Fundamentals Review

Pedro Vilanova

September 6, 2023

Contents

Sets

What is meant by a set is intuitively clear: "bag" of objects. Sets are denoted with uppercase, objects with lowercase.

- (Belongs). An object x belongs to a set A is denoted by $x \in A$. For instance, $3 \in \{1, 2, 3, 4, 5\}$ but $7 \notin \{1, 2, 3, 4, 5\}$.
- (Emptyset). A set which contains no elements is denoted $\emptyset$, i.e., for every object x we have $x \notin \emptyset$.
- (Sets are objects). If A is a set, then A is also an object. In particular, given two sets A and B, it is meaningful to ask $A \in B$?
- (Equality of sets). Two sets A and B are equal, $A = B$, iff [1] every element of A is an element of B and vice versa.

Sets

- (Subsets). Let A, B be sets. We say that A is a subset of B, denoted $A \subseteq B$, iff every element of A is also an element of B

  $$\forall x : (x \in A) \Rightarrow (x \in B) .$$

  We say that A is a proper subset of B, denoted $A \subsetneq B$, if $A \subseteq B$ and $A \neq B$.

- Given any set A, we always have $A \subseteq A$ (why?) and $\emptyset \subseteq A$ (why?). What about $A \in A$? Is 2 an element or a subset of $\{1, 2, 3\}$? Is $\{2\}$ an element or a subset of $\{1, 2, 3\}$? It is important to distinguish sets from their elements, as they can have different properties. Is it possible to have an infinite set consisting of finite numbers? Is it possible to have a finite set consisting of infinite objects?

- Examples: Is $\{\emptyset\} = \emptyset$ (why?). What about $\{\{\emptyset\}\}$ and $\{\emptyset, \{\emptyset\}\}$? Are those three sets equal? (why?)

- (Axiom of specification). Let A be a set, and for each $x \in A$, let $P(x)$ be a property pertaining to x that is either true or false. Then there exists the set

$$\{x \in A : P(x) \text{ is true}\}.$$

- (Pairwise union). Given any two sets A, B, there exists a set $A \cup B$, called the union of A and B:

$$\forall x : (x \in A \cup B) \iff (x \in A \text{ or } x \in B).$$

  Recall that "or" in mathematics is by default inclusive.

- (Intersection). Given any two sets A, B, there exists a set $A \cap B$, called the intersection of A and B:

$$\forall x : (x \in A \cap B) \iff (x \in A \text{ and } x \in B).$$

[1] if and only if

Functions

The concept of a function (or map) is central to all of mathematics. We begin with ordered pair using sets.

Definition (Ordered pair)

of elements x and y, written (x, y), is defined

$$(x, y) := \{\{x\}, \{x, y\}\}.$$

Two ordered pairs (a, b) and (c, d) are equal iff $a = c$ and $b = d$.

Definition (Cartesian product)

of sets A and B denoted $A \times B$ is

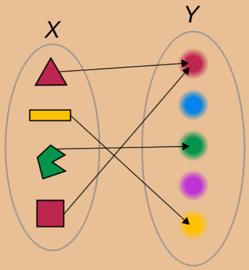
$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\}.$$

Definition (Function)

A function from set X into Y is a subset of $X \times Y$ denoted $f : X \to Y$ that satisfies

1. *If (x, y) and (x, y') belong to f, then $y = y'$.*
2. *If $x \in X$, then $(x, y) \in f$ for some $y \in Y$.*

The crucial property of a function is that with each (2) element x in X there is associated a unique (1) element y in Y.



X is called domain, Y is called codomain. In this course usually $X = \mathbb{R}^n$ and $Y = \mathbb{R}$.

Figure credit: wikipedia

Functions

Notation: If $(x, y) \in f$, we write $y = f(x)$ and call y the (direct) image of x under f.

Definition

Let $f : X \to Y$. The range of f is the set

$$\{f(x) : x \in X\}.$$

The range and the codomain are not necessary the same, the range is a subset of Y.

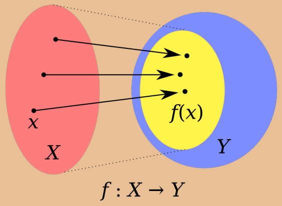


Figure credit: wikipedia

Definition

Let $f : X \to Y$. Let $A \subseteq X$ and $B \subseteq Y$.

1. *The image of A under f is the set $f(A) = \{f(x) : x \in A\}$.*
2. *The inverse image of B under f is the set $f^{-1}(B) = \{x : f(x) \in B\}$.*

Definition

Let $f : X \to Y$.

1. *f is surjective if*
$$f(X) = Y\,.$$
The range and codomain coincide.
2. *f is injective if*
$$\forall x, z \in X : (f(x) = f(z)) \Rightarrow (x = z)\,.$$
3. *f is bijective if*
$$\forall y \in Y : \exists! x \in X \text{ such that } y = f(x)\,.$$
where $\exists! x$ means "there exists exactly one x".

Functions

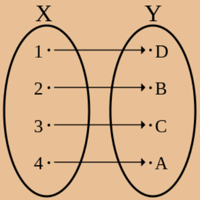
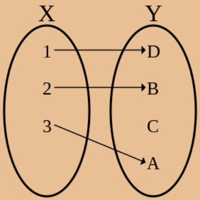
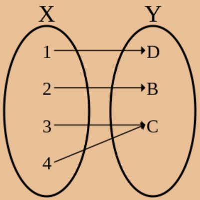
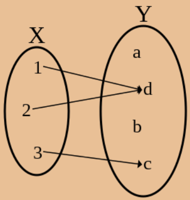
	surjective	non-surjective
injective	 bijective	 injective-only
non-injective	 surjective-only	 general

Figure credit: wikipedia

Definition (Inverse function)

If $f : X \to Y$ is injective, we may define the inverse function to f, denoted f^{-1}, from the range of f onto X by

$$(y, x) \in f^{-1} \text{ iff } (x, y) \in f\,.$$

Note: The inverse function f^{-1} is defined only if f is injective but the inverse image $f^{-1}(B)$ is defined for an arbitrary function f and for all sets $B \subseteq Y$.

Definition

If $f : X \to Y$ and $g : Y \to Z$, we define the composition $g \circ f : X \to Z$ by

$$(g \circ f)(x) = g(f(x)) \quad \text{for all } x \in X\,.$$

Functions

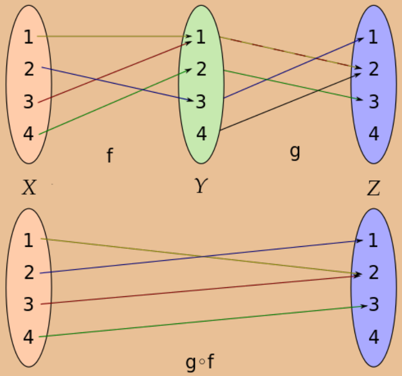


Figure: wikipedia

Example: Let

$$f = \{(1,1),(2,1),(3,4)\}, \quad A = \{1,2\}, \quad B = \{1\}\,.$$

The domain of f is $\{1,2,3\}$ and the range of f is $\{1,4\}$. The image of A under f is the set $f(A) = \{1\}$. The inverse image of B under f is the set $f^{-1}(B) = \{1,2\}$. If $Y = \{1,4\}$ then f is surjective. The function f is not injective since $f(1) = f(2)$.
Let $g = \{(1,1),(2,3)\}$. Then g is injective and the inverse function is

$$g^{-1} = \{(1,1),(3,2)\}\,.$$

The composition $g \circ f$ is the function

$$g \circ f = \{(1,1),(2,4)\}\,.$$

Quiz

Let $g = \{(1,2),(2,2),(3,1),(4,4)\}$, $f = \{(1,5),(2,7),(3,9),(4,17)\}$, and $A = \{1,2\}$. Determine

(a) The domain of g

(b) The range of g

(c) $g(A)$

(d) $g^{-1}(A)$

(e) $g \circ f$

(f) f^{-1}

Infimum and Supremum

Previous definitions (sets and functions) were very general. In this course we are interested in sets and functions on the real numbers:

Definition

The real numbers $\mathbb{R}$ *is a set of objects satisfying a set of algebraic axioms (addition and multiplication), a total order axiom, and the least upper bound axiom.*

We now discuss a fundamental property idea behind real numbers: Infimum and Supremum. First we recall the following: The *well-ordering principle* states that every non-empty subset of the natural numbers $\mathbb{N}$ has a least element (minimum). However, this property does not extend to arbitrary subsets of the real numbers $\mathbb{R}$:

A subset of $\mathbb{R}$ may or may not have a minimum element.

Infimum and Supremum

Example: interval $(0,1] \subset \mathbb{R}$. This set does not have a minimum element, as for any $x \in (0,1]$, you can always find another element in the set that is smaller than x but still greater than 0. On the other hand, $[0,1] \subset \mathbb{R}$ does have a minimum element, which is 0.

This issue is addressed through the concepts of the least upper bound (supremum) and the greatest lower bound (infimum).

The least upper bound and greatest lower bound provide a way to "complete" the real numbers, ensuring that there are no "gaps" in the number line. This completeness property is essential for the development of calculus and real analysis.

Then, while the open interval $(0,1)$ does not have a minimum or maximum, we can say that its infimum is 0 and its supremum is 1.

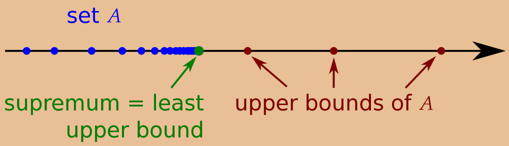
Infimum and Supremum

Let A be a nonempty subset of $\mathbb{R}$. The set of upper bounds of A, denoted $U(A)$, is defined as

$$U(A) = \{u \in \mathbb{R} \mid u \geq a \text{ for all } a \in A\}\,,$$

while the set of lower bounds of A, denoted $L(A)$, is given by

$$L(A) = \{l \in \mathbb{R} \mid l \leq a \text{ for all } a \in A\}\,.$$



Task: draw an analogous figure for infimum.

Infimum and Supremum

Remark: Any $U(A)$ and $L(A)$ could be empty. Examples:

- If $A = \mathbb{N}$ then $U(A)$ is empty.
- If $A = \mathbb{Z}$, the set of all integers, then both $U(A)$ and $L(A)$ are empty.
- If $U(A)$ is nonempty, then A is said to be bounded above.
- If $L(A)$ is nonempty, then A is said to be bounded below.

Definition

The supremum of A, denoted $\sup(A)$, is defined to be the least upper bound of A. Namely, if $U(A)$ is nonempty, then $\sup(A)$ is defined to be the unique point $a^* \in U(A)$ such that $a^* \leq u$ for all $u \in U(A)$. If $U(A)$ is empty, by convention, we set $\sup(A) = +\infty$.

Remark: Here we take an axiomatic approach to the real line. In this approach, the supremum of A is well defined, i.e., there is $a^* \in U(A)$ such that $a^* \leq u$ for all $u \in U(A)$ when $U(A)$ is non-empty. Other

Infimum and Supremum

authors adopt a constructive approach to the real line. In the constructive approach the following must be proved as a theorem: Let A be a non-empty subset of $\mathbb{R}$. If A has an upper bound, then it must have exactly one least upper bound.

Definition

The infimum of A, denoted $\inf(A)$, is defined to be the greatest lower bound of A. That is, when $L(A)$ is nonempty, then $\inf(A)$ is the unique point $\hat{a} \in L(A)$ such that $\hat{a} \geq l$ for all $l \in L(A)$. If $L(A)$ is empty, by convention we set $\inf(A) = -\infty$.

What if $A = \emptyset$? Then by convention we set $\inf(\emptyset) = +\infty$ and $\sup(\emptyset) = -\infty$.

Remark: $+\infty$ and $-\infty$ are just symbols and not real numbers. We may add $+\infty$ and $-\infty$ to the reals to form the extended real numbers $\bar{\mathbb{R}}$, but this is not as convenient to work with as the real numbers. Many of the laws of algebra break down: what is $+\infty + -\infty$?

Infimum and Supremum

Two concepts closely related to sup and inf are the maximum and the minimum of a nonempty set $A \subset \mathbb{R}$:

Definition (Maximum and minimum)

The maximum of A, denoted $\max(A)$, is defined as a point $z \in A$ such that $z \geq a$ for all $a \in A$. The minimum of A, denoted $\min(A)$, is defined as a point $w \in A$ such that $w < a$ for all $w \in A$.

By definition, the maximum must be an upper bound of A, and the minimum must be a lower bound of A. Therefore, we can equivalently define $\max(A) = A \cap U(A)$, and $\min(A) = A \cap L(A)$.

