PRINCETON UNIVERSITY
ORFE
Instructor:
Amir Ali Ahmadi
Fall 2015

TAs: G. Hall,
H. Hao,
J. Ye,
Z. Zhu

This lecture:

The goal of this lecture is to refresh your memory on some topics in linear algebra and multivariable calculus that will be relevant to this course. You can use this as a reference throughout the semester.

The topics that we cover are the following:

- Inner products and norms
  - Formal definitions
  - Euclidian inner product and orthogonality
  - Vector norms
  - Matrix norms
  - Cauchy-Schwarz inequality

- Eigenvalues and eigenvectors
  - Definitions
  - Positive definite and positive semidefinite matrices

- Elements of differential calculus
  - Continuity
  - Linear, affine and quadratic functions
  - Differentiability and useful rules for differentiation
  - Gradients and level sets
  - Hessians

- Taylor expansion
  - Little o and big O notation
  - Taylor expansion

Inner products and norms

Definition of an inner product

An inner product is a real-valued function $<\cdot\,,\,\cdot>: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

- Positivity: $< x, x > \geq 0$ and $< x, x > = 0$ iif x=0.
- Symmetry: $< x, y > =< y, x >$.
- Additivity: $< x + y, z > = < x, z > \; + < y, z >$.
- Homogeneity : $< rx, y > = r < x, y > \; \forall\, r \in \mathbb{R}$.

Examples in small dimension

Here are some examples in $\mathbb{R}$ and $\mathbb{R}^2$ that you are already familiar with.

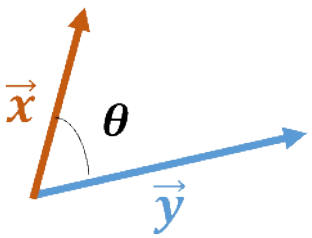
Example 1: Classical multiplication

$$<.,.>: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$
$$(x, y) \to x \cdot y$$

Check that this is indeed a inner product using the definition.

Example 2:

$<.,.>: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$
$\langle x, y\rangle = length(x) \cdot length(y) \cdot \cos(\theta)$



- This geometric definition is equivalent to the following algebraic one:
  $\langle x, y\rangle = x_1 y_1 + x_2 y_2$.

Notice that the inner product is positive when θ is smaller than 90 degrees, negative when it is greater than 90 degrees and zero when $\theta = 90$ degrees.

Euclidean inner product

The two previous examples are particular cases ($n = 1$ and $n = 2$) of the **Euclidean inner product**:

$$< x, y > = \sum_{i=1}^{n} x_i y_i = x^T y \quad where \ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

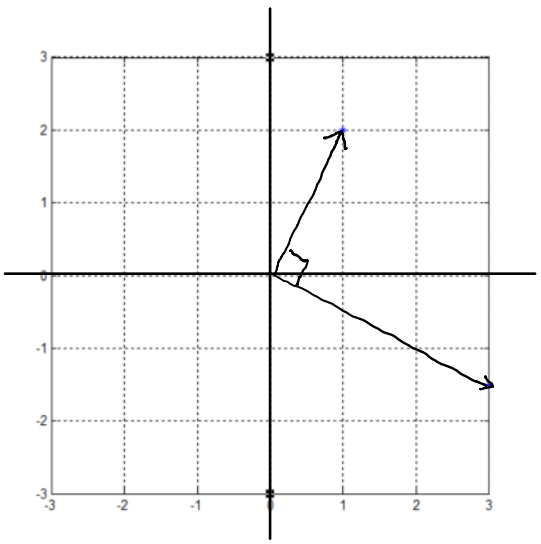
Check that this is an inner product using the definition.

Orthogonality

We say that two vectors x and y are orthogonal if $\langle x, y \rangle = 0$.

- Note that with this definition the zero vector is orthogonal to every other vector.
- But two nonzero vectors can also be orthogonal.
  - For example,

$$x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, y = \begin{pmatrix} 3 \\ -\frac{3}{2} \end{pmatrix}$$

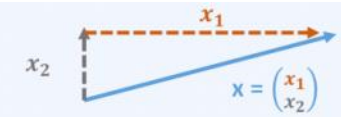


Norms

A vector norm is a real valued function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

- Positivity: $\|x\| \geq 0$ and $\|x\| = 0$ iif $x = 0$.
- Homogeneity: $\|rx\| = |r| \, \|x\|$ for all $r \in \mathbb{R}$.
- Triangle inequality : $\|x + y\| \leq \|x\| + \|y\|$.

