Preliminaries: Topics on differential multivariable calculus

Pedro Vilanova

January 25, 2024

Contents

- Point in $\mathbb{R}^n$
- Inner products and Norms
- Basic topological concepts
- Sequences and limits
- Partial derivatives and directional derivative
- Gradient
- Level sets
- Differentiability
- Taylor's theorem

Point in $\mathbb{R}^n$

A point in $\mathbb{R}^n$ is a vector $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ 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 $\alpha \in \mathbb{R}$:

$$
\begin{aligned}
x + y &= (x_1 + y_1, ..., x_n + y_n)^\top , \\
\alpha x &= (\alpha x_1, ..., \alpha x_n)^\top .
\end{aligned}
$$

Point in $\mathbb{R}^n$

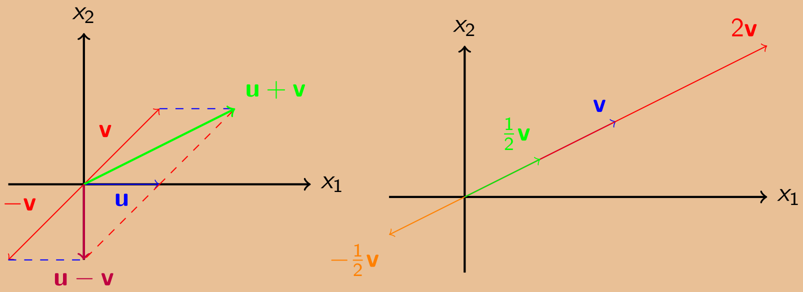


Figure: Vector Addition and Scalar Multiplication in $\mathbb{R}^2$.

Point in $\mathbb{R}^n$

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

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

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.

Euclidean inner product and vector 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 $\mathbb{R}^n$ measuring size of x,
- 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 vector 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 [1]

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

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

[1] 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

Proposition (Cauchy-Schwartz Inequality)

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

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

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

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

and the C-S inequality may be written as

$$|x^\top y| \leq \|x\| \|y\| \,.$$

Definition (Orthogonal vectors)

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

Euclidean inner product and norm

Theorem

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

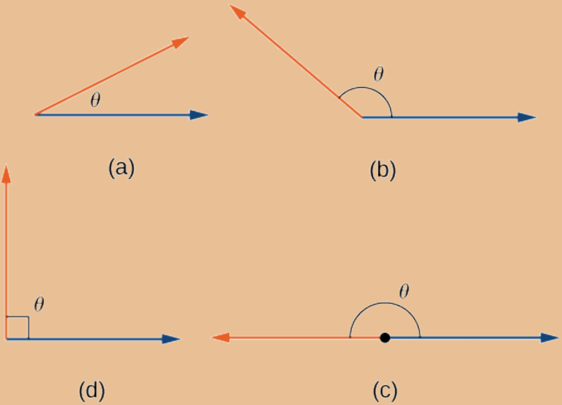
1. *Nonnegativity:* $\|x\| \geq 0$, *with equality iff* $x = 0$.
2. *Homogeneity:* $\|ax\| = |a| \, \|x\|$.
3. *Triangle Inequality:* $\|x + y\| \leq \|x\| + \|y\|$.

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

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

where θ is the angle between both vectors.

Euclidean inner product and norm



(a) An acute angle has $0 < \cos\theta < 1$.

(b) An obtuse angle has $-1 < \cos\theta < 0$.

(c) A straight line has $\cos\theta = -1$.

(d) A right angle (90 degrees) has $cos\theta = 0$.

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 $x \in \mathbb{R}^n$

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

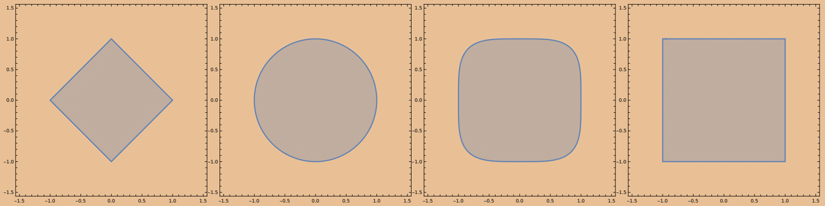
and

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

In particular, the Euclidean norm will be denoted

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

1-norm, 2-norm, 4-norm and infinite norm in $\mathbb{R}^2$:



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 and also focusing of the fact that a matrix represents a linear transformation.

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:

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

This essentially says that the spectral norm is the maximum factor by which A can stretch any unit vector.

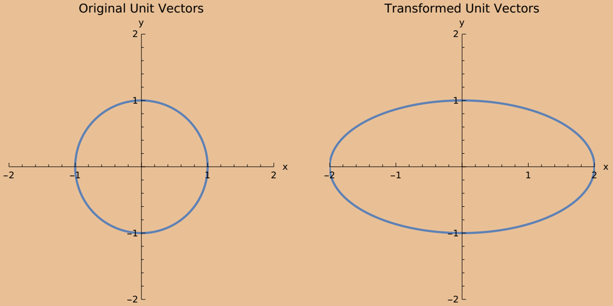
Note that the induced norm satisfies $\|Ax\|_{(m)} \leq \|A\|_{(n),(m)} \|x\|_{(n)}$.

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 $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$.

Matrix norms

Imagine a unit circle centered at the origin in $\mathbb{R}^2$. If you map every point on this circle through the matrix A the circle might get transformed into an ellipse. The length of the major axis of the resulting ellipse is $\|A\|_2$.