While $\sup(A)$ and $\inf(A)$ are always defined for any nonempty set A (they could be infinite), $A \cap U(A)$ and $A \cap L(A)$ could both be empty, so $\max(A)$ and $\min(A)$ not always exist. This is true even if $\sup(A)$ and $\inf(A)$ are both finite.
For example, neither $\mathbb{R}$ nor the interval $(0, 1)$ has a minimum, but $\inf \mathbb{R} = -\infty$ and $\inf(0, 1) = 0$.

inf and sup properties

Let the sets $A, B \subseteq \mathbb{R}$, and scalar $r \in \mathbb{R}$. Define (arithmetic operations over sets):

- $rA = \{ra : a \in A\}$; the scalar product of a set is just the scalar multiplied by every element in the set.
- $A + B = \{a + b : a \in A,\ \ b \in B\}$; called the Minkowski sum, it is the arithmetic sum of two sets is the sum of all possible pairs of numbers, one from each set.

Property

In those cases where the infima and suprema of the sets A and B exist, the following identities hold:

- *If* $A \subseteq B$ *then* $\inf(A) \geq \inf(B)$ *and* $\sup(A) \leq \sup(B)$*.*
- $A \neq \emptyset$ *iff* $\sup(A) \geq \inf(A)$*, and otherwise* $-\infty = \sup(\emptyset) < \inf(\emptyset) = \infty$*.*
- $p = \inf(A)$ *iff* p *is a lower bound and for every* $\epsilon > 0$ *there is an* $a(\epsilon) \in A$ *such that* $a(\epsilon) > p + \epsilon$*. (lower bound that can be approximated).*
- *If* $r \geq 0$ *then* $\inf(rA) = r\inf(A)$ *and* $\sup(rA) = r\sup(A)$*.*
- *If* $r \leq 0$ *then* $\inf(rA) = r\sup(A)$ *and* $\sup(rA) = r\inf(A)$*.*
- $\inf(A + B) = \inf(A) + \inf(B)$ *and* $\sup(A + B) = \sup(A) + \sup(B)$*.*

Examples

- *The infimum of the set of numbers* $\{2, 3, 4\}$ *is* 2. *The number* 1 *is a lower bound, but not the greatest lower bound, and hence not the infimum.*
- *More generally, if a set has a smallest element, then the smallest element is the infimum for the set (this element is called the minimum of the set).*
- $\inf\{1, 2, 3, \ldots\} = 1$.
- $\inf\{x \in \mathbb{R} : 0 < x < 1\} = 0$.
- $\inf\{x \in \mathbb{Q} : x^3 > 2\} = \sqrt[3]{2}$.
- $\inf\{(-1)^n + \frac{1}{n} : n = 1, 2, 3, \ldots\} = -1$.
- *If* $\{x_n\}_{n=1}^{\infty}$ *is a decreasing sequence with limit* x, *then* $\inf x_n = x$.

Complementary videos:

http://www.youtube.com/watch?v=8Cyvdv7Sm2s
http://www.youtube.com/watch?v=dY8aAP0JgkA
http://www.youtube.com/watch?v=3Z0B95lalpI

Quiz

- Q1 Supremum of a Bounded Set:

  Let $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$. Find the supremum and infimum of A, and determine whether the supremum and infimum are contained within the set.

- Q2 Infimum of a Sequence:

  Consider the sequence $\{a_n\}$ defined by $a_n = \frac{1}{n}$ for $n \in \mathbb{N}_+$. Determine the set $B = \{a_n \mid n \in \mathbb{N}_+\}$, and find the infimum and supremum of B.

Point in $\mathbb{R}^n$

A point in $\mathbb{R}^n$ is a vector $x = (x_1, ..., x_n)$ where each x_i is a real number, for $i = 1, .., n$.

Note: In Euclidean space, a vector is a geometric object that possesses a magnitude and a direction. A vector can be pictured as an arrow. Its magnitude is its length, and its direction is the direction to which the arrow points.

Vector addition and scalar multiplication are defined for $x, y \in \mathbb{R}^n$ and $a \in \mathbb{R}$:

$$
\begin{aligned}
x + y &= (x_1 + y_1, ..., x_n + y_n), \\
ax &= (ax_1, ..., ax_n).
\end{aligned}
$$

Point in $\mathbb{R}^n$

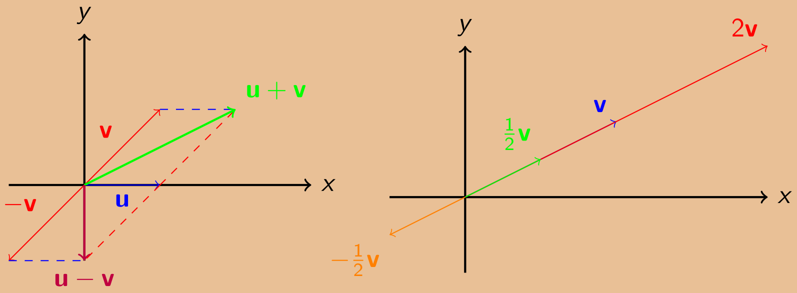


Figure: Vector Addition and Scalar Multiplication in $\mathbb{R}^2$.

Point in $\mathbb{R}^n$

Given any $x, y \in \mathbb{R}^n$ we write

$$\begin{aligned} x = y \quad & \text{if} \quad x_i = y_i \quad \text{for all } i \\ x \geq y \quad & \text{if} \quad x_i \geq y_i \quad \text{for all } i \\ x > y \quad & \text{if} \quad x \geq y \text{ and } x \neq y\,. \end{aligned}$$

Note: x and y may not be comparable under any of the categories above (example: $x = (2, 1)$ and $y = (1, 2)$ are not $=$ nor $x \geq y$ nor $y \geq x$). This is because $\mathbb{R}^n$ for $n > 1$ is not a total order. $\mathbb{R}$ is a total order.

Remark: The space $\mathbb{R}^n$ forms a vector space. In other words, it satisfies commutativity, associativity, and distributive properties, and it has an additive identity, an additive inverse, and a multiplicative identity.

A vector space is usually denoted by V. Functions $f : \mathbb{R}^n \to \mathbb{R}$ also form a vector space.

Linear combination and Span

Linear algebra focuses on finite-dimensional vector spaces like $\mathbb{R}^n$.

Definition (Linear combination)

A linear combination of a set of vectors $\{v_1, \ldots, v_m\}$, *is a vector of the form*

$$\alpha_1 v_1 + \cdots + \alpha_m v_m,$$

where $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$.

Examples: in $\mathbb{R}^3$

- $(17, -4, 2)$ is a linear combination of $(2, 1, -3), (1, -2, 4)$ because

$$(17, -4, 2) = 6(2, 1, -3) + 5(1, -2, 4)\,.$$

Linear combination and Span

- $(17, -4, 5)$ is not a linear combination of $(2, 1, -3), (1, -2, 4)$ because there do not exist scalars α_1, α_2 such that

$$(17, -4, 5) = \alpha_1(2, 1, -3) + \alpha_2(1, -2, 4)\,.$$

  In other words, the system of equations

$$\begin{aligned} 17 &= 2\alpha_1 + \alpha_2 \\ -4 &= \alpha_1 - 2\alpha_2 \\ 5 &= -3\alpha_1 + 4\alpha_2 \end{aligned}$$

  has no solutions (verify!).

Note: To "quickly" check that a set of n vectors in $\mathbb{R}^n$ are $\ell.\imath.$ compute the determinant of the matrix formed by placing the vectors as columns. If determinant is not zero then vectors are $\ell.\imath.$. For example, check that $(1, 4)^\top$, $(7, 5)^\top$ are $\ell.\imath.$.

Linear combination and Span

Definition (Span)

The set of all linear combinations of a set of vectors $\{v_1, \ldots, v_m\}$*, is called the span and denoted* $\operatorname{span}(v_1, \ldots, v_m)$*. In other words,*

$$\operatorname{span}(v_1, \ldots, v_m) = \{a_1 v_1 + \cdots + a_m v_m : \alpha_1, \ldots, \alpha_m \in \mathbb{R}\} .$$

The span of the empty set $\{\}$ *is defined to be* $\{0\}$*.*

Example: The previous example shows that

- $(17, -4, 2) \in \operatorname{span}((2, 1, -3), (1, -2, 4))$;
- $(17, -4, 5) \notin \operatorname{span}((2, 1, -3), (1, -2, 4))$.

Linear independence ($\ell.\imath.$)

A set of vectors $\{v_1, ..., v_k\}$ is said to be linearly independent ($\ell.\imath.$) if the equality

$$\alpha_1 v_1 + \alpha_2 v_2 + ... + \alpha_k v_k = 0$$

implies that $\alpha_i = 0$ for all i. A set of the vectors $\{v_1, ..., v_k\}$ is linearly dependent if it is not $\ell.\imath.$.

Examples:

- the set composed of the single vector 0 is linearly dependent, for if $\alpha \neq 0$ then $\alpha 0 = 0$. In fact, any set of vectors containing the vector 0 is linearly dependent.
- A set composed of a single nonzero vector $a \neq 0$ is $\ell.\imath.$ since $\alpha a = 0$ implies $\alpha = 0$. Note a singleton set of vectors maybe $\ell.\imath.$ or l.d.
- The set $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$ is $\ell.\imath.$ in $\mathbb{R}^4$.

Examples

Examples of linearly dependent sets:

- $\{(2,3,1),(1,-1,2),(7,3,8)\}$ is linearly dependent in $\mathbb{R}^3$ because $2(2,3,1)+3(1,-1,2)+(-1)(7,3,8)=(0,0,0)$.
- $\{(2,3,1),(1,-1,2),(7,3,c)\}$ is linearly dependent in $\mathbb{R}^3$ iff $c=8$ (verify!).

Property

A set of vectors $\{v_1,....,v_k\}$ is l.d. iff one of the vectors from the set is a linear combination of the remaining vectors.

- $\{(2,3,1),(1,-1,2),(7,3,8)\}$ is linearly dependent in $\mathbb{R}^3$ because $2(2,3,1)+3(1,-1,2)=(7,3,8)$.

Property

Let $\{u_1,..,u_m\}$ be $\ell.\imath.$ in a vector space V. Suppose also that $\{w_1,..,w_n\}$ spans V. Then, that $m \le n$.

- $\{(1,2,3),(4,5,8),(9,6,7),(-3,2,8)\}$ cannot be $\ell.\imath.$ in $\mathbb{R}^3$ because $\{(1,0,0),(0,1,0),(0,0,1)\}$ spans $\mathbb{R}^3$.

Bases

We discussed linearly independent sets of vectors and span. Now we bring these concepts together.

Definition (Basis)

A basis of a vector space V is a $\ell.\imath.$ set of vectors in V that spans V.

Here we are interesed when V represents $\mathbb{R}^n$ for $n \in \mathbb{N}$, or any subset that forms a vector space.

Examples:

- The set $\{(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$ is a basis of $\mathbb{R}^n$, and called the canonical (or standard) basis.
- The set $\{(1,2),(3,5)\}$ is a basis of $\mathbb{R}^2$.
- The set $\{(1,2,-4),(7,-5,6)\}$ is $\ell.\imath.$ in $\mathbb{R}^3$ but is not a basis of $\mathbb{R}^3$.
- The set $\{(1,2),(3,5),(4,13)\}$ spans $\mathbb{R}^2$ but is not a basis of $\mathbb{R}^2$ because it is not $\ell.\imath.$.
- The set $\{(1,1,0),(0,0,1)\}$ is a basis of $\left\{(x,x,y)\in\mathbb{R}^3 : x,y\in\mathbb{R}\right\}$.

- The set $\{(1,-1,0),(1,0,-1)\}$ is a basis of $\left\{(x,y,z)\in\mathbb{R}^3 : x+y+z=0\right\}$.
- The set $\{1,x,\ldots,x^m\}$ is a basis of the space of polynomials of degree m.

Property (Criterion for basis)

A set of vectors $\{v_1,\ldots,v_n\}$ in V is a basis of V iff every $v\in V$ can be written <u>uniquely</u> in the form

$$v=\alpha_1 v_1+\cdots+\alpha_n v_n,$$

where $\alpha_1,\ldots,\alpha_n\in\mathbb{R}$.

Definition (Dimension)

The dimension of a finite-dimensional vector space is the length of any basis of the vector space.

Linear transformations

A function $\mathcal{L} : V \to W$ is called a linear transformation if

1. $\mathcal{L}(\alpha x) = \alpha \mathcal{L}(x)$, for $x \in V$, and $\alpha \in \mathbb{R}$.
2. $\mathcal{L}(x + y) = \mathcal{L}(x) + \mathcal{L}(y)$, for $x, y \in V$.

Note: Linear transformations are essential in linear algebra and multivariable calculus. In this course we focus on transformations from $V = \mathbb{R}^n$ to $W = \mathbb{R}^m$, and in particular when $m = 1$.