Basic examples of vector norms

Figure			
Name	**1-norm or $\|\cdot\|_1$**	**2-norm or $\|\cdot\|_2$ or Euclidean norm**	**∞-norm or $\|\cdot\|_\infty$**
Definition	$\|x\|_1 = \|x_1\| + \cdots + \|x_n\|$	$\|x\|_2 = \sqrt{< x, x >} = \sqrt{x_1^2 + \cdots + x_n^2}$	$\|x\|_\infty = \max_i \|x_i\|$
On the figure	$\|x\|_1 = \|x_1\| + \|x_2\|$ = → + ↑	$\|x\|_2 =$ $\sqrt{x_1^2 + x_2^2} =$ ↗	$\|x\|_\infty = x_1 =$ →

- Check that these are norms using the definition!
- When no index is specified on a norm (e.g., $||.||$) this is considered to be the Euclidean norm.
- For the three norms above, we have the relation $\left|\left|x\right|\right|_1 \geq \left|\left|x\right|\right|_2 \geq \left|\left|x\right|\right|_\infty$.
- Given any inner product $\langle x, y \rangle$, one can construct a norm given by $\left|\left|x\right|\right| = \sqrt{\langle x, x \rangle}$. But not every norm comes from an inner product. (For example, one can show that the $||.||_1$ norm above doesn't.)

Cauchy Schwarz Inequality

For any two vectors x and y in $\mathbb{R}^n$, we have the so-called Cauchy-Schwarz inequality:

$$|< x, y >| \leq \|x\| \cdot \|y\|.$$

Furthermore, equality holds iif $x = \alpha \, y$ for some $\alpha \in \mathbb{R}$.

Matrix norms (We skipped this topic in lecture. We'll come back to it as we need to.)

Similar to vector norms, one can define norms on matrices. These are functions $\|.\|: \mathbb{R}^{m\times n} \to \mathbb{R}$, that satisfy exactly the same properties as in the definition of a vector norm (see page 36 of [CZ13]).

Induced norms

Consider any vector norm $\|.\|_*: \mathbb{R}^k \to \mathbb{R}$. The induced norm $\|.\|_*: \mathbb{R}^{m\times n} \to \mathbb{R}$ on the space of $m \times n$ matrices is defined as:

$$\|A\|_* = \max\{\|Ax\|_* : x \in \mathbb{R}^n \text{ and } \|x\|_* = 1\}$$

Notice that the vector norm and the matrix norm have the same notation; it is for you to know which one we are talking about depending on the context.

One can check that $\|A\|_*$ satisfies all properties of a norm.

Frobenius norm

The Frobenius norm $\|.\|_F: \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined by:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} {a_{ij}}^2}$$

The Frobenius norm is an example of a matrix norm that is not induced by a vector norm. Indeed, $\|I_{n\times n}\|_* = 1$ for any induced norm $\|.\|_*$ (why?) but $\|I_{n\times n}\|_F = n$.

Submultiplicative norms

A matrix norm is submultiplicative if it satisfies the following inequality:

$$\|A\,B\| \leq \|A\| \cdot \|B\|$$

- All induced norms are submultiplicative.
- The Frobenius norm is submultiplicative.
- Not every matrix norm is submultiplicative: $||.||: A \to \max_{i,j} |a_{i,j}|$

Take $A = B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Then $\|A\| \cdot \|B\| = 1 \cdot 1 = 1$.
But $AB = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$. Hence $\|AB\| = 2$ and $\|A\| \cdot \|B\| < \|AB\|$

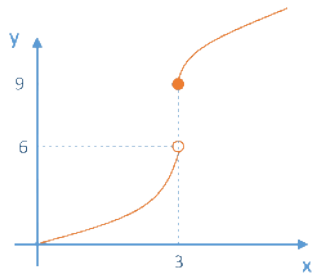
Continuity

- We first give the definition for a univariate function and then see that it generalizes in a straightforward fashion to multiple dimensions using the concept of a vector norm.