Thus, if A has large eigenvalues, this circle gets significantly stretched.



Let $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$. Largest eva of $A^\top A$ is 4. Spectral norm is 2.

Basic topological concepts

In the next slides we formalize the concept of “closeness”.

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 (or 2-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)

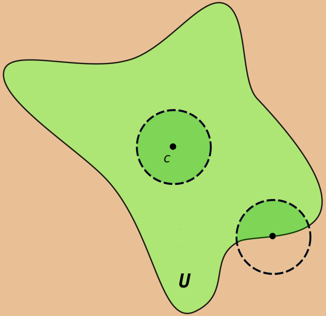
Given a set $U \subseteq \mathbb{R}^n$, a point $c \in U$ is an interior point of U if there exists $r > 0$ for which $B(c, r) \subseteq U$.

The set of all interior points of U:

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

Examples:

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



Basic topological concepts

Open set is a set that contains only interior points.

Definition (open sets)

$U \subseteq \mathbb{R}^n$ 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 $\mathbb{R}^n_{>0}$.

Basic topological concepts

Proposition

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 $x^* \in U$.

Proposition

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 $B[c, r]$.

Basic topological concepts

- 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:

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

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:

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

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.*

Basic topological concepts

Examples of compact sets:

- $[0, 1]$ in general $[x, y]$ for $x, y \in \mathbb{R}$, $x \neq y$.
- 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:

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

Exercise

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.

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

$$x_1, x_2, \dots$$

or simply $\{x_k\}$ [2]..

A sequence $\{x_k\}$ is *decreasing* if $x_1 > x_2 > \dots > x_k \dots$, 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

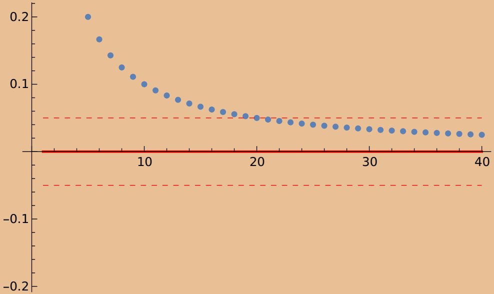
$$x^* = \lim_{k \to \infty} x_k \,, \qquad \text{or} \qquad x_k \to x^* \,.$$

[2] 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 $d\left(x_k, 0\right) = d(1/k, 0) = 1/k < 1/k(\epsilon) < \epsilon$, 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.

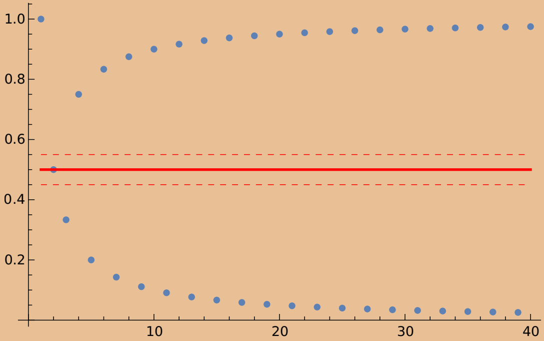
Note 1: This theorem implies that a sequence may have no limit. Indeed if the sequence is unbounded, then the sequence cannot converge (contrapositive).

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

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



Sequences and limits

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.

Sequences and limits

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

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 $\lim_{(x,y)\to(0,0)} \frac{3xy}{x^2+y^2}$. (try the sequence $\{(\frac{c}{k}, \frac{c}{k})\}$ for c constant, or equivalently lines $y = mx$).
   https://www.geogebra.org/material/iframe/id/e2pAxbeP
2. For $f : \mathbb{R}^2 \backslash \{(0,0)\} \to \mathbb{R}$, compute $\lim_{(x,y)\to(0,0)} \frac{xy}{\sqrt{x^2+y^2}}$.
3. Show that $\lim_{(x,y)\to(0,0)} \frac{6x^2y}{x^4+y^2}$ does not exist.
   https://www.geogebra.org/material/iframe/id/E3jUp27u

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 $x_0 \in D$.

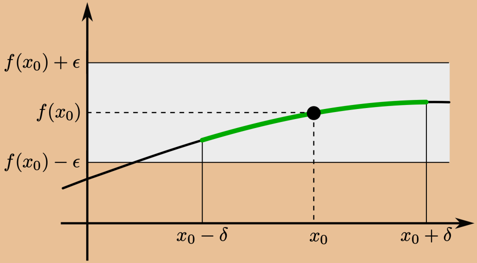
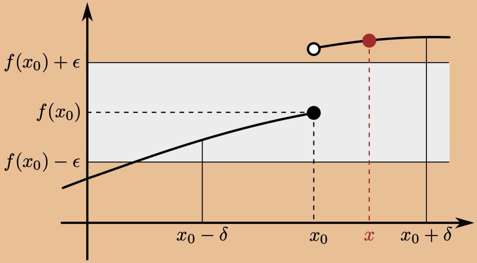


Image credit wikipedia.

Continuity

"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.

Directional derivatives

Recall: for a function $\varphi : \mathbb{R} \to \mathbb{R}$, the derivative at a point c, that is,

$$\varphi'(c) = \lim_{h \to 0} \frac{\varphi(c+h) - \varphi(c)}{h},$$

is the slope of the best affine approximation to φ at c. We may also regard it as the slope of the graph of φ at $(c, \varphi(c))$, or as the instantaneous rate of change of $\varphi(x)$ with respect to x when $x = c$.

