the kernel is often called the nullspace. Some examples are visualized below.



We define this formally below.

Kernel & Image of Matrix

Definition 37. *Given a matrix $A \in \mathbb{R}^{n \times m}$, we define its kernel (nullspace) and image (column space) as:*

1. $ker(A) = \{x \in \mathbb{R}^n | Ax = 0\}$
2. $im(A) = \{y \in \mathbb{R}^m | \exists x \in \mathbb{R}^n s.t. Ax = y\}$

6.4 Invertibility & Rank-Nullity Theorem

Given the discussion above, the following facts should not be surprising.

Proposition 28. *Let $n, m \in \mathbb{N}^*$, $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ be a linear map, and $\{e_1, \dots, e_n\}$ be a basis for $\mathbb{R}^n$. Then we have:*

1. *f is injective if and only if $\ker(f) = \{0\}$.*
2. *f is injective if and only if $\{f(e_1), \dots f(e_n)\}$ is linearly independent*
3. *f is surjective if and only if $\{f(e_1), \dots, f(e_n)\}$ is a spanning set for $\mathbb{R}^m$.*
4. *f is bijective if and only if $\{f(e_1), \dots, f(e_n)\}$ is a basis for $\mathbb{R}^m$.*

Proof.

1. **f is injective if and only if $\ker(f) = \{0\}$:** ($\Rightarrow$) If f is injective, then $f(x) = 0 = f(0)$ implies $x = 0$. Thus, the only element mapped to 0 is 0 itself, so $\ker(f) = \{0\}$.

   ($\Leftarrow$) If $\ker(f) = \{0\}$, suppose $f(x) = f(y)$. Then $f(x - y) = f(x) - f(y) = 0$, so $x - y \in \ker(f)$. Hence, $x - y = 0$, i.e., $x = y$, and f is injective.

2. **f is injective if and only if $\{f(e_1), \dots, f(e_n)\}$ is linearly independent:**

   Since $\{e_1, \dots, e_n\}$ is a basis for $\mathbb{R}^n$, any vector $x \in \mathbb{R}^n$ can be uniquely written as a linear combination $x = \sum_{i=1}^n \lambda_i e_i$. By linearity,

   $$f(x) = \sum_{i=1}^n \lambda_i f(e_i).$$

   So the image of x is a linear combination of the vectors $f(e_i)$. Now:

   ($\Rightarrow$) Suppose f is injective. Consider a linear combination $\sum_{i=1}^n \lambda_i f(e_i) = 0$. Then

   $$f\left(\sum_{i=1}^n \lambda_i e_i\right) = 0,$$

   so the vector $x := \sum_{i=1}^n \lambda_i e_i$ lies in the kernel of f. Since f is injective, $\ker(f) = \{0\}$, hence $x = 0$, which implies that all $\lambda_i = 0$. Thus, the family $\{f(e_1), \dots, f(e_n)\}$ is linearly independent.

   ($\Leftarrow$) Conversely, suppose that $\{f(e_1), \dots, f(e_n)\}$ is linearly independent. Suppose $f(x) = 0$ for some $x \in \mathbb{R}^n$. Write $x = \sum_{i=1}^n \lambda_i e_i$. Then

   $$0 = f(x) = \sum_{i=1}^n \lambda_i f(e_i),$$

   so a linear combination of $f(e_i)$ gives zero. Since they are linearly independent, it must be that all $\lambda_i = 0$, so $x = 0$, and therefore $\ker(f) = \{0\}$. By part (1), this implies f is injective.

3. **f is surjective if and only if $\{f(e_1), \dots, f(e_n)\}$ spans $\mathbb{R}^m$:**

   Every $x \in \mathbb{R}^n$ can be written as $x = \sum_{i=1}^n \lambda_i e_i$, and then

   $$f(x) = \sum_{i=1}^n \lambda_i f(e_i).$$

Thus, the image of f is the span of $\{f(e_1), \ldots, f(e_n)\}$:

$$\mathrm{Im}(f) = \mathrm{span}(f(e_1), \ldots, f(e_n)).$$

Therefore, f is surjective (i.e., $\mathrm{Im}(f) = \mathbb{R}^m$) if and only if this family spans $\mathbb{R}^m$.

4. **f is bijective if and only if $\{f(e_1), \ldots, f(e_n)\}$ is a basis of $\mathbb{R}^m$:**

   f is bijective if and only if it is both injective and surjective. By parts (2) and (3), this holds if and only if $\{f(e_1), \ldots, f(e_n)\}$ is both linearly independent and spanning in $\mathbb{R}^m$, which means it is a basis for $\mathbb{R}^m$.

$\square$

Proposition 29 (Rank-nullity theorem for isomorphisms)**.** *Let $\mathbb{R}^n$ and $\mathbb{R}^m$ be two vector spaces and $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ be an isomorphism. Then $n = m$.*

Proof. Let $B = \{e_1, \ldots, e_n\}$ be a basis for $\mathbb{R}^n$ by definition of the dimension. By Proposition 28, f is bijective meaning that $\{f(e_1), \ldots, f(e_n)\}$ is a basis for $\mathbb{R}^m$. By definition of the dimension, $n = \dim(\mathbb{R}^n) = \dim(\mathbb{R}^m) = m$. $\square$

Rank–Nullity Theorem

Below, we will prove perhaps the most important theorem of linear algebra, the rank-nullity theorem. It links the (dimension of the) image and the kernel of a linear map.

Theorem 3 (Rank-Nullity Theorem)**.** *Let $n, m \in \mathbb{N}^*$, and $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ a linear map. Then we have:*

$$rank(f) + dim(ker(f)) = n$$

Using the machinery that we have developed so far, proving this beautiful identity is actually surprisingly simple! To do so, we first show the following intermediary result

Proposition 30. *Let $n, m \in \mathbb{N}^*$, and $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ be a linear map. Let H be a complementary subspace of $ker(f)$ in $\mathbb{R}^n$. Define g as*

$$g : \begin{cases} H & \longmapsto \mathrm{Im}(f) \\ x & \longmapsto f(x) \end{cases}$$

Then g is an isomorphism.

Proof. To prove this result, it suffices to make sure that g verifies the definition of an isomorphism, e.g., it is a surjection and an injection.

- g is surjective. Let $y \in \mathrm{Im}(f)$, then there exists some $x \in \mathbb{R}^n$ such that $y = f(x)$. Moreover, since H is a complementary subspace of $\ker(f)$ in $\mathbb{R}^n$, there exists a unique pair $x_H \in H$, $x_k \in \ker(f)$ such that $x = x_H + x_k$. Then we have:

$$f(x_H) = f(x_H) + 0 = f(x_H) + f(x_k) = f(x_H + x_k) = f(x) = y$$

Thus, we have found some x_H in H such that $f(x_H) = y$, and g is surjective.

- g is injective. Let $x_1, x_2 \in H$ such that $f(x_1) = f(x_2)$, then we have that $f(x_1 - x_2) = 0$, and thus $x_1 - x_2 \in \ker(f)$. Because x_1 and x_2 belong (by definition) to H, $x_1 - x_2 \in H$, and $x_1 - x_2 \in H \cap \ker(f)$, thus, $x_1 - x_2 = 0$, as H and $\ker(f)$ are complementary.

$\square$

Proof of Theorem 3. Since g is an isomorphism from H to $\mathrm{Im}(f)$, we have that $\dim(\mathrm{Im}(f)) = \dim(H)$. Because $H \oplus \ker(f) = \mathbb{R}^n$, we have that $\dim H = n - \dim(\ker(E))$ (from our results on complementary vector spaces). Combining these two equalities, we obtain the desired result! $\square$

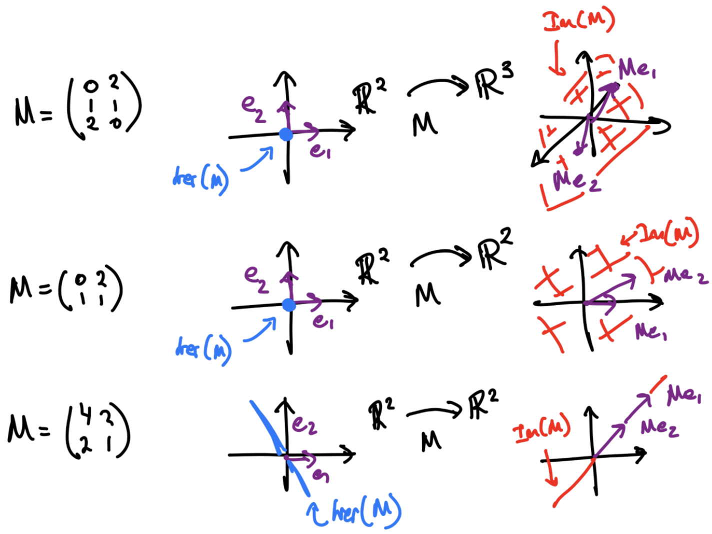
From this theorem, we derive this very nice corollary:

Corollary 2. *Let $n \in \mathbb{N}^*$, and let $f \in \mathcal{L}(\mathbb{R}^n)$. Then we have:*

$$f \textit{injective} \iff f \textit{surjective} \iff f \textit{bijective}$$

Invertibility of a Matrix

Using our link between linear maps and a matrix acting on a vector in the canonical basis, all these statements about bijectivity also apply similarly to matrices.
Have a think about which of the matrices below are invertible and why



Invertibility of Matrices

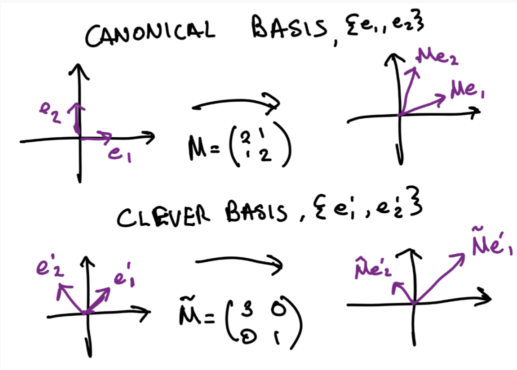
Proposition 31. *A matrix, $M \in \mathbb{R}^{n\times m}$ is a linear map $f : \mathbb{R}^n \to \mathbb{R}^m$, $f : x \to Mx$. Therefore, it is invertible if and only if:*

1. *$n = m$, i.e. the matrix is square*
2. *$rank(M) = n$, i.e. the matrix is 'full-rank'*

6.5 Changing Basis & Similar Matrices

Basis Dependence of Matrices

We saw before that the matrix representation of a linear map depends on the basis you choose:



Now we will make this dependence explicit, and consider the matrix representation of a linear map in an arbitrary basis, rather than the canonical one.

Definition 38 (Matrix representation of a linear map)**.** *Let $n, m \in \mathbb{N}^*$. Let $B := \{e_1, \dots, e_n\}$ (resp. $B' := \{f_1, \dots, f_m\}$), be a basis for $\mathbb{R}^n$ (resp. for $\mathbb{R}^m$). Let T be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the matrix of T in the bases B, B', denoted $Mat_{B,B'}(T)$ is defined as:*

$$Mat_B^{B'}(T)_{ij} := \begin{bmatrix} [T(e_1)]_{B'} & \dots & [T(e_n)]_{B'} \end{bmatrix}$$

When $n = m$ and $B = B'$, we will denote $M_B^B(T) = M_B(T)$ to simplify the notation.

Proposition 32. *Let $n, m \in \mathbb{N}^*$, and let T be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let B, B' be bases of $\mathbb{R}^n, \mathbb{R}^m$ respectively. Let $x \in \mathbb{R}^n$. Then we have: Then we have, for all $x \in \mathbb{R}^n$*

$$[T(x)]_{B'} = Mat_B^{B'}(T)[x]_B$$

Proof. let $B = \{e_1, \ldots, e_n\}$ and $B' = \{f_1, \ldots, f_m\}$ be bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $x \in \mathbb{R}^n$. Then we have:

$$\begin{aligned} T(x) &= T\left(\sum_{i=1}^{n}([x]_B)_i e_i\right) \\ &= \sum_{i=1}^{n}([x]_B)_i T(e_i) \\ &= \sum_{i=1}^{n}([x]_B)_i \sum_{j=1}^{m}([T(e_i)]_{B'})_j f_j \\ &= \sum_{j=1}^{m}\left(\sum_{i=1}^{n}([T(e_i)]_{B'})_j([x]_B)_i\right) f_j \quad = \sum_{j=1}^{m}(\mathrm{Mat}_B^{B'}(T)[x]_B)_j f_j \end{aligned}$$

By definition of coordinates, we thus have $[T(x)]_{B'} = \mathrm{Mat}_B^{B'}(T)[x]_B$. □

Impact of the basis on the matrix (visually explained)

Let $T \in \mathcal{L}(\mathbb{R}^2)$ represented in the canonical basis by the matrix

$$\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}$$

Then, show that

$$\mathrm{Mat}_B(T) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

for $B = \{(1,1),(1,0)\}$.
Thus, while T seems to operate on the vectors x in a complex manner when using the canonical basis, T simply scales the coordinates of x in the basis B by 2 and 1 respectively! The simplicity of T's action in the basis B is well captured visually by its actions on isolines in different basis:

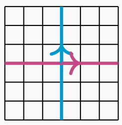


Figure 6.1: Isolines of $\mathbb{R}^2$ for the canonical basis $\{(1,0),(0,1)\}$

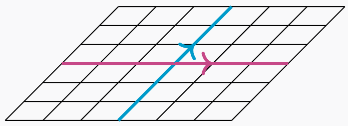


Figure 6.2: Action of T on the canonical basis isolines: T sheafes the isolines in a complex manner

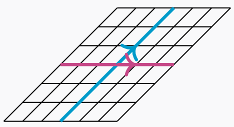


Figure 6.3: Isolines of $\mathbb{R}^2$ for the basis $\{(1,1),(0,1)\}$

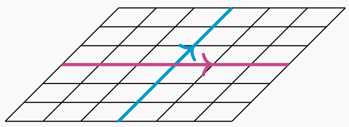


Figure 6.4: Action of T on the isolines of B: T only dilates the first isolines by a factor of 2, and leaves the second isolines unchanged

Operations on Map Representations

Furthering the analogy between linear maps and matrices, the following statement holds:

Proposition 33 (Composition as matrix multiplication)**.** *Let $n, m, p > 0$, consider B, B', B'', bases of $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$ respectively. Let $T : \mathbb{R}^n \longmapsto \mathbb{R}^m$ and $U : \mathbb{R}^m \longmapsto \mathbb{R}^p$ be two linear maps. Then we have that*

$$Mat_B^{B''}(U \circ T) = Mat_{B'}^{B''}(U) Mat_B^{B'}(T)$$

Proof. Let $B = \{e_1, \ldots, e_n\}$, $B' = \{f_1, \ldots, f_m\}$, and $B'' = \{g_1, \ldots, g_p\}$. Then by definition of matrix representations, we have that

$$\begin{aligned}
U \circ T(e_i) &= U\left(\sum_{j=1}^{m} \mathrm{Mat}_B^{B'}(T)_{ji} f_j\right) \\
&= \sum_{j=1}^{m} \mathrm{Mat}_B^{B'}(T)_{ji} U(f_j) \\
&= \sum_{j=1}^{m} \mathrm{Mat}_B^{B'}(T)_{ji} \sum_{k=1}^{p} \mathrm{Mat}_{B'}^{B''}(U)_{kj} g_k \\
&= \sum_{k=1}^{p} \left(\sum_{j=1}^{m} \mathrm{Mat}_B^{B'}(T)_{ji} \mathrm{Mat}_{B'}^{B''}(U)_{kj}\right) g_k \\
&= \sum_{k=1}^{p} (\mathrm{Mat}_{B'}^{B''}(U)\mathrm{Mat}_B^{B'}(T))_{ki}) g_k
\end{aligned}$$

□

Proposition 34 (Representation of change of basis matrices)**.** *Let $n > 0$ be a finite-dimensional vector space, and*

$$\begin{aligned}
id : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\
x &\longmapsto id(x) = x
\end{aligned}$$

Let B, B' be two bases of $\mathbb{R}^n$. Then we have:

$$P_B^{B'} = Mat_B^{B'}(id)$$

Change of Basis Rule, Equivalent Matrices, Similar Matrices

The set of tools developed in the previous section allows us to easily prove the following formula that relates the representations of a linear map T for two different choice of pairs of input-output bases:

Proposition 35 (Change of basis formula)**.** *Let $n, m \geq 0$. Let B_n, B'_n (resp. B_m, B'_m) be two bases for $\mathbb{R}^n$ (resp. $\mathbb{R}^m$), and let $T : \mathbb{R}^n \longmapsto \mathbb{R}^m$ be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then we have that:*

$$Mat(T)_{B'_n}^{B'_m} = P_{B'_m}^{B_m} Mat(T)_{B_n}^{B_m} P_{B_n}^{B'_n}$$

Proof. First note that by the representation of change of bases given by Proposition 34, the proposition is equivalent to:

$$\mathrm{Mat}(T)_{B'_n}^{B'_m} = \mathrm{Mat}(id_{\mathbb{R}^m})_{B_m}^{B'_m} \mathrm{Mat}(T)_{B_n}^{B_m} \mathrm{Mat}(id_{\mathbb{R}^n})_{B'_n}^{B_n}.$$

We have that

$$T = id_{\mathbb{R}^m} \circ T \circ id_{\mathbb{R}^n}.$$

By representing

- $id_{\mathbb{R}^n}$ using B'_n as the input basis and B_n as the output basis
- T using B_n as the input basis and B_m as the output basis
- $id_{\mathbb{R}^m}$ using B_m as the input basis and B'_m as the output basis

from the chain rule for matrix representations given by Proposition 15, we have that:

$$\mathrm{Mat}(T)_{B'_n}^{B'_m} = \mathrm{Mat}(id_{\mathbb{R}^m})_{B_m}^{B'_m} \mathrm{Mat}(T)_{B_n}^{B_m} \mathrm{Mat}(id_{\mathbb{R}^n})_{B'_n}^{B_n},$$

proving the desired result. □

Equivalent Matrices

Definition 39. *Two matrices $M, N \in \mathbb{R}^{n\times m}$ are said to be equivalent if there exist two invertible matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$ such that*

$$N = BMA.$$

Similar Matrices

Definition 40. *Two matrices $M, N \in M_n(\mathbb{R})$ are said to be similar if there exists an invertible matrix $P \in \mathbb{R}^{n\times n}$ such that*

$$N = P^{-1}MP.$$

If two matrices are similar, then they represent the same endomorphism in different bases.

6.6 Exercise Sheet 7: Maps and Matrices

6.6.1 Linear Functions

Exercise 44. *Let $n, m \in \mathbb{N}^*$. Show that a map $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ is linear if and only if*

$$f(x + \lambda y) = f(x) + \lambda f(y), \quad \forall x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}.$$

Solution. It is easy to show that if f is linear this statement holds: by the first property $f(x + \lambda y) = f(x) + f(\lambda y)$, and by the second $f(x) + f(\lambda y) = f(x) + \lambda f(y)$.
Let's then do the other direction, showing that if this statement holds, f is linear. Put in $\lambda = 1$ and we get $f(x + y) = f(x) + f(y)$, i.e. the first linear property. Then choosing $x = 0$ we get $f(\lambda y) = \lambda f(y)$, i.e. the second linear property. Hence, f is linear. ■

Exercise 45. *Which of these functions are linear?*

1. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x$
2. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$
3. $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = x_1 + x_2$, *where* $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
4. $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = x_1 x_2$, *where* $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Solution. TO DO ■

Exercise 46. *Farmer Aethelred's runs three shops selling apples, 'Farmer Aethelred's sublinear, linear and supralinear apples'^{TM}. In each shop the price of your basket of x apples is determined via the price function, $price(x)$ with the following properties:*

- *Sublinear:* $price(\lambda x) < \lambda price(x)$
- *Linear:* $price(\lambda x) = \lambda price(x)$
- *Supralinear:* $price(\lambda x) > \lambda price(x)$

1. *If I want to buy three apples what is the best tactic for buying from each shop to minimize their price?*
2. *After King Wigmore's decision to use apples as currency, Farmer Aethelred decides to act as an exchange, and will now buy apples at the same price as they sell them (It's early days, and this currency exchange has not worked out how to charge a fee). How should you buy from one shop and sell to another to make guaranteed profit??? Describe a tactic for buying and selling from the same shop that also generates guaranteed profit.*

Solution. TO DO ■

6.6.2 Kernel & Image

Exercise 47. *Show that the kernel of a linear map, $ker(f)$, is a subspace.*

Solution. TO DO ■

Exercise 48. *Describe the images and kernels of the following linear maps:*

1. $(x, y) \to (x, 0)$
2. $(x, y) \to (x + y, x + y)$
3. $(x, y) \to (x + y, x - y)$
4. $(x, y) \to (0, 0)$

Solution. TO DO ■

Exercise 49. *Given $A \in \mathbb{R}^{n\times m}$ show that the following 3 matrices all have the same rank:*

$$\begin{bmatrix} A \end{bmatrix} \qquad \begin{bmatrix} A \\ A \end{bmatrix} \qquad \begin{bmatrix} A & A \\ A & A \end{bmatrix} \tag{6.3}$$

Solution. TO DO ■

Exercise 50. *I tell you that $A \in \mathbb{R}^{n\times m}$ is rank r, and that the equation $Ax = b$ has no solutions for some choices of b (i.e. for some b there are no x that verify $Ax = b$). What inequalities must be true relating n, m, r.*

Solution. TO DO ■

Exercise 51. *Show that the rank, kernel, and image, of A and $-A$ are the same.*

Solution. TO DO ■

Exercise 52. *Suppose $A = uv^T + wz^T$ for vectors $u, w \in \mathbb{R}^n$, $v, z \in \mathbb{R}^m$. Which vectors span the column space of A, what about the column space of A^T. When is the rank less than 2?*

Solution. TO DO ■

Exercise 53. *If you were given $a, b, c \in \mathbb{R}$ all non-zero, how would you choose d such that the following matrix is rank 1:*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Solution. TO DO ■

Exercise 54 (Harder: Row Space Theorem). *For a matrix* $A \in \mathbb{R}^{n \times m}$*,*

1. *Show that* $Ker(A^T) = Im(A)^\perp$
2. *Similarly show that* $Im(A^T) = Ker(A)^\perp$
3. *Use these results and the rank-nullity theorem to show that* $rank(A) = rank(A^T)$
4. *Given two matrices* A *and* B *we construct* $C = AB$*. Using these results, show that the rank of* C *is at most* $\min(rank(A), rank(B))$

Solution. TO DO ∎

6.6.3 Invertibility & Rank-Nullity

Exercise 55. *Let* $A, B \in \mathcal{L}(\mathbb{R}^n)$*.*

1. *Show that* $Ax = Bx$ *for all* $x \in \mathbb{R}^n$ *implies* $A = B$*.*
2. *Use this (or anything else) to show that the inverse is unique.*

Solution. TO DO ∎

Exercise 56. *Given* $A, B \in \mathcal{L}(\mathbb{R}^n)$ *you would need to show both* $AB = I$ *and* $BA = I$ *to show* $A = A^{-1}$*. Show that* $AB = I_n \implies BA = I_n$*, therefore you only need to check one condition.*

Solution. **Step 1: A is surjective.**
From $AB = I_n$, we know that for any $x \in \mathbb{R}^n$,

$$ABx = x.$$

Setting $y = Bx$, we have $Ay = x$, so A maps some vector $y \in \mathbb{R}^n$ to every $x \in \mathbb{R}^n$. Thus, A is surjective.

Step 2: Surjectivity implies bijectivity.
Since $A \in \mathcal{L}(\mathbb{R}^n)$ by the rank-nullity theorem, it is also injective. Hence, A is bijective, and thus invertible.

Step 3: Conclude $BA = I$.
Because A is invertible, we can multiply both sides of $AB = I$ on the left by A^{-1}, yielding:

$$B = A^{-1} \quad \Rightarrow \quad BA = I_n.$$

Conclusion: If $AB = I_n$, then $BA = I_n$. Therefore, to show that A and B are inverses of each other, it is sufficient to check just one of the identities.