A linear transformation $\mathcal{L}$ can be represented by a matrix:

Definition (Matrix)

Let m and n denote positive integers. An m-by-n matrix A is a rectangular array of elements in $\mathbb{R}$ with m rows and n columns:

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{n,1} \\ \vdots & & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{pmatrix},$$

first index refers to the row number and the second index refers to the column number.

Linear transformations

Note: Elements of a matrix can be functions $\mathbb{R}^n \to \mathbb{R}$.

Assume the bases for $\mathbb{R}^n$ and $\mathbb{R}^m$ are the canonical ones, then we can write

$$\mathcal{L}(x) = Ax.$$

Recall: The *canonical* or standard basis is the set of vectors

$$e_1 \equiv \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e_2 \equiv \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad, e_n \equiv \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Important notation: From now on, an element of $\mathbb{R}^n$ (or a vector) is represented by a *column* array. That is: $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$, so it can be seen as a matrix of size $n \times 1$.

Note: $Ax + b$ is not a linear transformation, but an *affine* one.

Matrix operations

Marix sum: The sum of two matrices of the same size is the matrix obtained by adding corresponding entries in the matrices.

Matrix scalar product: The product of a scalar and a matrix is the matrix obtained by multiplying each entry in the matrix by the scalar.

The most interesting operation is the **matrix multiplication:**

Definition

Suppose A is an m-by-n matrix and C is an n-by-p matrix. Then AC is defined to be the m-by-p matrix whose entry in row j, column k, is given by the following equation:

$$(AC)_{j,k} = \sum_{r=1}^{n} a_{j,r} c_{r,k} \, .$$

In other words, the entry in row j, column k, of AC is computed by taking row j of A and column k of C, multiplying together corresponding entries, and then summing.

Note: that we define the product of two matrices only when the number of columns of the first matrix equals the number of rows of the second matrix.

Why the product of matrices is defined in this way?

Answer: Consider the linear transformations $\mathcal{L} : U \to V$ and $\mathcal{S} : V \to W$ with associated matrices C and A respectively.

Now observe the composition $\mathcal{S} \circ \mathcal{L}$ is a linear map from U to W. Suppose the associated matrix of this linear transformation is B.

Now the question is

$$CA \stackrel{?}{=} B\,.$$

Yes, since matrix multiplication definition makes this happen (see next slide).

Matrix operations

For $1 \leq k \leq p$, we have [2]

$$\begin{aligned}(\mathcal{S}\mathcal{L})u_k &= \mathcal{S}\left(\sum_{r=1}^{n} c_{r,k} v_r\right) \\ &= \sum_{r=1}^{n} c_{r,k} \mathcal{S} v_r \\ &= \sum_{r=1}^{n} c_{r,k} \sum_{j=1}^{m} a_{j,r} w_j \\ &= \sum_{j=1}^{m} \left(\sum_{r=1}^{n} a_{j,r} c_{r,k}\right) w_j \,.\end{aligned}$$

Thus

$$B = \sum_{r=1}^{n} a_{j,r} c_{r,k} \,. \qquad \text{(check the derivation)}$$

[2] u_k, v_k, and w_k are the elements of the basis of U, V and W.

Definition (Linear combination of columns)

Suppose A is an m-by- n matrix and $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ is an n-by-1 matrix (or vector). Then

$$Ax = x_1 a_{\cdot,1} + \cdots + x_n a_{\cdot,n}\,,$$

where $a_{\cdot,k}$ denotes the h-th column of A.

In other words, Ax is a linear combination of the columns of A, with the scalars that multiply the columns coming from x.

Linear system

Crucial feature of matrices: The linear system of equations

$$\begin{pmatrix} a_{1,1}x_1 & + & a_{1,2}x_2 & + & \dots & + & a_{1,n}x_n & = & b_1 \\ \vdots & & \vdots & & \dots & & \vdots & = & \vdots \\ a_{m,1}x_1 & + & a_{m,2}x_2 & + & \dots & + & a_{m,n}x_n & = & b_m \end{pmatrix}$$

can be compactly written as

$$Ax = b\,,$$

for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

Inverses

Definition (Invertible linear transformation)

- *A linear transformation* $\mathcal{L} : V \to W$ *is called invertible if there exists a linear transformation* $\mathcal{S} : W \to V$ *such that* $\mathcal{S}\mathcal{L}$ *is the identity map on* V *and* $\mathcal{L}\mathcal{S}$ *equals the identity map on* W*.*
- *A linear transformation* $\mathcal{S} : W \to V$ *satisfying* $\mathcal{S}\mathcal{L} = I$ *and* $\mathcal{L}\mathcal{S} = I$ *is called an inverse of* $\mathcal{L}$*.*

Property (Inverse is unique)

An invertible linear transformation has a unique inverse.

Notation: $\mathcal{L}^{-1}$.

The following result characterizes the invertible linear maps.

Property (Invertibility is equivalent to injectivity and surjectivity)

A linear transformation is invertible iff it is injective and surjective.

Definition (Operator: $\mathbb{R}^n \to \mathbb{R}^n$ transformation)

An important particular case of linear transformations are the ones from $\mathbb{R}^n \to \mathbb{R}^n$. Those are called operators and represented by square matrices.

Definition (Invertible qmatrix)

An n-by-n square matrix A is called invertible (also nonsingular) if there exists an n-by-n square matrix B such that

$$AB = BA = I_n$$

where I_n denotes the n-by-n identity matrix.

Inverses

In this case, the matrix B is uniquely determined by A.

Task: Discuss solving the n-by-n system.

Invertible matrix

Let $A \in \mathbb{R}^{n \times n}$. The following statements are equivalent:

- A is invertible, i.e. there exists a B such that $AB = I_n = BA$.
- The linear transformation that maps x to Ax is invertible (and bijective).
- A is row-equivalent and column-equivalent to the n-by-n identity matrix I_n.
- A has full rank: $\operatorname{rank}(A) = n$. (more on this later)
- A has trivial kernel: $\ker(A) = 0$. (more on this later)
- The columns of A are $\ell.\imath.$.
- The rows of A are $\ell.\imath.$.
- The determinant of A is nonzero.
- The number 0 is not an eigenvalue of A. (more on this later)
- The transpose $A^\top$ is an invertible matrix.

Similar matrices

Definition (Transformation matrix)

Let $\{b_1, b_2, ..., b_n\}$ and $\{\tilde{b}_1, \tilde{b}_2, ..., \tilde{b}_n\}$ be two bases for $\mathbb{R}^n$. Then the matrix

$$T = [\tilde{b}_1, \tilde{b}_2, ..., \tilde{b}_n]^{-1}[b_1, b_2, ..., b_n]$$

is called a transformation matrix.

Given a vector in $\mathbb{R}^n$, let x be its representation in the basis $\{b_1, b_2, ..., b_n\}$ and $\tilde{x}$ its representation in the basis $\{\tilde{b}_1, \tilde{b}_2, ..., \tilde{b}_n\}$, then $\tilde{x} = Tx$.

Note: This result allows a linear transformation to be represented by different matrices. This is crucial in applications since matrix A is usually given by the problem and depending on the applications alternative representations can be found.

Similar matrices

Consider a linear transformation $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^n$ and let A be its matrix representation in terms of the basis $\{b_1, b_2, ..., b_n\}$, and B is the matrix representation with respect to $\{\tilde{b}_1, \tilde{b}_2, ..., \tilde{b}_n\}$. Let $y = Ax$ and $\tilde{y} = B\tilde{x}$. We have

$$\tilde{y} = Ty = TAx = B\tilde{x} = BTx$$

and therefore

$$A = T^{-1}BT\,.$$

Then we say A is *similar* to B.

When A is similar to B, both matrices correspond to the same linear transformation but expressed in different bases.

A very important case is when a linear transformation (from $\mathbb{R}^n$ to $\mathbb{R}^n$) can be represented by a diagonal matrix. This is called the *eigen representation*.

Eigenvalues (evas) and Eigenvectors (eves)

Let A be an $n \times n$ **square** matrix. A scalar λ (possibly complex) and a *nonzero* vector v satisfying the equation

$$Av = \lambda v$$

are said to be, respectively, an eigenvalue and an eigenvector of A.

For λ to be an eigenvalue it is necessary and sufficient for the matrix $\lambda I - A$ to be *singular*, that is, $det(\lambda I - A) = 0$. This leads to an nth-order polynomial equation

$$det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + ... + a_1\lambda + a_0 = 0.$$

According to the fundamental theorem of algebra, the characteristic equation must have n (possibly nondistinct) roots that are the eigenvalues of A.

Eigenvalues and Eigenvectors

Theorem (A is diagonalizable)

Suppose the characteristic equation $\det(\lambda I - A) = 0$ *has* n *distinct roots* $\lambda_1, \lambda_2, ..., \lambda_n$*. Then, there exist* n *linearly independent (*$\ell.\imath.$*) vectors* $v_1, v_2, ..., v_n$ *such that*

$$Av_i = \lambda_i v_i, \quad i{=}1, 2, ..., n.$$

Consider a basis formed by these $\ell.\imath.$ vectors $v_1, v_2, ..., v_n$. Then, using this basis we can represent A using a diagonal matrix. Let

$$T = [v_1, v_2, ..., v_n].$$

Then

$$\begin{aligned} T^{-1}AT &= T^{-1}A[v_1, v_2, ..., v_n] \\ &= T^{-1}[Av_1, Av_2, ..., Av_n] \\ &= T^{-1}[\lambda_1 v_1, \lambda_2 v_2, ..., \lambda_n v_n] \end{aligned}$$

Eigenvalues and Eigenvectors

Finally,

$$T^{-1}AT = T^{-1}T\begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}.$$

Real symmetric matrices play a special role in this course (in particular, because they are diagonalizable).

Theorem (good news)

All eigenvalues of a symmetric matrix are real.

Diagonalizable

An operator $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^n$ is called diagonalizable if the operator has a diagonal matrix with respect to some basis.

Example: Let $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}^2$ by $\mathcal{L}(x, y) = (41x + 7y, -20x + 74y)$.
The associated matrix with respect to the standard basis of $\mathbb{R}^2$ is

$$\begin{pmatrix} 41 & 7 \\ -20 & 74 \end{pmatrix}$$

which is not a diagonal matrix. However, $\mathcal{L}$ is diagonalizable, because the associated matrix with respect to the basis $(1, 4)^\top$, $(7, 5)^\top$ is

$$\begin{pmatrix} 69 & 0 \\ 0 & 46 \end{pmatrix}.$$

(verify).

Eigenvalues and Eigenvectors

Theorem

Any real symmetric $n \times n$ matrix has a set of n eigenvectors that are mutually orthogonal.

If A is symmetric, then a set of its eigenvectors forms an orthogonal basis for $\mathbb{R}^n$. If the basis $\{v_1, v_2, ..., v_n\}$ is normalized so that each element has norm 1, then for

$$T = [v_1, v_2, ..., v_n],$$

we have

$$T^\top T = I,$$

and hence

$$T^\top = T^{-1}.$$

A matrix whose transpose is equal to its inverse is said to be an orthogonal matrix (extremely convenient: compare the complexity of inverting a matrix versus transpose).

Eigenvalues and Eigenvectors

MATLAB/Octave experiment.

```
n = 8;
Q = randn(n);
% make Q symmetric
Q = (Q+Q')/2;
% since Q is randomly generated, all columns are l.i.
rank(Q)

[X, D] = eig(Q);
% check eigendecomposition (all entries are numerical zeros)
Q - X*D*inv(X);
% since Q is real symmetric, evectors form an orthogonal matrix
id = X'*X;

% clean up numerical zeros
id(abs(id)<1E-14)=0
```

Rank of a matrix

Let $A \in \mathbb{R}^{m \times n}$ be the matrix associated with the linear transformation $\mathcal{L}(x) = Ax$.

Column-rank of A is the maximal number of $\ell.\imath.$ columns of A.

Row-rank of A is the maximal number of $\ell.\imath.$ rows of A.

The rank of A is

$$\operatorname{rank}(A) \leq \min(m, n) \,.$$

A has full row-rank (column-rank), if $\operatorname{rank}(A) = m$ ($\operatorname{rank}(A) = n$).

A matrix that has rank $\min(m, n)$ is said to have full rank; otherwise, the matrix is rank deficient.

Rank of a matrix

Only the zero matrix has rank zero.

$\mathcal{L}$ is injective (or "one-to-one") iff A has rank n (in this case, we say that A has full column rank).

$\mathcal{L}$ is injective (or "one-to-one") iff ($Ax = 0$ implies $x = 0$).

Nullspace of A is

$$\{x \in \mathbb{R}^n : Ax = 0\}.$$

$\mathcal{L}$ is injective (or "one-to-one") iff Nullspace of A contains only the 0 vector.

