CIS 515: Math for Machine Learning

Motivations: Fitting Data (Regressions)

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Rank: # of dimensions in output of transformation

Motivations

Two central problems in machine learning are

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).

(2.) *Data classification*.

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).

(2.) *Data classification*.

For this introduction we focus on the more classical problem of data fitting.

Fitting Points in the Plane

Assume we have some data points in the plane given as a list of m coordinates

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i, y_i \in \mathbb{R}.$$

Fitting Points in the Plane

Assume we have some data points in the plane given as a list of m coordinates

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i, y_i \in \mathbb{R}.$$

The figure on the next slide shows an example of 100 points in the plane.

fit data = guess function

Fitting Points in the Plane

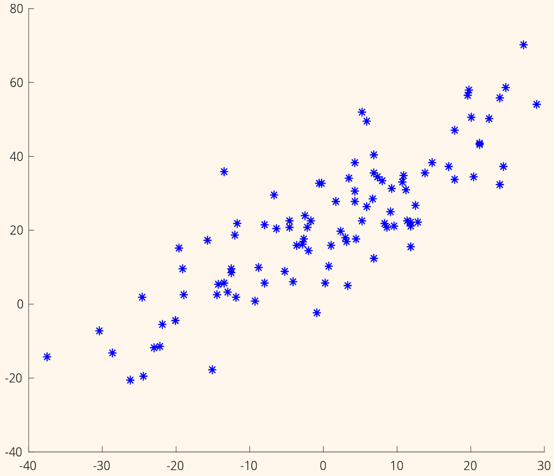


Figure 1: A data set of 100 points in the plane.

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

find a function that fits the data

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

go for simplicity

linear

The simplest kind of function is an *affine map*, that is, a map of the form

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \dots, 100$.

The simplest kind of function is an *affine map*, that is, a map of the form

$$f(x) = \overset{\text{weight}}{w}x + \overset{\text{intercept}}{b},$$

for some real numbers w, b. The number w is called a *weight*.

Learning an Affine Map

We are looking for a function $f \colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

The simplest kind of function is an *affine map*, that is, a map of the form

$$f(x) = wx + b,$$

for some real numbers w, b. The number w is called a *weight*.

The numbers w and b must satisfy the 100 (affine) equations

$$y_i = f(x_i) = wx_i + b.$$ for 100 points

- Very unlikely all points are on the same line

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

But what is the error?

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

But what is the error?

Gauss and Legendre proposed a method over 200 years ago: the *least squares method*.

least squares method

What is the Error?

Every equation $y_i = wx_i + b$ can be written as

$$y_i - wx_i - b = 0.$$ if no error

Think of $y_i - wx_i - b$ as an *error*.

not exactly 0, so error

What is the Error?

Every equation $y_i = wx_i + b$ can be written as

$$y_i - wx_i - b = 0.$$

Think of $y_i - wx_i - b$ as an *error*.

In the method of least squares, *the error (or loss) is the sum of the squares of the errors*:

$$\sum_{i=1}^{100} (y_i - wx_i - b)^2.$$

to ignore signs
compute derivatives

Least Squares Solution

Here the least squares solution for our data set of 100 points.

2D plane

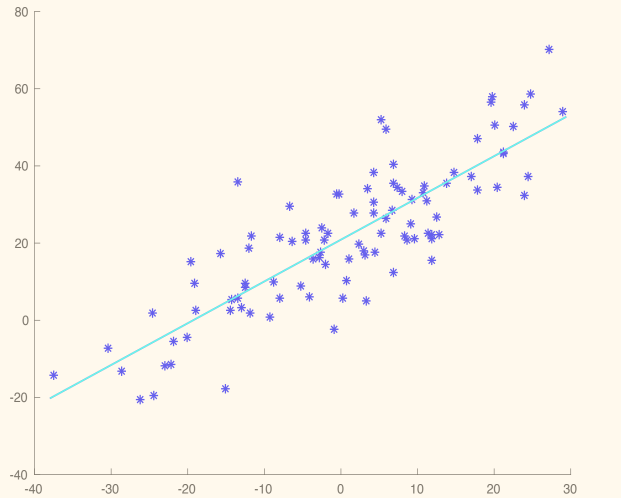


Figure 2: The least squares best fit.

Fitting Points in $\mathbb{R}^n$ $\longrightarrow$ allows for multiple dimensions

We can generalize the problem to data in $\mathbb{R}^n$.

Fitting Points in $\mathbb{R}^n$

We can generalize the problem to data in $\mathbb{R}^n$.
Assume we have some data given as a list of m pairs

$$((x_1, y_1), \dots, (x_m, y_m)), \quad x_i \in \mathbb{R}^n, y_i \in \mathbb{R}.$$

Fitting Points in $\mathbb{R}^n$

We can generalize the problem to data in $\mathbb{R}^n$.
Assume we have some data given as a list of m pairs

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i \in \mathbb{R}^n, y_i \in \mathbb{R}.$$

We wish to learn an affine map $f\colon \mathbb{R}^n \to \mathbb{R}$ of the form

combine coordinates
z_1, z_n w/ weights
$+ b$

$$f(z) = w_1 z_1 + \cdots + w_n z_n + b,$$

multiple weights, 1 per each input

with $z = (z_1, \ldots, z_n)$ and where $w_1, \ldots, w_n \in \mathbb{R}$ are *weights*.

It is convenient to denote the quantity $w_1 z_1 + \cdots + w_n z_n$ (an inner product) as $z^\top w$.

transpose

The Euclidean Norm (or ℓ^2-Norm)

The *Euclidean norm* (or *ℓ^2-norm*) of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_2 = (z_1^2 + \cdots + z_n^2)^{1/2} = (z^\top z)^{1/2}.$$

The Euclidean Norm (or ℓ^2-Norm)

The *Euclidean norm* (or *ℓ^2-norm*) of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_2 = (z_1^2 + \cdots + z_n^2)^{1/2} = (z^\top z)^{1/2}.$$

The *least squares problem* is find $w \in \mathbb{R}^n$ that minimizes

$$\|\xi\|_2^2,$$

square of the error

where $\xi = (\xi_1, \ldots, \xi_m)$ is the vector given by

$$\xi_i = y_i - x_i^\top w - b.$$

Specification for optimization problem.

not obvious

Pseudo-Inverse

It turns out that there is a *unique solution* w^+ *of least* ℓ^2*-norm*.

Pseudo-Inverse

It turns out that there is a *unique solution w^+ of least ℓ^2-norm*.

Furthermore, this solution w^+ is expressed in terms of something called a *pseudo-inverse*.

Pseudo-Inverse

It turns out that there is a *unique solution w^+ of least ℓ^2-norm*.

Furthermore, this solution w^+ is expressed in terms of something called a *pseudo-inverse*.

In our case

$$w^+ = A^+ y,$$

where A^+ is the pseudo-inverse of the matrix

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix}.$$

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*.)

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*.)

The SVD and the pseudo-inverse will be discussed extensively later.

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*.)

The SVD and the pseudo-inverse will be discussed extensively later.

The solution given by the pseudo-inverse is not always desirable or too expensive to compute.

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*.)

The SVD and the pseudo-inverse will be discussed extensively later.

The solution given by the pseudo-inverse is not always desirable or too expensive to compute.

Another method is to *penalize the ℓ^2-norm of w*.

add to minimization of error term a weight term

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

$$\begin{aligned} &\text{minimize} \quad \|\xi\|_2^2 + K\|w\|_2^2 \\ &\text{subject to} \\ &\qquad\qquad y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where K is positive constant.

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

$$\begin{aligned}
&\text{minimize} \quad \|\xi\|_2^2 + K\|w\|_2^2 \\
&\text{subject to} \\
&\qquad y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m
\end{aligned}$$

where K is positive constant.

This time there is a unique solution given in terms of the matrix X whose rows are the (row) vectors $x_i^\top$. For simplicity assume $b = 0$.

Ridge Regression

The unique minimizer is given by the expression

$$w = X^\top (XX^\top + KI_m)^{-1} y.$$

- inverting matrix
- symmetric

Ridge Regression

The unique minimizer is given by the expression

$$w = X^\top (XX^\top + KI_m)^{-1} y.$$

The matrix

$$XX^\top + KI_m$$

is particularly nice because it is *symmetric positive definite*. There are more efficient methods for solving linear system involving SPD matrices. We will study such matrices extensively.

ℓ^1-Norm and lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

do this by changing penalty term

ℓ^1-Norm and lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

A remedy to this problem is to *penalize the ℓ^1-norm $\|w\|_1$ of w* instead of its ℓ^2-norm $\|w\|_2^2$.

ℓ^1-Norm and lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

A remedy to this problem is to *penalize the ℓ^1-norm $\|w\|_1$ of w* instead of its ℓ^2-norm $\|w\|_2^2$.

The *ℓ^1-norm* of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_1 = |z_1| + \cdots + |z_n|.$$

Lasso Regression

Lasso regression is the following minimization problem:

Lasso Regression

Lasso regression is the following minimization problem:

$$\begin{aligned} &\text{minimize} \quad \|\xi\|_2^2 + \tau \underbrace{\|w\|_1}_{\text{penalty term}} \\ &\text{subject to} \\ &\qquad y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where τ is positive constant.

Lasso Regression

Lasso regression is the following minimization problem:

$$\begin{aligned} &\text{minimize} \quad \|\xi\|_2^2 + \tau \|w\|_1 \\ &\text{subject to} \\ &\qquad\qquad y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where τ is positive constant.

This time, there is no closed-form solution. However a solution can be computed using an iterative process (ADMM) which solves a sequence of linear systems involving SPD matrices.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

A way to retain the best features of ridge regression and lasso is to *penalize both the ℓ^1-norm and the ℓ^2-norm of w*.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

A way to retain the best features of ridge regression and lasso is to *penalize both the ℓ^1-norm and the ℓ^2-norm of w*.

Elastic net regression is the following minimization problem:

$$\begin{aligned} &\text{minimize} && \|\xi\|_2^2 + K\|w\|_2^2 + \tau\|w\|_1 \\ &\text{subject to} && \\ & && y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where K and τ are positive constants.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Remarkably, least squares, ridge regression, lasso, and elastic net, all rely on *solving linear systems involving SPD matrices*.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Remarkably, least squares, ridge regression, lasso, and elastic net, all rely on *solving linear systems involving SPD matrices*.

This is why most of this course will be devoted to these topics! The notion of *orthogonality* also play a crucial role.

CIS 515: Math for Machine Learning

$(+,-)$ $\qquad$ $(+,-,\cdot)$

Groups, Rings and Fields

$(+,-,\cdot,\div)$

Professors Jean Gallier & Jocelyn Quaintance

Groups

check

1) $a+(b+c)=(a+b)+c$ - grouping
2) $a+0=0+a=a$ - identity element
3) $a+(-a)=(-a)+a=0$ - inverse

A (real) vector space is a set E together with two operations, $+ \colon E \times E \to E$ and $\cdot \colon \mathbb{R} \times E \to E$, called *addition* and *scalar multiplication*, that satisfy some simple properties. First of all, E under addition, has to be a commutative (or abelian) group.