∎

Exercise 57. *Show that we always have* $\text{Ker}(B) \subseteq \text{Ker}(AB)$ *for two matrices* A *and* B*.*

Solution. TO DO ∎

Exercise 58. *We will show the following property step by step: Let $A \in \mathcal{M}_{n,n}(\mathbb{R})$ and $B \in \mathcal{M}_{n,p}(\mathbb{R})$, if A is invertible, then $\operatorname{rank}(AB) = \operatorname{rank}(B)$.*

1. *Show that if A is invertible then $\operatorname{Ker}(AB) \subseteq \operatorname{Ker}(B)$;*
2. *Conclude using the previous exercise that if A is invertible then $\operatorname{Ker}(AB) = \operatorname{Ker}(B)$;*
3. *Conclude that if A is invertible then $\operatorname{rank}(AB) = \operatorname{rank}(B)$. Hint: use the rank-nullity theorem.*

Solution. TO DO ■

Exercise 59. *Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$ such that $AB = I_n$.*

1. *Show that we necessarily have $n \leq m$. Hint: show that $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$, then show that $\operatorname{rank}(B) \leq \min\{n, m\}$ using Lemma* **??**, *conclude;*
2. *Find an example of such matrices;*
3. *Show that if $n \neq p$, we can never have $BA = I_p$.*

Solution. TO DO ■

Sherman-Morrison Formula

Exercise 60. *We're going to derive a handy result. First, consider this weird matrix $A \in \mathbb{R}^{N\times N}$, $A = \alpha\mathbf{1}\mathbf{1}^T + I$, where I is the identity matrix and $\mathbf{1}$ is a vector of ones.*

1. *This is actually not such a crazy matrix. Imagine you had a vector x, and wanted to remove the mean of the elements from each element. This gives a vector with elements $\bar{x}_i = x_i - \frac{1}{N}\sum_i x_i$. Convince yourself that performing this operation in matrix notation can be written as $\bar{x} = Ax$ for some value of α, and find the appropriate value of α.*
2. *Now demeaning is actually not invertible (many datasets that differ only by the mean are mapped to the same place!) But in this question we're going to be interested in inverses, which for this matrix will be matrices that add some kind of constant. Surely this can be expressed in the same form: $A^{-1} = \gamma\mathbf{1}\mathbf{1}^T + \delta I$, i.e. just add a constant. Show this guess is correct for $A = \alpha\mathbf{1}\mathbf{1}^T + I$ and derive expressions for γ and δ in terms of α.*
3. *We're going to consider a slight generalization to this, now $A = I + uv^T$, you could think of this as doing some kind of weighted demeaning operation using weights v, but only along with a different weighting on the output, u (I'm not sure that's the best interpretation). Anyway, guess a form for it's inverse in the style above and find it.*
4. *Now consider another generalization, $A = B + uv^T$, where B is an invertible matrix. Use your answer to the previous question to derive its inverse in terms of B.*

This is a useful expression, known as the Sherman-Morrison formula! Computing inverses can be very slow, especially if the matrices are very big. Let's pretend you invested all that effort and calculated B^{-1}, then it the matrix gets updated, but only by a small update uv^T.

This equation gives you a way to find the inverse of A using simple matrix multiplications and scaling, without having to do a full inverse calculation, nice!

Solution. TO DO ■

6.6.4 Change of Basis

Familiarisation Exercise

Exercise 61. *Compute $Mat_B^{B'}(T)$ for the following cases:*

- *Case 1: $E = F = \mathbb{R}^2$, and let $T : \mathbb{R}^2 \longmapsto \mathbb{R}^2$ such that $T([1,0]) = [1,2]$ and $T([0,1]) = [2,1]$, and*
  - *Case 1.1: $B = \{[1,0],[0,1]\}$, and $B' = \{[1,0],[0,1]\}$*
  - *Case 1.2: $B = \{[1,0],[0,1]\}$, and $B' = \{[1,2],[2,1]\}$*
- *Case 2: $E = \mathbb{R}^3$, $F = \mathbb{R}^2$, and $T : \mathbb{R}^3 \longmapsto \mathbb{R}^2$ such that $T([1,0,0]) = [1,2]$, $T([0,1,0]) = [2,1]$, and $T([0,0,1]) = [1,1]$.*
  - *Case 2.1: $B = \{[1,0,0],[0,1,0],[0,0,1]\}$, and $B' = \{[1,0],[0,1]\}$*
  - *Case 2.2: $B = \{[1,0,0],[0,1,0],[0,0,1]\}$, and $B' = \{[1,2],[2,1]\}$*

Solution. TO DO ■

6.6.5 Three Difficult Exercises

Exercise 62. *Consider $u \in \mathcal{L}(\mathbb{R}^n)$. For some $x \in \mathbb{R}^n$, we note $u^0(x) = x$, and $u^{k+1}(x) = u(u^k(x))$. We say that u is cyclic if there exists some $x_0 \in E$ such that*

$$span\left\{u^k(x_0); k \in \mathbb{N}\right\} = \mathbb{R}^n$$

Show that there exists a basis B of $\mathbb{R}^n$ such that:

$$Mat_{B_E}(u) = \begin{bmatrix} 0 & 0 & \dots & 0 & a_1 \\ 1 & 0 & \dots & 0 & a_2 \\ 0 & 1 & \dots & 0 & a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & a_n \end{bmatrix}$$

for some $a_1, \dots, a_n \in \mathbb{R}$. Such a matrix is called a companion matrix.

Solution. Since u is cyclic, we know that

$$\text{span}\left\{u^k(x_0) \mid k \in \mathbb{N}\right\} = \mathbb{R}^n.$$

Define the subspace

$$W := \text{span}\{x_0, u(x_0), \dots, u^{n-1}(x_0)\}.$$

Since $W \subseteq \mathbb{R}^n$, to prove that $W = \mathbb{R}^n$, it suffices to show that $\dim(W) = n$.

Let m be the smallest non-negative integer such that the set $\{x_0, u(x_0), \dots, u^m(x_0)\}$ is linearly dependent. Such an m must exist because if all $u^k(x_0)$ were linearly independent, we would obtain an infinite linearly independent set in $\mathbb{R}^n$, contradicting the finite dimension of the space.
So there exist scalars $c_0, \dots, c_m$, not all zero, such that:

$$c_0 x_0 + c_1 u(x_0) + \cdots + c_m u^m(x_0) = 0.$$

Since m is the smallest such integer, the vectors $\{x_0, u(x_0), \dots, u^{m-1}(x_0)\}$ are linearly independent. Moreover, since $c_m \neq 0$, we can write:

$$u^m(x_0) = -\frac{1}{c_m}(c_0 x_0 + c_1 u(x_0) + \cdots + c_{m-1} u^{m-1}(x_0)).$$

Now apply u to both sides:

$$u^{m+1}(x_0) = u\left(u^m(x_0)\right) = -\frac{1}{c_m}(c_0 u(x_0) + c_1 u^2(x_0) + \cdots + c_{m-1} u^m(x_0)).$$

So $u^{m+1}(x_0) \in \operatorname{span}\{x_0, \dots, u^m(x_0)\}$. By induction, this implies:

$$\forall k \geq m, \quad u^k(x_0) \in \operatorname{span}\{x_0, \dots, u^{m-1}(x_0)\}.$$

Therefore,

$$\operatorname{span}\{u^k(x_0) \mid k \in \mathbb{N}\} = \operatorname{span}\{x_0, \dots, u^{m-1}(x_0)\}.$$

But this span equals $\mathbb{R}^n$, so:

$$\mathbb{R}^n = \operatorname{span}\{x_0, u(x_0), \dots, u^{m-1}(x_0)\}.$$

Since these vectors are linearly independent and span $\mathbb{R}^n$, they form a basis. Hence, $m = n$.

In this basis $B = \{x_0, x_1, \dots, x_{n-1}\}$ with $x_k := u^k(x_0)$, we compute how u acts:

$$u(x_0) = x_1, \quad u(x_1) = x_2, \quad \dots, \quad u(x_{n-2}) = x_{n-1}.$$

And since $x_n = u(x_{n-1}) \in \operatorname{span}\{x_0, \dots, x_{n-1}\}$, write:

$$x_n = a_1 x_0 + a_2 x_1 + \cdots + a_n x_{n-1}.$$

Therefore, in basis B, the matrix of u is:

$$\operatorname{Mat}_B(u) = \begin{bmatrix} 0 & 0 & \dots & 0 & a_1 \\ 1 & 0 & \dots & 0 & a_2 \\ 0 & 1 & \dots & 0 & a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & a_n \end{bmatrix},$$

■

Exercise 63. *By showing that for the correct choice of B and A in Definition 39 all matrices can be transformed into a diagonal matrix with only zeros and ones on the diagonal, show that all matrices of the same rank are equivalent.*

Solution. TO DO ■

Exercise 64. *$f \in \mathcal{L}(\mathbb{R}^3)$ and $f \circ f = f^2 = 0$ (where 0 means the function that maps everything to zero), while $f \neq 0$. Show that the rank of f is 1. Write a matrix that represents such an f.*

Solution. **Step 1: Consequence of $f^2 = 0$**
The condition $f^2 = 0$ means:

$$\forall x \in \mathbb{R}^3, \quad f(f(x)) = 0.$$

That is, the image of f is contained in the kernel of f:

$$\mathrm{Im}(f) \subseteq \ker(f).$$

Step 2: Apply the Rank–Nullity Theorem
We know that:

$$3 = \dim(\mathbb{R}^3) = \mathrm{rank}(f) + \dim(\ker(f)).$$

Also from Step 1:

$$\mathrm{rank}(f) = \dim(\mathrm{Im}(f)) \leq \dim(\ker(f)).$$

Let $r = \mathrm{rank}(f)$, so:

$$3 = r + \dim(\ker(f)) \geq r + r = 2r \quad \Rightarrow \quad 2r \leq 3 \quad \Rightarrow \quad r \leq 1.5.$$

Since the rank is an integer, we conclude $r \leq 1$.

Step 3: Use $f \neq 0$
We are told that $f \neq 0$, so its rank cannot be 0. Hence:

$$\mathrm{rank}(f) = 1.$$

Step 4: Structure of rank 1 linear maps
We now prove that any linear map $f : \mathbb{R}^3 \to \mathbb{R}^3$ of rank 1 must be of the form $f = vu^\top$ for some nonzero vectors $u, v \in \mathbb{R}^3$.
Since $\mathrm{rank}(f) = 1$, the image of f is a one-dimensional subspace of $\mathbb{R}^3$, i.e., $\mathrm{Im}(f) = \mathrm{span}(v)$ for some nonzero vector $v \in \mathbb{R}^3$.
Hence, there exists a linear form $\phi : \mathbb{R}^3 \to \mathbb{R}$ such that:

$$f(x) = \phi(x) \cdot v.$$

As ϕ is linear,

$$\phi(x) = u^\top x,$$

for some $u \in \mathbb{R}^3$ (convince yourself). Therefore:

$$f(x) = (u^\top x)v = vu^\top x = (vu^\top)(x).$$

That is, $f = vu^\top$.

Conclusion: Any rank 1 linear map $f \in \mathcal{L}(\mathbb{R}^3)$ can be written as:

$$\boxed{f = vu^\top \quad \text{for some } u, v \in \mathbb{R}^3 \setminus \{0\}}.$$

Step 4: Use the condition $f^2 = 0$
We compute:

$$f^2 = (vu^\top)(vu^\top) = v(u^\top v)u^\top.$$

So:

$$f^2 = 0 \quad \Leftrightarrow \quad u^\top v = 0.$$

That is, the vectors u and v must be orthogonal.

Conclusion:
All such linear maps f with $f \neq 0$ and $f^2 = 0$ must be of the form:

$$f = vu^\top \quad \text{with } v, u \in \mathbb{R}^3 \setminus \{0\}, \text{ and } u^\top v = 0.$$

■

6.7 Neuro Q3: Nullspace and Motor Cortex

The motor cortex, especially in primates, appears to be involved in directly controlling your body, especially dexterous voluntary movements (On the other hand, in mice its hard to spot the difference between control mice and those who've had their motor cortex removed... leading people to suppose cortex is mainly there in mice to keep the basal ganglia warm). Classic models of motor cortex describe the neural activity and its effect on muscles as a linear dynamical system. This just means that both the neural activity at timepoint $t+1$ and the signal sent to the muscles at each time are a linear (really affine) function of the activity of timepoint t. We'll brush over the dynamics within motor cortex (though you'll cover it in your dynamical systems class), and focus instead on the linear map that relates neural activity to motor output. Call the neural activity at each timepoint $g_t \in \mathbb{R}^N$, where N is the number of neurons, and call the signal it sends the muscles at each timepoint $s_t \in \mathbb{R}^M$ where M is the dimensionality of your muscle control space (an example signal your motor cortex might send your muscles would be 'flex your deltoid you slovenely rascal!'). Then, the (common) linear assumption is that:

$$s_t = Rg_t \quad R \in \mathbb{R}^{M \times N} \tag{6.4}$$

1. You have many more neurons (N) than muscles (M). What can you infer about the dimensionality of the kernel of R?

2. Let's simplify, you're a small lunar dwelling pogo stick that subsists on regolith, buried ice, and the abundant supply of American flags. You have two neurons in your motor cortex ($N = 2$) that together control your pogo stick activating muscle ($M = 1$). Let's choose to work in a basis in which $R = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Now describe the kernel again?

3. You're lounging around with your pogostick buddies when you're surprised to see your friend's motor cortex activating, even though their pogo is perfectly still. Given that you know equation 6.4 is true, what must be happening in their motor cortex? (Describe the neural activity up to an unknown time dependent scalar)

4. These kinds of ponderings about real neural systems have led people to break motor cortical neural activity into output-potent and output-null subspaces. Neural activity in the output-potent subspace appears to drive muscle behaviour, conversely, neural activity in the output-null space has no effect on current muscle activities. Relate these terms (output-potent, output-null) to the kernel of R.

Why do animals have an output-null space? When you are preparing to perform an action your motor cortex activates in stereotypical ways. For example, if you're preparing to extend or contract your pogo stick, the neural activity will move to two different points in the output-null space. When you then initiate the action the neural activity spins off into the output-potent space with a trajectory that is determined by its starting point in the output-null space.

6.8 Neuro Q4: Invertibility & Trichromatic Theory of Colour

Colour matching experiments, first done in the early 19th century, demonstrated a suprising phenomenon. You are shown a beam of coloured light. You are given three other beams of differently coloured light to play with. It turns out you can mix together your three beams with appropriate intensities to recreate the original colour, for many choices of the colours of your three beams. Further, this is not possible with only 2 beams (for most people), and there is only one mixture that works for each demonstrated colour. This is very good news for digital screen makers, since now they can make a pixel with three colours (RGB) and, by changing their intensities, they can generate most colours. This is also very surprising - why is this true? The resolution of this surprise is the trichromatic theory of colour, first developed by Thomas Young in 1802.

A light's colour is its spectral content, i.e. how much of each wavelength is present in the light. You are only able to see wavelengths in the visible range, between 400nm and 700nm, therefore let's write the spectral content of some incident light as a vector $I \in \mathbb{R}^{300}$: I_i encodes the intensity of the incident light between $400 + i$ and $400 + i + 1$ nanometres. You can see things, therefore you must contain a receptor that activates when there is light. How it responds will vary depending on the wavelength of light (most simply, you don't respond to light beyond the visible range). Denote with $t \in \mathbb{R}^{300}$ the tuning of a receptor, how much it responds to one unit of intensity at each wavelength. Finally, assume the response of a receptor to broadband light is a linear function of the light intensity: $r = t^T I$.

1. Use linear algebra and invertibility to explain why the colour matching experiments are evidence for the trichromatic theory of colour? (This exercise in linear algebra was first done by James Clark Maxwell in the 1850s, who then used these ideas to develop the first ever colour photograph - go linear algebra!)

2. Metamers are objects that are different, but perceptually the same. Call the three receptor tunings t_1, t_2 and t_3, and stack them into a matrix T:

$$T = \begin{bmatrix} - & t_1^T & - \\ - & t_2^T & - \\ - & t_3^T & - \end{bmatrix}$$

   Given two metamer's spectral contents, I_1 and I_2, derive an expression relating T, I_1, and I_2. What space does $I_1 - I_2$ live in?

3. The colour of an object depends both on its reflectance properties, and the spectrum of the illuminating light. Explain why two colours might look the same in bright sunlight, but different under a halogen light?

4. The spatial denisty of your cone cells is very high in the fovea, but much lower in the rest of the retina. Use this to explain metameric failures: colours that appear the same when viewed in the periphery, but not when centrally fixated. Why does the same not happen when going from fovea to periphery?

5. I made a mistake, I am actually a honey-bee with 3 cones, two like humans, and one in the ultraviolet. Let's pretend their response properties don't overlap (i.e. the UV cone responds only to light $>$ 700nm, the other two to light $<$ 700nm). How would you test this with colour matching and what would the result be if I was or wasn't a bee.

6. If we had more cones we would be able to see more colours. Think biologically, why might we not have more cones?

7. At dawn and dusk both your rods and your cones play an important role in vision, how could this change the experimental results?

8. Computer screens effectively do the colour matching experiment on you. However, it is not possible to emit negative amounts of light at a particular wavelength. Explain why this means that our RGB system does not explore the full range of colours.

9. If our receptors were nonlinear would the original experiments still suggest we have three cones?

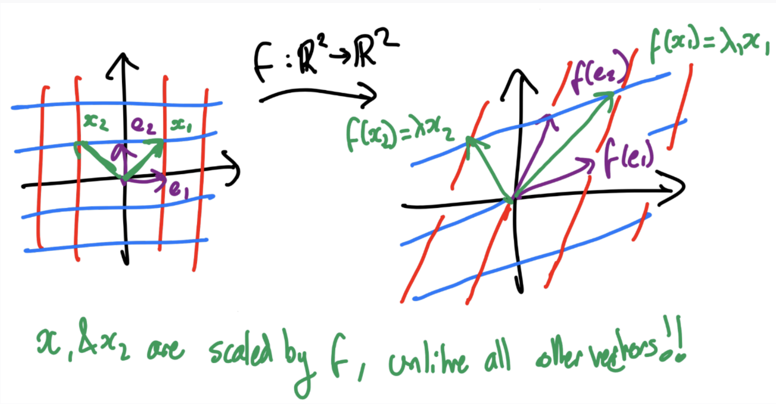
The last question feels like evidence that linear encoding has other advantages. You might be able to make some guesses about what these advantages are when you've finished the linear algebra content!

Chapter 7

Eigendecomposition

Motivation

In the previous section we saw that, sometimes, when we change the basis the matrix became very simple. Sometimes we could even make it diagonal. This section explores describes the ideas that perform exactly this diagonalisation.



7.1 Eigenvectors, Eigenvalues, Spectrums, Eigenspaces

First Definitions

Definition 41. *let $u \in \mathcal{L}(\mathbb{R}^n)$, and $\lambda \in \mathbb{R}$.*

1. *If there exists some $\exists x \neq 0$ s.t. $u(x) = \lambda x$, we say that λ is an* ***eigenvalue*** *of u, and x is an* ***eigenvector****.*

2. *Given an eigenvalue λ, the set of all eigenvectors associated with λ, e.g $\{x \in \mathbb{R}^n,\ u(x) = \lambda x\}$, is called the* ***eigenspace*** *associated with λ, denoted E_λ.*

3. *The (possibly empty) set $\{\lambda_1, \dots, \lambda_k\}$ of eigenvalues of u is called the* ***spectrum*** *of u, and is denoted $sp(u)$.*

Casually Explained

Eigenvectors are directions in which the endomorphism u simply scales the vector by a factor λ, without changing its direction. Eigenvectors uncover sometimes hidden simplicity in what may otherwise look like complicated mappings, as we see in the exercise sheet.

7.1.1 Properties of Eigenvalues and Eigenvectors

Eigenvectors make life much easier. Now we try and show that if we can find enough eigenvectors we can perform exactly the diagonalisation we hinted towards at the beginning.

Proposition 36. *Let $\lambda \in \mathbb{R}$ be an eigenvalue of some $u \in \mathcal{L}(\mathbb{R}^n)$. Then its corresponding eigenspace, denoted E_λ, is a vector subspace of $\mathbb{R}^n$.*

Eigenspaces are *stable*

Stability of a subspace by an endomorphism

Definition 42. *Let $u \in \mathcal{L}(\mathbb{R}^n)$ and F be subspaces of $\mathbb{R}^n$. We say that F is stable by u if $u(F) \subset F$.*

Stability of eigenspaces

Proposition 37. *Let $u \in \mathcal{L}(\mathbb{R}^n)$. Let $\lambda \in \mathbb{R}$ be an eigenvalue of u. Then the eigenspace associated with λ is stable under u, i.e. $u(E_\lambda) \subset E_\lambda$.*

Proof. Let $x \in E_\lambda$. Then $u(x) = \lambda x \in E_\lambda$. Thus, $u(x) \in E_\lambda$, and E_λ is stable under u. □

Eigenvectors are linearly independent

Proposition 38. *Let $u \in \mathcal{L}(\mathbb{R}^n)$. Let $\{e_1, \dots, e_k\}$ a family of k eigenvectors, associated with their respective eigenvalues $\lambda_1, \dots, \lambda_k$. Assume that $\lambda_i \neq \lambda_j$ for $i \neq j$. Then $\{e_1, \dots, e_k\}$ is linearly independent.*

Proof

Proof. By induction on $k \in \mathbb{N}$.

1. If $k = 1$, then $\{e_1\}$ is linearly independent.
2. Assume the result holds for some $k \geq 1$. Let us prove that it also holds for $k+1$. Let $\{e_1, \dots, e_{k+1}\}$ be a family of considered eigenvectors. Let $\mu_1, \dots, \mu_{k+1} \in \mathbb{R}$ such that
$$\sum_{i=1}^{k+1} \mu_i e_i = 0$$

applying u to both sides of the last equality, we get:

$$u\left(\sum_{i=1}^{k+1}\mu_i e_i\right) = \sum_{i=1}^{k+1}\lambda_i\mu_i e_i = 0$$

Subtracting λ_{k+1} times the first equality we get:

$$0 = \sum_{i=1}^{k}(\lambda_i - \lambda_{k+1})\mu_i e_i = 0$$

Which implies $\mu_i = 0$ for all $i \leq k$ since $\lambda_i \neq \lambda_{k+1}$ for all $0 \leq i \leq k$ and $\{e_1, \dots, e_k\}$ is linearly independent. Consequently:

$$\mu_{k+1}e_{k+1} = 0$$

meaning that $\mu_{k+1} = 0$, and $\{e_1, \dots, e_{k+1}\}$ is linearly independent.

□

This proposition shows us how eigenvectors with different eigenvalues are linearly independent. What if we have repeated eigenvalues? (Multiple linearly independent eigenvectors with the same eigenvalue). We show that we can create a basis for each of these eigenspaces and combine them to create a big set of linearly independent vectors in the same way we can combine distinct eigenvectors.

Proposition 39. *Let $u \in \mathcal{L}(\mathbb{R}^n)$, and let $B_i = \{e_{i1}, \dots e_{in_i}\}$ a basis of E_{λ_i} (of dimension n_i) where $(\lambda_i)_{1\leq i\leq k}$ are the eigenvalues of u. Then the vectors $\{e_{11}, \dots, e_{1n_1}, \dots, e_{k1}, \dots, e_{kn_k}\}$ are linearly independent.*

Proof

Proof. Let $\mu_{11}, \dots, \mu_{kn_k} \in \mathbb{R}$ such that: $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\mu_{ij}e_{ij} = 0$. Since

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}e_{ij} = \sum_{i=1}^{k} 1 \times (\sum_{j=1}^{n_i}\mu_{ij}e_{ij}) = 0$$

For all i, $\sum\limits_{j=1}^{n_i}\mu_{ij}e_{ij} \in E_{\lambda_i}$, and non-0 eigenvectors are linearly independent. Therefore, we must have that $\sum\limits_{j=1}^{n_i}\mu_{ij}e_{ij} = 0$ for all i. Since $\{e_{i1}, \dots, e_{in_i}\}$ are linearly independent, we must have that $\mu_{ij} = 0$ for all i, j. □

The two last results show that it is possible to create linearly independent families of vectors by concatenating bases of eigenspaces. It is natural to ask: can this family span the whole of $\mathbb{R}^n$? In other words, can it be a basis of $\mathbb{R}^n$? This question gives rises to the notion of diagonalizability.