$\mathcal{L}$ is surjective (or "onto") iff A has rank m (in this case, we say that A has full row rank).

If A is a square matrix ($m = n$), then A is invertible iff A has rank n (that is, A has full rank).

Euclidean inner product and norm

We now focus on 3 fundamental concepts on $\mathbb{R}^n$:

- the Euclidean inner product of two vectors x and y in $\mathbb{R}^n$,
- the Euclidean norm of a vector x in $\mathbb{R}^n$,
- the Euclidean metric measuring the distance between two points x and y in $\mathbb{R}^n$.

Each generalizes a familiar concept from $\mathbb{R}$:

- the Euclidean inner product of x and y is just the product xy of the numbers x and y;
- the Euclidean norm of x is simply the absolute value $|x|$ of x;
- the Euclidean distance between x and y is the absolute value $|x - y|$ of their difference.

Euclidean inner product and norm

Given $x, y \in \mathbb{R}^n$, the Euclidean inner product (or dot product) of x and y is defined as:

$$x \cdot y = \sum_{i=1}^{n} x_i y_i \,.$$

The dot product can also be written as a matrix product [3]

$$x \cdot y \equiv x^\top y$$

where $x^\top$ denotes the transpose of x. Recall the transpose of a matrix is the matrix obtained by interchanging the rows and columns.

[3] Recalling that a vector is a column matrix

Euclidean inner product and norm

Theorem

The dot product satisfies the following properties for all $x, y, z \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$

1. *Nonnegativity:* $x \cdot x \geq 0$, *with equality iff* $x = 0$.
2. *Symmetry:* $x \cdot y = y \cdot x$.
3. *Bilinearity:* $(ax + by) \cdot z = ax \cdot z + by \cdot z$ *and* $x \cdot (ay + bz) = x \cdot ay + x \cdot bz$.

An inner product is a generalization of the dot product and satisfies similar properties.

Euclidean inner product and norm

Property (Cauchy-Schwartz Inequality)

For any $x, y \in \mathbb{R}^n$ we have

$$|x \cdot y| \leq (x \cdot x)^{1/2} (y \cdot y)^{1/2} .$$

The Euclidean norm (or magnitude) of a vector $x \in \mathbb{R}^n$, denoted $\|x\|$, is defined as

$$\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} .$$

The norm is related to the inner product through the identity

$$\|x\| = (x \cdot x)^{1/2} ,$$

and the C-S inequality may be written as

$$|x \cdot y| \leq \|x\| \|y\| .$$

Definition

Two vectors x, y are called orthogonal if $x \cdot y = 0$

Euclidean inner product and norm

Theorem

The Euclidean norm satisfies the following properties for all $x, y \in \mathbb{R}^n$ and $a \in \mathbb{R}$

1. *Nonnegativity:* $\|x\| \geq 0$, *with equality iff* $x = 0$.
2. *Homogeneity:* $\|ax\| = |a| \, \|x\|$.
3. *Triangle Inequality:* $\|x + y\| \leq \|x\| + \|y\|$.

The dot product of two Euclidean vectors can be also expressed as

$$x \cdot y = \|x\| \|y\| \cos \theta \,,$$

where θ is the angle between both vectors.

Euclidean inner product and norm

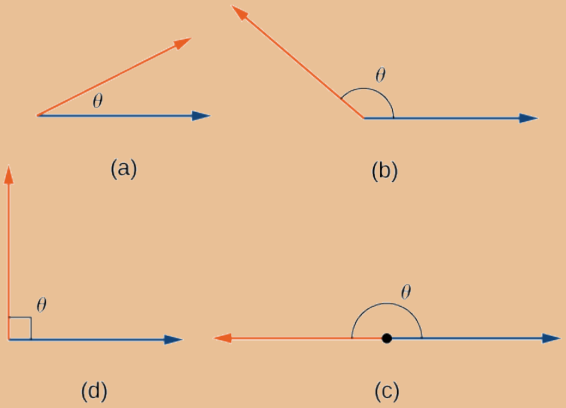


Figure: (a) An acute angle has $0 < \cos\theta < 1$. (b) An obtuse angle has $-1 < \cos\theta < 0$. (c) A straight line has $\cos\theta = -1$.

Euclidean distance

The Euclidean distance $d(x,y)$ between two vectors x and y in $\mathbb{R}^n$ is given by

$$d(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2} .$$

The distance function d is called a metric, and is related to the norm $\|\cdot\|$ through the identity

$$d(x,y) = \|x - y\| .$$

for all $x, y \in \mathbb{R}^n$.

Theorem

The metric d satisfies the following for all $x, y, z \in \mathbb{R}^n$ and $a \in \mathbb{R}$

1. *Nonnegativity: $d(x,y) \geq 0$ with equality iff $x = y$.*
2. *Symmetry: $d(x,y) = d(y,x)$.*
3. *Triangle Inequality: $d(x,z) \leq d(x,y) + d(y,z)$.*

More norms

In general, the p-norm of a vector $x \in \mathbb{R}^n$

$$\|x\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p}\,, \qquad \text{for } 1 \le p < \infty$$

and

$$\max\{|x_1|, \dots, |x_n|\}\,, \qquad \text{if } p = \infty\,.$$

In particular, the Euclidean norm will be denoted

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}\,, \qquad \text{for } x \in \mathbb{R}^n\,.$$

The norm of a matrix may be chosen in a variety of ways. Because the set of matrices $\mathbb{R}^{m \times n}$ can be viewed as the real vector space $\mathbb{R}^{m\,n}$, matrix norms should be no different from regular vector norms.

Matrix Norms

We define the norm of a matrix A, denoted $\|A\|$, to be any function $\|\cdot\|$ that satisfies the conditions:

1. Nonnegativity: $\|A\| > 0$ if $A \neq 0$, and $\|O\| = 0$, where O is the matrix with all entries equal to zero;
2. Homogeneity: $\|cA\| = |c|\|A\|$, for any $c \in \mathbb{R}$;
3. Triangle inequality: $\|A + B\| \leq \|A\| + \|B\|$.

An example of a matrix norm is the Frobenius norm, defined as

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2},$$

where $A \in \mathbb{R}^{m \times n}$.

The Frobenius norm is equivalent to the Euclidean norm on $\mathbb{R}^{m \times n}$.

Note that the Frobenius norm satisfies $\|AB\| \leq \|A\|\|B\|$.

Matrix Norms

It is convenient to construct the norm of a matrix in such a way that it is related with vector norms.

Induced norms: Let $\|\cdot\|_{(n)}$ and $\|\cdot\|_{(m)}$ be vector norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.

We define an induced matrix norm as:

$$\|A\|_{(n),(m)} = \max_{\|x\|_{(n)}=1} \|Ax\|_{(m)} \, .$$

Note that the induced norm satisfies $\|Ax\|_{(m)} \leq \|A\|_{(n),(m)} \|x\|_{(n)}$.

If $\|x\|_{(n)} = \|x\|_{(m)} = \|x\|_2$ then the induced norm of a matrix $A \in \mathbb{R}^{m \times n}$ is called the spectral norm $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$.

Basic topological concepts

We begin with the definition of a ball.

Definition (open ball, closed ball)

The open ball with center $c \in \mathbb{R}^n$ *and radius* r*:*

$$B(c, r) = \{x \in \mathbb{R}^n : \|x - c\| < r\} \ .$$

The closed ball with center c *and radius* r*:*

$$B[c, r] = \{x \in \mathbb{R}^n : \|x - c\| \leq r\} \ .$$

Note that the norm used in the definition of the ball may be any norm. If the norm is not specified, we assume is the Euclidean norm. The ball $B(c, r)$ for some arbitrary $r > 0$ is also referred to as a neighborhood of c.

Basic topological concepts

Interior point of a set: A point which has a neighborhood contained in the set.

Definition (interior points)

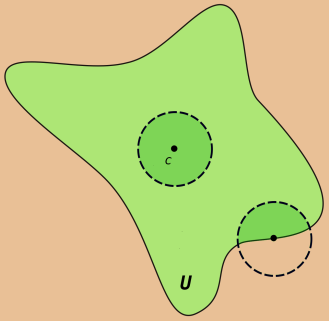
Given a set $U \subseteq \mathbb{R}^n$, a point $c \in U$ is an interior point of U if there exists $r > 0$ for which $B(c, r) \subseteq U$.

The set of all interior points of U:

$$\operatorname{int}(U) = \{x \in U : B(x, r) \subseteq U \text{ for some } r > 0\}.$$

Examples:

$$\begin{aligned} \operatorname{int}\left(\mathbb{R}^n_{\geq 0}\right) &= \mathbb{R}^n_{>0}, \\ \operatorname{int}(B[c, r]) &= B(c, r), \quad c \in \mathbb{R}^n,\ r > 0, \\ \operatorname{int}([x, y]) &= (x, y), \quad x, y \in \mathbb{R}^n,\ x \neq y. \end{aligned}$$



Basic topological concepts

Definition (open sets)

A set that contains only interior points: $U \subseteq \mathbb{R}^n$ is an open set if

$$\forall x \in U : \exists r > 0 \text{ such that } B(x, r) \subseteq U.$$

Examples:

- $(0, 1)$, $\{x \in \mathbb{R} : x > 0\}$.
- $\mathbb{R}^n$, $\emptyset$.
- open balls.
- the positive orthant $\mathbb{R}^n_{>0}$.

Property

A union of any number of open sets is an open set and the intersection of a finite number of open sets is open.

Definition (closed sets)

A set $U \subseteq \mathbb{R}^n$ is said to be closed if it contains all the limits of convergent sequences of points in U; that is, U is closed if for every sequence of points $\{x_i\}_{i \geq 1} \subseteq U$ satisfying $x_i \to x^*$ as $i \to \infty$, it holds that $x^* \in U$.

Property

A set U is closed iff its complement U^c is open.

Examples:

- $[0, 1]$, $\{x \in \mathbb{R} : x \geq 0\}$.
- $\mathbb{R}^n$, $\emptyset$.
- closed ball $B[c, r]$.

- the positive orthant $\mathbb{R}^n_{\geq 0}$.

What about $[0, 1)$?

Note: A set is not like a door because a set can be open, closed, both or neither.
Note 2: An important and useful result states that level sets, as well as contour sets, of continuous functions are closed (more on this later).

Basic topological concepts

Definition (boundary points)

Given a set $U \subseteq \mathbb{R}^n$, a boundary point of U is a point $x \in \mathbb{R}^n$ satisfying the following: any neighborbood of x contains at least one point in U and at least one point in its complement U^c.

The set of all boundary points of a set U is denoted by $\mathrm{bd}(U)$ (some authors ∂U) and is called the boundary of U.

Examples:

$$\begin{aligned}
\mathrm{bd}([0,1]) &= ? \\
\mathrm{bd}((0,1)) &= ? \\
\mathrm{bd}(B(c,r)) &= \mathrm{bd}(B[c,r]) = \{x \in \mathbb{R}^n : \|x - c\| = r\}\ ,\ c \in \mathbb{R}^n,\ r > 0\,, \\
\mathrm{bd}\left(\mathbb{R}^n_{>0}\right) &= \mathrm{bd}\left(\mathbb{R}^n_{\geq 0}\right) = \left\{x \in \mathbb{R}^n_{\geq 0} : \exists i : x_i = 0\right\}, \\
\mathrm{bd}\left(\mathbb{R}^n\right) &= \emptyset.
\end{aligned}$$

Basic topological concepts

The closure of a set $U \subseteq \mathbb{R}^n$ is denoted by $\mathrm{cl}(U)$ (some authors $\bar{U}$) is defined as

$$\mathrm{cl}(U) = U \cup \mathrm{bd}(U)\,.$$

Examples:

$$\begin{aligned}
&\mathrm{cl}\left(\mathbb{R}^n_{>0}\right) = \mathbb{R}^n_{\geq 0},\\
&\mathrm{cl}(B(c,r)) = B[c,r]\,, \quad c \in \mathbb{R}^n\,,\ r \in \mathbb{R}_{\geq 0}\,,\\
&\mathrm{cl}((x,y)) = [x,y]\,, \quad x, y \in \mathbb{R}^n,\ x \neq y\,.
\end{aligned}$$

Definition (boundedness and compactness)

1. *A set $U \subseteq \mathbb{R}^n$ is called bounded if there exists a real number $M > 0$ for which $U \subseteq B(0, M)$.*
2. *A set $U \subseteq \mathbb{R}^n$ is called compact if it is closed and bounded.*

Examples of compact sets:

- $[0, 1]$.
- Closed balls.
- $\emptyset$.

The positive orthant is not compact since it is unbounded, and open balls are not compact since they are not closed.