Directional derivatives

Definition (Directional derivative)

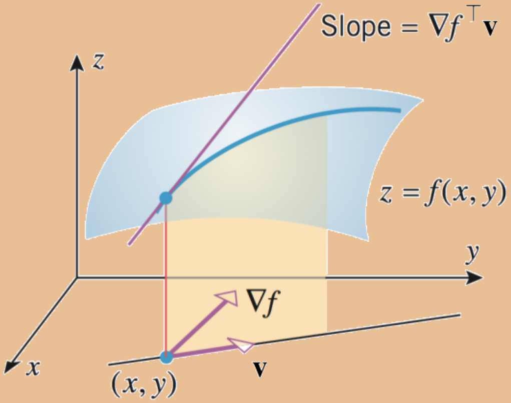
Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is defined on an open ball about a point c. Given a unit vector u, we call

$$\partial_u f(c) = \lim_{h \to 0} \frac{f(c + hu) - f(c)}{h}, \tag{1}$$

provided the limit exists, the directional derivative of f in the direction of u at c.

The graph of f has a slope of $\partial_u f(c)$ if we start “moving” from c and head in the (straight) direction of u.

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`

Partial derivatives

Directional derivatives in the direction of the standard basis vectors will be of special importance.

Definition (Partial derivative)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is defined on an open ball about a point c. If we consider f as a function of $x = (x_1, x_2, \ldots, x_n)$ and let e_k be the k-th standard basis vector, $k = 1, 2, \ldots, n$, then we call $\partial_{e_k} f(c)$, if it exists, the partial derivative of f with respect to x_k at c.

Notation:

$$\frac{\partial}{\partial x_k} f(x_1, x_2, \ldots, x_n).$$

Partial derivatives

Now suppose $f : \mathbb{R}^n \to \mathbb{R}$ and, for fixed $x = (x_1, x_2, \ldots, x_n)$, define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(t) = f(t, x_2, \ldots, x_n).$$

Then

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x_1, x_2, \ldots, x_n) &= \lim_{h \to 0} \frac{f((x_1, x_2, \ldots, x_n) + he_1) - f(x_1, x_2, \ldots, x_n)}{h} \\
&= \lim_{h \to 0} \frac{f((x_1, x_2, \ldots, x_n) + (h, 0, \ldots, 0)) - f(x_1, x_2, \ldots, x_n)}{h} \\
&= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h} \\
&= \lim_{h \to 0} \frac{g(x_1 + h) - g(x_1)}{h} \\
&= g'(x_1).
\end{aligned}$$