7.2 Diagonalizability

Diagonalizability

Definition 43. *We say that some $u \in \mathcal{L}(\mathbb{R}^n)$ is* ***diagonalizable*** *if there exists a basis B of $\mathbb{R}^n$ comprised only of eigenvectors of u.*

Diagonalizable endomorphisms are represented by diagonal matrices

This is a very important remark: from the definition of diagonalizablity, we have that if some $u \in \mathcal{L}(\mathbb{R}^n)$ is diagonalizable, we have by definition:

$$\mathrm{Mat}_B^B(u) = D := \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

where $\lambda_1, \dots, \lambda_n$ are the respective eigenvalues of $e_1, \dots, e_n \in B$. This equality holds from the fact that $u(e_i) = \lambda_i e_i$, e.g. $[u(e_i)]_B = (\lambda_i \delta_{ij})_{1 \leq j \leq n}$. Consequently, for the matrix M such that $u(x) = Mx$ (e.g. M represents u in the canonical basis), we have:

$$M = PDP^{-1}$$

for P the change of basis matrix from B to the canonical basis. The last equality is the traditional matrix definition of diagonalizability.

How diagonalizable endomorphism act on inputs

For some diagonalizable $u \in \mathcal{L}(\mathbb{R}^n)$, let $B = \{e_1, \dots, e_n\}$ its associated basis of eigenvectors, and let $\{\lambda_1, \dots, \lambda_n\}$ be their respective eigenvalues. Let $x = \sum_{i=1}^{n} \mu_i e_i$. Then we have the important formula:

$$u(x) = u\left(\sum_{i=1}^{n} \mu_i e_i\right) = \sum_{i=1}^{n} \lambda_i \mu_i e_i$$

7.2.1 Sufficient Condition for Diagonalization

In general finding whether a matrix is diagonalisable can be a challenge. We give one simple sufficient condition here.

Proposition 40. *Let $u \in \mathcal{L}(\mathbb{R}^n)$. Assume that u admits n distinct eigenvalues $\lambda_1, \dots, \lambda_n$ (e.g., $i \neq j \implies \lambda_i \neq \lambda_j$). Then u is diagonalizable.*

Proof. If $\lambda_1, \dots, \lambda_n$ are distinct, then the eigenvectors associated with these eigenvalues are linearly independent, meaning that we can create a basis of $\mathbb{R}^n$ by concatenating these eigenvectors, and that u is diagonalizable. $\square$

7.3 Symmetric Endomorphisms and the Spectral Theorem

We conclude this section on eigendecomposition by showing the diagonalizability of a very large class of matrices: symmetric matrices. Not only are these endomorphisms diagonalizable, but they are diagonalizable in an orthogonal basis, which is a very strong property. This result is so pervasive it has its own name: the Spectral Theorem. We start by recalling the definition of a symmetric matrix.

Definition 44. *A matrix $A \in \mathcal{L}(\mathbb{R}^n)$ is said to be symmetric if $A = A^T$.*

There exists an alternative functional definition of symmetric matrices, which is useful for proving the spectral theorem:

Proposition 41. *Let $u : x \longmapsto Ax \in \mathcal{L}(\mathbb{R}^n)$. Then we have that A is symmetric if and only if:*

$$\langle u(x), y\rangle = \langle x, u(y)\rangle$$

for all $x, y \in \mathbb{R}^n$. We say that u is a symmetric endomorphism

Proof.

- $\Longrightarrow$: Let $x, y \in \mathbb{R}^n$. The $\langle x, Ay\rangle = x^T Ay = (Ax)^T y = \langle Ax, y\rangle$.
- $\Longleftarrow$: Recall that $A = \mathrm{Mat}_B^B(u)$ where B is the canonical basis of $\mathbb{R}^n$. Then we have that: $A_{ij} = [u(e_j)]_{B,i} = \langle u(e_j), e_i\rangle = \langle e_j, u(e_i)\rangle = [u(e_i)]_{B,j} = A_{ji}$.

□

To prove the spectral theorem, we need the following lemma, which is admitted, but can be proved using calculus results.

Lemma 2. *Let $u \in \mathcal{L}(\mathbb{R}^n)$ be a symmetric endomorphism. Then u admits at least one real eigenvalue.*

From this result, it is possible to show the spectral theorem, which states that any symmetric endomorphism is diagonalizable in an orthonormal basis.

Lemma 3. *Let $u \in \mathcal{L}(\mathbb{R}^n)$ be a symmetric endomorphism. Then u is diagonalizable in an orthonormal basis of eigenvectors.*

Proof. The result is proven by induction on the dimension n of $u's$ domain.

- In the case $n = 1$, we have that $u(x) = x \times u(1)$ as x is a scalar.

  Taking $x/|x|$ as the basis vector, we have that u is diagonal in this basis, and that basis is orthonormal.
- Assume now that every symmetric endomorphism in dimension $n-1$ is diagonalizable in an orthonormal basis. Now $u \in \mathcal{L}(\mathbb{R}^n)$, from the lemma above, we have that u

admits at least one real eigenvalue λ_1. Let e_1 be an associated eigenvector, which we take of length 1. Let $E_1 = \text{span}(e_1)$. Then it turns out that $E_1^\perp$ is stable under u: for all $x \in E_1^\perp$, $\langle u(x), e_1 \rangle = \langle x, u(e_1) \rangle = \langle x, \lambda_1 e_1 \rangle = \lambda_1 \langle x, e_1 \rangle = 0$, and thus all $u(x) \in E_1^\perp$. Consequently, $E_1^\perp$ is stable under u.

- We can now define
$$\begin{aligned} v : E_1^\perp &\longrightarrow E_1^\perp \\ x &\longmapsto u(x) \end{aligned}$$
the restriction of u to $E_1^\perp$. Since $\dim(E_1^\perp) = n - 1$, we can use the induction hypothesis to get that v is diagonalizable in an orthonormal basis $B = \{e_2, \dots, e_n\}$. Since for all $u(e_i) = v(e_i) = \lambda_i e_i$, for all $i \geq 2$, we have that these vectors are also eigenvectors of u. To obtain an orthonormal basis of $\mathbb{R}^n$ comprised of eigenvectors of u, we can concatenate e_1 with B (we leave as a exercise the proof that this concatenation is indeed an orthonormal basis)

□

The spectral theorem admits an equivalent matrix version, which is very important and should be remembered:

Corollary 3. *Let $A \in \mathcal{L}(\mathbb{R}^n)$ be a symmetric matrix. Then there exists an orthogonal matrix P and a diagonal matrix D such that:*

$$A = PDP^T$$

e.g. A is similar to a diagonal matrix.

We will define properly what "orthogonal matrix" means and prove the corollary in the next chapter.

7.3.1 Positive Definite Matrices

Before closing we consider an interesting class of symmetric matrices.

Positive (Semi-)Definite Matrix

Definition 45. *A symmetric matrix, $A \in \mathbb{R}^{n \times n}$, is positive semidefinite if $\forall x \in \mathbb{R}^n$, $x^T A x \geq 0$. It is positive definite if the inequality is strict.*

The exercise sheet develops a few interesting properties of this class of matrix.

Dot Product Matrices

The exercise sheet tells us that we can see any positive semi-definite matrix as a matrix whose elements are the dot-products of a set of vectors, the columns of B, b_i:

$$A_{ij} = b_i^T b_j$$

These kinds of matrices pop up all over the place. For example in the next lecture on PCA.

7.4 Exercise Sheet 8: Eigendecomposition

Familiarizing Exercises

Exercise 65. *Consider the linear map $u : x \in \mathbb{R}^2 \longmapsto Mx \in \mathbb{R}^2$, with*

$$M = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

1. *Show that $(1, -1)$ and $(1, 1)$ are two eigenvectors of u, and find their eigenvalues.*
2. *Deduce a diagonal matrix similar to M.*
3. *Find a vector in $\mathbb{R}^2$ that is not an eigenvector of u.*
4. *Show that $M^n := \underbrace{M \times \cdots \times M}_{n \text{ times}}$ is similar to a diagonal matrix, for any $n \geq 0$.*

Solution.

1. We apply M to the given vectors:

$$M(1, -1)^\top = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 3 - (-1) \\ -1 - 3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 4 \\ -4 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 2(1, -1)^\top.$$

   So $(1, -1)$ is an eigenvector with eigenvalue $\lambda = 2$.
   Similarly,

$$M(1, 1)^\top = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 3 - 1 \\ -1 + 3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1(1, 1)^\top.$$

   So $(1, 1)$ is an eigenvector with eigenvalue $\lambda = 1$.

2. Let $P = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ be the matrix whose columns are the eigenvectors. Define the diagonal matrix

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

   Then $M = PDP^{-1}$, so M is similar to the diagonal matrix D.

3. Any vector that is not a scalar multiple of either $(1, -1)$ or $(1, 1)$ is not an eigenvector. For instance, $v = (1, 0)$ is not an eigenvector. To check:

$$M(1, 0)^\top = \frac{1}{2}\begin{bmatrix} 3 \\ -1 \end{bmatrix} = \left(\frac{3}{2}, -\frac{1}{2}\right)^\top,$$

   which is not a scalar multiple of $(1, 0)$, so $(1, 0)$ is not an eigenvector.

4. Since M is similar to the diagonal matrix D, for any $n \in \mathbb{N}$,

$$M^n = PD^nP^{-1}.$$

   Hence, M^n is similar to the diagonal matrix D^n, which is given by

$$D^n = \begin{bmatrix} 2^n & 0 \\ 0 & 1^n \end{bmatrix} = \begin{bmatrix} 2^n & 0 \\ 0 & 1 \end{bmatrix}.$$

∎

Exercise 66. *Let u be a non-injective endomorphism of $\mathbb{R}^n$. Show that u has at least one eigenvalue, and find its value.*

Solution. Since u is a non-injective endomorphism of $\mathbb{R}^n$, the kernel of u is nontrivial:

$$\exists x \in \mathbb{R}^n,\ x \neq 0 \quad \text{such that} \quad u(x) = 0.$$

This means that x is an eigenvector associated to the eigenvalue $\lambda = 0$, since

$$u(x) = 0 = 0 \cdot x.$$

Thus, 0 is an eigenvalue of u. Hence, any non-injective endomorphism has $\lambda = 0$ as an eigenvalue. ∎

Not all endomorphisms are diagonalizable!

Exercise 67.

- *Let $u : \mathbb{R}^2 \longmapsto \mathbb{R}^2$ such that $u(e_1) = e_2$ and $u(e_2) = -e_1$. Show that u does not admit any eigenvalues.*
- *Deduce an endomorphism of $\mathbb{R}^3$ that admits only one eigenvalue.*

Solution.

- We are told that the linear map $u : \mathbb{R}^2 \to \mathbb{R}^2$ satisfies:

$$u(e_1) = e_2, \quad u(e_2) = -e_1.$$

Let $x = ae_1 + be_2 \in \mathbb{R}^2$. Then:

$$u(x) = au(e_1) + bu(e_2) = ae_2 - be_1.$$

Suppose x is an eigenvector of u with eigenvalue λ. Then:

$$u(x) = \lambda x = \lambda(ae_1 + be_2) = \lambda ae_1 + \lambda be_2.$$

Comparing both expressions for $u(x)$, we get:

$$-be_1 + ae_2 = \lambda ae_1 + \lambda be_2.$$

Matching coefficients:

$$\begin{cases} -b = \lambda a, \\ a = \lambda b. \end{cases}$$

Substituting the second equation into the first:

$$-b = \lambda(\lambda b) \Rightarrow -b = \lambda^2 b.$$

If $b \neq 0$, we can divide both sides by b:

$$-1 = \lambda^2 \Rightarrow \lambda^2 = -1,$$

which has no solution in $\mathbb{R}$.

If $b = 0$, then from the second equation $a = \lambda b = 0$. So $x = 0$, which is not allowed for eigenvectors.

Therefore, u does not admit any eigenvalues in $\mathbb{R}$.

- Now we build an endomorphism $v \in \mathcal{L}(\mathbb{R}^3)$ that admits only one eigenvalue. Let:

$$v(e_1) = e_2, \quad v(e_2) = -e_1, \quad v(e_3) = e_3.$$

This extends the previous map to $\mathbb{R}^3$. If $x = ae_1 + be_2 + ce_3$, then:

$$v(x) = ae_2 - be_1 + ce_3.$$

Suppose x is an eigenvector of v, so $v(x) = \lambda x$. Then:

$$ae_2 - be_1 + ce_3 = \lambda(ae_1 + be_2 + ce_3) = \lambda ae_1 + \lambda be_2 + \lambda ce_3.$$

Matching coefficients:

$$\begin{cases} -b = \lambda a, \\ a = \lambda b, \\ c = \lambda c. \end{cases}$$

From the third equation, if $c \neq 0$, then $\lambda = 1$.

For the first two equations, proceed as before:

$$-b = \lambda a, \quad a = \lambda b \Rightarrow -b = \lambda^2 b \Rightarrow \lambda^2 = -1,$$

which again has no real solution unless $a = b = 0$.

So the only real eigenvectors are scalar multiples of e_3, with eigenvalue $\lambda = 1$. Therefore, v admits only one eigenvalue over $\mathbb{R}$.

■

Important Exercise

Exercise 68. *Let $(x(t))_{t\in\mathbb{N}} \in (\mathbb{R}^n)^{\mathbb{N}}$ be a sequence of vectors governed by the equation:*

$$\begin{cases} x(0) = x_0 \\ x(t+1) = Wx(t) \end{cases}$$

For $W \in \mathcal{L}(\mathbb{R}^n)$ a diagonalizable matrix. $x(t)$ can be understood as modelling the evolution of the firing rate of a network of interconnected neurons.

- *Write the solution $x(t)$ as a function of the eigenvalues and eigenvectors of W, and the coordinates of x_0 in the basis of eigenvectors of W.*
- *Find a necessary condition on the eigenvalues of W for the coordinates of $x(t)$ to be bounded, for all t.*

Solution. Since $W \in \mathcal{L}(\mathbb{R}^n)$ is diagonalizable, there exists a basis $(v_1, \ldots, v_n)$ of eigenvectors of W, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, such that:

$$Wv_i = \lambda_i v_i \quad \text{for each } i = 1, \ldots, n.$$

1. Expression for $x(t)$ in the eigenbasis:

We express the initial condition $x_0 \in \mathbb{R}^n$ as a linear combination of the eigenvectors:

$$x_0 = \sum_{i=1}^{n} \alpha_i v_i,$$

for some scalars $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. This gives:

$$x(0) = x_0 = \sum_{i=1}^{n} \alpha_i v_i.$$

We now compute the iterates of $x(t)$ using the recurrence $x(t+1) = Wx(t)$. By induction:

$$x(1) = Wx(0) = \sum_{i=1}^{n} \alpha_i W v_i = \sum_{i=1}^{n} \alpha_i \lambda_i v_i,$$

$$x(2) = Wx(1) = \sum_{i=1}^{n} \alpha_i \lambda_i W v_i = \sum_{i=1}^{n} \alpha_i \lambda_i^2 v_i,$$

and in general:

$$x(t) = \sum_{i=1}^{n} \alpha_i \lambda_i^t v_i.$$

2. Condition for boundedness:
We now analyze under what conditions the sequence $x(t)$ remains bounded as $t \to \infty$. Since:

$$x(t) = \sum_{i=1}^{n} \alpha_i \lambda_i^t v_i,$$

the growth of $x(t)$ depends on the magnitude of each λ_i^t. For the term $\alpha_i \lambda_i^t v_i$ to remain bounded for all t, it is necessary that $|\lambda_i| \leq 1$.
In fact, if $|\lambda_i| > 1$ for some i with $\alpha_i \neq 0$, then $\lambda_i^t \to \infty$ and $x(t)$ will become unbounded. Therefore, a necessary (and sufficient) condition for $x(t)$ to remain bounded for all t is:

$$\boxed{|\lambda_i| \leq 1 \quad \text{for all } i = 1, \dots, n.}$$

■

Diagonalisable

Exercise 69. *Let $A \in \mathcal{L}(\mathbb{R}^n)$ be a diagonalizable matrix, i.e. there exist an invertible matrix U and a diagonal matrix D such that $A = UDU^{-1}$, show that the rank of A is equal to the number of non zero eigenvalues.*

Solution. Let $A \in \mathcal{L}(\mathbb{R}^n)$ be a diagonalizable linear map. By assumption, there exists an invertible matrix $U \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that

$$A = UDU^{-1}.$$

This means that A is the matrix representation of the same linear map as D, but in a different basis. Since a change of basis does not affect the rank of a linear map, we conclude:

$$\operatorname{rank}(A) = \operatorname{rank}(D).$$

Now, since D is diagonal, its rank is simply the number of nonzero entries on the diagonal, that is, the number of nonzero eigenvalues of A (counted without multiplicity).

$$\boxed{\operatorname{rank}(A) = \#\{i : \lambda_i \neq 0\}},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. ∎

Exercise 70. *Let $a \in \mathbb{R}^n, b \in \mathbb{R}^d$. What is the size of the matrix $ab^\top$? What are the entries of $ab^\top$? What is the rank of $ab^\top$? If $a = b$, what are the eigenvalues/eigenvectors of $aa^\top$? If a has norm 1, and $x \in \mathbb{R}^n$ what is $aa^\top x$ doing?*

Solution. Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^d$. We analyze each question in turn.
1. Size of the matrix $ab^\top$:
The vector a is of size $n \times 1$, and $b^\top$ is of size $1 \times d$, so the product $ab^\top$ is a matrix of size:

$$ab^\top \in \mathbb{R}^{n \times d}.$$

2. Entries of the matrix $ab^\top$:
Each entry of the matrix $ab^\top$ is given by:

$$(ab^\top)_{ij} = a_i b_j,$$

for all $1 \leq i \leq n$, $1 \leq j \leq d$. This means each row of $ab^\top$ is a scalar multiple of $b^\top$, and each column is a scalar multiple of a.
3. Rank of $ab^\top$:
We claim that $ab^\top$ has rank at most 1. Indeed, every column of $ab^\top$ is a scalar multiple of a. Explicitly, the j-th column of $ab^\top$ is:

$$(ab^\top)_{:,j} = ab_j.$$

So all columns lie in the one-dimensional subspace spanned by a. Therefore:

$$\operatorname{rank}(ab^\top) = \begin{cases} 1 & \text{if } a \neq 0 \text{ and } b \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

4. Eigenvalues/eigenvectors of $aa^\top$:
Now assume $a = b \in \mathbb{R}^n$, so that $A = aa^\top \in \mathbb{R}^{n \times n}$. We analyze the eigenvalues of A.
Step 1: Show a is an eigenvector.
Let us compute:

$$aa^\top a = a(a^\top a) = \|a\|^2 a.$$

So a is an eigenvector of A with eigenvalue $\|a\|^2$.
Step 2: Show all vectors orthogonal to a are eigenvectors with eigenvalue 0.
Let $x \in \mathbb{R}^n$ be such that $x \perp a$, i.e. $a^\top x = 0$. Then:

$$aa^\top x = a(a^\top x) = a \cdot 0 = 0.$$

So x is an eigenvector of A with eigenvalue 0.
Conclusion:

The eigenspace of eigenvalue $\|a\|^2$ is $\mathrm{Span}(a)$, and the eigenspace of eigenvalue 0 is $\{a\}^\perp$, which has dimension $n-1$. Therefore, the eigenvalues of $aa^\top \in \mathbb{R}^{n\times n}$ are:

$$\begin{cases} \|a\|^2 & \text{(multiplicity 1)}, \\ 0 & \text{(multiplicity } n-1). \end{cases}$$

5. Action of $aa^\top$ on x, when $\|a\| = 1$:
Assume $\|a\| = 1$, and let $x \in \mathbb{R}^n$. Then:

$$aa^\top x = a(a^\top x) = \langle a, x\rangle a.$$

This is the projection of x onto the direction of a. Therefore,

$$\boxed{aa^\top x = \mathrm{proj}_a(x)}.$$

So $aa^\top$ acts as the orthogonal projection matrix onto the subspace spanned by a, provided that a is a unit vector. ■

Symmetric Matrix

Exercise 71. *Show that a symmetric matrix $A \in \mathbb{R}^{n\times n}$ can be written:*

$$A = \sum_{i=1}^n \lambda_i u_i u_i^T \qquad \lambda_i \in \mathbb{R}, u_i \in \mathbb{R}^n$$

Solution. We assume $A \in \mathbb{R}^{n\times n}$ is symmetric. By the spectral theorem, there exists an orthonormal basis of eigenvectors $(u_1, \dots, u_n)$ of $\mathbb{R}^n$, with corresponding real eigenvalues $\lambda_1, \dots, \lambda_n$, such that:

$$Au_i = \lambda_i u_i, \quad \text{and} \quad u_i^T u_j = \delta_{ij} \quad \text{for all } i, j.$$

Let us now show that:

$$A = \sum_{i=1}^n \lambda_i u_i u_i^T.$$

We will do this by checking that this expression behaves the same as A when applied to an arbitrary vector $x \in \mathbb{R}^n$.
Since $(u_1, \dots, u_n)$ is an orthonormal basis, we can write:

$$x = \sum_{i=1}^n \langle x, u_i\rangle u_i = \sum_{i=1}^n (u_i^T x) u_i.$$

Now apply A to x, using linearity and the eigenvector property:

$$Ax = A\left(\sum_{i=1}^n (u_i^T x) u_i\right) = \sum_{i=1}^n (u_i^T x) A u_i = \sum_{i=1}^n (u_i^T x)\lambda_i u_i.$$

On the other hand, consider the expression:

$$\left(\sum_{i=1}^n \lambda_i u_i u_i^T\right) x = \sum_{i=1}^n \lambda_i u_i (u_i^T x) = \sum_{i=1}^n (u_i^T x)\lambda_i u_i,$$

which is exactly the same expression as above. Thus, for all $x \in \mathbb{R}^n$,

$$Ax = \left(\sum_{i=1}^{n} \lambda_i u_i u_i^T \right) x,$$

so the two matrices must be equal:

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T.$$

Interpretation: Each matrix $u_i u_i^T$ is a rank-1 symmetric projection matrix onto the direction u_i. The scalar λ_i tells how strongly A acts along the direction u_i. The matrix A is thus a weighted sum of these orthogonal projections, and this expression reveals both the spectral structure and the geometric action of A. ■

Positive Semi-Definite Matrix

Exercise 72. *Show that any positive definite matrix is invertible.*

Solution. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. By definition, this means:

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad x^T A x > 0.$$

Suppose for contradiction that A is not invertible. Then there exists $x \in \mathbb{R}^n \setminus \{0\}$ such that $Ax = 0$. But then:

$$x^T A x = x^T \cdot 0 = 0,$$

which contradicts the positive definiteness of A, since $x \neq 0$ implies $x^T A x > 0$.
Therefore, A must be invertible. ■

Exercise 73. *Let $A \in \mathcal{L}(\mathbb{R}^n)$ be a symmetric matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. Show that $\lambda_i \geq 0$ for all $i = 1, \ldots, n$ implies that A is positive semi definite. Show that $\lambda_i > 0$ for all $i = 1, \ldots, n$ implies that A is positive definite. Hint: consider the orthonormal basis of eigenvectors for A. For $x \in \mathbb{R}^n$, decompose x on this basis and develop $x^\top A x$.*

Solution. Since $A \in \mathcal{L}(\mathbb{R}^n)$ is symmetric, the spectral theorem applies: there exists an orthonormal basis $\{u_1, \ldots, u_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of A, with corresponding real eigenvalues $\lambda_1, \ldots, \lambda_n$, such that

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T.$$

Let $x \in \mathbb{R}^n$, and write its decomposition in the orthonormal basis:

$$x = \sum_{i=1}^{n} \alpha_i u_i, \quad \text{where } \alpha_i = u_i^T x.$$

We now compute the quadratic form:

$$x^T A x = x^T \left(\sum_{i=1}^{n} \lambda_i u_i u_i^T \right) x = \sum_{i=1}^{n} \lambda_i (u_i^T x)^2 = \sum_{i=1}^{n} \lambda_i \alpha_i^2.$$

Now analyze the conditions:
(1) If $\lambda_i \geq 0$ for all i, then:

$$x^T A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \geq 0 \quad \text{for all } x \in \mathbb{R}^n.$$

Hence, A is *positive semi-definite.*
(2) If $\lambda_i > 0$ for all i, then for any $x \neq 0$, there exists at least one i such that $\alpha_i \neq 0$, and so:

$$x^T A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 > 0.$$

Hence, A is *positive definite.* ■

Exercise 74. *The matrix $A \in \mathbb{R}^{n\times n}$ is constructed from $B \in \mathbb{R}^{m\times n}$ via $A = B^T B$. Show that A is positive semi-definite.*