Definition of a Group

• Set of elements
• operative
• closed
• associative
• inverse
• identity

Definition. A *group* is a set G equipped with a binary operation $\cdot \colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in G is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c.$ grouping (associativity);

(G2) $a \cdot e = e \cdot a = a.$ identity element (identity);

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e.$ (inverse).

$a + a^{-1} =$ identity element

A group G is *abelian* (or *commutative*) if

not required

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$

Examples of a Group

(1.) The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is an abelian group under addition, with identity element 0.

(2.) The set $\mathbb{Q}$ of rational numbers (fractions p/q with $p, q \in \mathbb{Z}$ and $q \neq 0$) is an abelian group under addition, with identity element 0. The set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ is also an abelian group under multiplication, with identity element 1.

nonzero rational #s

identity element: no effect when combined w/ other elements

Examples of a Group

(3.) The sets $\mathbb{R}$ of real numbers and $\mathbb{C}$ of complex numbers are abelian groups under addition (with identity element 0), and $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$ are abelian groups under multiplication (with identity element 1).

(4.) The sets $\mathbb{R}^n$ and $\mathbb{C}^n$ of n-tuples of real or complex numbers are abelian groups under componentwise addition:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$

with identity element $(0, \ldots, 0)$.

Examples of a Group

(5.) The set of $n \times n$ matrices with real (or complex) coefficients is an abelian group under addition of matrices, with identity element the null matrix. It is denoted by $\mathrm{M}_n(\mathbb{R})$ (or $\mathrm{M}_n(\mathbb{C})$).

(6.) The set $\mathbb{R}[X]$ of all polynomials in one variable X with real coefficients,

$$P(X) = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0,$$

(with $a_i \in \mathbb{R}$), is an abelian group under addition of polynomials. The identity element is the zero polynomial.

Examples of a Group

(7.) The set of $n \times n$ invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *general linear group* and is usually denoted by $\mathbf{GL}(n, \mathbb{R})$ (or $\mathbf{GL}(n, \mathbb{C})$).

(8.) The set of $n \times n$ invertible matrices with real (or complex) coefficients and determinant $+1$ is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *special linear group* and is usually denoted by $\mathbf{SL}(n, \mathbb{R})$ (or $\mathbf{SL}(n, \mathbb{C})$).

Examples of a Group

(9.) The set of $n \times n$ invertible matrices with real coefficients such that $RR^\top = R^\top R = I_n$ and of determinant $+1$ is a group (under matrix multiplication) called the *special orthogonal group* and is usually denoted by $\mathbf{SO}(n)$ (where $R^\top$ is the *transpose* of the matrix R, i.e., the rows of $R^\top$ are the columns of R). It corresponds to the rotations in $\mathbb{R}^n$.

(10.) Given an open interval (a, b), the set $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ is an abelian group under the operation $f + g$ defined such that

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in (a, b)$.

Groups

Self Check.

Is $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is an abelian group under multiplication?

Groups

addition $+$ inverse $-a$
multiplication $\cdot$ inverse a^{-1}

It is customary to denote the operation of an abelian group G by $+$, in which case the inverse a^{-1} of an element $a \in G$ is denoted by $-a$.

The identity element of a group is *unique*.

Every element in a group has a *unique inverse*.

Rings $(+, -, \cdot)$

A vector space is an abelian group E with an additional operation $\cdot\colon K \times E \to E$ called scalar multiplication that allows rescaling a vector in E by an element in K. The set K itself is an algebraic structure called a *field*. A field is a special kind of structure called a *ring*.

w.r.t. = w/ reference to

Definition of a Ring

- add, mult 2 elements in ring, you get another element in ring
- mult elements may not have inverses/abelian

and therefore subtraction

Definition. A *ring* is a set A equipped with two operations $+ \colon A \times A \to A$ (called *addition*) and $* \colon A \times A \to A$ (called *multiplication*) having the following properties:

(R1) A is an abelian group w.r.t. $+$; addition

(R2) $*$ is associative and has an identity element $1 \in A$;

(R3) $*$ is distributive w.r.t. $+$. abelian not required

The identity element for addition is denoted 0, and the additive inverse of $a \in A$ is denoted by $-a$.

Definition of a Ring

More explicitly, the axioms of a ring are the following equations which hold for all $a, b, c \in A$:

group	$a + (b + c) = (a + b) + c$	(associativity of $+$)	(1)
ring	$a + b = b + a$	(commutativity of $+$)	(2)
group	$a + 0 = 0 + a = a$	(zero) no mult. inverse	(3)
group	$a + (-a) = (-a) + a = 0$	(additive inverse)	(4)
ring	$a * (b * c) = (a * b) * c$	(associativity of $*$)	(5)
ring	$a * 1 = 1 * a = a$	(identity for $*$)	(6)
ring	$(a + b) * c = (a * c) + (b * c)$	(distributivity)	(7)
ring	$a * (b + c) = (a * b) + (a * c)$	(distributivity)	(8)

Commutative Ring

not required

The ring A is *commutative* if

$$a * b = b * a \quad \text{for all } a, b \in A.$$

Examples of a Ring

identity element

1. The abelian group $\mathbb{Z}$ is a commutative ring (with unit 1).

2. The abelian group $\mathbb{R}[X]$ of polynomials is also a commutative ring (also with unit 1).

Fields

additive inverse: $5 = -5$
mult. inverse: $5 = \frac{1}{5}$

- Subtraction becomes additive inverse
- division becomes mult. inverse

A field is a commutative ring K for which $K - \{0\}$ is a group under multiplication.

Definition of a Field

Definition. A set K is a *field* if it is a ring and the following properties hold:

(F1) $0 \neq 1$; identity values are different?

(F2) $K^* = K - \{0\}$ is a group w.r.t. $*$ (i.e., every $a \neq 0$ has an inverse w.r.t. $*$);

$A \cdot B = B \cdot A = 1$

(F3) $*$ is commutative. has multiplicative inverses

If $*$ is not commutative but (F1) and (F2) hold, we say that we have a *skew field* (or *noncommutative field*).

Examples of a Field

$\mathbb{Z}$ = ints

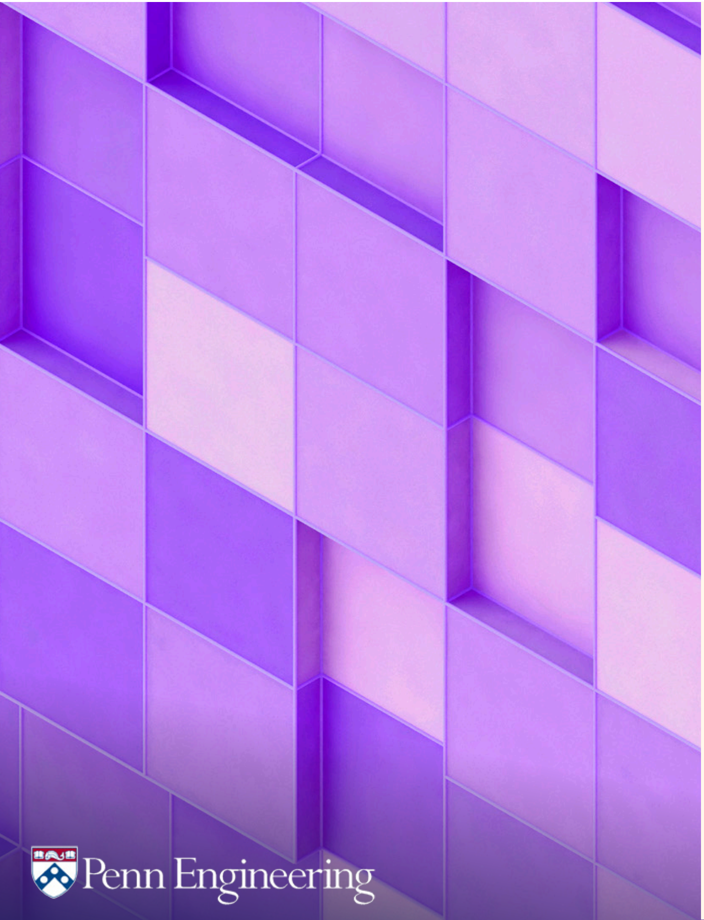
rational #s

real #s

complex #s ($a + bi$ $i = \sqrt{-1}$)

1. The rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields.

2. The set of (formal) fractions $f(X)/g(X)$ of polynomials $f(X), g(X) \in \mathbb{R}[X]$, where $g(X)$ is not the zero polynomial, is a field.



CIS 515: Math for Machine Learning

Vectors and Vector Spaces

Professors Jean Gallier & Jocelyn Quaintance

Vector Spaces commutative+ group

- add vectors together - commutative
- mult. vector · scalar

— don't consider mult. 2 vectors together

- 0 = identity element
- Inverses

Vectors play a crucial role in linear algebra! Thus we must define what is meant by a vector and vector space.

Definition of a Vector Space

Definition. A *real vector space* is a set E (of vectors) together with two operations $+ : E \times E \to E$ (called *vector addition*)and $\cdot : \mathbb{R} \times E \to E$ (called *scalar multiplication*) satisfying the following five conditions for all $\alpha, \beta \in \mathbb{R}$ reals and all $u, v \in E$; field

(V0) E is an abelian group w.r.t. $+$, with identity element 0
This means that E satisfies the following four axioms: For all $u, v, w \in E$

(G1) $u + (v + w) = (u + v) + w$; order the same, change group of two and get the same result (associativity)

(G2) There exists a vector $0 \in E$ such that $v + 0 = 0 + v = v$; (identity)

(G3) For every $v \in E$, there is some $-v \in E$ such that
$v + (-v) = -v + v = 0$; every element has an inverse (inverse)

(G4) $v + u = u + v$. change order of addition (commutativity)

Definition of a Vector Space

(V0) abelian group implies:

(V1) $\alpha \cdot (u + v) = (\alpha \cdot u) + (\alpha \cdot v)$; distribute α though ( )

(V2) $(\alpha + \beta) \cdot u = (\alpha \cdot u) + (\beta \cdot u)$; distribute u through ( )

(V3) $(\alpha * \beta) \cdot u = \alpha \cdot (\beta \cdot u)$, where $*$ denotes multiplication in $\mathbb{R}$;

- either multiply scalars together and apply them to u
- apply scalars sequentially from right to left to u

(V4) $1 \cdot u = u$. $1 \cdot u$ is always itself

$1v = v$
$a(bv) = (ab)v$
$a(u+v) = au + av$
$(a+b)v = av + bv$

Vectors V = commutative $+$
scalars F = field under $\cdot$

Vector Spaces

reals

Given $\alpha \in \mathbb{R}$ and $v \in E$, the element $\alpha \cdot v$ is also denoted by αv. The field $\mathbb{R}$ is often called the field of scalars.