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

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)
```

Gradient

Definition (Gradient)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is defined on an open ball containing the point c and $\frac{\partial}{\partial x_k} f(c)$ exists for $k = 1, 2, \ldots, n$. We call the vector

$$\nabla f(c) = \left(\frac{\partial}{\partial x_1} f(c), \frac{\partial}{\partial x_2} f(c), \ldots, \frac{\partial}{\partial x_n} f(c) \right), \tag{2}$$

the gradient of f at c.

Example: If $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$f(x, y) = 3x^2 - 4xy^2,$$

then

$$\nabla f(x, y) = (6x - 4y^2, -8xy).$$

Thus, for example, at the point $c = (2, -1)$ the gradient is the arrow $\nabla f(2, -1) = (8, 16)$.

Directional derivative as gradient

Definition (C^1 function (or continuously differentiable))

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 on an open set U if f is continuous on U and, $\frac{\partial f}{\partial x_k}$ is continuous on U for $k = 1, 2, \ldots, n$.

Theorem (Directional derivative for C^1 function can be expressed using the gradient)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 on an open ball containing the point c. Then for any unit vector u, $\partial_u f(c)$ exists and

$$\partial_u f(c) = \nabla f(c)^\top u \,.$$

Best affine approximation

Proposition

Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that $f \in C^1(U)$. Then

$$\lim_{h \to 0} \frac{f(c+h) - f(c) - \nabla f(c)^\top h}{\|h\|} = 0\,, \quad \textit{for all } c \in U\,.$$

Another way to write the above result is:

$$f(c+h) = f(c) + \nabla f(c)^\top h + r_c(h),$$

where $r_c(\cdot)$ is a function from $\mathbb{R}^n \to \mathbb{R}$ such that $\frac{r_c(h)}{\|h\|} \to 0$ when $h \to 0$. The function $r_c(\cdot)$ can be also written $r(\cdot; c)$.

This can be also written as:

$$f(c+h) = f(c) + \nabla f(c)^\top h + o(\|h\|)\,.$$

Gradient points in the direction of the maximum rate of increase of the function

An application of the Cauchy-Schwarz inequality shows us that

$$|\partial_u f(c)| = |\nabla f(c)^\top u| \le \|\nabla f(c)\|\|u\| = \|\nabla f(c)\|. \tag{3}$$

Thus, the magnitude of the rate of change of f in any direction at a given point never exceeds the length of the gradient vector at that point.

Moreover, in our discussion of the Cauchy-Schwarz inequality, we saw that we have equality in (3) if and only if u is parallel to $\nabla f(c)$. Indeed, supposing $\nabla f(c) \neq 0$, when

$$u = \frac{\nabla f(c)}{\|\nabla f(c)\|},$$

Gradient points in the direction of the maximum rate of increase of the function

we have

$$\begin{aligned}\partial_u f(c) &= \nabla f(c)^\top u \\ &= \frac{\nabla f(c)^\top \nabla f(c)}{\|\nabla f(c)\|} \\ &= \frac{\|\nabla f(c)\|^2}{\|\nabla f(c)\|} \\ &= \|\nabla f(c)\|,\end{aligned}$$

and

$$\partial_{-u} f(c) = -\|\nabla f(c)\|.$$

Hence we have the following result.

Gradient points in the direction of the maximum rate of increase of the function

Proposition

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 on an open ball containing the point c. Then $\partial_u f(c)$ has a maximum value of $\|\nabla f(c)\|$ when u is the direction of $\nabla f(c)$ and a minimum value of $-\|\nabla f(c)\|$ when u is the direction of $-\nabla f(c)$.

Hessian: second derivative for $\mathbb{R}^n \to \mathbb{R}$

If $f : \mathbb{R}^n \to \mathbb{R}$ has partial derivatives which exist on an open set U, then, for any $i = 1, 2, 3, \ldots, n$, $\frac{\partial f}{\partial x_i}$ is itself a function from $\mathbb{R}^n$ to $\mathbb{R}$.

The partial derivatives of $\frac{\partial f}{\partial x_i}$, if they exist, are called second-order partial derivatives of f.

We denote the partial derivative of $\frac{\partial f}{\partial x_i}$ with respect to x_j, for $j = 1, 2, 3, \ldots$ by

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

If $j = i$, we will write $\frac{\partial^2}{\partial x_i^2} f(x)$ for $\frac{\partial^2}{\partial x_i \partial x_i} f(x)$. It is, of course, possible to extend this notation to third, fourth, and higher-order derivatives.

Hessian: second derivative for $\mathbb{R}^n \to \mathbb{R}$

Definition (C^2 function)

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is C^2 on an open set U if $\frac{\partial^2}{\partial x_j \partial x_i} f$ is continuous on U for each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$.

Theorem

If f is C^2 on an open ball containing a point c, then

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

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$.

Hessian: second derivative for $\mathbb{R}^n \to \mathbb{R}$

Definition

Suppose the second-order partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$ all exist at the point c. We call the $n \times n$ matrix

$$\nabla^2 f(c) = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(c) & \frac{\partial^2}{\partial x_1 \partial x_2} f(c) & \frac{\partial^2}{\partial x_1 \partial x_3} f(c) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(c) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(c) & \frac{\partial^2}{\partial x_2^2} f(c) & \frac{\partial^2}{\partial x_2 \partial x_3} f(c) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} f(c) \\ \frac{\partial^2}{\partial x_3 \partial x_1} f(c) & \frac{\partial^2}{\partial x_3 \partial x_2} f(c) & \frac{\partial^2}{\partial x_3^2} f(c) & \cdots & \frac{\partial^2}{\partial x_3 \partial x_n} f(c) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(c) & \frac{\partial^2}{\partial x_n \partial x_2} f(c) & \frac{\partial^2}{\partial x_n \partial x_3} f(c) & \cdots & \frac{\partial^2}{\partial x_n^2} f(c) \end{pmatrix},$$