Further examples to think:

- $\bigcup_{n=1}^{\infty} \left(0, \frac{1}{n}\right) = (0, 1)$.
- $\bigcap_{n=1}^{\infty} \left(0, \frac{1}{n}\right) = \emptyset$.
- $\bigcup_{n=1}^{\infty} \left[0, \frac{n}{n+1}\right] = [0, 1)$.
- $\bigcap_{n=1}^{\infty} \left[1, 3 + \frac{1}{n}\right] = [1, 3]$.

Problem

For each of the following sets:

$$\mathcal{A} = \bigcup_{n>1} \left[\frac{1}{n}, \frac{n}{n+1}\right]$$

$$\mathcal{B} = \{[x_1, x_2, x_3]^\top \in \mathbb{R}^3 : \ \max\{\ |x_1|, |x_2|, |x_3|\} < 1\}$$

$$\mathcal{C} = \{[x_1, x_2]^\top \in \mathbb{R}^2 : \ 0 \leq x_1 < x_2\}$$

$$\mathcal{D} = \{[x_1, x_2, x_3]^\top \in \mathbb{R}^3 : \ |x_1| + |x_2| + |x_3| = 1, \ x_1, x_2, x_3 \geq 0\}$$

answer the questions

(a) Is the set closed, open, or neither?

(b) Describe the interior and the closure of the set.

(c) Is the set bounded? and if so, provide a bound.

Quadratic Forms

We now study the second most important function in this course.
A quadratic form $f : \mathbb{R}^n \to \mathbb{R}$ is a function

$$f(x) = x^\top Qx,$$

where Q is an $n \times n$ real matrix. There is no loss of generality in assuming Q to be symmetric, that is, $Q = Q^\top$. For if the matrix Q is not symmetric, we can always replace it with the symmetric matrix

$$Q_0 = Q_0^\top = \frac{1}{2}(Q + Q^\top).$$

Note that

$$x^\top Qx = x^\top Q_0 x = x^\top \left(\frac{1}{2}Q + \frac{1}{2}Q^\top\right) x.$$

A quadratic form $x^\top Qx$ for Q *symmetric* is said to be positive definite if $x^\top Qx > 0$ for all nonzero vectors x. It is positive semidefinite if $x^\top Qx \geq 0$ for all x.

We can define negative definite and semidefinite in a similar way.

Quadratic Forms

Recall that the minors of a matrix Q are the determinants of the matrices obtained by successively removing rows and columns from Q.

For $Q = \begin{pmatrix} q_{11} & q_{12} & \dots & q_{1n} \\ q_{21} & q_{22} & \dots & q_{2n} \\ \vdots & \vdots & & \vdots \\ q_{n1} & q_{n2} & \dots & q_{nn} \end{pmatrix}$ the leading principal minors are

$$\Delta_1 = q_{11}\,, \Delta_2 = \det \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}, \Delta_3 = \det \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix}, .., \Delta_n = \det Q$$

Theorem (Sylvester's Criterion)

A quadratic form $x^\top Qx$ for Q symmetric is positive definite iff the leading principal minors of Q are positive.

Note that if Q is not symmetric, Sylvester's criterion cannot be used to check positive definiteness of the quadratic form $x^\top Qx$.

Consider $Q = \begin{pmatrix} 1 & 0 \\ -4 & 1 \end{pmatrix}$. Minors are positive, however if $x = (1,1)^\top$, $x^\top Qx = -2$. However, by considering $Q_0 = \frac{1}{2}(Q + Q^\top)$, we can use the criterion.

Note: A necessary condition for a real quadratic form to be positive *semi*definite is that the leading principal minors be nonnegative. However, this is not a sufficient condition.

Quadratic Forms

A symmetric matrix Q is said to be positive definite (pd) if the quadratic form $x^\top Qx$ is positive definite.
Similarly, we define a symmetric matrix Q to be positive semidefinite (psd).

Note that the matrix Q is pd (or psd) iff the matrix $-Q$ is negative definite (or negative semidefinite).

The symmetric matrix Q is indefinite if it is neither psd nor negative semidefinite.

Theorem

A symmetric matrix Q is pd (or psd) iff all eigenvalues of Q are positive (or nonnegative).

Quadratic Forms

Through diagonalization, we can show that a symmetric psd matrix Q has a psd (symmetric) square root $Q^{1/2}$ satisfying $Q^{1/2}Q^{1/2} = Q$. For this, we use T as above and define

$$Q^{1/2} = T \begin{pmatrix} \lambda_1^{1/2} & & & 0 \\ & \lambda_2^{1/2} & & \\ & & \ddots & \\ 0 & & & \lambda_n^{1/2} \end{pmatrix} T^\top .$$

Note that the quadratic form $x^\top Qx$ can be expressed as $\|Q^{1/2}x\|^2$.

Some properties

- Let Q be pd. Then the diagonal elements are positive.
- Let Q be psd. Then the diagonal elements are nonnegative.
- Let Q be a symmetric matrix. If there exist positive and negative elements in the diagonal, then Q is indefinite.
- Let $A \in \mathbb{R}^{m \times n}$. Then $A^\top A$ is pd iff A has rank n. (If rank is less than n then $A^\top A$ is psd and has the eigenvalue 0 and therefore not invertible). In both cases $A^T A$ is symmetric.
- Let Q be pd. Then Q has inverse and is pd.

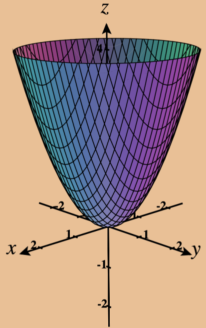
Quadratic form graphs

In the following slides we show the shape of “canonical” $\mathbb{R}^2$ quadratic form graphs. The graphs of other quadratic forms look similar, though they may be stretched in various directions. Later we will study quadratic functions.

Task: Visualize the following quadratic forms in your favourite plotting software. Also try examples where Q is non-diagonal.

Quadratic form graphs - pd - "basin" / "bowl"

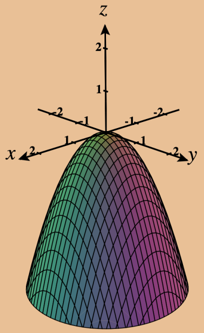
The quadratic form $z \equiv f(x, y) = x^2 + y^2$ is pd. Since $Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, its eigenvalue is 1 (duplicated). Recall the e-vas of a diagonal matrix is the diagonal.



Note that for a pd quadratic form, there is a strict (unique) minimum at the origin.

Quadratic form graphs - nd

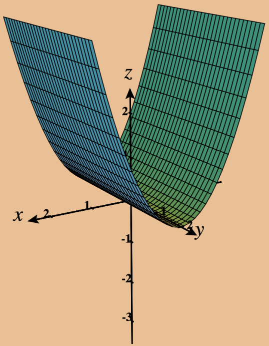
The quadratic form $f(x,y) = -x^2 - y^2$ is nd. Since $Q = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$, its eigenvalue is -1 (duplicated).



Note that for a nd quadratic form, there is a strict (unique) maximum at the origin.

Quadratic form graphs - psd - "creek"

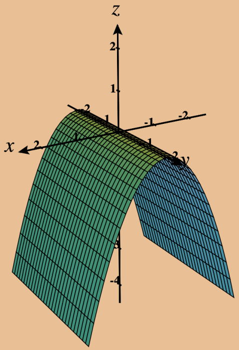
The quadratic form $f(x,y) = x^2$ is psd. Since $Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, its eigenvalues are 0 and 1.



Note that for a psd quadratic form, there is a non-strict (non-unique) minimum at the origin.

Quadratic form graphs - nsd

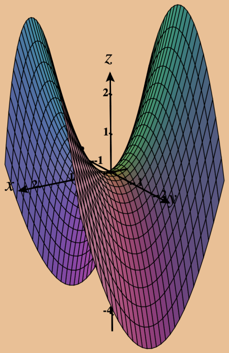
The quadratic form $f(x, y) = -x^2$ is nsd. Since $Q = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$, its eigenvalues are 0 and -1.



Note that for a nsd quadratic form, there is a non-strict (non-unique) maximum at the origin.

Quadratic form graphs - indefinite - "pringle chip"

The quadratic form $p(x, y) = x^2 - y^2$ is indefinite. Since $Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, its eigenvalues are 1 and -1.



Note in this case there is no maximum nor minimum at the origin. This type of point is called saddle.

Sequences and limits

Sequences are important in this course since in many cases, to solve a problem we require a sequence of approximations to be constructed.

A sequence in $\mathbb{R}$ is a function from $\mathbb{N} \to \mathbb{R}$, that is

$$x_1, x_2, \ldots$$

or simply $\{x_k\}$ [4]..

A sequence $\{x_k\}$ is *decreasing* if $x_1 > x_2 > \ldots > x_k \ldots$, that is, $x_k > x_{k+1} \, \forall k$.

A number $x^* \in \mathbb{R}$ is called the *limit* of the sequence $\{x_k\}$ if for any positive ϵ there is a number K (which may depend on ϵ) such that

$$\forall k > K \,, \quad |x_k - x^*| < \epsilon \,,$$

that is, x_k lies between $x^* - \epsilon$ and $x^* + \epsilon$ for all $k > K$. In this case, we write

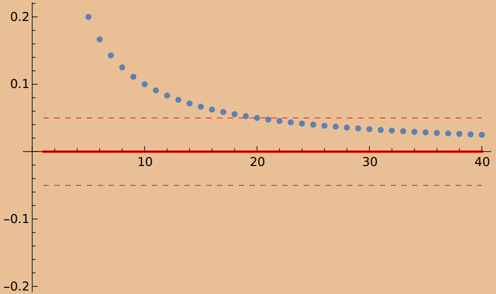
$$x^* = \lim_{k \to \infty} x_k \,, \quad \text{or} \quad x_k \to x^* \,.$$

[4] Occasionally we will use superscripts instead of subscripts, and denote the sequence by $\left\{x^k\right\}$

Sequences and limits

A sequence that has a limit is called a convergent sequence.

Example: The sequence $\{x_k\}$ in $\mathbb{R}$ defined by $x_k = 1/k$ for all k is a convergent sequence, with limit $x = 0$. Let any $\epsilon > 0$ be given. Let $k(\epsilon)$ be any integer such that $k(\epsilon) > 1/\epsilon$. Then, for all $k > k(\epsilon)$, we have $d(x_k, 0) = d(1/k, 0) = 1/k < 1/k(\epsilon) < \epsilon$, so indeed, $x_k \to 0$.



Note: The notion of a sequence in $\mathbb{R}$ can be extended to sequences with elements in $\mathbb{R}^n$ replacing absolute values with norms.

Sequences and limits

Theorem

A sequence can have at most one limit. That is, if $\{x_k\}$ is a sequence in $\mathbb{R}^n$ converging to a point $x \in \mathbb{R}^n$, it cannot also converge to a point $y \in \mathbb{R}^n$ for $y \neq x$.

A sequence $\{x_k\} \in \mathbb{R}^n$ is *bounded* if there exists a real number $B \geq 0$ such that $\|x_k\| \leq B$, for all k.

Theorem

Every convergent sequence in $\mathbb{R}^n$ is bounded.

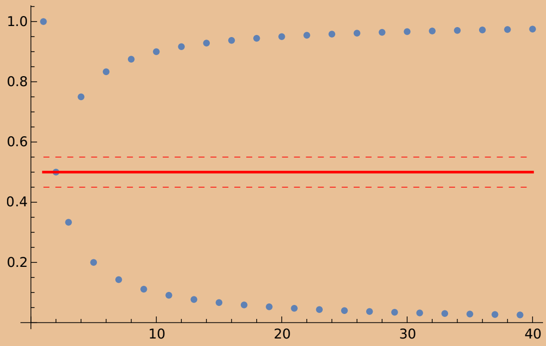
Note 1: This theorem implies that a sequence may have no limit. Indeed if the sequence is unbounded, then the sequence cannot converge (contrapositive).

For example, the sequence $x_k = k$ for all k is a non-convergent sequence.

Sequences and limits

Note 2: Unboundedness is not the only reason a sequence may fail to converge. Consider

$$x_k = \begin{cases} \frac{1}{k}, & k = 1, 3, 5, \ldots \\ 1 - \frac{1}{k}, & k = 2, 4, 6, \ldots \end{cases}$$



This sequence is bounded since we have $|x_k| \leq 1$ for all k. However, it does no possess a limit. The reason here is that the odd terms of the sequence are converging to the point 0, while the even terms are converging to the point 1. Since a sequence can have only one limit, this sequence does not converge.

Sequences and limits

Subsequence: Suppose we are given a sequence $\{x_k\}$ and an increasing sequence of natural numbers $\{m_k\}$. The sequence

$$\{x_{m_k}\} = \{x_{m_1}, x_{m_2}, ...\},$$

is called a subsequence of the sequence $\{x_k\}$. A subsequence of a given sequence can thus be obtained by discarding a finite number of elements of the given sequence.