In the definition of a vector space, the field $\mathbb{R}$ may be replaced by the field of complex numbers $\mathbb{C}$, in which case we have a *complex* vector space

Examples of a Vector Space

$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$

1. The fields $\mathbb{R}$ and $\mathbb{C}$ are vector spaces over $\mathbb{R}$.
2. The groups $\mathbb{R}^n$ and $\mathbb{C}^n$ are vector spaces over $\mathbb{R}$, and $\mathbb{C}^n$ is a vector space over $\mathbb{C}$.
3. The ring $\mathbb{R}[X]_n$ of polynomials of degree at most n with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]_n$ of polynomials of degree at most n with complex coefficients is a vector space over $\mathbb{C}$.
4. The ring $\mathbb{R}[X]$ of all polynomials with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]$ of all polynomials with complex coefficients is a vector space over $\mathbb{C}$. w/o deg(n) restriction
5. The ring of $n \times n$ matrices $\mathrm{M}_n(\mathbb{R})$ is a vector space over $\mathbb{R}$.

   rectangular
6. The ring of $m \times n$ matrices $\mathrm{M}_{m,n}(\mathbb{R})$ is a vector space over $\mathbb{R}$.

Vector Spaces

add two continuous f(x)s and get a continuous f(x)

real

Self Check.
Is the ring $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ a vector space over $\mathbb{R}$?

yes, under pointwise addition and pointwise scalar multiplication

CIS 515: Math for Machine Learning

Indexed Families

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Indexed Families

Given a set A, a *sequence* is an ordered n-tuple $(a_1, \ldots, a_n) \in A^n$ of elements from A, for some natural number n. The elements of a sequence need not be distinct and the order is important. For example, (a_1, a_2, a_1) and (a_2, a_1, a_1) are two distinct sequences in A^3. Their underlying set is $\{a_1, a_2\}$.

Indexed Families

What we just defined are *finite* sequences, which can also be viewed as functions from $\{1, 2, \ldots, n\}$ to the set A; the ith element of the sequence $(a_1, \ldots, a_n)$ is the image of i under the function. This viewpoint is fruitful, because it allows us to define (countably) infinite sequences as functions $s\colon \mathbb{N} \to A$. But then, why limit ourselves to ordered sets such as $\{1, \ldots, n\}$ or $\mathbb{N}$ as index sets?

Definition of an Indexed Family

The main role of the index set is to tag each element uniquely, and the order of the tags is not crucial, although convenient. Thus, it is natural to define an *I-indexed family of elements of A*, for short a *family*, as a function $a\colon I \to A$ where I is any set viewed as an index set.

Since the function a is determined by its graph

capital i

$$\{(i, a(i)) \mid i \in I\},$$

the family a can be viewed as the set of pairs $a = \{(i, a(i)) \mid i \in I\}$.

For notational simplicity, we write a_i instead of $a(i)$, and denote the family $a = \{(i, a(i)) \mid i \in I\}$ by $(a_i)_{i \in I}$

Example of an Indexed Family

a set for each element in I

For example, if $I = \{r, g, b, y\}$ and $A = \mathbb{N}$, the set of pairs

$$a = \{(r, 2), (g, 3), (b, 2), (y, 11)\}$$

is an indexed family. The element 2 appears twice in the family with the two distinct tags r and b.

CIS 515: Math for Machine Learning

Linear Combinations & Linear Independence

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Linear Combinations

Now that we have a definition of a vector space E, we may ask ourselves if there is a way to generate all of E from a "nice", perhaps even finite, subset of E. The answer is yes, namely we use a "basis" for E, but in order to define a basis, we need to define the notions of *linear combinations*, *linear independence*, and *spanning sets*.

Given a set A, recall that an I-indexed family $(a_i)_{i \in I}$ of elements of A (for short, a *family*) is a function $a\colon I \to A$, or equivalently a set of pairs $\{(i, a_i) \mid i \in I\}$. We agree that when $I = \emptyset$, $(a_i)_{i \in I} = \emptyset$. A family $(a_i)_{i \in I}$ is finite if I is finite.

index set empty set

Unless otherwise specified, we assume all families of scalars are *finite*.

Definition of a Linear Combination

Definition. Let E be a vector space. A vector $v \in E$ is a *linear combination of a family $(u_i)_{i \in I}$ of elements of E* iff there is a family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$ such that

$$v = \sum_{i \in I} \lambda_i u_i.$$

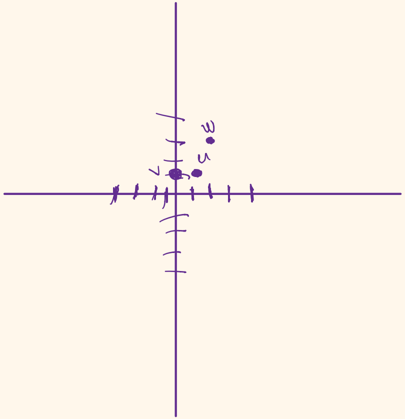
say

When $I = \emptyset$, we stipulate that $v = 0$.

Example of a Linear Combination

In $\mathbb{R}^2$, $w = (2, 3)$ is a linear combination of $u = (1, 1)$, $v = (0, 1)$ since

$$w = 2u + v.$$



Example of a Linear Combination

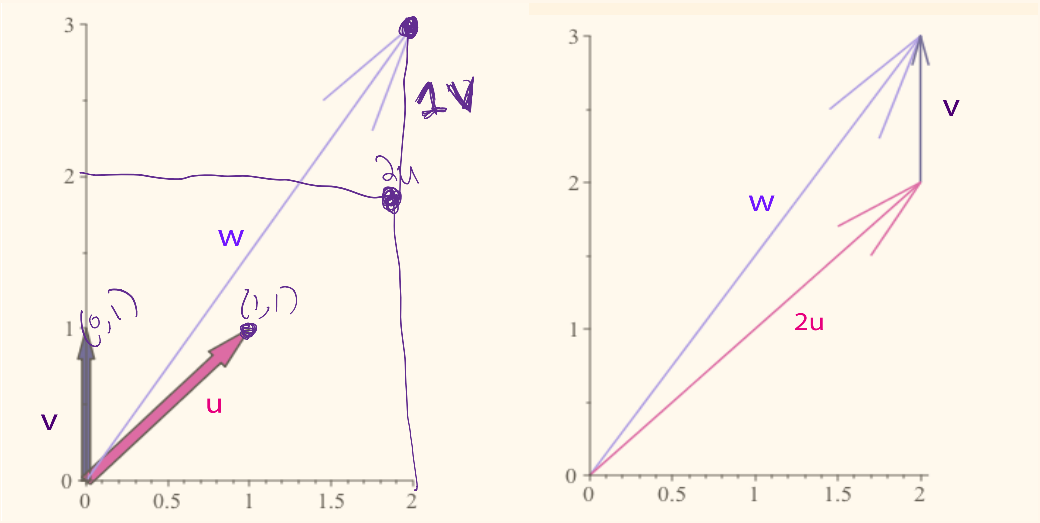


Figure 1: A visual depiction of the pink vector $u = (1, 1)$, the dark purple vector $v = (0, 1)$, and the vector sum $w = 2u + v$.

Definition of Linear Independence

span?, Cannot be represented as a linear combination of that span

Definition. Let E be a vector space. We say that a family $(u_i)_{i\in I}$ is *linearly independent* iff for every family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$,

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{implies that} \quad \lambda_i = 0 \;\text{ for all } i \in I.$$

only way to get 0 is if scalars are 0

b/c cannot mult. any constant scaler by 0 and get another number so its not linearly combined.

Equivalently, a family $(u_i)_{i\in I}$ is *linearly dependent* iff there is some family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$ such that

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{and} \quad \lambda_j \neq 0 \;\text{ for some } j \in I.$$

We agree that when $I = \emptyset$, the family $\emptyset$ is linearly independent.

Examples of Linear Independent Families

lambda   mu

1. Any two distinct scalars $\lambda, \mu \neq 0$ in $\mathbb{R}$ are linearly dependent.
2. In $\mathbb{R}^3$, the vectors $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ are linearly independent.

dimensions

Standard Basis

$$\begin{bmatrix}1\\0\\0\end{bmatrix} \begin{bmatrix}0\\1\\0\end{bmatrix} \begin{bmatrix}0\\0\\1\end{bmatrix}$$

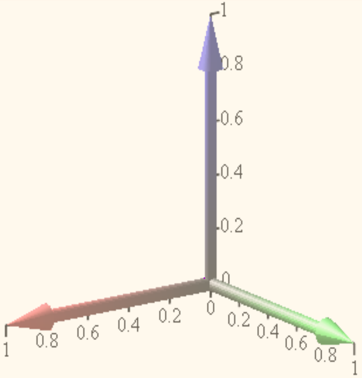


Figure 2: A visual depiction of the red vector $(1,0,0)$, the green vector $(0,1,0)$, and the blue vector $(0,0,1)$ in $\mathbb{R}^3$.

Linear Independence

$a(1,1,1,1) + b(0,1,1,1) + c(0,0,1,1) + d(0,0,0,1)$

$$\begin{bmatrix}1\\1\\1\\1\end{bmatrix} + \begin{bmatrix}0\\1\\1\\1\end{bmatrix} + \begin{bmatrix}0\\0\\1\\1\end{bmatrix} + \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

Self Check.

In $\mathbb{R}^4$, are the vectors $(1,1,1,1)$, $(0,1,1,1)$, $(0,0,1,1)$, $(0,0,0,1)$ are linearly independent?