the Hessian of f at c.

Notation: Hessian is also denoted $H_f(c)$.

2nd order Taylor expansion for $f : \mathbb{R}^n \to \mathbb{R}$

Let $f : \mathbb{R}^n \to \mathbb{R}$ where f is C^2 on an open ball containing the point c. The 2nd order Taylor expansion about c is

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

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

A similar expression can be obtained taking $h = x - c$.

2nd order Taylor expansion for $f : \mathbb{R}^n \to \mathbb{R}$ no remainder

The 2nd order Taylor expansion can be also written in the following way:

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is C^2 on an open ball $B_n(c, r)$ and let h be a point with $\|h\| < r$. Then there exists a real number $s \in [0, 1]$ such that

$$f(c+h) = f(c) + \nabla f(c)^\top h + \frac{1}{2} h^\top \nabla^2 f(c+sh) h.$$

Proof: Using Mean Value Theorem as follows. We show for the case $n = 2$, but the generalization to any n is analogous.

Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is C^2 on an open ball $B_2(c, r)$ and let $h = (h_1, h_2)$ be a point with $\|h\| < r$.

If we define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(t) = f(c + th)$, then $\phi(0) = f(c)$ and $\phi(1) = f(c + h)$.

2nd order Taylor expansion for $f : \mathbb{R}^n \to \mathbb{R}$ no remainder

From Taylor's theorem (one variable), we know that

$$\phi(1) = \phi(0) + \phi'(0) + \frac{1}{2}\phi''(s),$$

where s is a real number between 0 and 1. Using the chain rule in $\mathbb{R}$, we have

$$\begin{aligned}
\phi'(t) &= \nabla f(c+th)^\top \frac{d}{dt}(c+th) \\
&= \nabla f(c+th)^\top h \\
&= \frac{\partial f}{\partial x}(c+th)h_1 + \frac{\partial f}{\partial y}(c+th)h_2,
\end{aligned}$$

2nd order Taylor expansion for $f : \mathbb{R}^n \to \mathbb{R}$ no remainder

Lets use the notation $f_x \equiv \frac{\partial f}{\partial x}$ and $f_y \equiv \frac{\partial f}{\partial y}$. Recall $f_x : \mathbb{R}^2 \to \mathbb{R}$ and $f_x : \mathbb{R}^2 \to \mathbb{R}$. We have

$$\begin{aligned}
\phi''(t) &= h_1 \nabla f_x(c+th)^\top h + h_2 \nabla f_y(c+th)^\top h \\
&= (h_1 \nabla f_x(c+th) + h_2 \nabla f_y(c+th))^\top h \\
&= \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}^\top \begin{pmatrix} f_{xx}(c+th) & f_{yx}(c+th) \\ f_{xy}(c+th) & f_{yy}(c+th) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \\
&= h^\top \nabla^2 f(c+th) h.
\end{aligned}$$

Side note: Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

Theorem

If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at c, then the derivative of f at c is uniquely determined and is represented by the $m \times n$ derivative matrix $Df(c)$. Then

$$f(x) = f(c) + Df(c)(x - c) + o(\|x - c\|)\,.$$

Here

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

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

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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 in $\mathbb{R}^n \to \mathbb{R}^m$

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.

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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 in $\mathbb{R}^n \to \mathbb{R}^m$

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

Let's recast the definition for $\mathbb{R}$: 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 $\mathcal{L}(x) = ax$.

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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 in $\mathbb{R}^n \to \mathbb{R}^m$

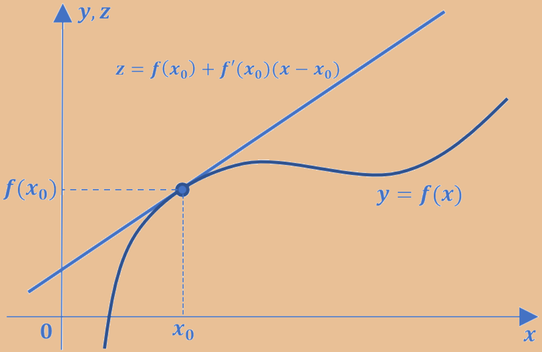


Figure: Illustration of the notion of the derivative.

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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 $j{=}1, ..., n$.

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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

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

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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

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

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

Differentiability in $\mathbb{R}^n \to \mathbb{R}^m$

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 in $\mathbb{R}^n \to \mathbb{R}^m$

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 $f \in C^p$.

Differentiability 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 [3]:
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.
iii) $f(x, y) = \frac{x^2 y}{\sqrt{x^2+y^2}}$ if $(x, y) \neq (0, 0)$, and 0 otherwise.

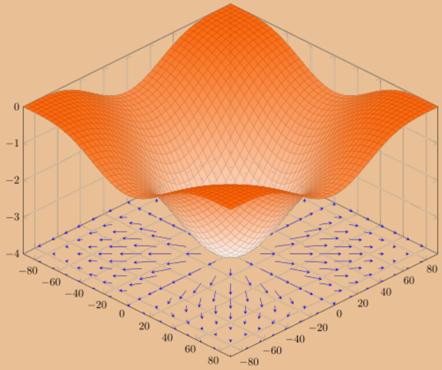
[3] use definition to compute partial derivatives

Gradient - Geometric view

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

$$\nabla f(c) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(c) \\ \vdots \\ \frac{\partial f}{\partial x_n}(c) \end{pmatrix}$$

The gradient defines a vector field, by the arrow $\nabla f(c)$ starting at the point c:



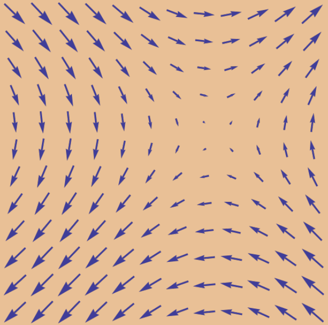
$f(x, y) = -(\cos^2 x + \cos^2 y)^2$ with the vector field given by its gradient.