Why subsequences? Even if a sequence $\{x_k\}$ is not convergent, it may contain subsequences that converge. For instance, the sequence $0, 1, 0, 1, 0, 1, \ldots$ has no limit, but the subsequences $0, 0, 0, \ldots$ and $1, 1, 1, \ldots$ which are obtained from the original sequence by selecting the odd and even elements, respectively, are both convergent.

If a sequence contains a convergent subsequence, the limit of the convergent subsequence is called a limit point of the original sequence. Thus, the sequence $0, 1, 0, 1, 0, 1, \ldots$ has two limit points 0 and 1.

Theorem

Consider a convergent sequence $\{x_k\}$ with limit x^*. Then, any subsequence of $\{x_k\}$ also converges to x^*.

Theorem (Bolzano-Weierstrass)

Any bounded sequence in $\mathbb{R}^n$ contains a convergent subsequence.

(fundamental property of real numbers). See previous example.

Limit of a function: Consider a function $f : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$ and a point $x_0 \in D$. Suppose that there exists f^* such that for any convergent sequence $\{x_k\} \subseteq D$ with limit x_0, we have

$$\lim_{k\to\infty} f(x_k) = f^*.$$

Then, we use the notation

$$\lim_{x\to x_0} f(x)$$

to represent the limit f^*.

Continuity of f in terms of limits of sequences: $f : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$ is continuous at $x_0 \in D$ iff, for any convergent sequence $\{x_k\}$ with limit x_0, we have

$$\lim_{k\to\infty} f(x_k) = f\left(\lim_{k\to\infty} x_k\right) = f(x_0).$$

Therefore, using the notation introduced above, the function f is continuous at x_0 iff

$$\lim_{x\to x_0} f(x) = f(x_0).$$

Examples:

1. For $f : \mathbb{R}^2 \backslash \{(0,0)\} \to \mathbb{R}$, compute $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$. (try the sequence $\{(\frac{c}{k}, \frac{c}{k})\}$ for c constant).
2. For $f : \mathbb{R}^2 \backslash \{(0,0)\} \to \mathbb{R}$, compute $\lim_{(x,y)\to(0,0)} \frac{xy}{\sqrt{x^2+y^2}}$.

Continuity

Epsilon-delta definition: An equivalent way of saying that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at the point $x_0 \in D \subseteq \mathbb{R}^n$ is: For every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D$:

$$\|x - x_0\| < \delta \quad \text{implies} \quad \|f(x) - f(x_0)\| < \epsilon\,.$$

Recall δ depends on ϵ.

We say f is continuous in a set D if it is continuous for all $x_0 \in D$.

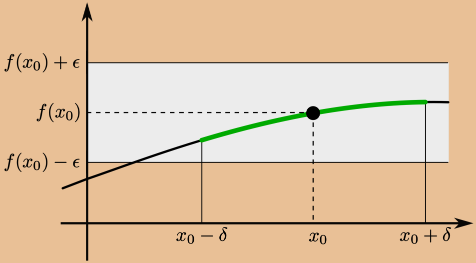
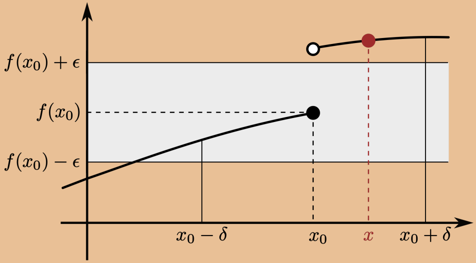


Image credit wikipedia.

"Zoom" interpretation: For all rectangle heights $\epsilon > 0$, there is a sufficiently small rectangle width $\delta > 0$, such that the graph of f is entirely inside the rectangle (or zoom area).

Discontinuous function:



No matter how small we choose δ, there will be an argument x with a distance of less than δ to x_0, such that the function value $f(x)$ differs by more than ϵ from $f(x_0$.

Weierstrass theorem

Theorem (Weierstrass extreme value theorem)

Let $f : [a, b] \to \mathbb{R}$ continuous. Then there exists $x_m, x_M \in [a, b]$ such that

$$f(x_m) \leq f(x) \leq f(x_M), \quad \forall x \in [a, b].$$

This theorem does not only says the function is bounded on the interval but also states the function attains maximum and minimum on the interval.

Note the two essential requirements: interval $[a, b]$ is *compact*, and f is continuous.

Examples where the theorem does **not** apply and the function fails to attain a maximum:

1. $f(x) = x$ defined on $[0, \infty)$ is not bounded from above.
2. $f(x) = \frac{x}{1+x}$ defined on $[0, \infty)$ is bounded but does not attain its least upper bound 1.
3. $f(x) = \frac{1}{x}$ defined on $(0, 1]$ is not bounded from above.
4. $f(x) = 1-x$ defined on $(0, 1]$ is bounded but does not attain its least upper bound 1.

Differentiability

Differential calculus is based on the idea of approximating an arbitrary function by an affine function.

A function $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is affine if there exists a linear transformation $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ and a $y_0 \in \mathbb{R}^m$ such that

$$\mathcal{A}(x) = \mathcal{L}(x) + y_0$$

for every $x \in \mathbb{R}^n$. Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$, and a point $x_0 \in \mathbb{R}^n$. We wish to find an affine function $\mathcal{A}$ that approximates f near the point x_0. First, it is natural to impose the condition

$$\mathcal{A}(x_0) = f(x_0).$$

Because $\mathcal{A}(x) = \mathcal{L}(x) + y_0$, we obtain $y_0 = f(x_0) - \mathcal{L}(x_0)$.

By the linearity of $\mathcal{L}$,

$$\mathcal{A}(x) = \mathcal{L}(x) + y_0 = \mathcal{L}(x) - \mathcal{L}(x_0) + f(x_0) = \mathcal{L}(x - x_0) + f(x_0).$$

Differentiability

Hence, we write

$$\mathcal{A}(x) = \mathcal{L}(x - x_0) + f(x_0).$$

We also require that $\mathcal{A}(x)$ approaches $f(x)$ faster than x approaches x_0, that is,

$$\lim_{x \to x_0} \frac{\|f(x) - \mathcal{A}(x)\|}{\|x - x_0\|} = 0$$

The above conditions ensure that $\mathcal{A}$ approximates f near x_0 in the sense that the error in the approximation at a given point is small compared with the distance of the point from x_0.

Definition (Differentiable function at x_0)

A function $f : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ open, is said to be differentiable at $x_0 \in U$ if there is an affine function that approximates f near x_0, that is, there exists a linear transformation $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\substack{x \to x_0 \\ x \in U}} \frac{\|f(x) - (\mathcal{L}(x - x_0) + f(x_0))\|}{\|x - x_0\|} = 0.$$

The linear transformation $\mathcal{L}$ above is uniquely determined by f and x_0, and is called the derivative of f at x_0.

Definition

A function $f : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ open is said to be differentiable on U if f is differentiable at every point of its domain U.

Differentiability

An alternative way of writing the above expression is

$$f(x) = f(x_0) + \mathcal{L}(x - x_0) + r(x - x_0)\,,$$

where $\frac{\|r(x-x_0)\|}{\|x-x_0\|} \to 0$ when $x \to x_0$.

Sometimes is useful to express the limit using an offset $h = x - x_0$, that is,

$$\lim_{\substack{h\to 0\\ x_0+h\in U}} \frac{\|f(x_0+h) - (\mathcal{L}_{\mathbf{x}_0}(h) + f(x_0))\|}{\|h\|} = 0\,,$$

and thus $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x_0 \in \mathbb{R}^n$ iff there exists a linear transformation $\mathcal{L}_{\mathbf{x}_0}$ such that

$$f(x_0 + h) = f(x_0) + \mathcal{L}_{\mathbf{x}_0}(h) + r_{x_0}(h)\,,$$

where $\frac{\|r_{x_0}(h)\|}{\|h\|} \to 0$ when $h \to 0$. Here $\mathcal{L}_{\mathbf{x}_0}$ denotes the fact that the linear transformation depends on x_0, the same for the remainder r.

Now the question is: Who is the linear transformation $\mathcal{L}$?

Lets see in $\mathbb{R}$: an affine function has the form $ax + b$, with $a, b \in \mathbb{R}$. Hence, a real-valued function $f(x)$ of a real variable x that is differentiable at x_0 can be approximated near x_0 by a function

$$\mathcal{A}(x) = ax + b.$$

Because $f(x_0) = \mathcal{A}(x_0) = ax_0 + b$, we obtain

$$\mathcal{A}(x) = ax + b = a(x - x_0) + f(x_0).$$

So the linear transformation is $\mathcal{L}(x) = ax$.

The norm of a real number is its absolute value, so by the definition of differentiability

$$\lim_{x \to x_0} \frac{|f(x) - (a(x - x_0) + f(x_0))|}{|x - x_0|} = 0,$$

which is equivalent to

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = a.$$

The number a is commonly denoted $f'(x_0)$, and is the derivative of f at x_0. The affine function $\mathcal{A}$ is therefore given by

$$\mathcal{A}(x) = f(x_0) + f'(x_0)(x - x_0).$$

This affine function is tangent to f at x_0 (see Figure).

Differentiability

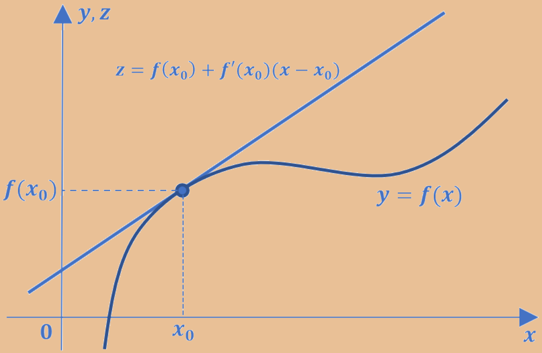


Figure: Illustration of the notion of the derivative.

To extend this idea to $\mathbb{R}^n$, we need to discuss partial derivatives.

Partial derivatives

Like ordinary derivatives, partial derivatives are defined as limits: Let $f : U \to \mathbb{R}$, $U \subseteq \mathbb{R}^n$. The partial derivative of f at the point $a = (a_1, \dots, a_n) \in U$ with respect to the i-th variable is defined as

$$\begin{aligned}\frac{\partial}{\partial x_i} f(a) &= \lim_{t \to 0} \frac{f(a_1, \dots, a_{i-1}, a_i + t, a_{i+1}, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{h} \\ &= \lim_{t \to 0} \frac{f(a + te_i) - f(a)}{h},\end{aligned}$$

where e_i is the i-th canonical basis vector.

Rate of change of the function at a in the i-th direction only.

If all partial derivatives $\partial f / \partial x_i(a)$ exist at a given point a, we say the function is *partially* differentiable at a. This *does not* imply that the function is differentiable at a.

If all partial derivatives exist in a neighborhood of *a and are continuous* there, then f is differentiable in that neighborhood.

Note: The above arguments can be generalized to vector valued functions, $f : U \to \mathbb{R}^m$, by using a componentwise argument.

Example

Let

$$z \equiv f(x, y) = x^2 + xy + y^2.$$

The graph of this function defines a surface.

To every point on this surface, there are an infinite number of tangent lines. Partial differentiation is the act of choosing one of these lines and finding its slope. Usually, the lines of most interest are those that are parallel to the xz-plane, and those that are parallel to the yz-plane (which result from holding either y or x constant, respectively).

By finding the derivative of the equation while assuming that y is a constant, we find that the slope of f at the point (x, y) is:

$$\frac{\partial f}{\partial x} = 2x + y.$$

Partial derivatives

Example (continued)

The function f *can be reinterpreted as a family of functions of one variable indexed by the other variables:*

$$f(x, y) = f_y(x) = x^2 + xy + y^2.$$

In other words, every value of y *defines a function, denoted* f_y*, which is a function of one variable* x*. That is,*

$$f_y(x) = x^2 + xy + y^2 \equiv f(x; y).$$

Once a value of y *is chosen, say* a*, then* $f(x, y)$ *determines a function* f_a *which traces a curve* $x^2 + ax + a^2$ *on the xz-plane:*

$$f(x; a) = f_a(x) = x^2 + ax + a^2.$$

In this expression, a *is a constant, not a variable, so* f_a *is a function of only* x*. Consequently, the definition of the derivative for a function of one variable applies.*

MATLAB/Octave code

```
f = @(x) cos(x);
x = 1;

d = [];
for h = 10.^-[1:14]
    d = [d (f(x+h)-f(x))./h];
end
semilogy(abs(-sin(1)-d),'*-')

d = [];
for h = 10.^-[1:14]
    d = [d (f(x+h)-f(x-h))./(2*h)];
end
hold on;
semilogy(abs(-sin(1)-d),'*-')
hold off;

syms x
f = cos(x)
diff(f)
```

Partial derivatives

Complementary videos:

http://www.youtube.com/watch?v=AXqhWeUEtQU

http://www.youtube.com/watch?v=dfvnCHqzK54

http://www.youtube.com/watch?v=kdMep5GUOBw

Directional derivative

What if we wanted to find the rate of change of f in a given direction v other than the canonical directions e_i?