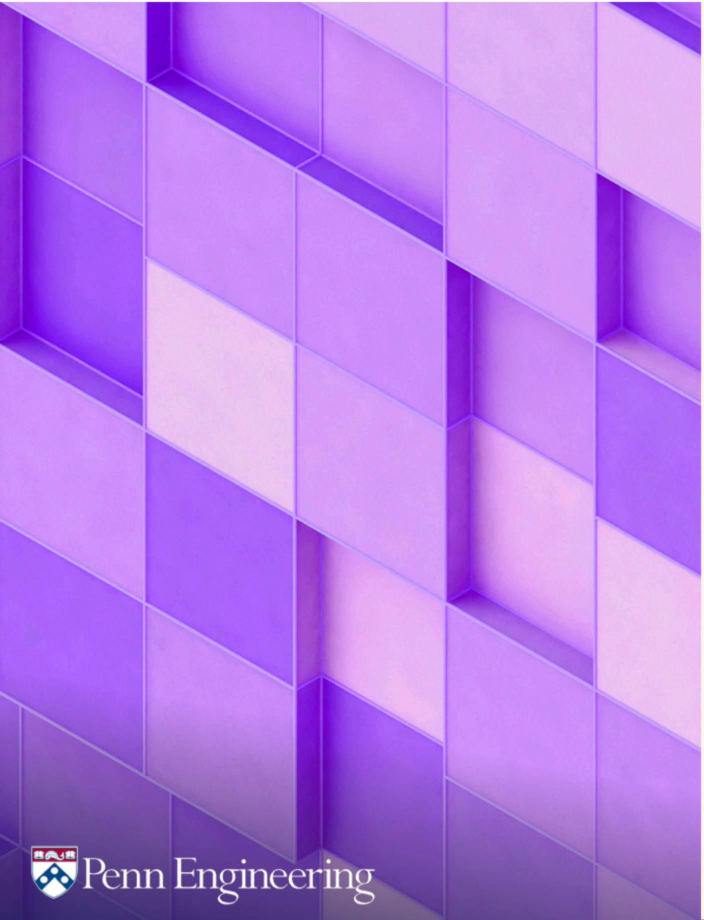
yes

Self Check.

In $\mathbb{R}^2$, are the vectors $u = (1,1)$, $v = (0,1)$ and $w = (2,3)$ linearly independent?

$$\begin{bmatrix}1\\1\end{bmatrix} \begin{bmatrix}0\\1\end{bmatrix} \begin{bmatrix}2\\3\end{bmatrix}$$

no



CIS 515: Math for Machine Learning

Linear Subspaces

Professors Jean Gallier & Jocelyn Quaintance

Linear Subspace

1) $\vec{x} \in F$ iff $c\vec{x} \in F$
scalar

2) contains 0 vector

Given a vector space E, linear combinations provide a way of obtaining vector subspaces of E.

Definition. Given a vector space E, a subset F of E is a *linear subspace* (or *subspace*) of E iff F is nonempty and $\lambda u + \mu v \in F$, for all $u, v \in F$, and all $\lambda, \mu \in \mathbb{R}$.

3) add two vectors F, still end up w/ a vector in F

Note that these axioms imply that $0 \in F$.

Examples of a Linear Subspace

1. In $\mathbb{R}^2$, the set of vectors $u = (x, y)$ such that

$$x + y = 0$$ $y = -x$

is the subspace illustrated by Figure 1.

$m = -1$

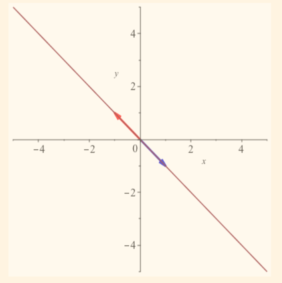


Figure 1: The subspace $x + y = 0$ is the line through the origin with slope -1. It consists of all vectors of the form $\lambda(-1, 1)$.

Examples of a Linear Subspace

2. In $\mathbb{R}^3$, the set of vectors $u = (x, y, z)$ such that

$$x + y + z = 0$$

is the subspace illustrated by Figure 2.

$z = -x - y$

coefficients = 1, 1, 1

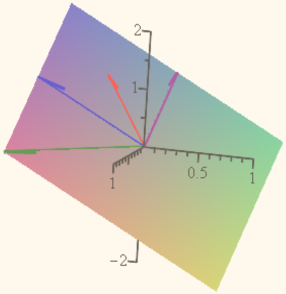


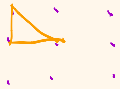
Figure 2: The subspace $x + y + z = 0$ is the plane through the origin with red normal $(1, 1, 1)$.

Examples of a Linear Subspace

larger examples

3. For any $n \geq 0$, the set of polynomials $f(X) \in \mathbb{R}[X]$ of degree at most n is a subspace of $\mathbb{R}[X]$.

4. The set of upper triangular $n \times n$ matrices is a subspace of the space of $n \times n$ matrices.



Sub-space

Span of a Set

To see how a set of vectors in E generates a subspace, we use the following proposition.

Proposition. Given any vector space E, if S is any nonempty subset of E, then the smallest subspace $\langle S\rangle$ (or $\mathrm{Span}(S)$) of E containing S is the set of all (finite) linear combinations of elements from S.

Span of a Set

For an illustration of this proposition, let $E = \mathbb{R}^3$ and $S = \{(1,1,0),(0,1,1))\}$. The $\mathrm{Span}(S)$ is all vectors of the form

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

where $\alpha, \beta \in \mathbb{R}$.

Note $\mathrm{Span}(S)$ is a plane through the origin with equation

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

CIS 515: Math for Machine Learning

Bases

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Bases = nice set of Vectors

Now that we understand the notion of linear combinations and how finite linear combinations generate vector subspaces, we can extend these notions to find a "nice" set of vectors S whose linear combinations generate the *entire* vector space. Of course, this "nice" set S of *generating* vectors should avoid *unnecessary* repetition. Then S is a basis (i.e an efficient construction set) of the vector space. The intuition of this paragraph is captured by the following definition.

Definition of a Basis

Definition. Given a vector space E and a subspace V of E, a family $(v_i)_{i \in I}$ of vectors $v_i \in V$ *spans V* or *generates V* iff for every $v \in V$, there is some family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$ such that

$$v = \sum_{i \in I} \lambda_i v_i.$$

We also say that the elements of $(v_i)_{i \in I}$ are *generators* of V and that V is *spanned by* $(v_i)_{i \in I}$, or *generated by* $(v_i)_{i \in I}$. If a subspace V of E is generated by a finite family $(v_i)_{i \in I}$, we say that V is *finitely generated*. A family $(u_i)_{i \in I}$ that spans V and is linearly independent is called a *basis* of V.

Examples of a Basis

1. In $\mathbb{R}^3$, the vectors $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ form a basis.
2. The vectors $(1,1,1,1), (1,1,-1,-1), (1,-1,0,0), (0,0,1,-1)$ form a basis of $\mathbb{R}^4$ known as the *Haar basis*. This basis and its generalization to dimension 2^n are crucial in wavelet theory.
3. In the subspace of polynomials in $\mathbb{R}[X]$ of degree at most n, the polynomials $1, X, X^2, \ldots, X^n$ form a basis.
4. The *Bernstein polynomials* $\binom{n}{k}(1-X)^{n-k}X^k$ for $k = 0, \ldots, n$, also form a basis of that space. These polynomials play a major role in the theory of spline curves.

Existence of a Basis

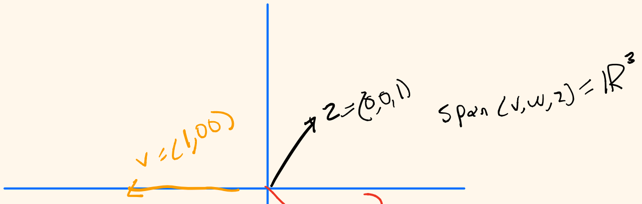
The first key result of linear algebra is that *every vector space E has a basis*!

Existence of a Basis

you start w/ subspace E

Intuitively, the construction of a basis is done by starting with a single vector v, taking the span of v, finding a vector not in the span of v, say w, calculating the $\mathrm{Span}(\{v, w\})$, and continuing this process recursively until we span the entire vector space. add new vectors not in span until all of E

For example, let $E = \mathbb{R}^3$ and $v = (1, 0, 0)$. The span of $(1, 0, 0)$ is the line $t(1, 0, 0)$. Let $w = (0, 1, 0)$. Then $\mathrm{Span}(\{v, w\})$ is the plane $z = 0$. Note $u = (0, 0, 1)$ is not contained in this plane. Form $\mathrm{Span}(\{v, w, u\})$ and check that it generates E.



span $\{v, w\}$ = x, y plane
invite all linear combinations
of this $\alpha(1,0,0) + \beta(0,1,0)$
$= (\alpha, \beta, 0)$

$w = (0,1,0)$
$w \notin \mathrm{span}(v)$

Existence of a Basis

The preceding incremental process is recorded in the following lemma.

Lemma. Given a linearly independent family $(u_i)_{i \in I}$ of elements of a vector space E, if $v \in E$ is not a linear combination of $(u_i)_{i \in I}$, then the family $(u_i)_{i \in I} \cup_k (v)$ obtained by adding v to the family $(u_i)_{i \in I}$ is linearly independent (where $k \notin I$).

Definition of Basis Revisited

An alternative definition for basis is obtained by using the notions of maximal linear independence and minimal generation.

Definition. Let $(v_i)_{i \in I}$ be a family of vectors in a vector space E. We say that $(v_i)_{i \in I}$ a *maximal linearly independent family of E* if it is linearly independent, and if for any vector $w \in E$, the family $(v_i)_{i \in I} \cup_k \{w\}$ obtained by adding w to the family $(v_i)_{i \in I}$ is linearly dependent. We say that $(v_i)_{i \in I}$ a *minimal generating family of E* if it spans E, and if for any index $p \in I$, the family $(v_i)_{i \in I - \{p\}}$ obtained by removing v_p from the family $(v_i)_{i \in I}$ does not span E.

Example of Maximal Linearly Independent Family

Let $E = \mathbb{R}^3$ and (v_1, v_2, v_3) be

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

The reader should check that (v_1, v_2, v_3) is a linearly independent family. We claim (v_1, v_2, v_3) is a maximal linear independent family of $\mathbb{R}^3$.

To prove this claim, given any $w = (w_1, w_2, w_3) \in \mathbb{R}^3$, we must find $\underbrace{\alpha, \beta, \gamma}_{\text{scalars?}} \in \mathbb{R}$ (in terms of w_1, w_2, w_3) such that

$$\alpha v_1 + \beta v_2 + \gamma v_3 = \alpha(1, 1, 0) + \beta(0, 1, 1) + \gamma(1, -1, 1) = (w_1, w_2, w_3).$$

Example of Maximal Linearly Independent Family

The equality of the coefficients implies that

$$\begin{aligned} \alpha + \gamma &= w_1 \\ \alpha + \beta - \gamma &= w_2 \\ \beta + \gamma &= w_3. \end{aligned}$$

We use the first and third equations to obtain $\alpha = w_1 - \gamma$ and $\beta = w_3 - \gamma$. We put these equivalences into the second equation and find that

$$\gamma = \frac{w_1 + w_3 - w_2}{3} \qquad \alpha = \frac{2w_1 - w_3 + w_2}{3} \qquad \beta = \frac{2w_3 - w_1 + w_2}{3}.$$

Example of a Minimal Generating Family

We also claim that (v_1, v_2, v_3) is a minimal generating family for $\mathbb{R}^3$.

This means if remove any vector from (v_1, v_2, v_3), we will not obtain all vectors in $\mathbb{R}^3$.