Side note: 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`

Gradient - Examples

Example 1: Compute the gradient of $f(x) = \|x\|$, for $x \in \mathbb{R}^n$.

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

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

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 - Examples

Example 1 solution: Let

$$f(x) = \|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

The partial derivative with respect to x_i is given by:

$$\begin{aligned}\frac{\partial}{\partial x_i}\|x\| &= \frac{\partial}{\partial x_i}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{\frac{1}{2}} \\ &= \frac{1}{2}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{-\frac{1}{2}} \cdot (2x_i) \\ &= \frac{x_i}{\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{\frac{1}{2}}} \\ &= \frac{x_i}{\|x\|}.\end{aligned}$$

Collect all partial derivatives, and take out the common factor, to find

$$\nabla f(x) = \nabla\|x\| = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \frac{1}{\|x\|}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \frac{x}{\|x\|}.$$

Gradient - Examples

Example 2 solution:

$$\begin{aligned} f(x) &= \frac{1}{2}\|Ax-b\|^2 \\ &= \frac{1}{2}(Ax-b)^\top(Ax-b) \\ &= \frac{1}{2}(x^\top A^\top - b^\top)(Ax-b) \\ &= \frac{1}{2}\left(x^\top A^\top Ax - x^\top A^\top b - b^\top Ax + b^\top b\right) \\ &= \frac{1}{2}x^\top A^\top Ax - (A^\top b)^\top x + \frac{1}{2}b^\top b. \end{aligned}$$

In the last equalitiy we used the fact that the number $x^\top A^\top b = (x^\top A^\top b)^\top = b^\top Ax = (A^\top b)^\top x$.
By using simple gradient formulas, we get the following

$$\nabla_x \left(\frac{1}{2}x^\top A^\top Ax - (A^\top b)^\top x + \frac{1}{2}b^\top b\right) = A^\top Ax - A^\top b.$$

We also have

$$\nabla^2 f(x) = A^\top A.$$

Gradient - Examples

Another way to compute the gradient: We know $\frac{1}{2}\|Ax - b\|^2$ is C^1 (and more). Then

$$f(x + h) = f(x) + \nabla f(x)^\top h + r_x(h),$$

where $r_x(\cdot)$ is a function from $\mathbb{R}^n \to \mathbb{R}$ that converges to zero faster than $\|h\|$ when $h \to 0$.

The idea will be to compute $f(x + h)$ and then recognize the terms $f(x)$, $\nabla f(x)$ and the remainder.

Gradient - Examples

$$
\begin{aligned}
f(x+h) &= \frac{1}{2}\|A(x+h)-b\|^2 \\
&= \frac{1}{2}(A(x+h)-b)^\top(A(x+h)-b) \\
&= \frac{1}{2}(Ax+Ah-b)^\top(Ax+Ah-b) \\
&= \frac{1}{2}(Ax-b+Ah)^\top(Ax-b+Ah) \\
&= \frac{1}{2}((Ax-b)^\top+(Ah)^\top)(Ax-b+Ah) \\
&= \frac{1}{2}\left((Ax-b)^\top(Ax-b)+\underline{(Ax-b)^\top(Ah)}+\underline{(Ah)^\top(Ax-b)}\right) \\
&+ \frac{1}{2}\left((Ah)^\top(Ah)\right) \\
&= \frac{1}{2}\|Ax-b\|^2+(Ax-b)^\top(Ah)+\frac{1}{2}\|Ah\|^2 \\
&= f(x)+\underbrace{(A^\top(Ax-b))^\top}h+\frac{1}{2}\|Ah\|^2
\end{aligned}
$$

Some differentiation formulas

Let $A \in \mathbb{R}^{m\times n}$ be a fixed matrix, and $b \in \mathbb{R}^m$ a fixed vector. Then,

$$\nabla_x(x^\top Ax) = x^\top (A + A^\top).$$

It follows that if Q is a symmetric matrix, then

$$\nabla_x(x^\top Qx) = 2x^\top Q.$$

In particular,

$$\nabla_x(x^\top x) = 2x.$$

In general,

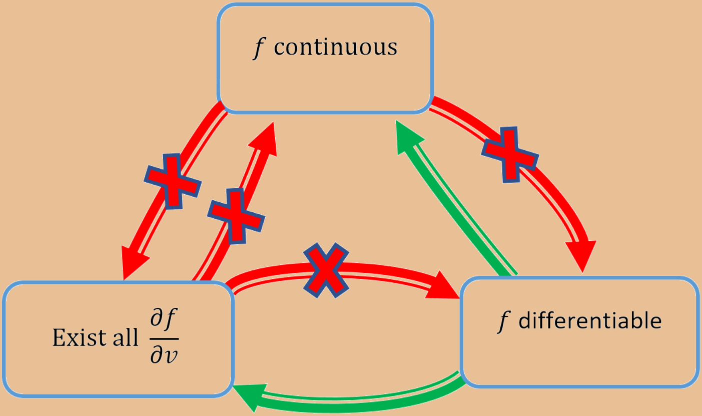
$$\nabla_x(x^\top Ax + bx) = (A + A^\top)x + b\,.$$

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

More formulas in Matrix Cookbook:

`http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

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.

Numerical gradient

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.

Recall: Chain rule in $\mathbb{R}^n$

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

Proposition

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

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

Here $D\phi(\cdot)$ generically denotes the derivative of a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$. In particular, if $\phi : \mathbb{R}^n \to \mathbb{R}$ then $D\phi(\cdot)$ denotes $\nabla\phi(\cdot)^\top$, if $\phi : \mathbb{R} \to \mathbb{R}$ then $D\phi(\cdot)$ denotes the usual derivative $\phi'(\cdot)$ or $\frac{d}{dx}\phi(\cdot)$. For the general case, denotes the Jacobian matrix (more on this later).

Recall: Chain rule in $\mathbb{R}^n$

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

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

Note 2: Formally, the same for $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^k \to \mathbb{R}^n$.

Examples:

1. $f(x) = x^2$, and $g(t) = 2t + 1$. Then $h(t) = f(g(t)) = (2t + 1)^2$. Also $f' = \frac{d}{dx} f = 2x$ and $g' = \frac{d}{dt} g = 2$. Finally, $f'(g(t)) = 2(2t + 1)$. Then $h' = 2(2t + 1)2 = 4(2t + 1)$.
2. $f(x, y, z) = xy - z^2$, and $g(t) = (\sin(t), \cos(t), e^t)$.
3. $f(x, y, z) = (x^2 y + e^z, \sin(x) + yz)$, 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`

Level sets

Example

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

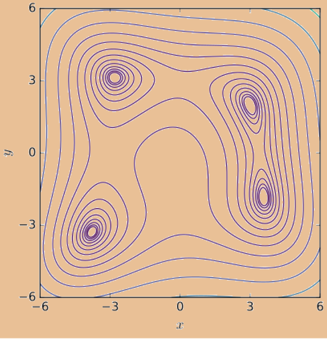
$$f(x,y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$

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

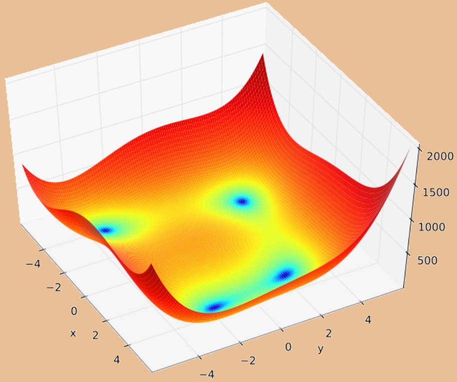
- $f(3.0, 2.0) = 0.0$,
- $f(-2.805118, 3.131312) = 0.0$,
- $f(-3.779310, -3.283186) = 0.0$,
- $f(3.584428, -1.848126) = 0.0$.

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

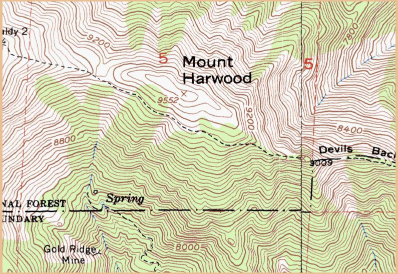


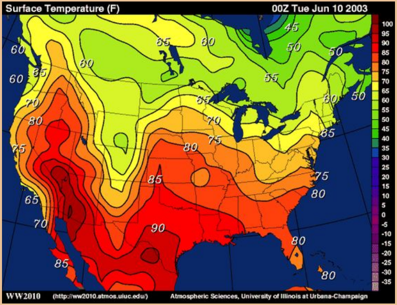
Level curve plot (contour) of Himmelblau's function