The directional derivative of $f : \mathbb{R}^n \to \mathbb{R}$ along a given vector v at a given point x is the function denoted by $\partial_v f$ and defined by the limit

$$\partial_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} \,,$$

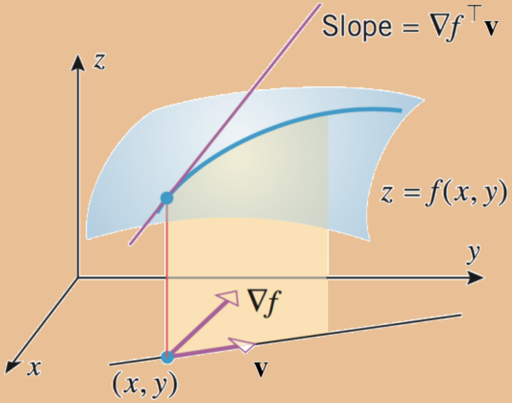
for $t \in \mathbb{R}$.

Intuitively represents the instantaneous rate of change of the function, moving through x with a velocity specified by v.

The partial derivative idea is thus generalized to the directional derivative, which will be useful in many contexts.

Some authors normalize v (i.e., scale it to have length 1). Then the directional derivative specifically measures the rate of change per unit distance in the direction of v.

Directional derivative



Complementary videos:

https://www.youtube.com/watch?v=N_ZRcLheNv0

https://www.youtube.com/watch?v=4RBkIJPG6Yo

https://www.youtube.com/watch?v=4tdyIGIEtNU

back to Differentiability

Any linear transformation from $\mathbb{R}^n \to \mathbb{R}^m$, and in particular the derivative $\mathcal{L}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$, can be represented by an $m \times n$ matrix.

To find the matrix representation L of the derivative $\mathcal{L}$ of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$, we use the canonical basis $\{e_1, ..., e_n\}$ for $\mathbb{R}^n$. Consider the directions

$$x_j = x_0 + te_j, \quad j=1, ..., n.$$

for $t \in \mathbb{R}$. By definition of differentiability:

$$\lim_{t \to 0} \frac{f(x_j) - (tLe_j + f(x_0))}{t} = 0$$

for $j=1, ..., n$.

This means that

$$\lim_{t\to 0} \frac{f(x_j) - f(x_0)}{t} = Le_j$$

for $j=1, ..., n$.

Observe Le_j is the jth column of the matrix L.

On the other hand, the vector x_j differs from x_0 only in the jth coordinate, and in that coordinate the difference is the number t.

Therefore, the left hand side of the last equation is the partial derivative

$$\frac{\partial f}{\partial x_j}(x_0).$$

Here

$$\frac{\partial f}{\partial x_j}(x_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_j}(x_0) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(x_0) \end{pmatrix},$$

and the matrix L (denoted $Df(x_0)$) is then

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0) \cdots \frac{\partial f}{\partial x_n}(x_0) \right] = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix}.$$

The matrix $Df(x_0)$ is called the Jacobian or derivative matrix, of f at x_0, usually denoted $\mathbb{J}_f(x_0)$.

Note: The columns of $Df(x_0)$ are *vector* partial derivatives.

Note: For convenience, we often refer to $Df(x_0)$ simply as the derivative of f at x_0.

We summarize the above discussion in the following theorem.

Theorem

If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at x_0, then the derivative of f at x_0 is uniquely determined and is represented by the $m \times n$ derivative matrix $Df(x_0)$. The best affine approximation to f near x_0 is then given by

$$\mathcal{A}_{\mathbf{x}_0}(x) = f(x_0) + Df(x_0)(x - x_0),$$

in the sense that

$$f(x) = \mathcal{A}_{\mathbf{x}_0}(x) + r_{x_0}(x)$$

where $\frac{\|r_{x_0}(x)\|}{\|x - x_0\|} \to 0$ *when* $x \to x_0$.

Differentiability

For the case $f : \mathbb{R}^n \to \mathbb{R}$ we have that

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} = Df(x)^\top$$

which is called the gradient of f.

The gradient can be pictured as a vector field, by drawing the arrow representing $\nabla f(x)$ so that its tail starts at x.

Definition (continuously differentiable)

A function $f : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ is said to be continuously differentiable on U if it is differentiable (on U), and its derivative $Df : U \to \mathbb{R}^{m \times n}$ is continuous, that is, the components of f have continuous partial derivatives.

In the case, we write $f \in C^1$. If the components of f have continuous partial derivatives of order p, then we write $f \in C^p$.

Examples in $\mathbb{R}^2$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ and $(x_0, y_0) \in \mathbb{R}^2$. We say f is differentiable at (x_0, y_0) iff there exist two real numbers a and b such that

$$f(x_0 + h_1, y_0 + h_2) = f(x_0, y_0) + ah_1 + bh_2 + r(h_1, h_2)\,,$$

where $r : \mathbb{R}^2 \to \mathbb{R}$ such that $\lim_{(h_1,h_2)\to(0,0)} \frac{r(h_1,h_2)}{\|(h_1,h_2)\|} = 0$.

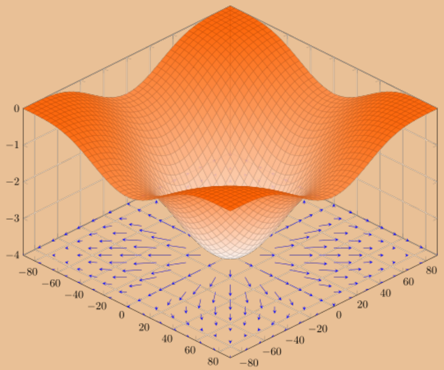
Here the increment is $h = (h_1, h_2)$, and the linear transformation is $ah_1 + bh_2$.

Moreover,

$$a = \frac{\partial f}{\partial x}(x_0, y_0)\,, \quad b = \frac{\partial f}{\partial y}(x_0, y_0)\,.$$

Study differentiability at $(0, 0)$ of some functions:
i) $f(x, y) = 2x + 3y + 4$.
ii) $f(x, y) = \frac{xy}{\sqrt{x^2+y^2}}$ if $(x, y) \neq (0, 0)$, and 0 otherwise. (not diff)
iii) $f(x, y) = \frac{x^2 y}{\sqrt{x^2+y^2}}$ if $(x, y) \neq (0, 0)$, and 0 otherwise. (diff).

Gradient



$f(x, y) = -(\cos^2 x + \cos^2 y)^2$ with the vector field given by the gradient.

The gradient is the zero vector at a point iff it is a stationary point (where the derivative vanishes). The gradient thus plays a fundamental role in optimization theory.

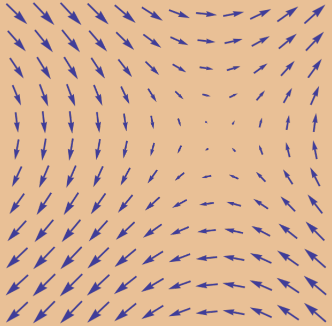
If the gradient is non-zero at a point x, points to the direction in which the function increases most quickly from x (more on this later).

Recall: Vector field

A vector field is an assignment of a vector to each point in a space.

A vector field in the plane can be visualised as a collection of arrows with a given magnitude and direction, each attached to a point in the plane.

Vector fields can be constructed out of scalar fields using the gradient.



The vector field of $(\sin y, \sin x)$ on a subset of $\mathbb{R}^2$.

Complementary videos:

`http://www.youtube.com/watch?v=5FWAVmwMXWg`
`http://www.youtube.com/watch?v=VJ2ZDLQk3IQ`

See Mathematica:

https://www.wolframcloud.com/obj/pguerra0/Published/MA576_Grads1.nb

Example 1: Compute the gradient of $f(x) = \|x\|_2$.

Example 2: Compute the gradient of $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, for $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Recall $\|x\|_2^2 = x^\top x$.

Check computation of the gradient in MATLAB/Octave for a given n.

Useful reference to avoid computing gradient formulas:
Matrix Cookbook:
http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
(see fmla. (84) pg 11).

Complementary videos:
https://www.youtube.com/watch?v=tIpKfDc295M
https://www.youtube.com/watch?v=_-02ze7tf08

Gradient: MATLAB/Octave code

We can use finite differences to compute partial derivatives and thus the gradient:

```
A = [1 2 3; 4 5 6; 7 8 9];
b = [10 11 12]';

x = [1 2 3]';
h = 0.001; %h = eps^(1/3); % cube root of machine epsilon
e = [0 0 1]';

f = @(x) norm(A*x-b)^2/2;
gf = @(x) A'*(A*x-b);

gg = gf(x)
(f(x+e*h)-f(x-e*h))./(2*h)
```

Note: "step-size dilemma": choose a small step size to minimize truncation error while avoiding the use of a step so small that errors due to subtractive cancellation become dominant.

Complex step derivative approximation

Complex step derivative approximation is an alternative method that is significantly more accurate since there are no subtractive cancellation errors because it does not involve a difference operation.

If the function is analytic (it is a stronger condition than being differentiable in $\mathbb{R}$) we have

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{\mathrm{Im}[f(x + ih)]}{h},$$

and thus

$$\frac{\partial f}{\partial x} \approx \frac{\mathrm{Im}[f(x + ih)]}{h},$$

for a small h.

An implementation is given in the next slide.

Gradient: Using complex step

```
function grad = grad_cs(f, x, h)
% f: function handle
% x: point at which to compute the gradient
% h: (optional) step size for complex step
    if nargin < 3
        h = eps^(1/3);
    end

    n = length(x);
    grad = zeros(n, 1); % initialize gradient vector

    for j = 1:n
        x_complex_step = x;
        % add a complex step to the j-th element
        x_complex_step(j) = x_complex_step(j) + 1i*h;
        % compute derivative using complex step
        grad(j) = imag(f(x_complex_step)) / h;
    end
end
```

Some differentiation formulas

In each case, we compute the gradient with respect to $x \in \mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$ be a fixed matrix, and $b \in \mathbb{R}^m$ a fixed vector. Then,

$$\frac{\partial}{\partial x}(x^\top Ax) = x^\top(A + A^\top).$$

It follows that if Q is a symmetric matrix, then

$$\frac{\partial}{\partial x}(x^\top Qx) = 2x^\top Q.$$

In particular,

$$\frac{\partial}{\partial x}(x^\top x) = 2x^\top.$$

In general,

$$\frac{\partial}{\partial x}(x^\top Ax + bx) = (A + A^\top)x + b\,.$$

Check the formulas computing $f(x+h)$ as we did in class.

All differentiation formulas you need: **Matrix Cookbook**:
`http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

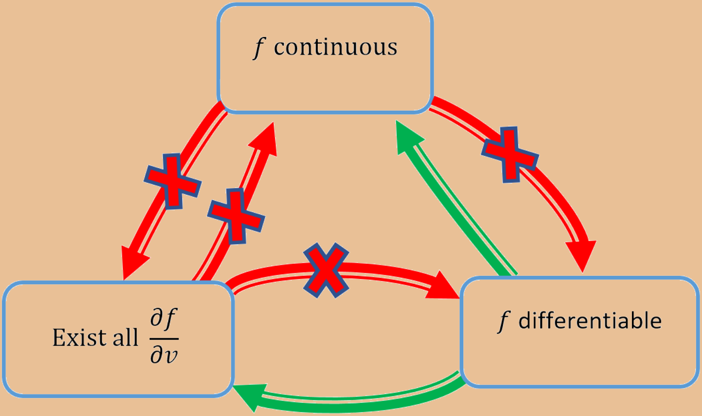
Directional derivative and gradient

If the function f is differentiable at x, then we can compute the directional derivative by using the gradient:

$$\partial_v f(x) = \nabla f(x)^\top v\,.$$

Intuitively, when f is differentiable at a point it is enough to know the partial derivatives to compute all directional derivatives. This is consistent with the geometric idea of a tangent plane that approximates the graph of the function: Knowing the plane (that is the partial derivatives) we know the local behaviour of the function.

Differentiability/Continuity/Directional derivatives



Example: $f(x,y) = \begin{cases} \frac{x^3}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$.

f not differentiable at $(0,0)$ but all directional derivatives exist.

Differentiability - Differentiation rules

Chain rule for differentiating the composition of functions $g : \mathbb{R} \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$.