Solution. We are given a matrix $A \in \mathbb{R}^{n\times n}$ defined as $A = B^T B$, where $B \in \mathbb{R}^{m\times n}$. A is symmetric. We aim to show that A is positive semi-definite, i.e., for all $x \in \mathbb{R}^n$,

$$x^T A x \geq 0.$$

For any $x \in \mathbb{R}^n$,

$$x^T A x = x^T B^T B x = (Bx)^T (Bx) = \|Bx\|^2 \geq 0.$$

Therefore, $A = B^T B$ is positive semi-definite. ■

Exercise 75. *Show that if rank$(B) \geq n$, A is positive definite.*

Solution. We are given that $A = B^T B$, where $B \in \mathbb{R}^{m\times n}$, and we are told that $\operatorname{rank}(B) \geq n$. Since $B \in \mathbb{R}^{m\times n}$, this implies $m \geq n$, and $\operatorname{rank}(B) = n$. That is, the columns of B are linearly independent.
We want to show that $A = B^T B$ is positive definite. That is, for all $x \in \mathbb{R}^n \setminus \{0\}$,

$$x^T A x = x^T B^T B x = \|Bx\|^2 > 0.$$

As $\operatorname{rank}(B) = n$, the kernel of B is $\{0\}$. So if $x \in \mathbb{R}^n \setminus \{0\}$, then $Bx \neq 0$. Thus,

$$\|Bx\|^2 > 0,$$

and hence

$$x^T A x = \|Bx\|^2 > 0.$$

For all $x \neq 0$, we have $x^T A x > 0$, so A is positive definite. ■

Exercise 76. *Show that all positive semi-definite matrices, A, can be decomposed as $A = B^T B$ for some matrix B.*

Solution. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive semi-definite matrix. By the **spectral theorem**, there exists an orthogonal matrix $P \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i \geq 0$ such that

$$A = PDP^T.$$

Let $\sqrt{D}$ be the diagonal matrix with entries $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$. Define

$$B := \sqrt{D} P^T.$$

Then:

$$B^T B = P\sqrt{D}\sqrt{D}P^T = PDP^T = A.$$

Conclusion: Every symmetric positive semi-definite matrix A can be written as $A = B^T B$ with $B = \sqrt{D}P^T$. ■

Exercise 77. *When A and B are positive definite matrices AB might not even be symmetric, but it's eigenvalues are positive. Show this, by starting from $ABx = \lambda x$ and taking dot products with Bx.*

Solution. Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric positive definite matrices. Suppose $x \in \mathbb{R}^n \setminus \{0\}$ and that $ABx = \lambda x$ for some eigenvalue $\lambda \in \mathbb{R}$.
We take the dot product of both sides of the equation with Bx:

$$(ABx)^T Bx = \lambda x^T Bx.$$

Note that $(ABx)^T Bx = x^T B^T A^T Bx = x^T BABx$, since A and B are symmetric. So we have:

$$x^T BABx = \lambda x^T Bx.$$

Now observe:
- $x^T BABx > 0$ because $Bx \neq 0$ (since B is positive definite), and A is positive definite, so:

$$x^T BABx = (Bx)^T A(Bx) > 0.$$

- $x^T Bx > 0$, again because B is positive definite.
Therefore, we conclude that

$$\lambda = \frac{x^T BABx}{x^T Bx} > 0.$$

Conclusion: Even though AB is not symmetric, its eigenvalues are strictly positive when A and B are positive definite. ■

7.5 Neuro Q5: Eigenspaces & Line Attractors

In this question we will unleash the power of linear algebra to design a network of neurons. Our goal is to use these neurons to encode a continuous variable and preserve memory of that variable over time. By encode we mean that there must be a mapping between the variable and the behaviour of the neurons so that the animal can use the neurons as a proxy for the variable. For example, the variable might be "the distance between me and the nearest supermarket bakery section". If you have a set of neurons that encode this variable you can use them to direct your behaviour: e.g. find the direction that your neurons tell you most decreases the distance, then walk forward. Further, we want these neurons to keep representing the variable even if you get no input telling you what the variable is, e.g. if you close your eyes. Sound good? Let's go.

Let's call the scalar variable $s \in \mathbb{R}$ and lets call the neural encoding $g : \mathbb{R} \to \mathbb{R}^N$ where N is the number of neurons, i.e. the encoding is a function that maps s to a vector of associated neural activities. We now specify our nice simple model of neurons (though that doesn't make it a trivial model, many interesting models of neural function are similar to this one).

We have a population of N neurons. Each of them receives inputs from all the other neurons. We will represent the activity of the neurons as a real number, roughly interpreted as the rate of spiking (minus a baseline so it can be negative too!). Each neuron projects its activity through its axon, which proceeds to hit the dendrites of the other neurons. In this way, one neuron influences all the others, mediated by the synapses. We will model this as the firing rate of each neuron at timepoint $t+1$, $g_{n,t+1}$ being a weighted sum of the others, weighted by the synaptic connections:

$$g_{n,t+1} = \sum_{n'=1}^{N} W_{n,n'} g_{n',t}$$

where $W_{n,n'}$ is the strength of the connection from neuron n' to neuron n. In other words, grouping the N firing rates into the vector g, their behaviour is given by the following linear equations:

$$g_{t+1} = W g_t$$

The goal of this population is to encode a variable s. Say the initial activity at time $t = 0$ is $g_0 = g(s)$, perhaps via the activity of sensory inputs, and let's assume that W is diagonalisable. Recall: this means the behaviour of any vector of firing rates can be understood by thinking about the behaviour of the eigenvectors, and the coefficients of the vector in the eigenbasis.

1. First things first, we want to be able to leave the neurons on their own for a long time, come back, and find that they have not exploded. No neuron can fire with an arbitrarily large firing rate! How must we choose the eigenvalues of W to stop this?
2. Okay, but we also want the neural population to keep a memory of the variable s no matter how long we wait. What must at least one of the eigenvalues equal for this to be true?
3. Given your answers to the previous two questions, on which eigenspace will the activity eventually land? Call this the line attractor. (for reasons that should make sense)

Now, so far we have argued that the line attractor should exist, but nothing more, in particular we don't care about which direction it points.

4. Justify why pointing the line attractor along the first canonical basis vector would mean only a single neuron would encode the variable.

5. Let's assume you point the line attractor along a neuron's axis. Use biological reasoning to argue why this might not be a good idea?

We've made interesting conclusions about how biological constraints force us to construct our line attractor. We also think neurons are noisy (which just means 'they fire and we can't always work out why').
For example, a cosmic ray might skewer a set of neurons exciting them all, or the whole brain might get more excitable if you start running. Let's pretend that the following noise hits the system:

$$g'_{n,t} = g_{n,t} + \boldsymbol{\eta}$$

$$\boldsymbol{\eta} = \begin{cases} \mathbf{0}, & \text{with probability } 0.98 \\ \mathbf{1}, & \text{with probability } 0.01 \\ -\mathbf{1}, & \text{with probability } 0.01 \end{cases}$$

where $\mathbf{1}$ is the all-ones vector: either there is no noise, or it pushes all neurons either up or down.

6. Point the line attractor in the $\mathbf{1}$ direction. Assume the noise arrives, then the dynamics are run until the activity is stationary. What is the mean difference between the position the activity was at on the line attractor before we added the noise and after the activity is left to settle back to the line attractor? What about the variance?

7. How can I choose my eigenstructure to ensure this noise doesn't screw up my encoding?

Now, let's zoom into the behaviour along the line attractor. Denote the component of the neural activity within the line attractor at each time point as l_t. Say there's some noise along your line attractor that multiplicatively scales your neural activity up:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \qquad \eta = 1 + |\epsilon|$$

$$l_{t+1} = \lambda \eta l_t$$

9. Show the mean of η is now non-zero.

10. Explain why this is bad.

11. Let's say now the noise tends to arrive every timestep, a different sample each time, how can you change the eigenvalue of your line attractor so that on average $l_{t+1} = l_t$?

Sometimes it is good to forget. For example, let's say you close your eyes and are trying to keep track of the position of your toddler in the playground. At first the toddler is likely to be close to where you last saw them, but you know your toddler is full of energy and runs all over the place, eventually you will have no idea where your toddler is, and you might as well guess any position in the playground randomly. No need to waste effort remembering.

12. Let's say you want to remember something for T timesteps, but after that you don't care and would like to forget. Roughly what should the eigenvalue of your line attractor be?

Chapter 8

Singular Value Decomposition

8.1 Orthogonal Matrices

This section will introduce a very useful type of matrix, it pops up a lot, and if nothing else, the rest of this lecture should justify its usefulness.

Orthogonal Matrix

Definition 46. *A matrix, $M \in \mathbb{R}^{n\times n}$, is called orthogonal if its columns form an orthonormal basis.*

Note that we call the matrix "orthogonal" but its columns form an "orthonormal" basis. The columns of an orthogonal matrix always have norm 1!

Checking that a matrix is orthogonal

Proposition 42. *A matrix $M \in \mathbb{R}^{n\times n}$ is orthogonal if and only if: $M^T M = id_n$.*

Proof

Let's write the matrix in terms of its columns:

$$M = \begin{bmatrix} | & \dots & | \\ c_1 & \dots & c_n \\ | & \dots & | \end{bmatrix}$$

for $c_i \in \mathbb{R}^n$. Then we can use the statement that the columns form an orthonormal basis:

$$M^T M = \begin{bmatrix} - & c_1 & - \\ \vdots & \vdots & \vdots \\ - & c_n & - \end{bmatrix} \begin{bmatrix} | & \dots & | \\ c_1 & \dots & c_n \\ | & \dots & | \end{bmatrix} = \begin{bmatrix} \langle c_1, c_1\rangle & \langle c_1, c_2\rangle & \dots & \langle c_1, c_n\rangle \\ \langle c_1, c_2\rangle & \langle c_2, c_2\rangle & \dots & \langle c_2, c_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle c_n, c_1\rangle & \langle c_n, c_2\rangle & \dots & \langle c_n, c_n\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

So we can see that $\langle c_i, c_j\rangle = \delta_{ij}$ if and only if $M^T M = id_n$.

Example

An important example that illustrates these points is the following matrix:

$$M = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

This satisfies orthogonality:

$$M^T M = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \cos^2(\theta) + \sin^2(\theta) & 0 \\ 0 & \cos^2(\theta) + \sin^2(\theta) \end{bmatrix} = id_2$$

Further, this makes sense from geometrically looking at the matrix, it is the two dimensional rotation matrix!

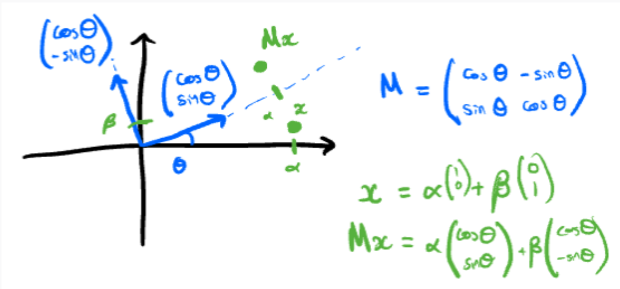


Figure 8.1: This matrix is orthogonal, and simply performs a rotation.

Equivalent Characterisations of Orthogonality

Proposition 43. *The following six propositions about a matrix $M \in \mathbb{R}^{n\times n}$ are equivalent:*

1. *M is called orthogonal, i.e. its columns form an orthonormal basis*
2. *$M^T M = id_n$*
3. *$M^T = M^{-1}$*
4. *$\forall x, y \in \mathbb{R}^n$, $\langle x, y\rangle = \langle Mx, My\rangle$*
5. *$\forall x \in \mathbb{R}^n$, $||Mx|| = ||x||$*
6. *If a set of n vectors, $\{e_1, ..., e_n\}$ form an orthogonal basis, then so too do $\{Me_1, ..., Me_n\}$.*

Proof. We prove the implications in a cycle: (1) ⇒ (2) ⇒ (3) ⇒ (4) ⇒ (5) ⇒ (6) ⇒ (1).
(1) ⇒ (2): See Proposition 42.
(2) ⇒ (3): See Exercise 56.
(3) ⇒ (4): Assume $M^T = M^{-1}$. Then for any $x, y \in \mathbb{R}^n$,

$$\langle Mx, My\rangle = (Mx)^T(My) = x^T M^T My = x^T Iy = \langle x, y\rangle.$$

(4) ⇒ (5): Take $x = y$ in the inner product preservation condition:

$$\|Mx\|^2 = \langle Mx, Mx\rangle = \langle x, x\rangle = \|x\|^2.$$

So $\|Mx\| = \|x\|$.
(5) $\Rightarrow$ (6): Let $\{e_1, \dots, e_n\}$ be an orthonormal basis. For any i, we have:

$$\|Me_i\| = \|e_i\| = 1.$$

For $i \neq j$, use the polarization identity:

$$\langle Me_i, Me_j \rangle = \frac{1}{2}\left(\|Me_i + Me_j\|^2 - \|Me_i\|^2 - \|Me_j\|^2\right) = \frac{1}{2}(\|e_i + e_j\|^2 - 1 - 1) = 0,$$

using the assumption that $\|Mx\| = \|x\|$ for all x. Hence, $\{Me_1, \dots, Me_n\}$ is also an orthonormal basis.
(6) $\Rightarrow$ (1): Apply (6) to the standard orthonormal basis $\{e_1, \dots, e_n\}$. Then $Me_1, \dots, Me_n$ form an orthonormal basis. These are precisely the columns of M, so the columns of M form an orthonormal basis. □

Proof of Corollary 3

Proof. Let us define an endomorphism $u : \mathbb{R}^n \to \mathbb{R}^n$ by $u(x) = Ax$. Since A is symmetric (i.e., $A^\top = A$), the map u is symmetric with respect to the standard inner product on $\mathbb{R}^n$. That is, for all $x, y \in \mathbb{R}^n$, we have:

$$\langle u(x), y \rangle = \langle Ax, y \rangle = x^\top A^\top y = x^\top Ay = \langle x, Ay \rangle = \langle x, u(y) \rangle.$$

We now apply Lemma 3 to u. We obtain an orthonormal basis $\mathcal{B} = \{v_1, \dots, v_n\}$ of eigenvectors of u, with corresponding real eigenvalues $\lambda_1, \dots, \lambda_n$, such that:

$$u(v_i) = \lambda_i v_i.$$

Let $P \in \mathbb{R}^{n \times n}$ be the change-of-basis matrix from the canonical basis to the basis $\mathcal{B}$. That is, the i-th column of P is the vector v_i. Since $\mathcal{B}$ is orthonormal, P is orthogonal and $P^\top = P^{-1}$.
Let $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Then D is the matrix of the endomorphism u in the eigenbasis $\mathcal{B}$, and A is its matrix in the canonical basis. By the change-of-basis formula for endomorphisms:

$$A = PDP^{-1} = PDP^\top.$$

□

8.2 Singular Value Decomposition

Motivation

When eigendecomposition and diagonalisation worked it gave us a very clean way to view matrices: in some basis they are just scaling the coordinate axes, what could be simpler!

However, we found that not all matrices could be diagonalised, and eigenvalues are only defined for square matrices. Today we will cover a more general decomposition that is nearly as clean as eigendecomposition, and works all the time.

Singular Value Decomposition

Theorem 4. *Every matrix, $A \in \mathbb{R}^{n\times m}$ can be written as the product of three simple matrices:*

$$A = USV^T$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices and $S \in \mathbb{R}^{n\times m}$ is diagonal.

Singular Values & Singular Vectors

Rewrite these three matrices as follows (for display purposes assume $m > n$, if not just reverse everything):

$$U = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}, \quad V = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & & | \end{bmatrix}, \quad S = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & & & \vdots \\ 0 & 0 & \dots & \sigma_n & 0 & \dots & 0 \end{bmatrix}$$

The vectors v_i are called the right singular vectors, u_i are the left singular vectors, and σ_i are the singular values. We can see they obey some eigentastic equations:

$$Av_i = \begin{cases} \sigma_i u_i & \forall i \leq n \\ 0 & \forall i > n \end{cases}, \quad A^T u_i = \sigma_i v_i$$

Casual Explanation

This is a stunning result. For rectangular matrices this is the best generalisation of eigen-decomposition we could have hoped for. Recall: eigenvectors are mapped to scaled versions of themselves. Further, in the nicest setting, symmetric matrices, the eigenvectors form an orthonormal basis. Since rectangular matrices map between spaces of different dimensionalities, it would never have been possible for exactly this kind of result to hold (a two dimensional vector can't be a scaled version of itself if it lives in three dimensions...). But we get the next best thing, there are a set of vectors in the input space, the right singular vectors, that map to scaled version of another set of vectors in the output space. Further, these two sets of vectors, appropriately completed if necessary, form an orthonormal basis for the two spaces. This means that under some orthogonal transformation, all rectangular matrices are just a pairing between dimensions in the input and output space, with a dimension-wise scaling! Finally, for square matrices, we are not able to diagonalise all matrices, and even if we can the eigenbasis might not be orthonormal. However, if we relax to singular vectors, these problems melt away, divine!

Singular value decomposition can be visually interpreted, therefore, as performing an orthogonal transform V^T, a scaling S, followed by an orthogonal transform in the output space U, as shown in this figure.

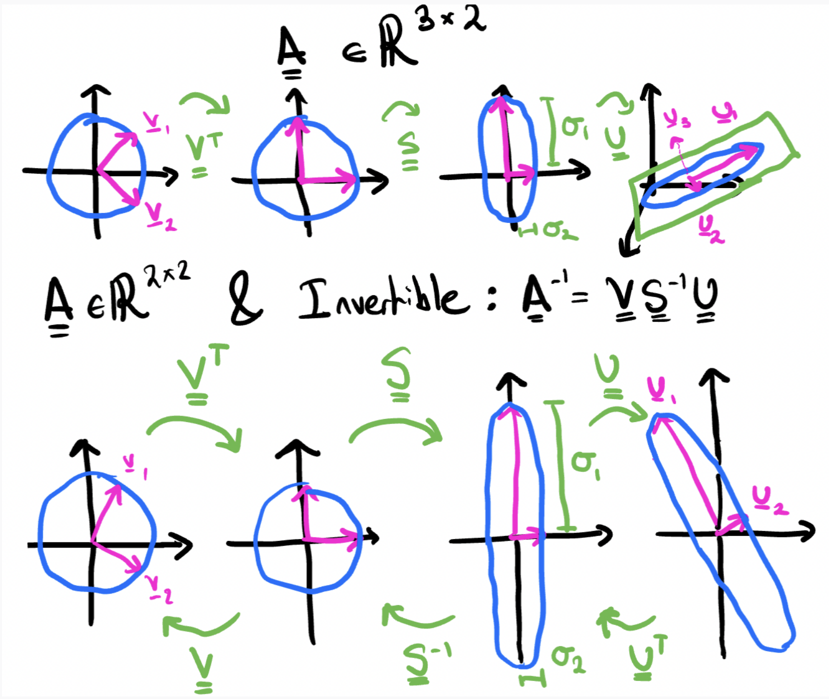


Figure 8.2: All matrices can be understood as the composition of an orthogonal transform, a dimension-wise scaling, and an orthogonal transform!

To close, let's prove the singular value decomposition exists, and find another characterization of the singular values and vectors en route.

Proof

To begin, recall the following facts from the eigendecomposition lecture:

- The spectral theorem: any symmetric matrix is diagonalizable in an orthonormal basis.
- Positive definite (resp. semi-definite) matrices A are those such that $x^\top Ax > 0$ (resp $x^\top Ax \geq 0$)
- Any positive definite matrix has only positive (resp. nonnegative) eigenvalues. Indeed, for symmetric A this proof is a one-liner: for any $x \in \mathbb{R}^n$:

$$x^\top Ax = \sum_{i=1}^{n} \lambda_i \langle x, f_i \rangle^2 \geq 0, \forall x \iff \lambda_i \geq 0$$

  where $\{f_1, \ldots, f_n\}$ is the orthonormal basis of eigenvectors of A.

We now return to prove the main result. $A \in \mathbb{R}^{n\times m}$ and we'll assume $n > m$, if not take the transpose, do the same thing, before transposing again at the end to get the same result.

1. First, realize that $A^T A \in \mathbb{R}^{m\times m}$ is symmetric positive semi-definite. Let us call $\{v_1,\dots,v_m\}$ the ONB of eigenvectors, with their associated eigenvalues $\lambda_1,\dots,\lambda_m$ s.t $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$.

2. Let us denote:
$$\begin{aligned}\sigma_i &= \|Av_i\| \\ &= \sqrt{\|Av_i\|^2} \\ &= \sqrt{v_i^\top A^\top A v_i} \\ &= \sqrt{\lambda_i}\end{aligned}$$
which will make up the entries for our S matrix. Since $\lambda_1 \geq \dots, \geq \lambda_m \geq 0$, $\sigma_1 \geq \dots, \geq \sigma_m \geq 0$.

3. Note that $\sigma_i = 0$ if and only if $\lambda_i = 0$, and for such λ_i's, by the previous derivations, we have $\|Av_i\| = 0$ and therefore $Av_i = 0$.

4. Let $r \leq m$ be the largest index such that $\sigma_r > 0$. For $i = 1,\dots,r$, let us denote $u_i = \frac{1}{\sigma_i} Av_i$, which is possible since for $i = 1,\dots,r$, $\sigma_i \neq 0$. For $i = 1,\dots,r$, $j = 1,\dots,r$ and $\lambda_i \neq \lambda_j$, we have that:
$$\begin{aligned}\langle u_i, u_j\rangle &= \frac{1}{\sigma_i\sigma_j}\langle Av_i, Av_j\rangle \\ &= \frac{1}{\sigma_i\sigma_j} v_i^\top A^\top A v_j \\ &= \frac{\lambda_i}{\sigma_i\sigma_j}\delta_{ij} = \delta_{ij}\end{aligned}$$
Thus, $\{u_1,\dots,u_r\}$ is an orthonormal system of $\mathbb{R}^n$. Note that it cannot be an orthonormal basis of $\mathbb{R}^n$, since it has only r vectors and $r \leq m < n$. We complete the orthonormal system $\{u_1,\dots,u_r\}$ into an orthonormal basis of $\mathbb{R}^n$ by adding $n-r$ vectors, which we denote by $u_{r+1},\dots,u_n$.

5. From the relationship $Av_i = \sigma_i u_i$ for $i = 1,\dots,r$ and $Av_i = 0 = \sigma_i u_i$ for $i = r+1,\dots,m$ (since $\sigma_i = 0$ for $i = r+1,\dots,m$), we deduce that:
$$AV = A\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_m u_m \\ | & | & & | \end{bmatrix} = \widehat{U}\widehat{S}$$
Where $\widehat{S} = \mathrm{diag}(\sigma_1,\dots,\sigma_m) \in \mathbb{R}^{m\times m}$ and
$$\widehat{U} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_m \\ | & | & & | \end{bmatrix}$$
Since V is orthogonal, and therefore $V^{-1} = V^T$ we can write: $AV = \widehat{U}\widehat{S} \implies A = \widehat{U}\widehat{S}V^\top$. This is called the reduced singular value decomposition of A: it is not the full svd of A because $\widehat{U}$ is not of size $n \times n$, but only of size $n \times m$ and $\widehat{S}$ is not of

size $n \times m$, but only of size $m \times m$. To complete $\widehat{U}$ into a matrix of size $n \times n$ we had the columns $u_{m+1}, \ldots, u_n$ to it:

$$U = \begin{bmatrix} \widehat{U} & U_\perp \end{bmatrix} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}$$

with $U_\perp \in \mathbb{R}^{n\times(n-m)}$, and to complete $\widehat{S}$ into a matrix of size $n \times m$, we just have to pad $\widehat{S}$ with zeros to get:

$$S = \begin{bmatrix} \widehat{S} \\ 0_{(n-m)\times n} \end{bmatrix} \in \mathbb{R}^{n\times m}$$

We now show that

$$A = \widehat{U}\widehat{S}V^\top = USV^\top$$

Indeed, we have:

$$US = \begin{bmatrix} \widehat{U} & U_\perp \end{bmatrix} \begin{bmatrix} \widehat{S} \\ 0 \end{bmatrix} = \widehat{U}\widehat{S},$$

which concludes the proof.

8.3 Exercise Sheet 9: Orthogonal Matrices and SVD

Exercises

Exercise 78. *Show that if two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ are orthogonal, so is AB.*

Solution. We are given that $A, B \in \mathbb{R}^{n\times n}$ are orthogonal matrices. By definition, this means:

$$A^T A = I_n \quad \text{and} \quad B^T B = I_n.$$

We want to show that the matrix product AB is also orthogonal, i.e., that:

$$(AB)^T (AB) = I_n.$$

We compute:

$$(AB)^T (AB) = B^T A^T AB = B^T I_n B = B^T B = I_n,$$

since $A^T A = I_n$ and $B^T B = I_n$ by assumption. Thus, $(AB)^T (AB) = I_n$, so AB is orthogonal. ■

Exercise 79. *A matrix $A \in \mathbb{R}^{n\times n}$ can be diagonalised using an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$: $A = QDQ^{-1}$ for some diagonal matrix D. Show that it is symmetric.*
Explain why this is like the converse of the spectral theorem.