We leave it to the reader to show that $\mathrm{Span}(\{v_1, v_2\})$ is the plane through the origin of equation $x - y + z = 0$. Since this plane has normal $(1, -1, 1)$, we see that v_3 is not contained in $\mathrm{Span}(\{v_1, v_2\})$.

Example of a Minimal Generating Family

removing v_2

We also claim $v_2 \notin \text{Span}(\{v_1, v_3\})$. continue by contradiction

Suppose $v_2 \in \text{Span}(\{v_1, v_3\})$. Then we can find $\alpha, \beta \in \mathbb{R}$ such that

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

This implies that

$2 + 1 \neq 0$, so contradiction

$$\alpha + \gamma = 0$$

$$\alpha - \gamma = 1, \alpha = 2$$

$$\gamma = 1.$$

Example of a Minimal Generating Family

Since $\gamma = 1$, the second equation implies $\alpha = 2$. But then $\alpha + \gamma \neq 0$, contradicting the first equation. Hence, the system is unsolvable, which means $v_2 \notin \mathrm{Span}(\{v_1, v_3\})$.

Definition of a Basis Revisited

As promised, the concepts of maximal independence and minimal generation provide the following alternative characterization of a basis.

Proposition. Given a vector space E, for any family $B = (v_i)_{i \in I}$ of vectors of E, the following properties are equivalent:

(1) B is a basis of E. (1), (2 or 3)

(2) B is a maximal linearly independent family of E.

or

(3) B is a minimal generating family of E.

Example of a Basis

Going back to the example of $E = \mathbb{R}^3$ and

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},$$

since (v_1, v_2, v_3) is a maximal linear independent family, it is also a *basis* for $\mathbb{R}^3$. We should compare this basis to the standard basis

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and observe that both basis have the same cardinality.

CIS 515: Math for Machine Learning

The Replacement Lemma

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Dimension of a Vector Space

The second key result of linear algebra is that for any two bases $(u_i)_{i \in I}$ and $(v_j)_{j \in J}$ of a vector space E, the index sets I and J have the same *cardinality*.

vector space

In particular, if E has a finite basis of n elements, every basis of E has n elements, and the integer n is called the *dimension* of the vector space E.

Replacement Lemma

To prove this result we can use the following *replacement lemma* due to Steinitz. This result shows the relationship between finite linearly independent families and finite families of generators of a vector space.

Replacement Lemma

Proposition (*Replacement Lemma*). Given a vector space E, let $(u_1, \ldots, u_m)$ be any finite linearly independent family in E, and let $(v_1, \ldots, v_n)$ be any finite family such that every u_i is a linear combination of $(v_1, \ldots, v_n)$. Then we must have $m \leq n$, and there is a replacement of m of the vectors v_j by $(u_1, \ldots, u_m)$, such that after renaming some of the indices of the v_js, the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace of E.

Any vector $V_{n+1} \ldots n+m$ has to be independent

Example of the Replacement Lemma

Here is an example illustrating the replacement lemma. Consider sequences (u_1, u_2, u_3) and $(v_1, v_2, v_3, v_4, v_5)$, where (u_1, u_2, u_3) is a linearly independent family and with the u_is expressed in terms of the v_js as follows:

$$\begin{aligned}
u_1 &= v_4 + v_5 \\
u_2 &= v_3 + v_4 - v_5 \\
u_3 &= v_1 + v_2 + v_3.
\end{aligned}$$

Example of the Replacement Lemma

From the first equation we get

$$v_4 = u_1 - v_5,$$

and by substituting in the second equation we have

$$u_2 = v_3 + v_4 - v_5 = v_3 + \overset{v_4}{(u_1 - v_5)} - v_5 = u_1 + v_3 - 2v_5.$$

Example of the Replacement Lemma

From the above equation we get

$$v_3 = -u_1 + u_2 + 2v_5,$$

and so

$$u_3 = v_1 + v_2 + v_3 = v_1 + v_2 - u_1 + u_2 + 2v_5.$$

Finally we get

$$v_1 = u_1 - u_2 + u_3 - v_2 - 2v_5.$$

Example of the Replacement Lemma

Therefore we have

$$\begin{aligned} v_1 &= u_1 - u_2 + u_3 - v_2 - 2v_5 \\ v_3 &= -u_1 + u_2 + 2v_5 \\ v_4 &= u_1 - v_5, \end{aligned}$$

which shows that $(u_1, u_2, u_3, v_2, v_5)$ spans the same subspace as $(v_1, v_2, v_3, v_4, v_5)$. The vectors (v_1, v_3, v_4) have been replaced by (u_1, u_2, u_3), and the vectors left over are (v_2, v_5). We can rename them (v_4, v_5).

Fundamental Basis Theorem

every Basis of E has same # of vectors

The replacement lemma is key to proving the following important theorem.

Theorem (*Fundamental Basis Theorem*). Let E be a finitely generated vector space. Any family $(u_i)_{i \in I}$ generating E contains a subfamily $(u_j)_{j \in J}$ which is a basis of E. Any linearly independent family $(u_i)_{i \in I}$ can be extended to a family $(u_j)_{j \in J}$ which is a basis of E (with $I \subseteq J$). Furthermore, for every two bases $(u_i)_{i \in I}$ and $(v_j)_{j \in J}$ of E, we have $|I| = |J| = n$ for some fixed integer $n \geq 0$.

Fundamental Basis Theorem

The fundamental basis theorem also holds for vector spaces that are not finitely generated.

Dimension of a Vector Space

When a vector space E is not finitely generated, we say that E is of infinite dimension.

Definition. The *dimension* of a finitely generated vector space E is the common dimension n of all of its bases and is denoted by $\dim(E)$.

Lines, Planes, Hyperplanes

If E is a vector space of dimension $n \geq 1$, for any subspace U of E, if $\dim(U) = 1$, then U is called a *line*; $ax + by + c = 0 \rightarrow 2D$

if $\dim(U) = 2$, then U is called a *plane*; $ax + by + cz + d = 0 \rightarrow 3D$

if $\dim(U) = n - 1$, then U is called a *hyperplane*. more than 3D

If $\dim(U) = k$, then U is sometimes called a *k-plane*.

CIS 515: Math for Machine Learning

Coordinates

Professors Jean Gallier & Jocelyn Quaintance

Coordinates of a Vector

Let $(u_i)_{i \in I}$ be a basis of a vector space E. For any vector $v \in E$, since the family $(u_i)_{i \in I}$ generates E, there is a family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$, such that

$$v = \sum_{i \in I} \lambda_i u_i.$$ linear combination

A very important fact is that the family $(\lambda_i)_{i \in I}$ is unique.

scalars

Coordinates of a Vector

linear combination

Proposition. Given a vector space E, let $(u_i)_{i \in I}$ be a family of vectors in E. Let $v \in E$, and assume that $v = \sum_{i \in I} \lambda_i u_i$. Then the family $(\lambda_i)_{i \in I}$ of scalars such that $v = \sum_{i \in I} \lambda_i u_i$ is unique iff $(u_i)_{i \in I}$ is linearly independent.

Coordinates of a Vector

Proof. First assume that $(u_i)_{i \in I}$ is linearly independent.
By hypothesis $v = \sum_{i \in I} \lambda_i u_i$. If $(\mu_i)_{i \in I}$ is another family of scalars in $\mathbb{R}$ such that $v = \sum_{i \in I} \mu_i u_i$, then by subtraction we have

$$\sum_{i \in I} \lambda_i u_i - \sum_{i \in I} \mu_i u_i = \sum_{i \in I} (\lambda_i - \mu_i) u_i = 0,$$

and since $(u_i)_{i \in I}$ is linearly independent, we must have $\lambda_i - \mu_i = 0$ for all $i \in I$, that is, $\lambda_i = \mu_i$ for all $i \in I$.

Coordinates of a Vector

The converse is shown by contradiction.

Coordinates of a Vector

The converse is shown by contradiction. If $(u_i)_{i\in I}$ was linearly dependent, there would be a family $(\mu_i)_{i\in I}$ of scalars not all zero such that

$$\sum_{i\in I} \mu_i u_i = 0$$

and $\mu_j \neq 0$ for some $j \in I$.

Coordinates of a Vector

The converse is shown by contradiction. If $(u_i)_{i\in I}$ was linearly dependent, there would be a family $(\mu_i)_{i\in I}$ of scalars not all zero such that

$$\sum_{i\in I} \mu_i u_i = 0$$

and $\mu_j \neq 0$ for some $j \in I$. But then

$$v = \sum_{i\in I} \lambda_i u_i + 0 = \sum_{i\in I} \lambda_i u_i + \sum_{i\in I} \mu_i u_i = \sum_{i\in I} (\lambda_i + \mu_i) u_i,$$

with $\lambda_j \neq \lambda_j + \mu_j$ since $\mu_j \neq 0$, contradicting the assumption that $(\lambda_i)_{i\in I}$ is the unique family such that $v = \sum_{i\in I} \lambda_i u_i$. $\square$

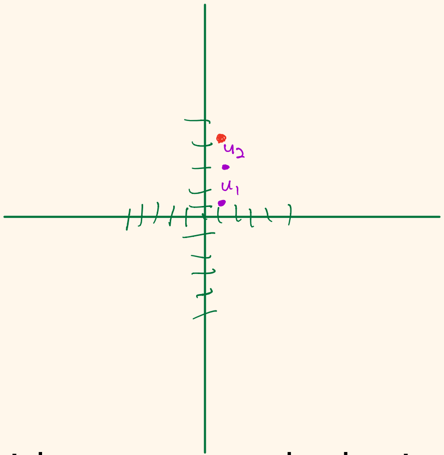
Coordinates of a Vector

Definition. If $(u_i)_{i \in I}$ is a basis of a vector space E, for any vector $v \in E$, if $(x_i)_{i \in I}$ is the unique family of scalars in $\mathbb{R}$ such that

$$v = \sum_{i \in I} x_i u_i,$$

each x_i is called the *component* (or *coordinate*) of index i of v with respect to the basis $(u_i)_{i \in I}$.

Coordinates of Vector



Self Check.

Find the coordinates of $(1, 4)$ with respect to the basis (u_1, u_2) where $u_1 = (1, 1)$ and $u_2 = (1, 3)$.

$$a(1,1) + b(1,3) = (1,4)$$

$x \rightarrow a + b = 1$

$y \rightarrow a + 3b = 4$

$a - a = 0$

$4 - 1 = 3$

$3b - b = 2b$

$2b = 3$

$b = \frac{3}{2}$

$a + \frac{3}{2} = 1, \quad a = -\frac{1}{2}$

$$-\frac{1}{2}(1,1) + \frac{3}{2}(1,3) = (1,4)$$

CIS 515: Math for Machine Learning

Coordinates

REPEAT

Professors Jean Gallier & Jocelyn Quaintance

Coordinates of a Vector

Let $(u_i)_{i\in I}$ be a basis of a vector space E. For any vector $v \in E$, since the family $(u_i)_{i\in I}$ generates E, there is a family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$, such that

$$v = \sum_{i\in I} \lambda_i u_i.$$

A very important fact is that the family $(\lambda_i)_{i\in I}$ is unique.