Theorem

Let $f : U \to \mathbb{R}$ be differentiable on a open set $U \subseteq \mathbb{R}^n$ and let $g : (a, b) \to U$ be differentiable on (a, b). Then, the composite function $h : (a, b) \to \mathbb{R}$ given by $h(t) = f(g(t))$ is differentiable on (a, b), and

$$\frac{d}{dt}h(t) = h'(t) = Df(g(t))Dg(t) = \nabla f(g(t))^\top \begin{pmatrix} g_1'(t) \\ \vdots \\ g_n'(t) \end{pmatrix}.$$

This expression reminds the chain rule in $\mathbb{R}$:

$$\frac{d}{dt}f(g(t)) = f'(g(t))g'(t).$$

Formally, the same for $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^k \to \mathbb{R}^n$.

Examples: $f(x, y, z) = xy - z^2$ and $g(t) = (\sin t, \cos t, e^t)$.
$f(x, y, z) = (x^2 y + e^z, \sin(x) + yz)$ and $g(u, v) = (uv, e^u, \cos(v))$

Level sets

The level set of a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is a set where the function takes on a given constant value c, that is:

$$S_c(f) = \{x \in \mathbb{R}^n | \ f(x) = c\}.$$

When $n = 2$, a level set is called a level curve, also known as contour line or isoline. When $n = 3$, a level set is called a level surface (or isosurface); so a level surface is the set of all real-valued roots of an equation in three variables x_1, x_2 and x_3. For higher values of n, the level set is a level hypersurface, the set of all real-valued roots of an equation in $n > 3$ variables.

Complementary video: `http://www.youtube.com/watch?v=WsZj5Rb6do8`

Example

Himmelblau's function Multi-modal function used to test the performance of optimization algorithms:

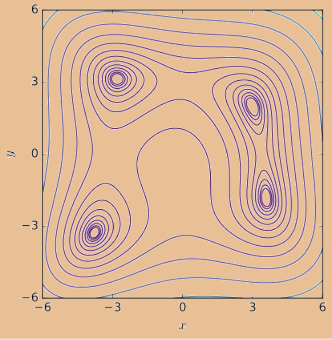
$$f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$

It has one local maximum at $x = -0.270845$ *and* $y = -0.923039$ *where* $f(x, y) = 181.617$, *and four identnical local minima:*

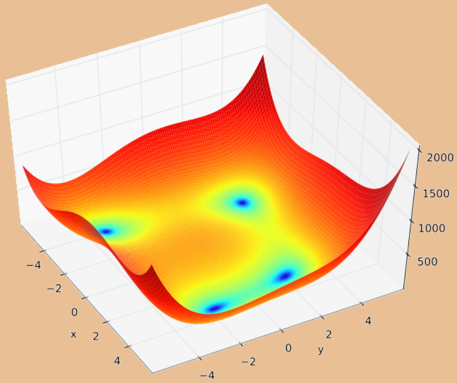
- $f(3.0, 2.0) = 0.0$,
- $f(-2.805118, 3.131312) = 0.0$,
- $f(-3.779310, -3.283186) = 0.0$,
- $f(3.584428, -1.848126) = 0.0$.

The locations of all the minima can be found analytically. However, because they are roots of cubic polynomials, when written in terms of radicals, the expressions are somewhat complicated.

Level sets



Level curve plot (contour) of Himmelblau's function




```
f[x_,y_]=(x^2+y-11)^2+(x+y^2-7)^2;
Plot3D[f[x,y],{x,-6,6},{y,-6,6},PlotRange->All]
ContourPlot[f[x,y],{x,-5,5},{y,-5,5},ContourShading->False,
ContourLabels->True,Contours->30]
```

Level sets

In $\mathbb{R}^3$: A level set of the form $f(x, y, z) = c$ is equivalent to defining a function of two variables, which may be plotted in a 3D diagram. Example: $f(x, y, z) = x^2 - y^2 + z^2$.

Sublevel and superlevel sets
A set of the form

$$S_c^-(f) = \{x \in \mathbb{R}^n | \ f(x) \leq c\}$$

is called a sublevel set of f (or, alternatively, a lower level set or trench of f). Similarly,

$$S_c^+(f) = \{x \in \mathbb{R}^n | \ f(x) \geq c\}$$

is called a superlevel set of f (or, alternatively, an upper level set of f).

Sublevel sets are important in optimization theory. By Weierstrass's theorem, the boundness of some non-empty sublevel set and the lower-semicontinuity of the function implies that a function attains its minimum. The convexity of all the sublevel sets characterizes quasiconvex functions.

Level sets and gradients

A point x_0 is on the level set S_c at level c means $f(x_0) = c$.

Now suppose that there is a curve γ lying on S_c and parameterized by a continuously differentiable function $g : \mathbb{R} \to \mathbb{R}^n$.

Suppose also that $g(t_0) = x_0$ and $Dg(t_0) = v \neq 0$, so that v is a tangent vector to γ at x_0 (see Figure).

Applying the chain rule to the function $h(t) = f(g(t))$ at t_0, gives

$$h'(t_0) = Df(g(t_0))Dg(t_0) = Df(x_0)v = \nabla f(x_0)^\top v.$$

since $f : \mathbb{R}^n \to \mathbb{R}$. Now γ lies on S_c, so $f = c$ and thus we have

$$h(t) = f(g(t)) = c,$$

that is, h is constant. Thus, $h'(t_0) = 0$ and

$$Df(x_0)v = \nabla f(x_0)^\top v = 0.$$

Hence, we have proved, assuming f continuously differentiable, the following theorem.

Level sets and gradients

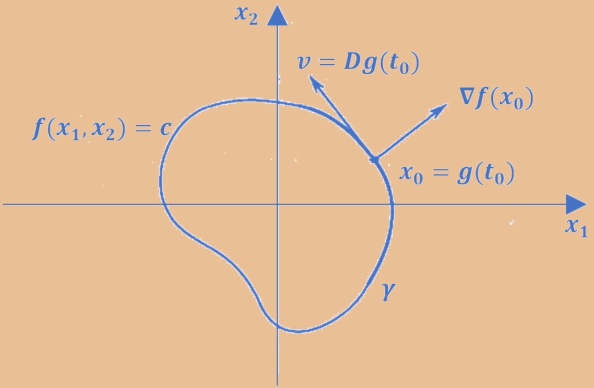


Figure: Orthogonality of the gradient to the level set

$\nabla f(x)$ is orthogonal to the level set through x

Theorem

The vector $\nabla f(x_0)$ is orthogonal to the tangent vector to an arbitrary smooth curve passing through x_0 on the level set determined by $f(x) = f(x_0)$

It is natural to say that $\nabla f(x_0)$ is orthogonal or normal to the level set S corresponding to x_0, and to take as the tangent plane (or line) to S at x_0 the set of all points x satisfying

$$\nabla f(x_0)^\top (x - x_0) = 0\,, \quad \text{if } \nabla f(x_0) \neq 0.$$

As we shall see in the next slide, $\nabla f(x_0)$ is the direction of maximum rate of increase of f at x_0.

Because $\nabla f(x_0)$ is orthogonal to the level set through x_0 determined by $f(x) = f(x_0)$, we deduce that the direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point.

∇f is the direction of maximum rate of increase

Standing at x, choose the direction v such that $\partial_v f$ is the largest possible.

By introducing φ as the angle between the vectors $\nabla f(x)$ and v, we have

$$\nabla f(x) \cdot v = \nabla f(x)^\top v = \|\nabla f(x)\| \|v\| \cos(\varphi) \,.$$

Recall that $\cos(\varphi) \leq 1$, for any φ and $\cos(0) = 1$.

So when $\|v\| = 1$, it follows that at points where $\nabla f(x) \neq 0$, the number $\nabla f^\top v$ is largest when $\varphi = 0$, i.e. when v points in the same direction as $\nabla f(x)$, while $\nabla f^\top v$ is smallest when $\varphi = \pi$ (and hence $\cos(\varphi) = -1$), i.e. when v points in the opposite direction of $\nabla f(x)$.

Note that $\|\nabla f(x)\|$ measures how fast the function increases in the direction of maximal increase.

Hessian

The partial derivative $\frac{\partial f}{\partial x_i}$ is another function and can be partially differentiated:

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial(\frac{\partial f}{\partial x_i})}{\partial x_j}.$$

Given $f : \mathbb{R}^n \to \mathbb{R}$ the matrix of all mixed second order partial derivatives is called the Hessian

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}.$$

If the second partial derivatives are all continuous, the Hessian matrix is a symmetric matrix by Schwarz's theorem and we say f is a C^2 function.

A useful second order differentiation formula (Task check it):

$$\frac{\partial^2}{\partial x \partial x^\top}(x^\top A x + bx) = A + A^\top.$$

Task: Code a snippet that computes the Hessian of a given function.

Taylor's theorem in $\mathbb{R}$

The basis for many numerical methods and models for optimization is Taylor's formula, which is given by Taylor's theorem below.

Recall theorem in $\mathbb{R}$: Assume $f : \mathbb{R} \to \mathbb{R}$, and C^m (continuously differentiable m times) on a closed interval. Let $h = x - x_0$. Then,

$$f(x_0+h) = f(x_0) + \frac{h}{1!} f^{(1)}(x_0) + \frac{h^2}{2!} f^{(2)}(a) + ... + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(x_0) + R_m$$

where

$$R_m = \frac{h^m}{m!} f^{(m)}(x_0 + \xi h)$$

with $\xi \in (0, 1)$.

Taylor's theorem in $\mathbb{R}^n$

In the case of $f : \mathbb{R}^n \to \mathbb{R}$ about the point $x_0 \in \mathbb{R}^n$ where $f \in C^2$ we have:

$$\begin{aligned} f(x) = f(x_0) &+ \nabla f(x_0)^\top (x - x_0) \\ &+ \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0) + r(x - x_0) \end{aligned}$$

where $\frac{r(x-x_0)}{\|x-x_0\|^2} \to 0$ when $x \to x_0$.

In h notation:

$$\begin{aligned} f(x_0+h) = f(x_0) &+ \nabla f(x_0)^\top h \\ &+ \frac{1}{2} h^\top \nabla^2 f(x_0) h + r(h)\,, \end{aligned}$$

where $\frac{r(h)}{\|h\|^2} \to 0$ when $h \to 0$.

This can also be written, for $\xi \in [0,1]$ as

$$f(x_0+h) = f(x_0) + \nabla f(x_0)^\top h + \frac{1}{2} h^\top \nabla^2 f(x_0 + \xi h) h\,.$$

Approximations

Local linearizaton / tangent plane is to approximate $f : \mathbb{R}^n \to \mathbb{R}$ near the point x_0 by means of

$$f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0)\,.$$

Quadratic approximation $f : \mathbb{R}^n \to \mathbb{R}$ near the point x_0 by means of

$$f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0)\,.$$

Note: Suppose x_0 is such that $\nabla f(x_0) = 0$ (stationary point). What happens if $\nabla^2 f$ is pd?

See Mathematica:

`https://www.wolframcloud.com/obj/pguerra0/Published/MA576_Grads2.nb`

Complementary videos:

Local linearization / Tangent plane:

http://www.youtube.com/watch?v=H2hOwKszKRo

http://www.youtube.com/watch?v=QL6qb1h65hg

http://www.youtube.com/watch?v=o7_zS7Bx2VA

Quadratic approx:

http://www.youtube.com/watch?v=8ObJA_tSbo4

http://www.youtube.com/watch?v=UV5yj5A3QIM

http://www.youtube.com/watch?v=szHMvVXxp-g

http://www.youtube.com/watch?v=fW3snxnCPEY

http://www.youtube.com/watch?v=LbBcuZukCAw

http://www.youtube.com/watch?v=0yEiCV-xEWQ

http://www.youtube.com/watch?v=ClFrIg0PpnM

References

- Axler, S. (2014).
  Linear Algebra Done Right.
  Springer.
- Bazaraa, M. S. and Shetty, C. M. (1979).
  Nonlinear Programming: Theory and Algorithms.
  Wiley, 1 edition.
- Beck, A. (2014).
  Introduction to Nonlinear Optimization.
  MOS-SIAM Series on Optimization. SIAM.
- Bertsekas, D. P. (1999).
  Nonlinear Programming.
  Athena Scientific, 2 edition.
- Boyd, S. and Vandenberghe, L. (2004).
  Convex Optimization.
  Cambridge University Press.
- Chong, E. K. P. and Żak, S. H. (2013).
  An Introduction to Optimization.
  Wiley, 4 edition.
- Fletcher, R. (1987).
  Practical Methods of Optimization.
  John Wiley & Sons, New York, NY, USA, second edition.
- Nocedal, J. and Wright, S. (2006).
  Numerical Optimization.
  Springer.
- Ruszczynski, A. (2006).
  Nonlinear Optimization.
  Princeton University Press.