Solution. We are given that a matrix $A \in \mathbb{R}^{n\times n}$ can be diagonalized using an orthogonal matrix Q, i.e., there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that

$$A = QDQ^{-1}.$$

Since Q is orthogonal, we have $Q^{-1} = Q^T$, so we can write:

$$A = QDQ^T.$$

We now compute the transpose of A:

$$A^T = (QDQ^T)^T = (Q^T)^T D^T Q^T = QDQ^T = A,$$

since $D^T = D$ (as D is diagonal) and $(Q^T)^T = Q$.
Hence, $A^T = A$, and therefore A is symmetric.

Remark (Converse of the Spectral Theorem): The spectral theorem states that any real symmetric matrix can be diagonalized by an orthogonal matrix. This exercise shows the converse: if a real matrix can be diagonalized by an orthogonal matrix, then it must be symmetric. So the two statements together characterize real symmetric matrices as precisely those that are orthogonally diagonalizable. ■

Exercise 80. *The matrix A can be written:*

$$A = \begin{bmatrix} 1 & 0 & x \\ 0 & \cos(\theta) & y \\ 0 & \sin(\theta) & z \end{bmatrix}$$

Describe the values of (x, y, z) that make the matrix A orthogonal.

Solution. We want to find the values of $(x, y, z) \in \mathbb{R}^3$ such that A is orthogonal, i.e.,

$$A^T A = I.$$

Let us compute $A^T A$. First, compute A^T:

$$A^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ x & y & z \end{bmatrix}$$

Then compute $A^T A$:

$$A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ x & y & z \end{bmatrix} \begin{bmatrix} 1 & 0 & x \\ 0 & \cos(\theta) & y \\ 0 & \sin(\theta) & z \end{bmatrix}$$

Compute each entry of the resulting matrix:
- First row:

$$(1, 0, 0) \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1, \quad (1, 0, 0) \cdot \begin{bmatrix} 0 \\ \cos(\theta) \\ \sin(\theta) \end{bmatrix} = 0, \quad (1, 0, 0) \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x$$

- Second row:

$$(0, \cos(\theta), \sin(\theta)) \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 0$$

$$(0, \cos(\theta), \sin(\theta)) \cdot \begin{bmatrix} 0 \\ \cos(\theta) \\ \sin(\theta) \end{bmatrix} = \cos^2(\theta) + \sin^2(\theta) = 1$$

$$(0, \cos(\theta), \sin(\theta)) \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \cos(\theta)y + \sin(\theta)z$$

- Third row:

$$(x, y, z) \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = x, \quad (x, y, z) \cdot \begin{bmatrix} 0 \\ \cos(\theta) \\ \sin(\theta) \end{bmatrix} = y\cos(\theta) + z\sin(\theta)$$

$$(x, y, z) \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x^2 + y^2 + z^2$$

Putting it all together:

$$A^T A = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y\cos(\theta) + z\sin(\theta) \\ x & y\cos(\theta) + z\sin(\theta) & x^2 + y^2 + z^2 \end{bmatrix}$$

In order for A to be orthogonal, we require $A^T A = I$, so we must have:

$$x = 0, \quad y\cos(\theta) + z\sin(\theta) = 0, \quad x^2 + y^2 + z^2 = 1.$$

Since $x = 0$, the last condition becomes:

$$y^2 + z^2 = 1, \quad \text{and} \quad y\cos(\theta) + z\sin(\theta) = 0.$$

This means $(y, z) \in \mathbb{R}^2$ is a unit vector orthogonal to $(\cos(\theta), \sin(\theta))$, i.e., it must lie along the vector $(-\sin(\theta), \cos(\theta))$ or its opposite.
So we conclude:

$$(x, y, z) = (0, -\sin(\theta), \cos(\theta)) \quad \text{or} \quad (0, \sin(\theta), -\cos(\theta)).$$

■

Exercise 81. *Show that an orthogonal matrix as orthonormal rows.*

Solution. Let $M \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. By definition, this means that its columns are orthonormal, we saw that it is equivalent to $M^\top = M^{-1}$, and therefore:

$$MM^\top = I_n.$$

This implies that $M^\top$ is also an orthogonal matrix. Therefore, the columns of $M^\top$ form an orthonormal basis of $\mathbb{R}^n$.
But the columns of $M^\top$ are exactly the rows of M. Hence, the rows of M are orthonormal.
■

Exercises

Exercise 82. *Let $A = USV^T$ be the SVD of $A \in \mathbb{R}^{n\times m}$. Let $\{u_1, \dots, u_n\}$ be the left singular vectors (columns of U) and $\{v_1, \dots, v_m\}$ be the right singular vectors (columns of V). Let $r \leq \min(n, m)$ be the number of non-zero singular values of A. Show that we can write:*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

Solution. Let $A \in \mathbb{R}^{n\times m}$, and let its singular value decomposition be given by:

$$A = U\Sigma V^\top,$$

where:
- $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns $u_1, \dots, u_n$ are the left singular vectors,
- $V \in \mathbb{R}^{m\times m}$ is an orthogonal matrix whose columns $v_1, \dots, v_m$ are the right singular vectors,
- $\Sigma \in \mathbb{R}^{n\times m}$ is a diagonal matrix with nonnegative real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ on the diagonal, and zeros elsewhere.
We can express the matrix product $A = U\Sigma V^\top$ in the following way:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$

To see why this is true, observe that:
- The matrix Σ has the singular values $\sigma_1, \dots, \sigma_r$ (with $r \leq \min(n, m)$) on the diagonal and zeros elsewhere.
- Therefore, we can write:

$$\Sigma = \sum_{i=1}^{r} \sigma_i e_i f_i^\top,$$

where $e_i \in \mathbb{R}^n$ and $f_i \in \mathbb{R}^m$ are the standard basis vectors, and each σ_i is in position (i, i).
Now consider the product:

$$A = U\Sigma V^\top = U \left(\sum_{i=1}^{r} \sigma_i e_i f_i^\top \right) V^\top = \sum_{i=1}^{r} \sigma_i (Ue_i)(Vf_i)^\top.$$

But $Ue_i = u_i$ and $Vf_i = v_i$, since u_i and v_i are the columns of U and V, respectively. Therefore:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

as desired. ■

Exercise 83. *Let $A \in \mathbb{R}^{n\times m}$. Show that the rank of A is equal to the number of non-zero singular values of A.*

Solution. Let $A \in \mathbb{R}^{n\times m}$ and consider its singular value decomposition (SVD):

$$A = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{n\times m}$ is a diagonal matrix (possibly rectangular) with non-negative real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ on the diagonal, followed by zeros.
Let r be the number of non-zero singular values. Then Σ has exactly r non-zero rows, and the matrix A can be written as:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

where u_i and v_i are the i-th columns of U and V, respectively.
Now, consider $x \in \ker A$. Then:

$$Ax = \sum_{i=1}^{r} \sigma_i u_i \langle v_i, x\rangle = 0.$$

Since the $\{u_i\}$ are linearly independent and $\sigma_i > 0$, this implies:

$$\langle v_i, x\rangle = 0 \quad \text{for all } i = 1, \dots, r.$$

Therefore, $x \in \operatorname{span}(v_{r+1}, \dots, v_m)$, and so:

$$\ker A = \operatorname{span}(v_{r+1}, \dots, v_m),$$

which has dimension $m - r$. Applying the **rank-nullity theorem**:

$$\operatorname{rank}(A) = m - \dim(\ker A) = m - (m - r) = r.$$

Hence, the rank of A is exactly the number of non-zero singular values. ■

Exercise 84. *Why are A^TA and AA^T symmetric positive definite?*

Solution. Let $A \in \mathbb{R}^{m\times n}$. We want to study the matrices $A^TA \in \mathbb{R}^{n\times n}$ and $AA^T \in \mathbb{R}^{m\times m}$. We'll show they are both symmetric and positive semi-definite.
1. Symmetry:
- $(A^TA)^T = A^T(A^T)^T = A^TA \Rightarrow A^TA$ is symmetric.
- $(AA^T)^T = (A^T)^TA^T = AA^T \Rightarrow AA^T$ is symmetric.
2. Positive semi-definiteness:
Let $x \in \mathbb{R}^n$. Then:

$$x^TA^TAx = (Ax)^T(Ax) = \|Ax\|^2 \geq 0.$$

So A^TA is positive semi-definite. Similarly, let $y \in \mathbb{R}^m$. Then:

$$y^TAA^Ty = (A^Ty)^T(A^Ty) = \|A^Ty\|^2 \geq 0,$$

so AA^T is also positive semi-definite. ■

Exercise 85. *When are the SVD and the eigendecomposition of the matrix the same?*

Solution. The Singular Value Decomposition (SVD) and the eigendecomposition of a matrix are generally distinct decompositions. However, they coincide under specific conditions.
Reminder:
- The *SVD* of a matrix $A \in \mathbb{R}^{n\times n}$ is:

$$A = U\Sigma V^T,$$

where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and Σ is a diagonal matrix with non-negative entries (the singular values).
- The *eigendecomposition* of a matrix $A \in \mathbb{R}^{n\times n}$ (when it exists) is:

$$A = Q\Lambda Q^{-1},$$

where Q is invertible (or orthogonal if A is symmetric) and Λ is diagonal (with eigenvalues on the diagonal).
When are they the same?
The SVD and the eigendecomposition of a matrix $A \in \mathbb{R}^{n\times n}$ are the same *if and only if* A is **symmetric and positive semi-definite**.
Why?
1. If A is symmetric, then its eigendecomposition exists:

$$A = Q\Lambda Q^T,$$

with Q orthogonal and Λ real diagonal.
2. If in addition A is positive semi-definite, then all eigenvalues $\lambda_i \geq 0$. Let $\Sigma = \sqrt{\Lambda}$, and define:

$$A = Q\Sigma\Sigma Q^T = (Q\Sigma)(Q)^T,$$

which is an SVD:

$$A = U\Sigma V^T,$$

with $U = Q$, $V = Q$, and Σ containing the singular values, which coincide with the eigenvalues in this case.

Conclusion:
The SVD and eigendecomposition of a real square matrix coincide if and only if the matrix is symmetric and positive semi-definite. ∎

Exercise 86. *Derive expression for AA^T and A^TA in terms of the SVD of A. In terms of the singular values, when are these matrices each invertible?*

Solution. Let $A \in \mathbb{R}^{n\times m}$, and let the Singular Value Decomposition (SVD) of A be:

$$A = U\Sigma V^T,$$

where:

- $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix (its columns are the left singular vectors),
- $V \in \mathbb{R}^{m\times m}$ is an orthogonal matrix (its columns are the right singular vectors),
- $\Sigma \in \mathbb{R}^{n\times m}$ is a rectangular diagonal matrix, with non-negative real numbers $\sigma_1, \dots, \sigma_r$ (the singular values) on the diagonal, where $r = \text{rank}(A)$.

1. Expression for AA^T:

$$AA^T = (U\Sigma V^T)(V\Sigma^T U^T) = U\Sigma\Sigma^T U^T.$$

Since $\Sigma\Sigma^T \in \mathbb{R}^{n\times n}$ is symmetric and diagonal with entries $\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{n-r}$, this shows:

$$AA^T = U(\Sigma\Sigma^T)U^T.$$

2. Expression for A^TA:

$$A^TA = (V\Sigma^T U^T)(U\Sigma V^T) = V\Sigma^T\Sigma V^T.$$

Again, $\Sigma^T\Sigma \in \mathbb{R}^{m\times m}$ is diagonal with entries $\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{m-r}$, so:

$$A^TA = V(\Sigma^T\Sigma)V^T.$$

3. When are AA^T and A^TA invertible?
- $AA^T \in \mathbb{R}^{n\times n}$ is invertible if and only if A has full row rank, i.e., $\text{rank}(A) = n$. This occurs when all $\sigma_i > 0$ for $i = 1, \dots, n$ and $n \leq m$.
- $A^TA \in \mathbb{R}^{m\times m}$ is invertible if and only if A has full column rank, i.e., $\text{rank}(A) = m$. This occurs when all $\sigma_i > 0$ for $i = 1, \dots, m$ and $m \leq n$.
In both cases, invertibility is equivalent to all the singular values σ_i being strictly positive. ∎

Exercise 87. *Show that the rank of a matrix A is equal to the rank of $A^\top A$ and $AA^\top$.*

Solution. Let $A \in \mathbb{R}^{n\times m}$. We want to show that

$$\text{rank}(A) = \text{rank}(A^\top A) = \text{rank}(AA^\top).$$

Step 1: Relate $\ker(A^\top A)$ **and** $\ker(A)$ Note that

$$A^\top Ax = 0 \iff x^\top A^\top Ax = \|Ax\|^2 = 0 \iff Ax = 0.$$

Therefore,

$$\ker(A^\top A) = \ker(A).$$

Applying the rank-nullity theorem to $A^\top A \in \mathbb{R}^{m\times m}$, we get

$$\operatorname{rank}(A^\top A) = m - \dim(\ker(A^\top A)) = m - \dim(\ker(A)) = \operatorname{rank}(A).$$

Step 2: Relate $\ker(AA^\top)$ **and** $\ker(A^\top)$ Similarly,

$$AA^\top y = 0 \implies y^\top AA^\top y = \|A^\top y\|^2 = 0 \implies A^\top y = 0,$$

which gives

$$\ker(AA^\top) = \ker(A^\top).$$

By the rank-nullity theorem for $AA^\top \in \mathbb{R}^{n\times n}$,

$$\operatorname{rank}(AA^\top) = n - \dim(\ker(AA^\top)) = n - \dim(\ker(A^\top)) = \operatorname{rank}(A).$$

Conclusion We conclude that

$$\operatorname{rank}(A) = \operatorname{rank}(A^\top A) = \operatorname{rank}(AA^\top).$$

■

Exercise 88. *Show that AA^T and A^TA have the same non-zero eigenvalues.*

Solution. Let $A \in \mathbb{R}^{n\times m}$. We aim to show that the matrices $AA^\top \in \mathbb{R}^{n\times n}$ and $A^\top A \in \mathbb{R}^{m\times m}$ have the same non-zero eigenvalues (with the same multiplicities).
Assume $\lambda \neq 0$ is an eigenvalue of $A^\top A$ with corresponding eigenvector $v \in \mathbb{R}^m$, i.e.

$$A^\top Av = \lambda v.$$

Multiplying both sides on the left by A, we get:

$$AA^\top(Av) = \lambda(Av).$$

If $Av \neq 0$, this shows that Av is an eigenvector of $AA^\top$ associated to the same eigenvalue $\lambda \neq 0$.
Conversely, assume $\lambda \neq 0$ is an eigenvalue of $AA^\top$ with eigenvector $u \in \mathbb{R}^n$, so

$$AA^\top u = \lambda u.$$

Multiplying both sides on the left by $A^\top$, we get:

$$A^\top A(A^\top u) = \lambda(A^\top u).$$

If $A^\top u \neq 0$, this shows that $A^\top u$ is an eigenvector of $A^\top A$ with eigenvalue λ.
Finally, observe that both $A^\top A$ and $AA^\top$ are symmetric and positive semi-definite, and both have rank equal to $\operatorname{rank}(A)$. Thus they have exactly the same number of non-zero eigenvalues (counted with multiplicity), and we have shown above that these eigenvalues must be the same.

■

Exercise 89. *In the proof of the SVD we showed that the columns of V (the right singular vectors) are the eigenvectors of $A^T A$. Show that the columns of U (the left singular vectors) are the eigenvectors of AA^T.*

Solution. Let $A \in \mathbb{R}^{n\times m}$, and suppose $A = U\Sigma V^\top$ is the singular value decomposition (SVD) of A, where:

- $U \in \mathbb{R}^{n\times n}$ has orthonormal columns (left singular vectors),
- $V \in \mathbb{R}^{m\times m}$ has orthonormal columns (right singular vectors),
- $\Sigma \in \mathbb{R}^{n\times m}$ is a diagonal matrix with non-negative entries $\sigma_1, \dots, \sigma_r$ (singular values), where $r = \text{rank}(A)$.

We now compute:

$$AA^\top = (U\Sigma V^\top)(V\Sigma^\top U^\top) = U\Sigma\Sigma^\top U^\top.$$

Since $\Sigma\Sigma^\top$ is a symmetric, positive semi-definite matrix of size $n\times n$, and U is orthogonal, it follows that $AA^\top = U(\Sigma\Sigma^\top)U^\top$ is the eigendecomposition of $AA^\top$. Therefore, the columns of U are the eigenvectors of $AA^\top$, and the eigenvalues are given by the squared singular values σ_i^2.

■

From these two exercises, notice that we have therefore obtained that the singular values of A are the square roots of the non-zero eigenvalues of AA^T and $A^T A$, and the right (left) singular vectors are the eigenvectors of $A^T A$ (AA^T).

Exercise 90. *Given a matrix A what parts of the SVD give an orthonormal basis for the column space of A, the row space of A, and the null space of A.*

Solution. Suppose A has rank r, and write

$$\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \quad \text{where} \quad \Sigma_r = \text{diag}(\sigma_1, \dots, \sigma_r), \quad \sigma_i > 0.$$

- **Column space of** A: The first r columns of U, i.e. $\{u_1, \dots, u_r\}$, form an orthonormal basis of $\text{Im}(A)$.
- **Row space of** A: The first r columns of V, i.e. $\{v_1, \dots, v_r\}$, form an orthonormal basis of the row space $\text{Im}(A^\top)$.
- **Null space of** A: The last $m - r$ columns of V, i.e. $\{v_{r+1}, \dots, v_m\}$, form an orthonormal basis of $\ker(A)$.

■

Exercise 91. *Show that the Frobenius norm of a matrix $A \in \mathbb{R}^{n\times m}$, defined as*

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2,$$

is equal to the sum of the squared singular values of A.

Solution. Let $A = U\Sigma V^\top$ be the singular value decomposition of A, where

$$\Sigma = \begin{bmatrix} \mathrm{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix}$$

with $r = \mathrm{rank}(A)$.
Recall that the Frobenius norm can be expressed as:

$$\|A\|_F^2 = \mathrm{trace}(A^\top A).$$

Now substitute the SVD of A:

$$\|A\|_F^2 = \mathrm{trace}\left((U\Sigma V^\top)^\top (U\Sigma V^\top)\right) = \mathrm{trace}\left(V\Sigma^\top U^\top U\Sigma V^\top\right).$$

Since U is orthogonal, $U^\top U = I$, so:

$$\|A\|_F^2 = \mathrm{trace}(V\Sigma^\top \Sigma V^\top).$$

Using the cyclic property of the trace:

$$\|A\|_F^2 = \mathrm{trace}(\Sigma^\top \Sigma V^\top V) = \mathrm{trace}(\Sigma^\top \Sigma),$$

because V is orthogonal, so $V^\top V = I$.
Since Σ is diagonal (except for zero blocks), $\Sigma^\top \Sigma$ is diagonal with entries σ_i^2, hence

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2.$$

This shows explicitly that the Frobenius norm is invariant under multiplication by orthogonal matrices, and that it equals the sum of the squares of the singular values. ∎

8.4 Neuro Q6: SVD & Deep Linear Networks

Motivation

Neural networks, either Chat-GPT or your brain, are hard to understand. They do many surprising things, and in general are very hard to analyse. A good route to understanding is proposing simple models and seeing how many of the observed phenomena can be seen in these understandable models.

We are going to analyse such a simple model, known as a deep linear network, introduced by our very own Andrew Saxe (Saxe et al. 2013). Despite their simplicity, it turns out a variety of surprises will pop out that seem to match the behaviour of both neural networks and animals.

In this question we will derive the learning dynamics of deep linear networks. In the process we will model cognitive findings, in particular the step-like emergence of concepts. When trying to understand SVD you rarely feel 50% good, then 49%, etc. Rather, you know nothing, you know a little bit, then it all clicks, and you basically understand the whole thing in one go.

Model Setup

Standard feedforward neural networks swallow an input vector and iteratively apply a transformation. This transformation does two things, (i) it projects the incoming vector through an affine transform, then (ii) it applies an element-wise nonlinearity:

$$f(x) = W_n\phi(W_{n-1}\phi(...\phi(W_2\phi(W_1x + b_1) + b_2)... + b_{n-1}) + b_n \tag{8.1}$$

An elementwise nonlinearity just means you take a function, $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$, e.g. the famous ReLU function, $\text{ReLU}(x) = \max(0, x)$, and apply it to a vector by applying it individually to each element of a vector, $\phi : \mathbb{R}^n \to \mathbb{R}^n$, $= \phi(x)$, $y_i = \tilde{\phi}(x_i)$. We're going to simplify this by removing the bias vectors, b, and the nonlinearities:

$$f(x) = W_nW_{n-1}...W_2W_1x \tag{8.2}$$

In fact, it turns out that analysing a two layer version of this will show most of the intracies required, generalising again back to many layers is not that hard, therefore we analyse the following model.

$$f(x) = W_2W_1x \tag{8.3}$$

So, neural networks are function approximators. They are given training data (example input-output pairs) and they have to change their parameters so that the function they embody (f(x)) matches the training data. This is just regression. The way this is usually done is through gradient descent: you measure the discrepancy between the true output and the neural network's guess, and you take the gradient with respect to the parameters. You then change the parameters to follow the gradient, decreasing the loss. (This should be a refresher, ask if you've never heard of these ideas)
We therefore assume access to a large dataset, $\{x_i, y_i\}_{i=1}^N$, that we will use to train the weight matrices W_1, W_2.

Question

1. Many people are surprised that deep linear networks are interesting (including initially Andrew), one reason for this is the nature of the input output mapping. What type of function does a deep linear network implement?

The training procedure is standard, for each datapoint the network's estimated output is calculated:

$$\hat{y}_i = W_2 W_1 x \tag{8.4}$$

Then the error is measured, the L2 distance between true and predicted output:

$$\mathcal{L}(W_1, W_2) = \frac{1}{2}||y_i - \hat{y}_i||_2^2 \tag{8.5}$$

Then the parameters are updated by following the gradient some distance:

$$\Delta W_1 = -\lambda \frac{\partial}{\partial W_1}\mathcal{L}(W_1, W_2) \tag{8.6}$$

$$\Delta W_2 = -\lambda \frac{\partial}{\partial W_2}\mathcal{L}(W_1, W_2) \tag{8.7}$$

2. Denoting $h_i = W_1 x_i$, the hidden layer activities when input x_i is provided, derive the following expressions for the weight update from equations 8.6 and 8.7:

$$\Delta_i W_1 = \lambda W_2^T (y_i - \hat{y}_i) x_i^T \tag{8.8}$$

$$\Delta_i W_2 = \lambda (y_i - \hat{y}_i) h_i^T \tag{8.9}$$

3. This was the update if one datapoint was provided. We will assume that the learning rate, λ, is very small so we can average the update equations over P datapoints where P is very large and approximately average over the distribution of the data. Show:

$$\Delta W_1 = \sum_{i=1}^{P} \Delta_i W_1 \approx P\mathbb{E}_i[\Delta_i W_1] = \lambda P W_2^T(\mathbf{\Sigma}_{xy} - W_2 W_1 \mathbf{\Sigma}_x) \tag{8.10}$$

$$\Delta W_2 = \sum_{i=1}^{P} \Delta_i W_2 \approx P\mathbb{E}_i[\Delta_i W_2] = \lambda P(\mathbf{\Sigma}_{xy} - W_2 W_1 \mathbf{\Sigma}_x) W_1^T \tag{8.11}$$

Where we have denoted $\mathbf{\Sigma}_{yx} = \mathbb{E}[yx^T]$, and $\mathbf{\Sigma}_x = \mathbb{E}[xx^T]$.

We take the continuum limit of this to turn these discrete update equations into continuous time differential equations, namely:

$$\tau \frac{d}{dt} W_1 = W_2^T(\mathbf{\Sigma}_{xy} - W_2 W_1 \mathbf{\Sigma}_x) \tag{8.12}$$

$$\tau \frac{d}{dt} W_2 = (\mathbf{\Sigma}_{xy} - W_2 W_1 \mathbf{\Sigma}_x) W_1^T \tag{8.13}$$

Where $\tau = \frac{1}{P\lambda}$.