Coordinates of a Vector

Proposition. Given a vector space E, let $(u_i)_{i \in I}$ be a family of vectors in E. Let $v \in E$, and assume that $v = \sum_{i \in I} \lambda_i u_i$. Then the family $(\lambda_i)_{i \in I}$ of scalars such that $v = \sum_{i \in I} \lambda_i u_i$ is unique iff $(u_i)_{i \in I}$ is linearly independent.

Coordinates of a Vector

Proof. First assume that $(u_i)_{i\in I}$ is linearly independent.
By hypothesis $v = \sum_{i\in I} \lambda_i u_i$. If $(\mu_i)_{i\in I}$ is another family of scalars in $\mathbb{R}$ such that $v = \sum_{i\in I} \mu_i u_i$, then by subtraction we have

$$\sum_{i\in I} \lambda_i u_i - \sum_{i\in I} \mu_i u_i = \sum_{i\in I} (\lambda_i - \mu_i) u_i = 0,$$

and since $(u_i)_{i\in I}$ is linearly independent, we must have $\lambda_i - \mu_i = 0$ for all $i \in I$, that is, $\lambda_i = \mu_i$ for all $i \in I$.

Coordinates of a Vector

The converse is shown by contradiction. If $(u_i)_{i \in I}$ was linearly dependent, there would be a family $(\mu_i)_{i \in I}$ of scalars not all zero such that

$$\sum_{i \in I} \mu_i u_i = 0$$

and $\mu_j \neq 0$ for some $j \in I$. But then

$$v = \sum_{i \in I} \lambda_i u_i + 0 = \sum_{i \in I} \lambda_i u_i + \sum_{i \in I} \mu_i u_i = \sum_{i \in I} (\lambda_i + \mu_i) u_i,$$

with $\lambda_j \neq \lambda_j + \mu_j$ since $\mu_j \neq 0$, contradicting the assumption that $(\lambda_i)_{i \in I}$ is the unique family such that $v = \sum_{i \in I} \lambda_i u_i$. $\square$

Coordinates of a Vector

Definition. If $(u_i)_{i \in I}$ is a basis of a vector space E, for any vector $v \in E$, if $(x_i)_{i \in I}$ is the unique family of scalars in $\mathbb{R}$ such that

$$v = \sum_{i \in I} x_i u_i,$$

each x_i is called the *component* (or *coordinate*) of index i of v with respect to the basis $(u_i)_{i \in I}$.

Coordinates of Vector

Self Check.

Find the coordinates of $(1, 4)$ with respect to the basis (u_1, u_2) where $u_1 = (1, 1)$ and $u_2 = (1, 3)$.

CIS 515: Math for Machine Learning

Matrix Algebra

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Matrices

We have already been introduced to the vector space consisting of the set of $m \times n$ matrices. This vector space is important since matrices play a crucial role in representing linear transformations. Therefore, it pays to take a moment to investigate some of their properties.

Definition of a Matrix

Definition. If $K = \mathbb{R}$ or $K = \mathbb{C}$, an *$m \times n$-matrix over K* is a family $(a_{i\,j})_{1 \leq i \leq m,\ 1 \leq j \leq n}$ of scalars in K, represented by an array

$$K= \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix}$$

Scalars

.

Row and Column Vectors

In the special case where $m = 1$, we have a *row vector*, represented by

$$(a_{1\,1} \cdots a_{1\,n}),$$

and in the special case where $n = 1$, we have a *column vector*, represented by

$$\begin{pmatrix} a_{1\,1} \\ \vdots \\ a_{m\,1} \end{pmatrix}.$$

Definition of $\mathrm{M}_{m,n}$

The set of all $m \times n$-matrices is denoted by $\mathrm{M}_{m,n}(K)$ or $\mathrm{M}_{m,n}$. An $n \times n$-matrix is called a *square matrix of dimension n*. The set of all square matrices of dimension n is denoted by $\mathrm{M}_n(K)$, or M_n.

Identity Matrix

The square matrix I_n of dimension n containing 1 on the diagonal and 0 everywhere else is called the *identity matrix*. It is denoted by

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Transpose of a Matrix

Given an $m \times n$ matrix $A = (a_{i\,j})$, its *transpose* $A^{\top} = (a^{\top}_{j\,i})$, is the $n \times m$-matrix such that $a^{\top}_{j\,i} = a_{i\,j}$, for all i, $1 \leq i \leq m$, and all j, $1 \leq j \leq n$.

In other words, the transpose operation interchanges the rows and columns.

Transpose of a Matrix

one clockwise 45° motion

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

then

$$A^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

Algebraic Operations on Matrices: Sums

Sum, subtraction has to be done on same size matrix

Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define their *sum* $A + B$ as the matrix $C = (c_{ij})$ such that $c_{ij} = a_{ij} + b_{ij}$; that is,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

Algebraic Operations on Matrices: Scalar Multiples

For any matrix $A = (a_{i\,j})$, we let $-A$ be the matrix $(-a_{i\,j})$. Given a scalar $\lambda \in K$, we define the matrix λA as the matrix $C = (c_{i\,j})$ such that $c_{i\,j} = \lambda a_{i\,j}$; that is

$$\lambda \begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} = \begin{pmatrix} \lambda a_{1\,1} & \lambda a_{1\,2} & \dots & \lambda a_{1\,n} \\ \lambda a_{2\,1} & \lambda a_{2\,2} & \dots & \lambda a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m\,1} & \lambda a_{m\,2} & \dots & \lambda a_{m\,n} \end{pmatrix}.$$

Algebraic Operations on Matrices: Products

Given an $m \times n$ matrix $A = (a_{i\,k})$ and an $n \times p$ matrix $B = (b_{k\,j})$, we define their *product* AB as the $m \times p$ matrix $C = (c_{i\,j})$ such that

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j},$$

for $1 \leq i \leq m$, and $1 \leq j \leq p$.

Algebraic Operations on Matrices: Products

In the product $AB = C$

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} \begin{pmatrix} b_{1\,1} & b_{1\,2} & \dots & b_{1\,p} \\ b_{2\,1} & b_{2\,2} & \dots & b_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n\,1} & b_{n\,2} & \dots & b_{n\,p} \end{pmatrix} = \begin{pmatrix} c_{1\,1} & c_{1\,2} & \dots & c_{1\,p} \\ c_{2\,1} & c_{2\,2} & \dots & c_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m\,1} & c_{m\,2} & \dots & c_{m\,p} \end{pmatrix},$$

the entry of index i and j of the matrix AB can be identified with the product of the row matrix corresponding to the i-th row of A with the column matrix corresponding to the j-column of B:

Algebraic Operations on Matrices: Products

that is

$$\begin{pmatrix} a_{i\,1} & \cdots & a_{i\,n} \end{pmatrix} \begin{pmatrix} b_{1\,j} \\ \vdots \\ b_{n\,j} \end{pmatrix} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j}.$$

Properties of Matrix Multiplication for Square Matrices

When working $\mathrm{M}_n(K)$, besides the operations of matrix addition and scalar multiplication, we have the operation of matrix multiplication, which obeys the four properties:

For all $A, B, C \in \mathrm{M}_n(K)$,

(R1) $A(BC) = (AB)C$; (associativity)

(R2) $(A+B)C = AB + AC$; (distributivity)

(R3) $A(B+C) = AB + BC$; (distributivity)

(R4) $AI_n = I_nA = A$. (identity)

Ring of Square Matrices

Then $\mathrm{M}_n(K)$, with the operations of matrix addition and multiplication, is a *noncommutative ring*.

We can show that $\mathrm{M}_n(K)$ has *zero divisors*. This means we can find two nonzero $A, B \in \mathrm{M}_n(K)$ such that $AB = 0$.

Ring of Square Matrices

Self Check.

Find two 2×2 nonzero matrices A and B such that $AB = 0$.

Algebra of Square Matices

We also observe that matrix multiplication in $\mathrm{M}_n(K)$, when combined with scalar multiplication, obeys the following two properties:

For all $A, B \in \mathrm{M}_n(K)$ and all $\lambda \in K$,

(S1) $(\lambda A)B = \lambda(AB)$;

(S2) $A(\lambda B) = \lambda(AB)$.

Then $\mathrm{M}_n(K)$, with the operations of addition, scalar multiplication, and matrix multiplication is a *K-algebra*.

Inverse of a Square Matrix

For any square matrix A of dimension n, if a matrix B such that $AB = BA = I_n$ exists, then it is unique, and it is called the *inverse* of A. The matrix B is also denoted by A^{-1}.

An invertible matrix is also called a *nonsingular* matrix, and a matrix that is not invertible is called a *singular* matrix.

Matrix Multiplication for Rectangular Matrices

We can generalize Properties (R1) through (R4) and Properties (S1) and (S2) to the situation of rectangular matrices.

Matrix Multiplication for Rectangular Matrices

Proposition. (1) Given any matrices $A \in \mathrm{M}_{m,n}(K)$, $B \in \mathrm{M}_{n,p}(K)$, and $C \in \mathrm{M}_{p,q}(K)$, we have

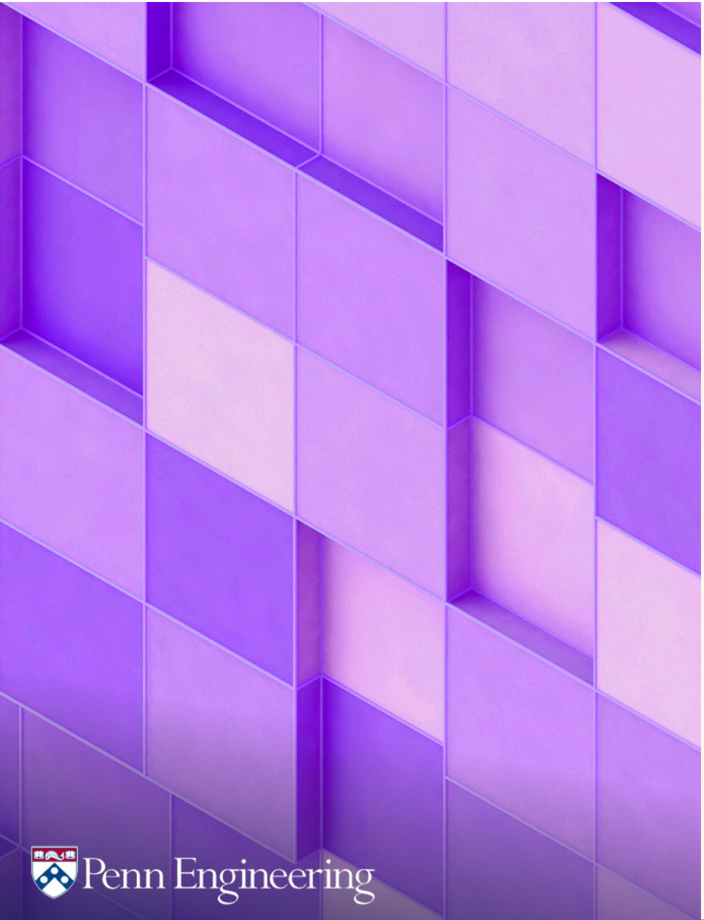
$$(AB)C = A(BC);$$

that is, matrix multiplication is associative.