```
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 - topography / isotherms





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: $f(x, y, z) = x^2 - y^2 + z^2$.

Definition (Sublevel set)

The set

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

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

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

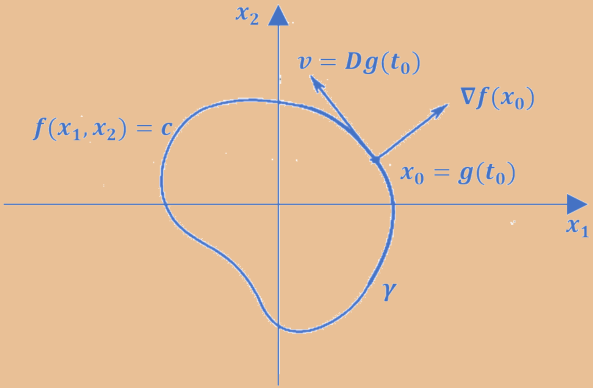


Figure: Orthogonality of the gradient to the level set

$\nabla f(x)$ 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

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

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.

Approximations

Local linearizaton / tangent plane is to approximate $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$ near the point c by means of

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

Quadratic approximation $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$ near the point c by means of

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

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

See Mathematica:

https://www.wolframcloud.com/env/deb27d6c-edaf-430d-a499-0682a7ac9a2c

Approximations

Complementary videos:

Local linearization / Tangent plane:

http://www.youtube.com/watch?v=H2h0wKszKRo
http://www.youtube.com/watch?v=QL6qb1h65hg
http://www.youtube.com/watch?v=o7_zS7Bx2VA

Quadratic approx:

http://www.youtube.com/watch?v=80bJA_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

Taylor expansion with integral remainder

Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$ that is differentiable at a point $c \in \mathbb{R}^n$, the Taylor expansion around c is:

$$f(x) = f(c) + Df(c)(x-c) + \int_0^1 [Df(c + t(x-c)) - Df(c)](x-c)\, dt\,.$$

Here, $Df(x)$ denotes the Jacobian matrix of f at $x \neq c$.

The integral term $\int_0^1 [Df(c + t(x-a)) - Df(c)](x-c)\, dt$ represents the remainder or error in the approximation, accounting for the difference between the linear approximation and the actual function value over the interval between a and x.

If a Lipschitz condition is satisfied

$$\|Df(y) - Df(x)\|_{\mathrm{op}} \leq L\|y-x\|\,.$$

then L can be interpreted as a measure of nonlinearity of f. Here, $\|\cdot\|_{\mathrm{op}}$ denotes the an operator norm.

Note on differentiability

The class C^0 consists of all continuous functions. The class C^1 consists of all differentiable functions whose derivative is continuous; such functions are called continuously differentiable.

Example: $f(x) = \max(0, x)$ is C^0 but not C^1.

Example: $f(x) = |x|$ is C^0 but not C^1 because is not differentiable at $x = 0$.

Example: $f(x) = |x|^2$ is C^1. In general, for each <u>even</u> integer k, the function $f(x) = |x|^{k+1}$ is continuous and k times differentiable at all x, so it is of class C^k.

Note on differentiability

Example: Let $f : \mathbb{R} \to \mathbb{R}$ continuous at $x = c$. Define

$$h(x) = (x - c)f(x).$$

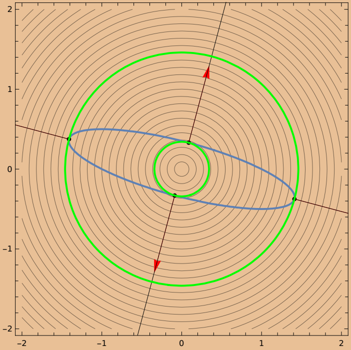
Is the function h differentiable at c? Yes, because

$$\begin{aligned} h'(c) &= \lim_{x \to x_0} \frac{h(x) - h(c)}{x - c} \\ &= \lim_{x \to c} \frac{(x - c)g(x) - 0}{x - c} \\ &= \lim_{x \to c} g(x) \\ &= g(c). \end{aligned}$$

In particular, taking $f(x) = |x|$ and $c = 0$ we have that $h(x) = x|x|$ is differentiable at 0, therefore we say is C^1. What about second derivative? It turns out that $x|x|$ is not twice differentiable at $x = 0$. Therefore is not C^2.