4. Assume $\mathbf{\Sigma}_x = I$, and write the SVD of the input-output correlation as: $\mathbf{\Sigma}_{xy} = USV^T$. Rotate the bases to write the weight matrices in these bases, U and V:

$$W_1 = R\bar{W}_1 V^T \tag{8.14}$$

$$W_2 = U\bar{W}_2 R^T \tag{8.15}$$

Where R is some arbitrary orthogonal matrix. Show that the dynamics of these rotated matrices are nice and simple (hence the rotation):

$$\tau \frac{d}{dt}\bar{W}_1 = \bar{W}_2^T(S - \bar{W}_2\bar{W}_1) \tag{8.16}$$

$$\tau \frac{d}{dt}\bar{W}_2 = (S - \bar{W}_2\bar{W}_1)\bar{W}_1^T \tag{8.17}$$

5. These equations are hard to solve because they describe the coupled dynamics of two matrices, however we actually get lucky. Explain why if the rotated weight matrices are diagonal at the beginning of training they will remain diagonal.

6. Empirically, if the weights are initialised to very small values, then the off-diagonal elements actually decay to zero. (I don't think this is currently theoretically well understood) This means we can just study the behaviour of all the diagonal elements, and it turns out those decouple, so we can understand their simple dynamics in isolation!

   Calling the diagonal elements $c_\alpha = \bar{W}_{1,\alpha\alpha}$ and $d_\alpha = W_{2,\alpha\alpha}$ show that their dynamics decouple, i.e. show that:

$$\tau \frac{d}{dt}c_\alpha = d_\alpha(s_\alpha - c_\alpha d_\alpha) \tag{8.18}$$

$$\tau \frac{d}{dt}d_\alpha = c_\alpha(s_\alpha - c_\alpha d_\alpha) \tag{8.19}$$

   Where s_α is the αth diagonal element of S.

7. It turns out that if you initialise the weights to be very small (or if you add a term to the loss that penalises the L2 norm of the weight matrices) then the solution the neural network chooses is balanced, meaning $c_\alpha = d_\alpha$ throughout training!

   We can then track the dynamics of just one variable per mode (i.e. per α), $a_\alpha = c_\alpha d_\alpha$. Show the dyanmics are:

$$\tau \frac{d}{dt}a_\alpha = 2a_\alpha(s_\alpha - a_\alpha) \tag{8.20}$$

8. This is a differential equation which you might be able to solve. Denote the initial value of a_α as a_α^0. Show that (it requires doing a funky integral, google if you are struggling):

$$a_\alpha(t) = \frac{s_\alpha e^{\frac{2s_\alpha t}{\tau}}}{e^{\frac{2s_\alpha t}{\tau}} - 1 + \frac{s_\alpha}{a_\alpha^0}} \tag{8.21}$$

9. Plot this curve on your favourite plotting programme, and play with the behaviour of the curve as you vary the singular value s_α and other things.

Great, we've derived the key dynamics! And look! You should have found that the dynamics are step-like! Each singular value is learnt as its own concept, and how quickly it is learnt is determined by the size of the singular value, which roughly encodes how important this 'concept' was for understanding the data.
Let's emerge from the hole we've fallen down, and recover the solution to the initial problem. Make the matrix $A(t)$ which is diagonal with diagonal elements $a_\alpha(t)$. Then $A(t) = \bar{W}_2\bar{W}_1(t)$. We invert the initial change of basis to finally uncover that:

$$W_2W_1(t) = U\bar{W}_2\bar{W}_1(t)V^T = UA(t)V^T \tag{8.22}$$

Let's recall this solution required:

- Very slow learning rate, $\lambda << 1$
- Balanced solution, $c_\alpha = d_\alpha$
- Decoupled dynamics, offdiagonal elements decay to 0

The last two conditions are generally achieved if the weights are initialised very small.
See Saxe et al. 2019 for some of the fun cognitive phenomena that pop out of this model. We will just highlight one here. Neural networks and animals are observed to go through step-like learning transitions.

10. Repeat the same analysis as above for a one layer linear neural network:

$$f(x) = Wx \tag{8.23}$$

(don't worry it's similar but much simpler).

Show that the dynamics is completely described by a sum of exponentially decaying modes.

This last question asks you to show that shallow networks never show step-like transitions, they exponentially decay from one state to another. Conversely, the deep dynamics you derived can show step-like changes, as in the figure below. Ain't that funky! Nonlinear neural networks often shown similar phenomena.

Figure 8.3: Mean squared error of the loss shows phase like transitions.

Chapter 9

Short Note: Determinants

Depending on the basis you're working in a matrix can look very different, figure 9.1. Yet, somehow, in all bases (cases), there are a few behaviours that seem to be the same. This lecture, we're going to discover one such behaviour for square matrices: in all bases, a square linear map expands the space by the same amount, and this amount is called the determinant. You can see this illustrated intuitively in figure 9.1. It is easily an calculated, intepretable, basis-independent function of a linear map that pops up surprisingly often.

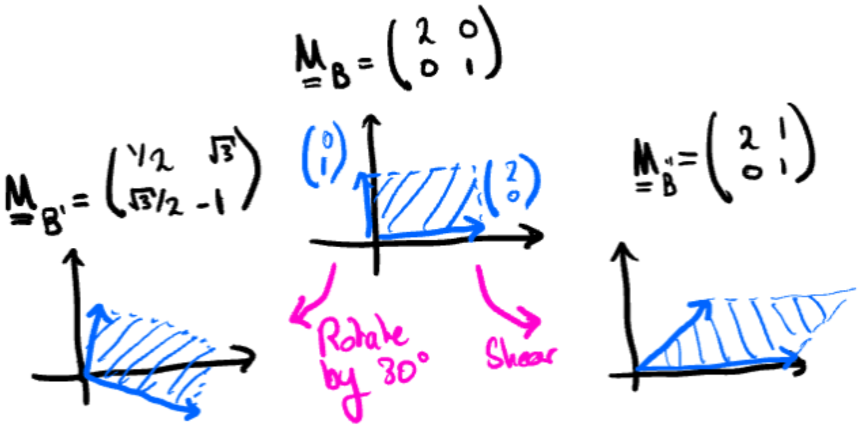


Figure 9.1: The matrix looks very different in these different bases. But take the basis vectors and map them through M, they will arrive at the blue arrows, the columns of M. The basis vectors define a unit square, the amount this square is expanded (the area of the blue shape) is the same in all bases!

To describe this object we're going to need a slightly weird definition of what we mean by volume, that is close but not the same as the normal daily-life notion. Our notion of volume will have a sign, i.e. it can be negative, and the sign will encode the orientation: e.g. two shapes of the same size but swapped orientation will have opposite signed volume, though their intuitive volume will be the same, figure. This quantity turns out to be the most natural way to characterise matrix volume expansions. One place where it will pop up is finding the eigenvalues of a network.

9.1 Determinant a Natural Expression of Volume

The determinant will be the unique function that satisfies these properties:

Theorem 5 (Existence and unicity of the determinant)**.** *Let B be basis of $\mathbb{R}^n$. Then there exists a unique map from $\mathbb{R}^{n\times n}$ (the space of n, n dimensional vectors) to $\mathbb{R}$ called det_B such that:*

- *det_B is $n-$linear - it is a linear function of each of the n input vectors;*
- *det_B is alternating, meaning that it is zero if two of the input vectors are the same, the determinant is zero;*
- *$det_B(B) = 1$.*

Proof. This theorem is admitted. □

We've described this as a function of n, n-dimensional vectors, measuring the volume traced out by these vectors. Often you'll see the determinant of a matrix, which

Informal Discussion

The determinant is supposed to embody the notion of volume contained within the interior of a basis, e.g., the volume of the set $\left\{\sum_{i=1}^{n} \lambda_i x_i : \lambda_i \in [0,1]\right\}$ and generalize it to arbitrary dimensions. It is an astonishing fact that there exists only one notion of volume that verify the two very basic desiderata for volume:

- (multi-)linearity: scaling a side by λ scales the volume by λ
- being alternated: if two sides are pointing in the same direction, then the volume is flat.

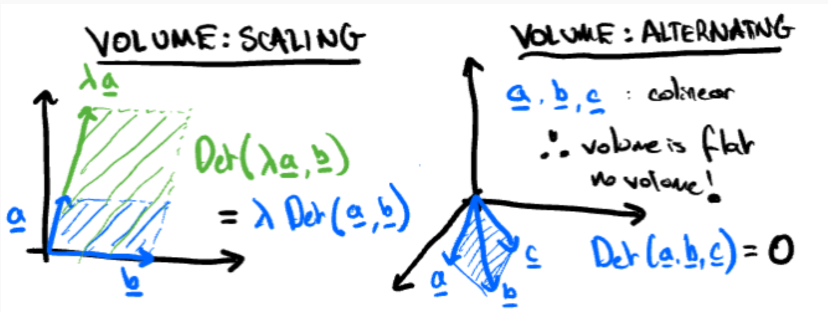


Figure 9.2: Two natural properties, that it scales linearly as you scale one side, and it is zero is the sides are co-linear, is enough to uniquely define volume!

Two remaining comments are that the last degeneracy, fixed by setting $det_B(B) = 1$, simply expresses the fact that we need to choose our units for the volume, once that is

done, the volume is unique! There is a different notion of volume if we work in metres, or in centimetres.
Finally, our linearity constraint is what makes this signed volume different to volume. If you multiply a vector by -1 the volume doesn't change, but the signed volume does. Since we care a lot about linearity, we'll stick with signed volume. It also pops up in some relatively surprising places. Often you'll see the absolute value of the determinant being used, this is more like the traditional notion of the volume.

A Series of Familiarizing Exercises

We'll now step through a series of examples. In each you have to use the properties of the determinant to show that it abides by the natural notions of volume. For all of them *det* should be taken to be with respect to the canonical basis.

Exercise 92 (Volume of a Cuboid). *Start in* $\mathbb{R}^3$*, let* $a_1, a_2, a_3 \in \mathbb{R}$*, show:*

$$det\left(\begin{bmatrix} a_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ a_2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ a_3 \end{bmatrix}\right) = a_1 a_2 a_3$$

Where det is taken to be the determinant in the canonical basis of $\mathbb{R}^3$*. Hence, it accords with the intuitive notion of volume.*

Exercise 93 (Volume of Flat Shape). *Let* $\{a_i\}_{i=1}^n$ *be a set of linearly dependent vectors in* $\mathbb{R}^n$*. Show that* $det(a_1, .., a_n) = 0$*.*

Exercise 94 (Volume unchanged by adding a column). *Now let* $\{a_i\}_{i=1}^n$ *be a set of vectors in* $\mathbb{R}^n$*. Show that if you add a multiple of one vector to another, it does not change the determinant.*

Exercise 95 (Volume of Sheared-Cuboid). *Now let* $\{a_i\}_{i=1}^n$ *be a set of vectors in* $\mathbb{R}^n$ *with the property that only the top i components of the* i^{th} *vector are non-zero. Show that the determinant is the product of the diagonal elements:*

$$det(a_1, ..., a_n) = (a_1)_1 ... (a_n)_n$$

This question is quite hard, try showing it in 2D using the previous exercise. Then try generalising. It also has a nice geometric intuition.

Exercise 96 (General Det in $\mathbb{R}^2$). *Show that:*

$$det(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix}) = ad - bc$$

There are 3 quite different ways to do this.

This last exercise is a useful property. Another one is the following, which we quote without proof:

Determinant of Product of Matrices

The determinant of a product of two square matrices is the product of the determinants of the matrix:

$$\det(AB) = \det(A)\det(B) \tag{9.1}$$

9.2 Finding eigenvalues and establishing diagonalizability

In the following, we investigate further how one can determine whether a matrix is diagonalizable, and how to find its eigenvalues. We first characterize eigenvalues of a matrix through the lens of (non-)injectivity and determinants. This is one of the major uses of the determinant.

Eigenvalues through non-injectivity

If λ is an eigenvector of u, then since there exists x such that $u(x) = \lambda x$, we have that

$$(u - \lambda \mathrm{Id})(x) = 0$$

e.g the function $u - \lambda \mathrm{id}_E$ is not injective.

Consequently, we can use the machinery of determinants to find eigenvalues, to characterize an eigenvalue:

Criterion for non-injectivity

Proposition 44. *Let $u \in \mathcal{L}(\mathbb{R}^n)$. Then we have:*

$$det(u) = 0 \iff u \text{ is not injective}$$

Worked Example

Let's find the eigenvalues of the following matrix using this approach $M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. We can do this by calculating the following determinant:

$$\det(M - \lambda I) = \det \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1) = 0 \tag{9.2}$$

Hence the eigenvalues are 1 and 3.

If we can solve this equation to find n distinct eigenvalues for an n dimensional square matrix, then it must be diagonalisable.

Chapter 10

Projections and Principal Component Analysis

10.1 Projections

10.1.1 Projection with complementary subspaces

Projection

Definition 47. *Let F and G be two subspaces of $\mathbb{R}^n$ that are complementary: $\mathbb{R}^n = F \oplus G$. $\forall x \in \mathbb{R}^n$, there is a unique pair $(x_F, x_G) \in F \times G$ such that $x = x_F + x_G$. The projection of a vector $x \in \mathbb{R}^n$ onto F along G, $p_{F,G}$, is defined as $p_{F,G}(x) = x_F$.*

Casually Explained

This is a very natural definition. Decomposing a space into two subspaces, the projection finds the component of any vector in one of those subspaces. Physically, you can think of it as mapping a vector to its shadow on F when illuminated by light aligned with G.

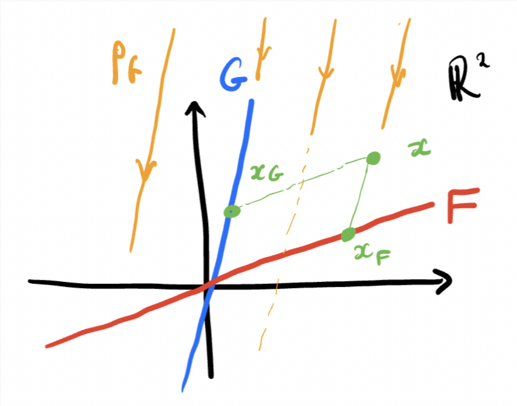


Figure 10.1: Under illumination by light aligned with G, the shadow of each vector x in the subspace F is the projection of x onto F, x_F.

Projections are super important and happen everywhere! In physics (measurement in quantum mechanics), in statistics (PCA, QR decomposition), and in general in the theory of Euclidean and Hilbert spaces.

Projection

Proposition 45. *Let F and G be complementary subspaces of $\mathbb{R}^n$, i.e., $\mathbb{R}^n = F \oplus G$, and let $p_{F,G} : \mathbb{R}^n \to F$ be the projection onto F along G as defined in Definition 47. Then $p_{F,G}$ is a linear map.*

Projection

Proof. Let $x, y \in \mathbb{R}^n$, and let $\lambda \in \mathbb{R}$. Since $\mathbb{R}^n = F \oplus G$, we can uniquely decompose each vector as

$$x = x_F + x_G, \quad y = y_F + y_G,$$

with $x_F, y_F \in F$ and $x_G, y_G \in G$. Then

$$x + y = (x_F + y_F) + (x_G + y_G),$$

where $x_F + y_F \in F$ and $x_G + y_G \in G$, so this is the unique decomposition of $x + y$ along $F \oplus G$. Therefore,

$$p_{F,G}(x + y) = x_F + y_F = p_{F,G}(x) + p_{F,G}(y).$$

Similarly, for scalar multiplication, we have

$$\lambda x = \lambda(x_F + x_G) = (\lambda x_F) + (\lambda x_G),$$

with $\lambda x_F \in F$, $\lambda x_G \in G$, so this is the unique decomposition of λx, and hence

$$p_{F,G}(\lambda x) = \lambda x_F = \lambda p_{F,G}(x).$$

Thus, $p_{F,G}$ preserves addition and scalar multiplication, and is therefore linear. $\square$

10.1.2 Orthogonal Projections

Projections require you to specify the two subspaces you are breaking the space into, p_F in definition 47 changes if you change G. Let's say you don't really care about G, really what you're interested in is F, then a very natural choice is to make $G = F^\perp$. This is the choice the orthogonal projection takes, and we will see it has nice properties that justify its naturalness. First we need to introduce the orthogonal complement of a subspace.

Orthogonal subspaces

Definition 48. *Let $F \subset \mathbb{R}^n$ be a vector subspace of $\mathbb{R}^n$. Then the space:*

$$\{y \in \mathbb{R}^n : \langle x, y\rangle = 0 \quad \forall x \in F\}$$

*is called the **orthogonal complement** of F, and is denoted by $F^\perp$.*

Proposition 46. *Let $F \subset \mathbb{R}^n$ be a vector subspace of $\mathbb{R}^n$. Then*

- *$F^\perp$ is a vector subspace of $\mathbb{R}^n$*
- *F and $F^\perp$ are in direct sum, and $F \oplus F^\perp = \mathbb{R}^n$.*
- *$(F^\perp)^\perp = F$*

Proof.

1. the proof for the first point is left as an exercise.
2. admitted.
3. the last point is left as an exercise.

□

Orthogonal Projection

Definition 49. *Let $F \subset \mathbb{R}^n$ be a subspace of $\mathbb{R}^n$, and $F^\perp$ be the orthogonal complement of F. For $x \in \mathbb{R}^n$, $x = x_F + x_\perp$, $x_F \in F$, $x_\perp \in F^\perp$, we define the orthogonal projection onto F, or just projection onto F as $p_F(x) = x_F$.*

Projection onto Orthogonal Basis

Proposition 47. *Let $F \subset \mathbb{R}^n$ be a subspace of $\mathbb{R}^n$ of dimension p. Let $\{u_1, ..., u_p\}$ be an orthonormal basis of F. Then the projection of x onto F is given by:*

$$p_F(x) = \sum_{i=1}^{P} \langle x, u_i\rangle u_i$$

Proof

Proof. Since $\{u_1, \dots, u_p\}$ is an orthonormal basis of F, any vector $y \in F$ can be uniquely written as $y = \sum_{i=1}^{p} \alpha_i u_i$. We now decompose $x \in \mathbb{R}^n$ as:

$$x = y + z, \quad \text{with } y \in F \text{ and } z \in F^{\perp}.$$

Then $y = p_F(x)$ by definition of orthogonal projection. We determine the coefficients α_i using orthonormality:

$$\langle x, u_i \rangle = \langle y + z, u_i \rangle = \langle y, u_i \rangle + \langle z, u_i \rangle = \alpha_i + 0 = \alpha_i,$$

since $z \in F^{\perp}$ and $u_i \in F$. Therefore,

$$p_F(x) = y = \sum_{i=1}^{p} \langle x, u_i \rangle u_i.$$

□

Another view of orthogonal projections in vector spaces is as finding the closest point within a subspace to the original vector.

Proposition 48. *Let F be a subspace of $\mathbb{R}^n$, and let p_F denotes the orthogonal projection on F. We have,*

$$\begin{aligned} p_F \colon \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto \arg\min_{y \in F} ||y - x||. \end{aligned}$$

Furthermore, $p_F(x)$ is the unique minimizer of $||y - x||$ on F.

Proof. Fix $x \in \mathbb{R}^n$ and let $x = p_F(x) + x_{F^{\perp}}$ be its orthogonal decomposition in $F \oplus F^{\perp}$. Then for all $y \in F$,

$$\|x - y\|^2 = \|(p_F(x) - y) + x_{F^{\perp}}\|^2 = \|p_F(x) - y\|^2 + \|x_{F^{\perp}}\|^2 \geq \|x_{F^{\perp}}\|^2.$$

Therefore, $y = p_F(x)$ is the unique minimizer of the minimization of $||y - x||$ on F. □

10.2 Principal Component Analysis - PCA

Principal Component Analysis is a workhorse of modern data analysis and particularly neuroscience. It is an application of all the ideas that we have covered so far.

Casually Explained

One motivation for PCA is the prevalence of high-dimensional data, which is very hard to interpret. A reasonable way to try to proceed is to find some subspace in which the phenomena of interest are happening, and to work in this lower-dimensional subspace.

Following this motivation, given a dataset $\{x_1, ..., x_n\}$ where $x_i \in \mathbb{R}^p$, PCA finds the vector subspace of dimensions $d \leq p$ such that the projection of $\{x_1, ..., x_n\}$ onto this subspace has maximal variance.

PCA Problem Formulation

- Given a dataset $\{x_1, ..., x_n\}$ of vectors in $\mathbb{R}^p$, let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ be their mean. We measure the variance of this set as:

$$\mathbb{V}(\{x_1, ..., x_n\}) = \frac{1}{n}\sum_{i=1}^n ||x_i - \bar{x}||^2$$

- Let $F \subseteq \mathbb{R}^n$ be a subspace of dimension d. PCA aims to find the F that maximises the projected variance, i.e.:

$$\underset{F \subset \mathbb{R}^n, \dim(F)=d}{\arg\max} \frac{1}{n}\sum_{i=1}^n ||p_F(x_i) - p_F(\bar{x})||^2$$

- Optimising over subspaces sounds hard, luckily we can map this to optimisation over vectors using a basis, i.e. we map this problem into a constrained Euclidean optimisation.

  As discussed earlier to any $F \subseteq \mathbb{R}^p$ of dimension d we can associate an orthonormal family of vectors $\{e_1, ..., e_d\}$ such that span$(\{e_1, ..., e_d\}) = F$. Then, $\forall x \in \mathbb{R}^p$, $p_F(x) = \sum_{i=1}^d \langle x, e_i\rangle e_i$. Therefore we can rewrite our PCA optimisation problem as:

$$\underset{\{e_1,...,e_d\} \text{ ONF}}{\arg\max} \frac{1}{n}\sum_{i=1}^n \left\| \sum_{j=1}^d \langle x_i, e_j\rangle e_j - \sum_{j=1}^d \langle \bar{x}, e_j\rangle e_j \right\|^2$$

  Where we have used the terminology ONF to denote an orthonormal family: a set of vectors $\{e_1, ..., e_d\}$ such that $\langle e_i, e_j\rangle = \delta_{ij}$.

- By bilinearity of the inner product, and orthonormality of the basis we can develop this:

$$\underset{\{e_1,...,e_d\}, \langle e_i, e_j\rangle = \delta_{ij}}{\arg\max} \frac{1}{n}\sum_{i=1}^n \left\| \sum_{j=1}^d \langle x_i - \bar{x}, e_j\rangle e_j \right\|^2 = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d \langle x_i - \bar{x}, e_j\rangle^2$$

- We are going to rewrite this by creating the matrix $X \in \mathbb{R}^{p\times n}$ that is formed from stacking the vectors x_i along each of the columns. We also denote the $p \times n$ matrix made by stacking n columns of the mean $\bar{x}$ as $\bar{X}$. With these, we can rewrite the previous expression as:

$$\begin{aligned}\underset{\{e_1,...,e_d\}, \langle e_i, e_j\rangle = \delta_{ij}}{\arg\max} \frac{1}{n}\sum_{j=1}^d \langle (X - \bar{X})^\top e_j, (X - \bar{X})^\top e_j\rangle^2 &= \frac{1}{n}\sum_{j=1}^d e_j^\top (X - \bar{X})(X - \bar{X})^\top e_j \\ &= \sum_{j=1}^d e_j^\top C_X e_j\end{aligned}$$

Where we have defined the covariance matrix as is typical, $C_X = \frac{1}{n}(X - \bar{X})(X - \bar{X})^\top$. Why is this the natural definition, well:

$$\begin{aligned}(C_X)_{ij} &= \frac{1}{n}\sum_{k=1}^{n}(X_{ik} - \bar{X}_{ik})(X_{jk} - \bar{X}_{jk}) \\ &= \frac{1}{n}\sum_{k=1}^{n}((x_k)_i - \bar{x}_i)((x_k)_j - \bar{x}_j)\end{aligned}$$

which is the covariance between the dimension i and the dimension j across the n vectors.

This has been a very interesting rewrite, as we have now arrived at a matrix with very convenient properties.

Lemma

C_X is a symmetric matrix with nonnegative eigenvalues.