(2) Given any matrices $A, B \in \mathrm{M}_{m,n}(K)$, and $C, D \in \mathrm{M}_{n,p}(K)$, for all $\lambda \in K$, we have

$$\begin{aligned} (A+B)C &= AC + BC & \qquad A(C+D) &= AC + AD \\ (\lambda A)C &= \lambda(AC) & A(\lambda C) &= \lambda(AC), \end{aligned}$$

so that matrix multiplication $\cdot \colon \mathrm{M}_{m,n}(K) \times \mathrm{M}_{n,p}(K) \to \mathrm{M}_{m,p}(K)$ is bilinear.



CIS 515: Math for Machine Learning

Linear Maps

Professors Jean Gallier & Jocelyn Quaintance

Linear Maps

Now that we understand vectors spaces, and how to generate them, we would like to perform various algebraic operations on them. In particular, we would like to be able to transform one vector space E into another vector space F. The tool for doing such a transformation is the linear map. Key to the notion of linear map is that it *preserves* the vector space structure.

Definition of a Linear Map

Definition. Given two vector spaces E and F, a *linear map* between E and F is a function $f\colon E \to F$ satisfying the following two conditions:

linear transformation

vector vector

$$f(x + y) = f(x) + f(y), \qquad \text{for all } x, y \in E;$$

Satisfies linear combinations

$$f(\lambda x) = \lambda f(x), \qquad \text{for all } \lambda \in \mathbb{R},\ x \in E.$$

Scalar

Properties of a Linear Map

Setting $x = y = 0$ in the first identity, we get $f(0) = 0$.

The basic property of linear maps is that they transform linear combinations into linear combinations. Given any finite family $(u_i)_{i \in I}$ of vectors in E, given any family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$, we have

$$f(\sum_{i \in I} \lambda_i u_i) = \sum_{i \in I} \lambda_i f(u_i).$$

Examples of a Linear Map

1. The map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined such that

$$\begin{aligned} x' &= x - y \\ y' &= x + y \end{aligned}$$

   is a linear map. The reader should check that it is the composition of a rotation by $\pi/4$ with a magnification of ratio $\sqrt{2}$.

2. For any vector space E, the *identity map* $\mathrm{id}\colon E \to E$ given by

$$\mathrm{id}(u) = u \quad \text{for all } u \in E$$

   is a linear map.

Examples of a Linear Map

Calculus

3. The map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ defined such that

$$D(f(X)) = f'(X),$$

   where $f'(X)$ is the derivative of the polynomial $f(X)$, is a linear map.

4. The map $\Phi\colon \mathcal{C}([a, b]) \to \mathbb{R}$ given by

$$\Phi(f) = \int_a^b f(t)\,dt,$$

   where $\mathcal{C}([a, b])$ is the set of continuous functions defined on the interval $[a, b]$, is a linear map.

Examples of a Linear Map

Calculus

5. The function $\langle -, - \rangle : \mathcal{C}([a,b]) \times \mathcal{C}([a,b]) \to \mathbb{R}$ given by

$$\langle f, g \rangle = \int_a^b f(t)g(t)\,dt,$$

is linear in each of the variable f, g. It also satisfies the properties $\langle f, g \rangle = \langle g, f \rangle$ and $\langle f, f \rangle = 0$ iff $f = 0$. It is an example of an *inner product*.

CIS 515: Math for Machine Learning

Kernels & Images

Professors Jean Gallier & Jocelyn Quaintance

Kernel and Image of a Linear Map

Every linear map determines two important subspaces, its kernel (or nullspace) and its image.

Definition of Kernel and Image

Definition. Given a linear map $f\colon E \to F$, we define its *image* (or *range*) $\operatorname{Im} f = f(E)$, as the set

target space

$$\operatorname{Im} f = \{y \in F \mid (\exists x \in E)(y = f(x))\},$$

and its *Kernel* (or *nullspace*) $\operatorname{Ker} f = f^{-1}(0)$, as the set

$$\operatorname{Ker} f = \{x \in E \mid f(x) = 0\}.$$

collection of all vectors
in source space mapped to zero by f

Examples of Kernel and Image

Recall that $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ is defined such that

$$D(f(X)) = f'(X),$$

where $f'(X)$ is the derivative of the polynomial $f(X)$.

The derivative map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ has kernel the constant polynomials, so $\operatorname{Ker} D = \mathbb{R}$.

The image of $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ is actually $\mathbb{R}[X]$ itself, because every polynomial $P(X) = a_0 X^n + \cdots + a_{n-1} X + a_n$ of degree n is the derivative of the polynomial $Q(X)$ of degree $n+1$ given by

$$Q(X) = a_0 \frac{X^{n+1}}{n+1} + \cdots + a_{n-1} \frac{X^2}{2} + a_n X.$$

Examples of Kernel and Image

Now consider the restriction of D to the vector space $\mathbb{R}[X]_n$ of polynomials of degree $\leq n$.

The kernel of D is still $\mathbb{R}$, but the image of D is the $\mathbb{R}[X]_{n-1}$, the vector space of polynomials of degree $\leq n-1$.

Kernel and Image of a Linear Map

Self Check.

Determine the kernel of the second derivative $D \circ D \colon \mathbb{R}[X] \to \mathbb{R}[X]$.

Kernel is the 2D subspace whose basis = 1 and x,
namely linear combinations of the constant maps
i.e. polynomials of form $ax+b$

Properties of Kernel and Image

The following proposition summarizes the important properties of the kernel and the image.

Proposition. Given a linear map $f \colon E \to F$, the set $\mathrm{Im}\, f$ is a subspace of F, and the set $\mathrm{Ker}\, f$ is a subspace of E. The linear map $f \colon E \to F$ is injective iff $\mathrm{Ker}\, f = (0)$ (where (0) is the trivial subspace $\{0\}$).

kernel and images are subspaces.

injective = nullspace reduced to 0 vector

Properties of Kernel and Image

nullspace not reduced to 0

The previous proposition shows that $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ is not injective.

Since the image $\operatorname{Im} f$ of a linear map f is a subspace of F, we can define the *rank* $\operatorname{rk}(f)$ of f as the dimension of $\operatorname{Im} f$.

Rank of a Linear Map

This happens in target space

Definition. Given a linear map $f\colon E \to F$, the *rank* $\mathrm{rk}(f)$ of f is the dimension of the image $\mathrm{Im}\, f$ of f.

CIS 515: Math for Machine Learning

Constructing Linear Maps

Professors Jean Gallier & Jocelyn Quaintance

Penn Engineering

Linear Maps

We know that every vector space has a basis and this basis generates the entire vector space. We will use the existence of a basis to *uniquely* determine the behavior a linear map. The idea is as follows. Suppose we have two vectors spaces E and F. Let $(u_i)_{i \in I}$ be a basis of E. If we know how a linear map "takes" each u_i into F, we will have constructed a *unique* linear map.

Example Construction of a Linear Map

Let $E = \mathbb{R}^3$ and $F = \mathbb{R}^2$. Let

vector vector vector

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad u_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

be a basis for E.

Example Construction of a Linear Map

Let $E = \mathbb{R}^3$ and $F = \mathbb{R}^2$. Let

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad u_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

be a basis for E.

We construct a linear map $f \colon E \to F$ by assigning each basis element of E an image in F.

Example Construction of a Linear Map

Suppose

$$f(u_1) = (1,0) = v_1 \qquad f(u_2) = (0,1) = v_2 \qquad f(u_3) = (1,1) = v_3.$$

Example Construction of a Linear Map

Suppose

$$f(u_1) = (1, 0) = v_1 \qquad f(u_2) = (0, 1) = v_2 \qquad f(u_3) = (1, 1) = v_3.$$

We claim f is uniquely determined since any $w \in \mathbb{R}^3$ may be written as

$$w = \lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3,$$

and the linearity of f ensures that

$$\begin{aligned} f(w) &= \lambda_1 f(u_1) + \lambda_2 f(u_2) + \lambda_3 f(u_3) = \lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3 \\ &= (\lambda_1 + \lambda_3, \lambda_2 + \lambda_3). \end{aligned}$$

Linear Map

This basis mapping construction for uniquely determining a linear map is recorded in the following proposition.

Basis Construction of Linear Map

Proposition. (*Basis Construction of Linear Map*) Given any two vector spaces E and F, given any basis $(u_i)_{i\in I}$ of E, given any other family of vectors $(v_i)_{i\in I}$ in F, there is a unique linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

Basis Construction of Linear Map

Proposition. (*Basis Construction of Linear Map*) Given any two vector spaces E and F, given any basis $(u_i)_{i \in I}$ of E, given any other family of vectors $(v_i)_{i \in I}$ in F, there is a unique linear map $f \colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$. Furthermore, f is injective iff $(v_i)_{i \in I}$ is linearly independent, and f is surjective iff $(v_i)_{i \in I}$ generates F.

Basis Construction of Linear Map

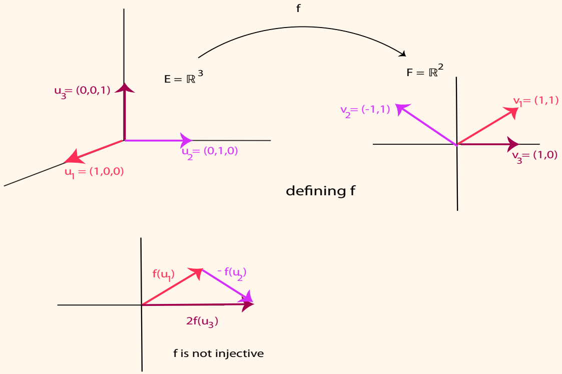


Figure 1: Given $u_1 = (1, 0, 0)$, $u_2 = (0, 1, 0)$, $u_3 = (0, 0, 1)$ and $v_1 = (1, 1)$, $v_2 = (-1, 1)$, $v_3 = (1, 0)$, define the unique linear map $f\colon \mathbb{R}^3 \to \mathbb{R}^2$ by $f(u_1) = v_1$, $f(u_2) = v_2$, and $f(u_3) = v_3$. This map is surjective but not injective since $f(u_1 - u_2) = f(u_1) - f(u_2) = (1, 1) - (-1, 1) = (2, 0) = 2f(u_3) = f(2u_3)$.