Nonlinear eigenvalue problem

Let $f(x) = 0$ be a n-dimensional surface. Goal: Find those points on the surface whose distance from the origin is stationary (that is the rate of change of the distance with respect to movement along the surface is zero). If f is smooth, this problem can be recasted as an eigenvalue problem involving ∇f and given a simple geometric interpretation.



Nonlinear eigenvalue problem

This idea is illustrated in two dimensions. A typical smooth curve $f(x) = 0$ is shown in the figure. At any point on the surface where a circle centered at the origin is tangent to $f(x) = 0$, we have that $\nabla f(x)$ must be parallel to the vector x. That is, there is some scalar λ such that

$$\nabla f(x) = \lambda x\,.$$

Values of λ satisfying the equation are called eigenvalues and the associated solutions for x, eigenvectors.

Note: If f is a quadratic function of x, then ∇f becomes a linear operator and then the equation reduces to the familiar eigenvalue problem.

Nonlinear eigenvalue problem

Example: Find the point(s) on the graph $y = \frac{1}{x^4}$ closest to the origin.

Solution: Let $f(x, y) = x^4 y - 1$. Then, we have the gradient $\nabla f = (4x^3 y, x^4)$. The nonlinear eigenvalue problem reduces to

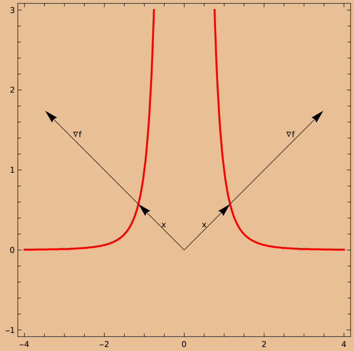
$$\begin{aligned} 4x^3 y &= \lambda x, \\ x^4 &= \lambda y. \end{aligned}$$

If $x \neq 0$, the first equation yields $\lambda = 4x^2 y$. Hence, from the second equation, $x^2 = 4y^2$ or $x = \pm 2y$.
The points we seek must lie on $f(x, y) = 0$. Hence $16y^5 - 1 = 0$ or $y = 16^{-1/5}$. Thus there is only one eigenvalue, $\lambda = 16^{2/5}$, but two associated eigenvectors,

$$\left(2 \cdot 16^{-1/5}, 16^{-1/5}\right)^\top, \quad \left(-2 \cdot 16^{-1/5}, 16^{-1/5}\right)^\top,$$

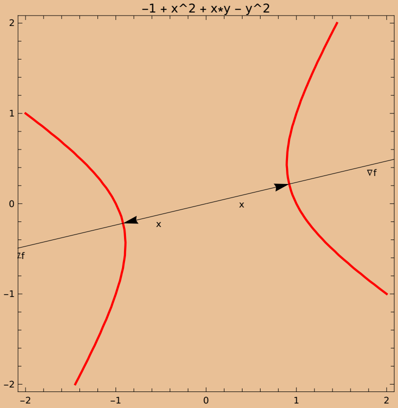
each lying the same distance from the origin.

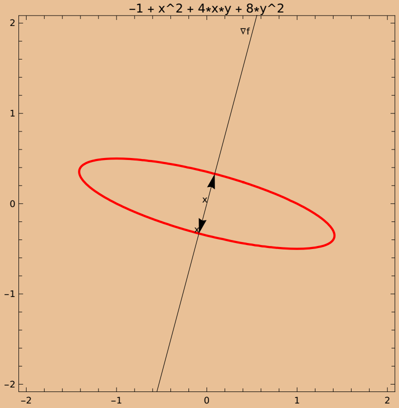
Nonlinear eigenvalue problem



Nonlinear eigenvalue problem

Other examples:





Task: Write a mathematica code that solves the nonlinear eigenvalue problem.

Nonlinear eva problem and Lagrange multipliers

The nonlinear eigenvalue problem is related to the method of Lagrange multipliers, which is a strategy used for finding local optima of a function subject to equality constraints.

The problem of finding points on the surface $f(x) = 0$ whose distance from the origin is stationary can be reformulated as a constrained optimization problem:

$$\text{minimize} \quad \|x\| \quad \text{s.t.} \quad f(x) = 0\,.$$

Using the method of Lagrange multipliers, we introduce a scalar λ (the Lagrange multiplier), and consider the Lagrangian function:

$$\mathcal{L}(x, \lambda) = \|x\| + \lambda f(x)\,.$$

Nonlinear eva problem and Lagrange multipliers

To find the stationary points, we take the derivative of $\mathcal{L}$ with respect to x and set it to zero:

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla(\|x\|) + \lambda \nabla f(x) = 0\,.$$

Assuming $\|x\|$ is differentiable away from the origin (which requires $x \neq 0$), we have $\nabla(\|x\|) = \frac{x}{\|x\|}$, and the equation becomes:

$$\frac{x}{\|x\|} + \lambda \nabla f(x) = 0\,.$$

This can be rewritten as:

$$\nabla f(x) = -\frac{1}{\lambda \|x\|} x\,.$$

Nonlinear eva problem and Lagrange multipliers

Comparing this with the equation $\nabla f(x) = \tilde{\lambda} x$, we can see that $\tilde{\lambda} = -\frac{1}{\lambda \|x\|}$, establishing a relationship between the Lagrange multiplier λ and the scalar $\tilde{\lambda}$ from the eigenvalue problem.

In geometric terms, at the points where the distance from the origin is stationary under the constraint $f(x) = 0$, the gradient of the constraint function $\nabla f(x)$ is aligned with the position vector x, satisfying the eigenvalue-like condition $\nabla f(x) = \lambda x$.

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.