Proof

Proof. Exercise. □

This result takes us to our final PCA theorem, which is a cracking result that should make you deeply happy:

PCA performs covariance eigendecomposition

Theorem 6. *The PCA problem:*

$$\underset{\{e_1,\ldots,e_d\},\langle e_i,e_j\rangle=\delta_{ij}}{\arg\max} \sum_{j=1}^{d} e_j^\top C_X e_j$$

attains its maximum for $\{e_1, \ldots, e_d\} = \{f_1, \ldots, f_d\}$ *when* $\{f_1, \ldots, f_d\}$ *are the eigenvectors of* C_X *associated with the* d *largest eigenvalues.*

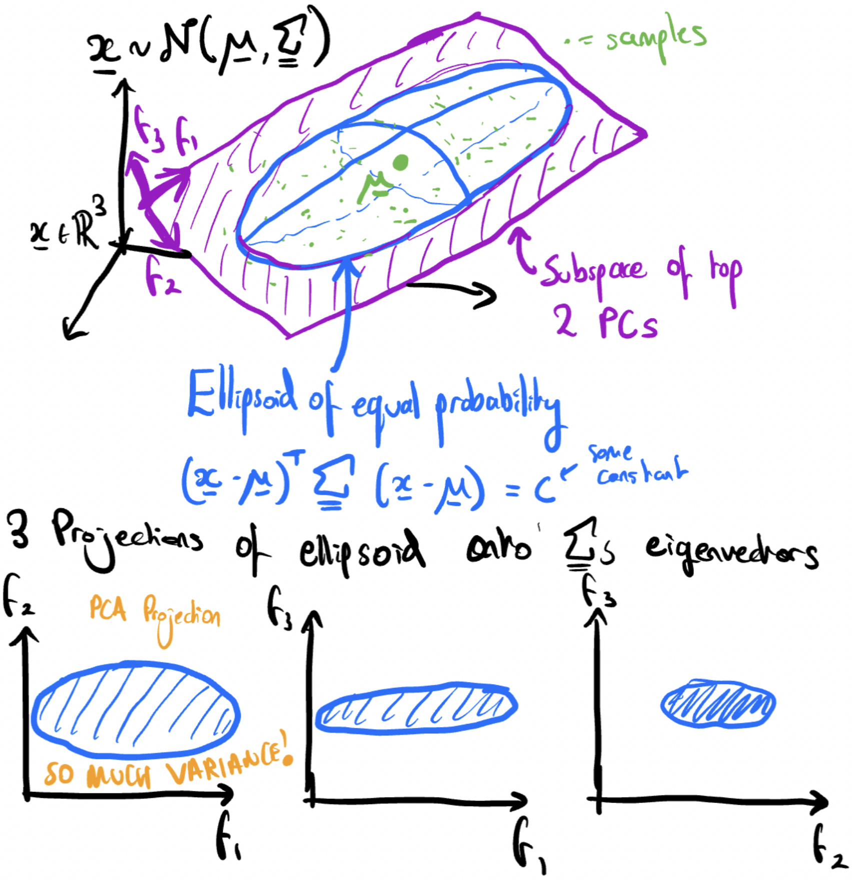


Figure 10.2: Data is normally distributed in 3D with mean μ and covariance Σ. Equally likely points under this distribution form an ellipsoid, the axes of this ellipsoid are the eigenvectors of Σ and their size is the associated eigenvalues. 2-dimensional PCA finds the plane onto which the projected data has maximal variance, intuitively, it makes sense that this is the first two eigenvectors of the distribution, given that they form the two largest axes of the ellipse.

Proof

Proof by iterative maximization. Let $C_X \in \mathbb{R}^{p \times p}$ be the covariance matrix of a centered dataset X, i.e., $C_X = \frac{1}{n} X^\top X$. Since C_X is symmetric and positive semi-definite, it admits an orthonormal basis of eigenvectors $\{f_1, \ldots, f_p\}$, with associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
We aim to solve the PCA optimization problem:

$$\max_{\substack{e_1, \ldots, e_d \in \mathbb{R}^p \\ \text{orthonormal}}} \sum_{j=1}^{d} \langle e_j, C_X e_j \rangle$$

We proceed by successively optimizing each vector e_j under orthogonality constraints with respect to the previous ones.

Step 1: Optimize e_1 We want to find:

$$e_1^* = \arg \max_{\|e_1\|=1} \langle e_1, C_X e_1 \rangle$$

Decomposing e_1 in the eigenbasis:

$$e_1 = \sum_{i=1}^{p} \alpha_i f_i, \quad \text{with } \sum_{i=1}^{p} \alpha_i^2 = 1$$

Then,

$$\langle e_1, C_X e_1 \rangle = \sum_{i=1}^{p} \lambda_i \alpha_i^2 \leq \lambda_1$$

Equality is achieved when $\alpha_1 = 1$, i.e., $e_1 = f_1$. So the optimal first direction is:

$$e_1^* = f_1$$

Step 2: Optimize $e_2 \perp e_1$ Now, we search for:

$$e_2^* = \arg \max_{\substack{\|e_2\|=1 \\ \langle e_2, e_1^* \rangle = 0}} \langle e_2, C_X e_2 \rangle$$

Since $e_1^* = f_1$, we restrict our search to the subspace orthogonal to f_1, spanned by $\{f_2, \ldots, f_p\}$. The same reasoning as in Step 1 gives:

$$\langle e_2, C_X e_2 \rangle = \sum_{i=2}^{p} \lambda_i \alpha_i^2 \leq \lambda_2$$

with equality if $e_2 = f_2$, hence:

$$e_2^* = f_2$$

General Step $j = 1, \ldots, d$ At step j, assume we have already found $e_1^* = f_1, \ldots, e_{j-1}^* = f_{j-1}$. We now look for:

$$e_j^* = \arg \max_{\substack{\|e_j\|=1 \\ \langle e_j, f_i \rangle = 0,\, i<j}} \langle e_j, C_X e_j \rangle$$

We restrict our search to the subspace orthogonal to $\mathrm{Span}(f_1, \ldots, f_{j-1})$, i.e., spanned by $\{f_j, \ldots, f_p\}$. The same argument applies:

$$\langle e_j, C_X e_j \rangle = \sum_{i=j}^{p} \lambda_i \alpha_i^2 \leq \lambda_j$$

with equality if $e_j = f_j$. So the optimal choice is:

$$e_j^* = f_j$$

Conclusion: The orthonormal family $\{e_1^*, \ldots, e_d^*\} = \{f_1, \ldots, f_d\}$ maximizes the PCA objective:

$$\sum_{j=1}^{d} \langle e_j^*, C_X e_j^* \rangle = \sum_{j=1}^{d} \lambda_j$$

This is the maximum total variance that can be captured by projecting onto a d-dimensional subspace. Therefore, the top d eigenvectors of the covariance matrix form the optimal principal directions. $\square$

10.3 Exercise Sheet 10: Projections and PCA

Important Exercises!

Exercise 97. *F and G are complementary subspaces of $\mathbb{R}^n$, $p_{F,G}$ is the projection onto F along G. Show that Im$(p_{F,G}) = F$, and ker$(p_{F,G}) = G$.*

Solution. Let $x \in \mathbb{R}^n$. Since $\mathbb{R}^n = F \oplus G$, there exists a unique decomposition $x = f + g$ with $f \in F$, $g \in G$. By definition, $p_{F,G}(x) = f$.
Image: For any $x \in \mathbb{R}^n$, we have $p_{F,G}(x) = f \in F$, so $\mathrm{Im}(p_{F,G}) \subseteq F$. Conversely, for any $f \in F$, write $f = f + 0 \in F + G$, and so $p_{F,G}(f) = f$. Hence, $F \subseteq \mathrm{Im}(p_F)$, and we conclude:

$$\mathrm{Im}(p_{F,G}) = F.$$

Kernel: Let $x \in \ker(p_{F,G})$. Then $p_{F,G}(x) = 0$, so $f = 0$, and $x = g \in G$, hence $\ker(p_{F,G}) \subseteq G$. Conversely, for any $g \in G$, write $g = 0 + g$, then $p_{F,G}(g) = 0$, so $g \in \ker(p_{F,G})$. Therefore,

$$\ker(p_{F,G}) = G.$$

■

Exercise 98. *Show that any projection, $u : \mathbb{R}^n \to \mathbb{R}^n$, satisfy $u \circ u = u$.*

Exercise 99. *Show that any linear map, $u : \mathbb{R}^n \to \mathbb{R}^n$, for which $u \circ u = u$ is a projection onto Im(u) along ker(u)*

Solution. Let $u : \mathbb{R}^n \to \mathbb{R}^n$ be a linear map such that $u \circ u = u$ (such a map is called idempotent). We claim that u is the projection onto $\mathrm{Im}(u)$ along $\ker(u)$. To prove this, we show that $\mathrm{Im}(u)$ and $\ker(u)$ are complementary subspaces of $\mathbb{R}^n$ by showing that every vector $x \in \mathbb{R}^n$ can be uniquely written as

$$x = u(x) + (x - u(x)),$$

where $u(x) \in \mathrm{Im}(u)$ and $x - u(x) \in \ker(u)$.
1. $u(x) \in \mathrm{Im}(u)$ by definition.
2. We check that $x - u(x) \in \ker(u)$:

$$u(x - u(x)) = u(x) - u(u(x)) = u(x) - u(x) = 0,$$

so $x - u(x) \in \ker(u)$.
3. This decomposition is unique: if $x = y + z$ with $y \in \mathrm{Im}(u)$ and $z \in \ker(u)$, then applying u gives

$$u(x) = u(y + z) = u(y) + u(z) = y + 0 = y,$$

so $y = u(x)$ and $z = x - u(x)$.
Therefore, $\mathbb{R}^n = \mathrm{Im}(u) \oplus \ker(u)$, and u acts as the projection onto $\mathrm{Im}(u)$ along $\ker(u)$. ■

Exercise 100. *Show that $F^\perp$ is a vector subspace of $\mathbb{R}^n$ and $(F^\perp)^\perp = F$.*

Exercise

Exercise 101. *Show that* $u : \mathbb{R}^n \to \mathbb{R}^n$ *such that* $u(x) = x - \langle a, x\rangle a$ *with* $a \in \mathbb{R}^n, \|a\| = 1$ *is a projection. Show it is an orthogonal projection. What is the corresponding subspace (that is being projected onto)? Hint: use the result of the previous exercise.*

Solution. **1.** u **is a projection:**
We compute $u(u(x))$:

$$u(u(x)) = u\left(x - \langle a, x\rangle a\right) = (x - \langle a, x\rangle a) - \langle a, x - \langle a, x\rangle a\rangle\, a.$$

We expand the inner product:

$$\langle a, x - \langle a, x\rangle a\rangle = \langle a, x\rangle - \langle a, x\rangle\langle a, a\rangle = \langle a, x\rangle(1 - \|a\|^2) = 0.$$

So:

$$u(u(x)) = x - \langle a, x\rangle a = u(x),$$

so u is a projection by the result of the previous exercise.
2. u **is an orthogonal projection:**
To show that u is orthogonal we show that $\mathrm{Ker}(u)$ and $\mathrm{Im}(u)$ are orthogonal to each other. That is, for all $x \in \mathbb{R}^n$ and all $y \in \mathrm{Ker}(u)$, we have $\langle u(x), y\rangle = 0$. Note that $y \in \mathrm{Ker}(u)$ implies

$$y = \langle a, y\rangle a \in \mathrm{span}(a),$$

Furthermore,

$$u(x) = x - \langle a, x\rangle a \in \{a\}^\perp,$$

since

$$\langle x - \langle a, x\rangle a, a\rangle = \langle a, x\rangle - \langle a, x\rangle\|a\|^2 = 0,$$

using $\|a\| = 1$. Thus $u(x) \in \{a\}^\perp$, and $y \in \mathrm{span}(a)$, implying $\langle u(x), y\rangle = 0$.
3. Subspace onto which u **projects:**
We conclude that u is the orthogonal projection onto the hyperplane:

$$\mathrm{Im}(u) = \{a\}^\perp = \{x \in \mathbb{R}^n \mid \langle x, a\rangle = 0\},$$

i.e., the subspace orthogonal to the vector a. ■

Exercise 102. *Let* $F \subset \mathbb{R}^n$ *be a subspace of* $\mathbb{R}^n$*, and* $p_{F,G}$ *the (not necessarily orthogonal) projection onto* F *along some subspace,* G*, that is complementary with* F*. Show that* $p_{F,G}$ *being orthogonal is equivalent to* $||p_{F,G}(x)|| \leq ||x||, \forall x \in \mathbb{R}^n$*.*

Solution. **(⇒) If** $p_{F,G}$ **is orthogonal, then** $\|p_{F,G}(x)\| \leq \|x\|$**:**
If $p_{F,G}$ is the orthogonal projection, then for each $x \in \mathbb{R}^n$, we have the decomposition $x = f + g$ with $f \in F$, $g \in F^\perp$, and $p_{F,G}(x) = f$. Since $f \perp g$, we have:

$$\|x\|^2 = \|f + g\|^2 = \|f\|^2 + \|g\|^2 \geq \|f\|^2 = \|p_{F,G}(x)\|^2.$$

Taking square roots yields $\|p_{F,G}(x)\| \leq \|x\|$, as required.
(⇐) If $\|p_{F,G}(x)\| \leq \|x\|$ **for all** x**, then** $p_{F,G}$ **is orthogonal:**
Assume $\{f_1, \ldots, f_p\}$ is an orthonormal basis of F.

Suppose the projection p_F is not orthogonal. Then, by definition, the subspace G along which we project is not the orthogonal complement of F. Hence, there exists a vector $x \in G$ and some $f_k \in F$ such that

$$\langle x, f_k \rangle \neq 0.$$

Let f_k be one such basis vector of F, and consider the vector

$$y = f_k + \lambda x,$$

for some scalar $\lambda \in \mathbb{R}$. Since p_F projects onto F along G, the projection of y is

$$p_F(y) = f_k,$$

because $x \in G$. Now, compute the norm squared:

$$\|y\|^2 = \|f_k + \lambda x\|^2 = \|f_k\|^2 + 2\lambda\langle f_k, x\rangle + \lambda^2\|x\|^2 = 1 + 2\lambda\langle f_k, x\rangle + \lambda^2\|x\|^2,$$

using that f_k is unit norm. Because $\langle f_k, x\rangle \neq 0$, we can choose λ with the opposite sign to $\langle f_k, x\rangle$ and sufficiently large magnitude such that

$$\|y\|^2 < 1,$$

meaning

$$\|y\| < \|p_F(y)\|.$$

This contradicts the assumption that

$$\|p_F(y)\| \leq \|y\| \quad \text{for all } y.$$

Therefore, if the norm inequality

$$\|p_F(x)\| \leq \|x\| \quad \forall x$$

holds, then necessarily

$$\langle x, f_i \rangle = 0 \quad \forall x \in G, \forall f_i \in F,$$

which means

$$G = F^{\perp},$$

and the projection p_F is orthogonal. ■

Exercise: PCA

Exercise 103. *Let $A \in \mathcal{L}(\mathbb{R}^n)$ be a symmetric matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. In Exercise 73 we showed that $\lambda_i \geq 0$ for all $i = 1, \ldots, n$ implies that A is positive semi-definite (if you have not done this exercise, now is a good time). Show the reverse: if A is positive semi-definite, then $\lambda_i \geq 0$ for all $i = 1, \ldots, n$.*

Exercise 104. *Show that C_X is a symmetric matrix. Use the previous exercise to show that C_X only admits nonnegative eigenvalues.*

Exercise 105. *Let* $\{u_1, \ldots, u_d\}$ *be a set of orthogonal vectors in* $\mathbb{R}^n$*, show that:*

$$\left\| \sum_{i=1}^{d} u_i \right\|^2 = \sum_{i=1}^{d} \|u_i\|^2.$$

If $\{u_1, \ldots, u_d\}$ *are orthonormal, what does the sum reduce to?*

Exercise 106. *Load up Neural_Code.npy . Hidden in there is a message that a set of neurons are trying to send you. Use PCA to find this message.*

10.4 Neuro Q7: PCA & Synaptic Learning Rules

The brain is a consummate learner; how is it that this mush of tissue is able to meaningfully adapt to changes in the world around it? Seminal work has discovered that one biochemical basis of learning is changes in the strength of synapses between neurons. You're all familiar with how synapses work: the pre-synaptic neuron emits a spike, this triggers a chemical cascade to cross the synapse, which in turn triggers a voltage change in the postsynaptic neuron. Biology can exploit this to learn. Through changing many detailed features of the synapse, the same presynaptic spike can cause a larger or smaller effect on the postsynaptic neuron. Following a common approach, we will wrap all of this synpatic gobbledegook into the idea of synaptic weight, a single number that stands in for the strength of a connection between neurons, and changing this will be our model of learning. Let's think about a simple neuron with firing rate $y_t \in \mathbb{R}$ connected to a sensory population with firing rates $x_t \in \mathbb{R}^N$, via weight vector $w \in \mathbb{R}^N$:

$$y_t = w_t^T x_t \tag{10.1}$$

x_t represents the firing of the input neurons at time t, and w_t represents the synaptic weights. When you're born your neurons don't know how to arrange themselves (even if you're a fly, which you're not), so each of these synpatic weights needs to change to make the system as a whole do something interesting. A classic approach was suggested by Donald Hebb in the 40s: if two neurons are co-firing a lot they should be connected ("Fire together, wire together!") Let's formalise that in our model:

$$\Delta w_t = \eta x_t y_t \tag{10.2}$$

i.e. the change in a synaptic weight at time t is equal to a learning rate, η, times by the product of pre and post-synaptic firing rates. This rule says that if two neurons activate together, they should become more connected!

1. Assume that η is very small. This means that before the weights change very much you will have likely seen many different input-output pairs (x_t, y_t). Rather than computing the weight change for a particular stream of pairs, you can take the average over the distribution. Further, assume the input data are mean zero. Show that, on average, Hebbian weight updates are given by:

   $$\langle \Delta w_t \rangle = \eta C_{xx} w_t \tag{10.3}$$

   Where C_{xx} is the input data covariance matrix.

2. You know a lot about the eigenstructure of covariance matrices. Use this to argue that Hebb's rule is actually a bad rule as it will lead weights to do uninteresting things, in particular, exploding.

3. Despite this, what interesting subspace will the weight vector increasingly lie in as it follows Hebb's rule for a long time?

Okay, so Hebb's rule is out. Thankfully a chap called Oja had a think about this and came up with a better rule. He proposed:

$$\Delta w_t = \eta(x_t y_t - y_t^2 w_t) \tag{10.4}$$

You can see the first term is just Hebb's rule, but clever old Oja added a second term. This second term is going to stop the Hebbian bit from exploding, as you might already be able to guess from its form (it shrinks you along the current weight vector - good for stability!). It also obeys the key tenet of synaptic learning rules - locality! A synapse is a physical object, a mass of chemicals somewhere in your brain. It only has information about what is happening nearby, i.e. what the presynaptic neuron is doing, what the postsynaptic neuron is doing, and what happening within the synapse. You can't ask a synapse in your visual cortex to change weights according to a rule that depends on the synapses in your cerebellum. (Meditate and justify to yourself that Oja's rule is indeed local)

4. Do the averaging over datapoints again and show that the averaged learning rule can be written:

$$\Delta w_t = \eta(C_{xx}w_t - (w_t^T C_{xx} w_t)w_t) \tag{10.5}$$

5. First lets show this solves exploding weight problem. Study the behaviour of $||w_{t+1}||_2 - ||w_t||_2$. Since the learning rate is low only consider terms up to order η (i.e. because η is so small treat all the terms multiplied by η^2, η^3 etc. as 0). Show that the dynamics cause the length of the weight vector to converge to 1.

6. Study the dynamics of the components of w in the eigenbasis. If all the eigenvalues of C_{xx} are distinct show that only eigenvectors are fixed points under these dynamics. (What happens if C_{xx} has repeated eigenvalues?)

Wow, this is pretty cool, these random neurons are extracting principal components! But which principal component...?
You showed that all the eigenvectors were fixed points of the dynamics, i.e. if you start at an eigenvector you will stay there. Now what happens if you are perturbed slightly from one of these eigenvectors? Does the dynamics push you back to where you started (called a stable fixed point) or does it push you away (called unstable). The best analogy here is something like a freely swinging pendulum. This has two fixed points: when the pendulum is pointing straight up, or straight down. In either situation the forces balance, and it is a fixed point of the dynamics. But, if you perturb slightly from these two fixed points, very different things occur. For obvious reasons, we only really care about the stable fixed points, the system will never settle on the unstable ones. Let's find the stable fixed points of Oja's rule.

7. Assume the weight vector is sitting on an eigenvector, and is perturbed slightly by some noise vector of small magnitude. Decompose the noise vector in the eigenbasis and study the dynamics of the weight vector to first order in the noise. Show that the fixed point is only stable if the weight vector is sitting on the largest eigenvector.

Look at that, that neuron is extracting the first principal component of the data!! And from such a simple local rule!
Some interesting extensions to this are networks of neurons that extract many principal components, or how to choose η based on the rate of change of the distribution of data. Both of these were the topics of theoretical neuroscience papers from the last 20 years. There are learning rules that make the network do all sorts of things. That said, our understanding is very nascent, and all the best neural network models of the brain do not use such unsupervised learning rules.

Chapter 11

Least Squares, Pseudoinverses, & Regression

11.1 Least Squares

Motivation

During this class, we often came across the following linear relationship between two vectors x and y, of not necessarily the same dimensionality:

$$y = Ax \tag{11.1}$$

If you know A and x then getting y is a simple matrix-vector multiplication.

If you know A and y then, as we've discussed, there might be no x (an over-determined problem, more constraints than variables), a unique x, or infinitely many x that solve this equation (an under-determined problem, fewer constraints than variables), depending on A.

However, in many practical cases we might want to relax the kinds of solutions we're looking for. Rather than finding the perfect choice of x for which $Ax = y$, which might not exist, perhaps we can find a good choice of x, one for which Ax is close to y?

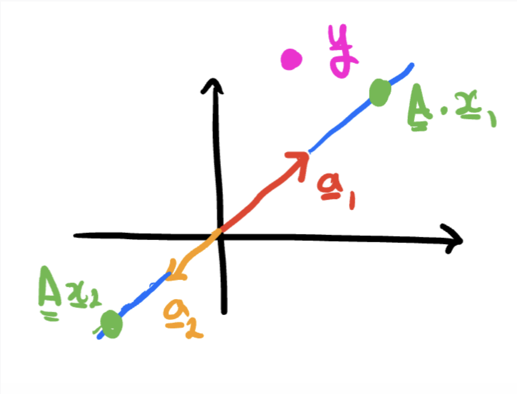


Figure 11.1: $y, x \in \mathbb{R}^2$, $A \in \mathbb{R}^{2\times 2}$, and we're failing to find an x such that $Ax = y$. This is because A is rank 1, and y is outside the image of A. However, despite this, there is still a meaningful way in which x_1 seems a much better guess than x_2, even if it is not perfect, because Ax_1 gets closer to y than Ax_2.

Let's find the input, x, whose prediction Ax, is closest to y. Measuring the distance between y and Ax requires choosing a norm. A reasonable choice is the L2 norm, hence, this the name of this section, Least Squares.

Least Squares Problem Formulation

For a given x, we define the *residual* as :

$$r(x) = y - Ax \tag{11.2}$$

Which characterizes the discrepancy between the y yielded by our current estimate x of the solution to Eq. (11.1) and the true y. Our goal is to minimize the length of r, yielding the following optimization problem:

$$\begin{aligned}&\underset{x\in\mathbb{R}^n}{\text{minimize}}\,\|r(x)\|^2\\ &=\underset{x\in\mathbb{R}^n}{\text{minimize}}\,\|y-Ax\|^2\end{aligned}$$

This is called the least squares problem. By analogy with Proposition 48, we see that the least squares problem perform a projection onto $\text{Im}(A)$:

$$\underset{z\in\text{Im(A)}}{\text{minimize}}\,\|y-z\|^2$$

Casual Explanation

Intuitively: y can be broken down into two orthogonal components, $y_{\text{Im}(A)} \in \text{Im}(A)$ and $y_{\text{Im}(A)^\perp} \in \text{Im}(A)^\perp$:

$$y = y_{\text{Im}(A)} + y_{\text{Im}(A)^\perp}$$

The best you can do is when the residual $r = Ax - y$ is orthogonal to $\text{Im}(A)$, and hence $y_{\text{Im}(A)} - Ax = 0$. You can see this in the figures below, if this is not the case you can always move closer by reducing the component of the residual within $\mathrm{Im}(A)$. We'll now prove this formally, and use it to get an explicit formula for the best fitting x.