Injective Linear Map Properties

An injective linear map $f\colon E \to F$ sends a basis $(u_i)_{i\in I}$ to a linearly independent family $(f(u_i))_{i\in I}$ of F, which is also a basis when f is bijective.

Injective Linear Map Properties

An injective linear map $f\colon E \to F$ sends a basis $(u_i)_{i\in I}$ to a linearly independent family $(f(u_i))_{i\in I}$ of F, which is also a basis when f is bijective.

When E and F have the same finite dimension n, $(u_i)_{i\in I}$ is a basis of E, and $f\colon E \to F$ is injective, then $(f(u_i))_{i\in I}$ is a basis of F.

Proposition

Proposition. Given any two vector spaces E and F, with F nontrivial, given any family $(u_i)_{i \in I}$ of vectors in E, the following properties hold:

Proposition

Proposition. Given any two vector spaces E and F, with F nontrivial, given any family $(u_i)_{i\in I}$ of vectors in E, the following properties hold:

(1) The family $(u_i)_{i\in I}$ generates E iff for every family of vectors $(v_i)_{i\in I}$ in F, there is at most one linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$. See Figure 2.

Proposition

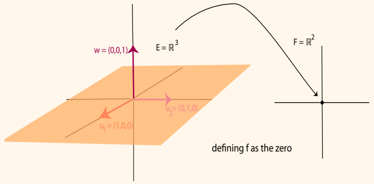
Proposition. Given any two vector spaces E and F, with F nontrivial, given any family $(u_i)_{i \in I}$ of vectors in E, the following properties hold:

(1) The family $(u_i)_{i \in I}$ generates E iff for every family of vectors $(v_i)_{i \in I}$ in F, there is at most one linear map $f \colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$. See Figure 2.

(2) The family $(u_i)_{i \in I}$ is linearly independent iff for every family of vectors $(v_i)_{i \in I}$ in F, there is some linear map $f \colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

Illustration of Proposition

2 linear maps can be constructed from $\mathbb{R}^3$ to $\mathbb{R}^2$



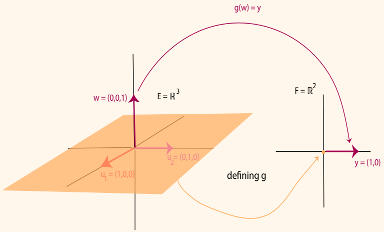
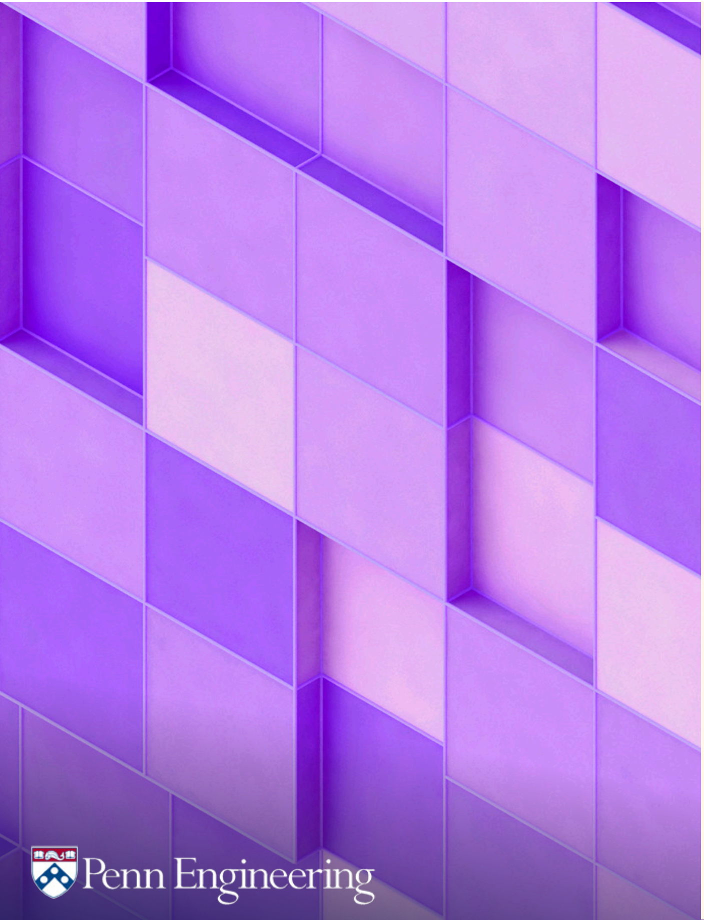


Figure 2: Let $E = \mathbb{R}^3$ and $F = \mathbb{R}^2$. The vectors $u_1 = (1, 0, 0)$, $u_2 = (0, 1, 0)$ do not generate $\mathbb{R}^3$ since both the zero map and the map g, where $g(u_1) = 0$, $g(u_2) = 0$ and $g(0, 0, 1) = (1, 0)$, send the peach xy-plane to the origin.



CIS 515: Math for Machine Learning

Isomorphisms and Endomorphisms of Vector Spaces

Professors Jean Gallier & Jocelyn Quaintance

Composition of Linear Maps

So far we have been working with only a single linear map $f\colon E \to F$. But the operation of composition allows us to combine two linear maps to obtain a new linear map.

Composition of Linear Maps

Given vector spaces E, F, and G, and linear maps $f\colon E \to F$ and $g\colon F \to G$, it is easily verified that the composition $g \circ f\colon E \to G$ of f and g is a linear map.

Here, as usual in mathematics

$$(g \circ f)(x) = g(f(x)).$$

Inverse of a Linear Map

Definition. A linear map $f\colon E \to F$ is an *isomorphism* iff there is a linear map $g\colon F \to E$, such that

$$g \circ f = \mathrm{id}_E \quad \text{and} \quad f \circ g = \mathrm{id}_F. \qquad (*)$$

Compose Compose

Inverse of Linear Map

The map g in the previous definition is unique.

Inverse of Linear Map

The map g in the previous definition is unique.

This is because if g and h both satisfy

$$g \circ f = \mathrm{id}_E, \ f \circ g = \mathrm{id}_F, \ h \circ f = \mathrm{id}_E, \text{ and } f \circ h = \mathrm{id}_F,$$

then

$$g = g \circ \mathrm{id}_F = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_E \circ h = h.$$

Inverse of Linear Map

The map g satisfying $(*)$ is called the *inverse* of f and it is also denoted by f^{-1}.

Isomorphisms between Vector Spaces

If $F = \mathbb{R}^n$, then we get an isomorphism between any vector space E of dimension n and $\mathbb{R}^n$.

Isomorphisms between Vector Spaces

If $F = \mathbb{R}^n$, then we get an isomorphism between any vector space E of dimension n and $\mathbb{R}^n$.

The basis construction proposition implies that if E and F are two vector spaces, $(u_i)_{i \in I}$ is a basis of E, and $f \colon E \to F$ is a linear map which is an isomorphism, then the family $(f(u_i))_{i \in I}$ is a basis of F.

Isomorphisms between Vector Spaces

One can verify that if $f\colon E \to F$ is a bijective linear map, then its inverse $f^{-1}\colon F \to E$, as a function, is also a linear map, and thus f is an isomorphism.

Hom Space

Definition. The set of all linear maps between two vector spaces E and F is denoted by $\mathrm{Hom}(E, F)$ or by $\mathcal{L}(E, F)$ (the notation $\mathcal{L}(E, F)$ is usually reserved to the set of continuous linear maps, where E and F are normed vector spaces).

Vector Space Structure of $\mathrm{Hom}(E, F)$

The set $\mathrm{Hom}(E, F)$ is a vector space under the operations defined below:

Vector Space Structure of $\mathrm{Hom}(E, F)$

The set $\mathrm{Hom}(E, F)$ is a vector space under the operations defined below:

Pointwise Addition:

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in E$,

Vector Space Structure of $\mathrm{Hom}(E, F)$

The set $\mathrm{Hom}(E, F)$ is a vector space under the operations defined below:

Pointwise Addition:

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in E$,

Pointwise Scalar Multiplication:

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in E$.

Endomorphisms

Definition. When $E = F$, a linear map $f \colon E \to E$ is also called an *endomorphism*.

remains inside

Endomorphisms

Definition. When $E = F$, a linear map $f \colon E \to E$ is also called an *endomorphism*. The space $\mathrm{Hom}(E, E)$ is also denoted by $\mathrm{End}(E)$.

Ring structure of $\mathrm{End}(E)$

Composition confers to $\mathrm{Hom}(E, E)$ a ring structure. and vector space

Ring structure of $\mathrm{End}(E)$

Composition confers to $\mathrm{Hom}(E, E)$ a ring structure.

Composition is an operation $\circ\colon \mathrm{Hom}(E, E) \times \mathrm{Hom}(E, E) \to \mathrm{Hom}(E, E)$, which is associative and has an identity id_E, and the distributivity properties hold:

Ring structure of $\mathrm{End}(E)$

Composition confers to $\mathrm{Hom}(E, E)$ a ring structure.

Composition is an operation $\circ\colon \mathrm{Hom}(E, E) \times \mathrm{Hom}(E, E) \to \mathrm{Hom}(E, E)$, which is associative and has an identity id_E, and the distributivity properties hold:
w/ respect to addition

$$(g_1 + g_2) \circ f = g_1 \circ f + g_2 \circ f;$$
$$g \circ (f_1 + f_2) = g \circ f_1 + g \circ f_2.$$

Ring structure of $\mathrm{End}(E)$

Composition confers to $\mathrm{Hom}(E, E)$ a ring structure.

Composition is an operation $\circ \colon \mathrm{Hom}(E, E) \times \mathrm{Hom}(E, E) \to \mathrm{Hom}(E, E)$, which is associative and has an identity id_E, and the distributivity properties hold:

$$(g_1 + g_2) \circ f = g_1 \circ f + g_2 \circ f;$$
$$g \circ (f_1 + f_2) = g \circ f_1 + g \circ f_2.$$

The ring $\mathrm{Hom}(E, E)$ is an example of a noncommutative ring.

Automorphisms

It is easily seen that the set of bijective linear maps $f\colon E \to E$ is a group under composition.

Automorphisms

Definition. Bijective linear maps $f \colon E \to E$ are also called *automorphisms*.

Automorphisms

Definition. Bijective linear maps $f \colon E \to E$ are also called *automorphisms*. The group of automorphisms of E is called the *general linear group (of E)*, and it is denoted by $\mathbf{GL}(E)$, or by $\mathrm{Aut}(E)$, or when $E = \mathbb{R}^n$, by $\mathbf{GL}(n, \mathbb{R})$, or even by $\mathbf{GL}(n)$.