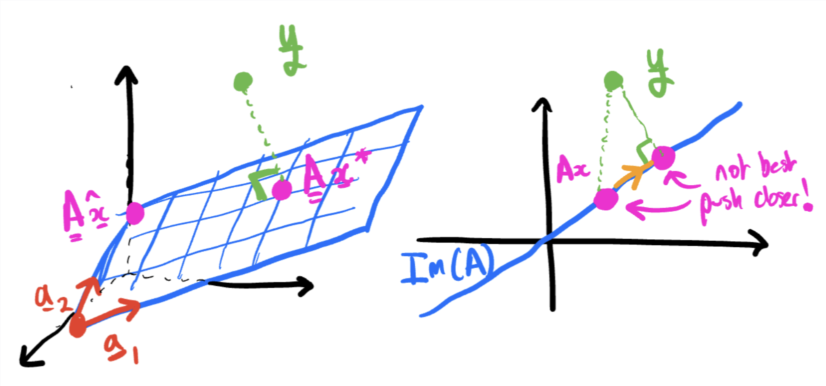


Figure 11.2: Two examples, in 3D and 2D, to illustrate that the closest you can get is when the residual is orthogonal to the image of A. In the left example a_1 and a_2 are the columns of A, hence they span the image of A.

Theorem 7. *A point $x^\star$ is a minimizer of the least squares problem if and only if $r(x^*) \in \mathit{Im}(A)^\perp$. Call p_A the orthogonal projection onto Im(A). Then:*

$$p_A(y) = Ax^*$$

Proof

The fact that any minimizer x^* of the least squares problem satisfies $p_A(y) = Ax^*$ is obtained from Proposition 48. We have:

$$r(x) = y - Ax = \underbrace{(y_{\mathrm{Im}(A)} - Ax)}_{\in \mathrm{Im}(A)} + \underbrace{y_{\mathrm{Im}(A)^\perp}}_{\in \mathrm{Im}(A)^\perp}$$

Therefore, $p_A(r(x))) = p_A(y) - Ax$. Let x^* be a minimizer of the least squares problem, since it satisfies $p_A(y) = Ax^*$, we have $p_A(r(x^*))) = 0$ and therefore $r(x^*) \in \mathrm{Im}(A)^\perp$. On the other hand, assume that x^* is such that $r(x^*) \in \mathrm{Im}(A)^\perp$, then $p_A(r(x^*))) = 0$ and therefore $p_A(y) = Ax^*$ which proves that x^* is a minimizer.

Uniqueness

There might be several x^* such that $p_A(y) = Ax^*$, therefore the least squares problem does not necessarily admit a unique solution. However, if A has linearly independent columns (i.e. A is injective), then for all y there exists a unique x^* such that $p_A(y) = Ax^*$.

11.2 Pseudoinverse Solution to Least Squares Problem

Having characterized minimizers of the least squares problem, we will now find a formula for it. We first need the following result.

The "Row Space Theorem"

Theorem 8. *Let $M \in \mathcal{M}_{p,n}(\mathbb{R})$. Then we have that:*

1. *$Ker(M^\top) = Im(M)^\perp$*
2. *$Im(M^\top) = Ker(M)^\perp$*

Proof.
1. Let $x \in \ker(M^\top)$. Then, for all $y \in \mathbb{R}^n$, we have $\langle x, My\rangle = \langle M^\top x, y\rangle = 0$, therefore $x \in \mathrm{Im}(M)^\perp$. Now, let $x \in \mathrm{Im}(M)^\perp$, therefore, for all $y \in \mathbb{R}^n$, we have $0 = \langle x, My\rangle = \langle M^\top x, y\rangle$. Take $y = M^\top x$, we obtain $0 = \langle M^\top x, M^\top x\rangle = \|M^\top x\|^2$, which implies $M^\top x = 0$, i.e. $x \in \ker(M^\top)$. By the principle of double inclusion, we have proved $\mathrm{Ker}(M^\top) = \mathrm{Im}(M)^\perp$.
2. Apply point 1 to $M := M^\top$, remembering that $(M^\top)^\top = M$.

□

Least Squares Problem

Proposition 49. *If A has linearly independent columns, then:*

$$x^\star = (A^\top A)^{-1} A^\top y$$

Proof

Proof. Since A has linearly independent columns, we have explained before that the solution x^* is unique. Since $r(x^*) \in \mathrm{Im}(A)^\perp$, then, from Theorem 8, $r(x^*) \in \mathrm{Ker}(A^\top)$, which means

$$A^\top r(x^*) = A^\top(y - Ax^*) = 0$$

And therefore:

$$A^\top A x^* = A^\top y$$

To finish the proof, we have to show that $A^\top A$ is invertible. If it is singular then $A^\top Ax = 0$ for some non-zero x. But then $A^\top Ax = 0 = x^\top A^\top Ax = ||Ax||^2$, and the only way that $||Ax|| = 0$ given that A has linearly independent columns is for x to be 0. Therefore $A^\top A$ is injective, and thus invertible since $A^\top A$ is a square matrix. Therefore:

$$x^\star = (A^\top A)^{-1} A^\top y.$$

□

The Pseudoinverse Matrix

The matrix we've uncovered is called the pseudoinverse, or the Moore-Penrose Pseudoinverse. For a matrix $A \in \mathcal{M}_{n,m}(\mathbb{R})$ with linearly independent columns, the pseudoinverse can be calculated as:

$$A^\dagger = (A^\top A)^{-1} A^\top \in \mathbb{R}^{m\times n}$$

It is a very useful object. It's a bit like the inverse, but performing $A^\dagger y$, instead of returning to you the x such that $Ax = y$ like A^{-1} would do, it returns you the $x^\star$ such that $Ax^\star$ is the nearest to y. You can find a pseudoinverse for all A, even if they don't have linearly independent columns, though then it doesn't take this particular form (since then $A^\top A$ is no longer invertible).

11.3 Linear Regression

Motivation

We'll now consider a very common problem setting that arises if you have two sets of variables that you think are related, but you don't know exactly how they're related. For example, you might have a set of stimuli and associated neural firing rates, they're related, but how is unclear a priori. Instead you want to learn this relationship from data. In general it's often very hard to know how these variables are related, but a good first guess is that the relationship might be linear. This is like fitting a line of best fit to your data. Given a dataset of pairs of variables, finding the best fitting relationship between the two is called regression. If you assume the relationship is linear its called linear regression. If you use an L2 norm to measure distance then this is the famous ordinary least squares problem.

This is the third version of missing variables in the linear equation, $y = Ax$, now we know the pair (x, y), but A is unknown. Unlike the other two settings, given one pair (x, y) you can say very little about A (Ax' for all x' orthogonal to x can be varied arbitrarily while ensuring that $Ax = y$, this shows us that there are many equally good A matrices). But often we have a large dataset of pairs that we can use, allowing us to find a reasonable estimate of A. We'll now formalize this.

Ordinary Least Squares

We are given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^p$ of pairs of vectors $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$, we want to find the linear relationship, $A \in \mathbb{R}^{n\times m}$, that minimizes the distance between all y_i and their associated Ax_i:

$$\min_{A\in\mathbb{R}^{n\times m}} \sum_{i=1}^{p} \|y_i - Ax_i\|^2$$

Ordinary Least Squares Solution

Theorem 9. *Define the data matrices:*

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_p \\ | & | & & | \end{bmatrix}, \quad Y = \begin{bmatrix} | & | & & | \\ y_1 & y_2 & \cdots & y_p \\ | & | & & | \end{bmatrix}$$

Then, if $X^\top$ has linearly independent columns, the ordinary least squares solution is:

$$A^\top = (X^\top)^\dagger Y^\top$$

Proof. Let's show that this problem decomposes in terms of the rows of A (by 'decomposes' we mean that you can break the optimisation problem over A into completely separate, smaller, and therefore easier, optimisation problems over each row of A). Label the j^{th} row of A as r_j, then this problem is equivalent to:

$$\begin{aligned}
&\min_{\{r_j\}_{j=1}^n, r_j \in \mathbb{R}^m} \sum_{i=1}^p \sum_{j=1}^n (y_{ij} - \langle r_j, x_i \rangle)^2 \\
=& \min_{\{r_j\}_{j=1}^n, r_j \in \mathbb{R}^m} \sum_{j=1}^n \sum_{i=1}^p (y_{ij} - \langle r_j, x_i \rangle)^2 \\
=& \min_{\{r_j\}_{j=1}^n, r_i \in \mathbb{R}^m} \sum_{j=1}^n \ell(r_j, \mathcal{D})
\end{aligned}$$

Where $\ell(r_j, \mathcal{D})$ depends only on the j-th row of A. This means that each row can be chosen independently to minimize $\ell(r_j, \mathcal{D})$. Let us show that each $\ell(r_j)$ is a hidden least-squares problem. This is one of nicest tricks of maths, we re-frame our problem and show it is equivalent to one we've already solved, so we can just use that solution!
Define the output vector for dimension j, i.e. the rows of Y:

$$\tilde{y}_j = \begin{bmatrix} y_{1j} & \dots & y_{pj} \end{bmatrix}^\top$$

Then we have:

$$\begin{aligned}
\ell(r_j, \mathcal{D}) &= \sum_{i=1}^p (y_{ij} - \langle r_j, x_i \rangle)^2 \\
&= \left\| \tilde{y}_j - X^\top r_j \right\|_2^2
\end{aligned}$$

This is a least-squares problem but where $X^\top$ plays the role of A, $\tilde{y}_j$ of y, and r_j of x. Therefore, if $X^\top$ has linearly independent columns (which requires $m \leq p$), then we can find the least-squares r_j using the pseudoinverse:

$$r_j = (X^\top)^\dagger = (XX^\top)^{-1} X \tilde{y}_j$$

Stacking each of these solutions we derive that:

$$A^\top = (X^\top)^\dagger Y^\top$$

So we can find the best A with a simple matrix pseudoinverse and product! □

11.4 Ridge Regression

Motivation

In the previous section we found a solution if $X^\top$ has linearly independent columns. Roughly speaking, this tends to happen if the number of pairs of data is larger than the input dimensionality. This is morally like solving an overdetermined linear equation $y = Ax$ for x given y and A, in which there is no good solution for x. You therefore find the one that does best.

However, often we're in the underdetermined setting. For example, imagine you're measuring the neural representation of some variable $x \in \mathbb{R}$, call the neural response function $g : \mathbb{R} \to \mathbb{R}^N$, i.e. when the variable takes value x', the neurons fire with rates $g(x')$. Now you would like to build a linear decoder, A, that tells you the most likely value of x given a measurement of neural activities: $\hat{x} = Ag(x)$. You're going to build this decoder using measured pairs of $\{x_i, g(x_i)\}_{i=1}^p$. This is exactly the ordinary least squares problem, however, since you might be recording 1000s of neurons, you're very unlikely to record enough trials to ensure that $p > n$. This means the columns of $X^\top$ won't be linearly independent and our previous solution won't work.

Intuitively, the solution stops working because there are subspaces of $\mathbb{R}^n$ that the data does not explore. As a result, how Ay behaves for y orthogonal to the span of $\{g(x_i)\}_{i=1}^p$ is completely unconstrained: Ay can take whatever values you want, and it won't effect how well A fits the data. This means even the pseudoinverse as advertised above can't work: how can the pseudoinverse know which A to output? There are infinitely many that behave optimally!

Our reasonable hunch to choose between them will be that if we don't know how the data behaves in some subspace, the decoder should guess that the output is 0. We will do this by adding a term to the objective that penalizes the learnt fit, A, based on its size. Then, as you'll show in the exercises, this term breaks the equivalence between all the solutions and chooses the one that outputs zero on all the unconstrained spaces. For a mixture of mathematical convenience and it's beautiful interpretation as a Gaussian prior on A (see probability lectures), we'll choose the term to be $||A||_F^2$. This leads us to the ridge regression problem.

Ridge Regression

We are given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^p$ of pairs of vectors $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$, we want to find the linear relationship, $A \in \mathbb{R}^{n\times m}$, that minimizes a mixture prediction error between Ax_i and the data y_i, and the size of A:

$$\min_{A\in\mathbb{R}^{n\times m}} \sum_{i=1}^p \|y_i - Ax_i\|^2 + \lambda ||A||_F^2$$

$\lambda > 0$ is the ridge parameter that controls the tradeoff between the two terms.

Ridge Regression Solution

Theorem 10. *The ridge regression solution is:*

$$A = (XX^\top + \lambda I)^{-1} XY^\top$$

Proof. Again we can decompose the problem according to the rows of A:

$$\begin{aligned}
&\min_{\{r_i\}_{i=1}^m, r_i \in \mathbb{R}^n} \sum_{i=1}^p \sum_{j=1}^m (y_{ij} - \langle r_j, x_i \rangle) + \lambda ||r_j||^2 \\
&= \sum_{j=1}^m \min_{r_j \in \mathbb{R}^n} \sum_{i=1}^p (y_{ij} - \langle r_j, x_i \rangle) + \lambda ||r_j||^2 = \sum_{j=1}^m \min_{r_j \in \mathbb{R}^n} \ell(r_j, \mathcal{D})
\end{aligned}$$

Then we're going to do the clever maths thing again and show how this problem is like a modified version of the ordinary least squares problem we're already solved. In the previously defined stacked dataset variables:

$$\ell(r_j, \mathcal{D}) = ||\tilde{y}_j - X^\top r_j||^2 + \lambda ||r_j||^2$$

We create the stacked objects, $\bar{X} \in \mathbb{R}^{n \times (n+p)}$, $\bar{y}_j \in \mathbb{R}^{n+p}$:

$$\bar{X} = \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \qquad \bar{y}_j = \begin{bmatrix} \tilde{y}_j \\ 0 \end{bmatrix}$$

Now it is easy to verify that:

$$||\bar{y}_j - \bar{X}^\top r_j||^2 = ||\tilde{y}_j - X^\top r_j||^2 + \lambda ||r_j||^2 = \ell(r_j, \mathcal{D})$$

But then this is just another ordinary least squares problem! The minima is therefore:

$$r_j = (\bar{X}\bar{X}^\top)^{-1} \bar{X} \bar{y}_j = (XX^\top + \lambda I)^{-1} X y_j$$

And stacking each of these rows of the optimal A, we recover the stated result. □

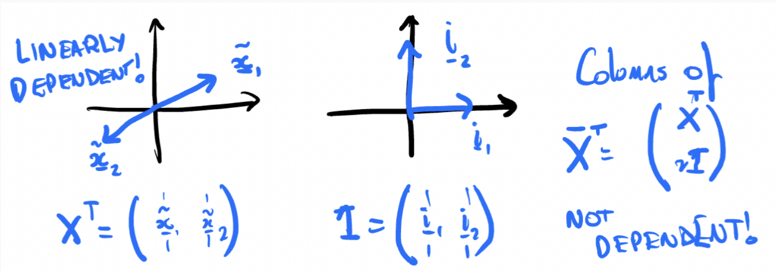
Comments

Trade-off Between Terms This procedure works for any $\lambda > 0$. If λ is very large the optimal A will be very close to 0, and won't fit the data very well. If it is very small the optimal A will fit the data nearly as well as it can. Choosing the best λ, and other similar question, are hot questions in machine learning and statistics, related to the idea of overfitting.

Ridge Regression as Orthogonalising Aligned Datapoints We motivated ridge regression in the setting where $n > p$, where parts of A are unconstrained. However, another setting where $XX^\top$ is singular and therefore ordinary least squares doesn't work is when $p > n$ but the columns of $X^\top$ are linearly dependent. We can interpret the ridge regression solution in this setting in the following way. The ridge regression problem is performed on the stacked input data vectors:

$$\bar{X}^\top = \begin{bmatrix} X^\top \\ \nu I \end{bmatrix}$$

This is like taking each column of $X^\top$, $\tilde{x}_j$, and adding on a new vector on the end, i_j, making a bigger, $n+p$ dimensional vector. In the p dimensional subspace you've added each of these vectors points in an orthogonal direction, so even if the columns of X were linearly dependent, the columns of $\bar{X}$ never will be, ensuring the ridge regression solution exists!



This exercises justifies some of the motivating claims we made.

11.5 Exercise Sheet 11: Regression

Exercises: Pseudoinverse

Exercise 107. *Show that, if A is invertible, $A^{\dagger} = A^{-1}$.*

Exercise 108. *What is the pseudoinverse of a vector?*

Exercise 109. *Show that $AA^{\dagger}$ is idempotent (i.e. $(AA^{\dagger})^2 = AA^{\dagger}$). What space does $AA^{\dagger}$ project onto?*

Exercise 110. *What is the pseudoinverse of A in terms of the components of its singular value decompositions, U, V, and S? Interpret what the pseudoinverse is doing using a set of diagrams like those in figure 8.2*

Important Exercise!

Exercise 111. *Why is $(XX^{\top} + \lambda I)$ invertible for any $\lambda > 0$?*

11.6 Neuro Q8: fMRI & Correlated Regressors

What is fMRI? Linear regression pops up literally everywhere. One common setting is in fMRI - functional Magnetic Resonance Imaging. fMRI involves placing animals, including famously humans (and even some famous humans), in enormous magnetic fields - 10^6 times bigger than the earth's magnetic field. It turns out that oxygenated blood and deoxygenated blood have different magnetic response properties: deoxygenated blood is diamagnetic, meaning it antialigns its magnetic dipole with the applied magnetic field, locally damping it, while oxygenated blood is paramagnetic, meaning it aligns, causing local amplification. If you suddenly start using one brain area more, your neurons will fire using up all the oxygen available locally, causing the blood to become diamagnetic and the local magnetic field to reduce. Then cells in that brain area realise there's less oxygen and tell the capillaries to swell, this causes much more oxygenated blood to rush in, enhancing the magnetic field. If the resolution is high enough, you can watch these changes happen in the local magnetic field (by measuring the magnetic field around the participant and solving Maxwell's equations to recover what the magnetic field must be inside the animal's head). This is very cool! It means you can measure which areas of a person's brain are activating, and people have used this to test all sorts of interesting brain hypotheses (and some garbage ones).

Regressing Voxels For example, you might use fMRI to measure how the brain encodes a set of images. To that end, you present the images to participants in the scanner and measure how each little part of the brain responds (the discrete units you chunk the brain into are called voxels, the 3D volume generalisation of the pixel). To draw interesting conclusions you might want to understand how each brain area encodes some features of the image, such as the size and colour. One way to examine this is to predict how a particular voxel will respond from these features, if it is predictable then the brain area encodes them. To do this you could build a linear model of each voxel's response:

$$v_{ij} = w_{cj} c_i + w_{sj} s_i + \eta_{ij}$$

Where v_{ij} is voxel j's response to image i, c_i and s_i are the colour and size of image i, and η_{ij} is noise encoding that this model will never be a perfect. w_{cj} and w_{sj} therefore represent the jth voxel's tuning to the two features, colour and size. Here these are the quantities of scientific interest.

Linear Regression There are I images, stack the responses and noise in each voxel j and the size and colour of each image into vectors, $v_j, \eta_j, c, s \in \mathbb{R}^I$; then our model of each voxel's response is:

$$v_j = w_{cj} c + w_{sj} s + \eta_j$$

We use ordinary least squares to predict the regression weights, $\hat{w}_{cj}$ and $\hat{w}_{sj}$, creating the optimal predicted response: $\hat{v}_j = \hat{w}_{cj} c + \hat{w}_{sj} s$.

A Result! You'll be famous! You notice something strange: $\hat{w}_j$ and $\hat{v}_j$ are positively correlated ($\sum_j \hat{v}_j \hat{w}_j > 0$)! If a voxel has high tuning to shape it also has high tuning to colour and vice versa. Perhaps this is a fundamental claim about the way the brain encodes these two concepts? You run excitedly to the nearest soapbox and loudly proclaim your newest finding to all and sundry. As you are signing sponsership deals with your favourite magnet firms, a wizened character emerges from the gawping crowd. He walks with a limp and the support of staff, and his countenance is wrinkled like a walnut with long years of thought over the intricacies of linear regression. He pauses, the crowd withdraws in anticipation, and he begins a series of riddles:

1. "As all practioner's of the art of linear regression know, the key space to consider is the span of the stimulus vectors" he begins,

   Thoughts of hot summers spent sweating over linear algebra wing through your head. Meditate and justify why the residuals of the optimal predicted response $v_j - \hat{v}_j$ live in the orthogonal complement of the span of the stimulus and colour vectors: $span(\{c, s\})^{\perp}$.

2. He pauses, the crowd waits with baited breath, he scratches his tangled beard causing a flurry of birds to escape.

   "What does it mean for my brain to respond to a scratch? I think it is that the more I scratch the more it responds. However, if my brain is always responding at a non-zero rate regardless of the scratch, that is not really a response, no?". You hastily agree. "Yet if your data has a mean this might be the case."

   The ground drops beneath you: you realise you forgot to de-mean your regressors!

   (a) Explain why, if the response and stimuli all have non-zero means, then v_j, s, and c all have a non-zero component along the all-ones vector, $\mathbf{1} \in \mathbb{R}^I$.

   (b) Take the extreme case where one stimulus is constant $s = \alpha\mathbf{1}$. Explain why, if the jth voxel's response has non-zero mean, its regression weight, w_{sj}, will be non-zero.

   (c) Take another extreme case, let's say the voxel's true response doesn't depend on either stimulus, but has a non-zero mean, for example $v_j = \beta\mathbf{1}$. Explain why the regression weights will again be non-zero.

   You realise the old man is right. Even if the voxel is unrelated to the stimulus, if both the voxel response and the regressor have non-zero means the regression weights will be non-zero.

   (d) Explain why de-meaning either the voxel response or the stimulus before performing regression will fix this problem. (i.e. you do regression not with v_j, c and s, but with the same vectors after you subtract their mean)

   "Hence our first rule of thumb: life is easier if you de-mean your regressors" He expounds.

   The crowd tuts in the background and your sponsors return to their waiting limousine. You panic, and quickly de-mean your regressors (assume from this point that all variables are mean-zero), you re-run the regression and breathe a sigh of relief:

   "It's okay! you shout, the regressors are still correlated, but after subtracting the mean the regressors are negatively correlated now! I'm still clever and important!!"

3. The crowd turns to the old man. "The world is often noisy. A small amount of noise is not a problem, but too much and problems arise." Let's take the simplest case: orthogonal regressors and just one voxel.

   (a) With no noise we can imagine the true voxel response as lying in the span of the stimulus: $v_1 = w_{1c}c + w_{1s}s$. In this plane, draw some of the lines of constant $w_{1c}w_{1s}$.

   (b) The true data is $w_{1c}c + w_{1s}s + \eta$, $\eta \sim \mathcal{N}(0, \sigma^2 I)$. The projection of Guassian noise is also Gaussian noise. Use this to sketch on the same plot the distribution of measured $\hat{w}_{1c}$ and $\hat{w}_{1s}$ given a few different true w_{1c} and w_{1s}.

(c) An estimator $\hat{\alpha}$ for a quantity α is unbiased if the average estimator, averaged over the noise in the world, equals the true value: $\mathbb{E}[\hat{\alpha}] = \alpha$.

Explain why, for small but non-zero σ, $\hat{w}_{1c}\hat{w}_{1s}$ is an unbiased estimator for $w_{1c}w_{1s}$.

(d) Now consider larger σ. Using the curvature of the lines of equal $w_{1c}w_{1s}$ that you drew earlier, justify that, no matter the value of $w_{1c}w_{1s}$: $|\hat{w}_{1c}\hat{w}_{1s}| < w_{1s}w_{1s}$, i.e. the estimator makes systematic errors towards 0.

4. "Now let's return to the case of correlated regressors", he says, sucking on his enormous twizzly-twirly tobacco pipe, which emits ellipsoidal balls of smoke from a series of holes. "If the two variables are correlated, does this not mean that c and s are aligned?"

Justify that this is true. "Of course," you say.

"But then this changes the problem"

(a) Again draw lines of constant $w_{1c}w_{1s}$, they are not the same right?

(b) Assume s and c are positively aligned. Imagine you have two voxels with $w_{1c}w_{1s} = \beta$ and $w_{2c}w_{2s} = -\beta$ for some positive β. Argue from the same sort of diagrams that, for a fixed amount of noise, the noise will reduce $w_{1c}w_{1s}$ more than $w_{2c}w_{2s}$.

(c) Use this to argue that, even if the true regression weights are uncorrelated $\sum_j w_{jc}w_{js} = 0$, then with correlated regressors the noise will make the mean of your estimator, $\sum_j \hat{w}_{jc}\hat{w}_{js}$ positive.

"This leads us to our second rule of thumb for regression, correlated regressors can produce uncorrelated regression weights, even if the true regression weights are uncorrelated!"