Definition in $\mathbb{R}$

A function $f: \mathbb{R} \to \mathbb{R}$ is continuous at a point $a \in \mathbb{R}$ if $\forall\, \epsilon > 0,\ \exists\, \delta > 0\ s.t.$ for all x with $|x - a| < \delta$ we have $|f(x) - f(a)| < \epsilon$.

A function $f: \mathbb{R} \to \mathbb{R}$ is said to be continuous if it is continuous at every point over its domain.



A function that is not continuous

Definition in $\mathbb{R}^n$

A function $f: \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $a \in \mathbb{R}^n$ if :

$\forall\, \epsilon > 0, \exists\, \delta > 0\ s.t.\ \ \|x - a\| < \delta \Rightarrow \|f(x) - f(a)\| < \epsilon.$

Once again, if f is continuous at all points in its domain, then f is said to be continuous.

Remarks.

- If in the above definition we change the 2-norm with any other vector norm, the class of continuous functions would not change.
  - This is because of "equivalence of norms in finite dimensions", a result we didn't prove.
- A function $f: \mathbb{R}^n \to \mathbb{R}^m$ given as $f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$ is continuous if and only if each entry $f_i: \mathbb{R}^n \to \mathbb{R}$ is continuous.

Linear, Affine and Quadratic functions

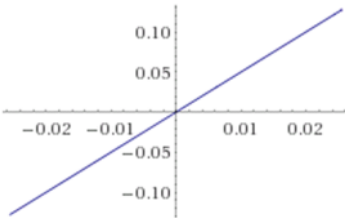
Linear functions

A function $L\colon \mathbb{R}^n \to \mathbb{R}^m$ is called a linear if:

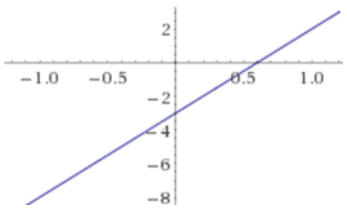
- $L(\alpha\, x) = \alpha\, L(x)\ \ \forall\, x \in \mathbb{R}^n$ and $\forall\, \alpha \in \mathbb{R}$
- $L(x + y) = L(x) + L(y)\ \forall\, x, y \in \mathbb{R}^n$

- Any linear function can be represented as
  $$L(x) = Ax,$$
  where A is an $m \times n$ matrix.

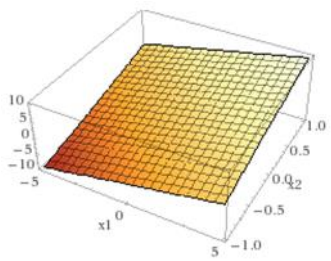
- The special case where $m = 1$ will be encountered a lot. In this case, linear functions take the form $L(x) = a^T x$ for some vector $a \in \mathbb{R}^n$.



Linear $m = n = 1$



Affine $m = n = 1$



Linear $n = 2, m = 1$

Affine functions

A function $f\colon \mathbb{R}^m \to \mathbb{R}^n$ is affine if there exists a linear function $L\colon \mathbb{R}^n \to \mathbb{R}^m$ and a vector $y \in \mathbb{R}^m$ such that:

$$f(x) = L(x) + y \qquad \forall\, x \in \mathbb{R}^n$$

When $m = 1$, affine functions are functions of the form
$f(x) = a^T x + b$ where $a \in \mathbb{R}^n, b \in \mathbb{R}$.

Quadratic functions

A *quadratic form* $f: \mathbb{R}^n \to \mathbb{R}$ is a function that can be represented as

$f(x) = x^T\, Q\, x$

where Q is a $n \times n$ matrix that we can assume to be symmetric without loss of generality (i.e., $Q = Q^T$).

Why can we assume this without loss of generality?

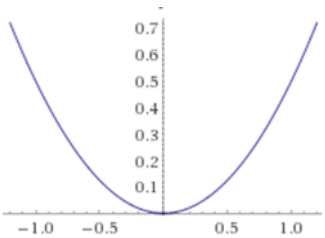
If Q is not symmetric, then we can define $Q_0 = \frac{1}{2}(Q + Q^T)$ which is a symmetric matrix (why?) and we would have $x^T Q x = x^T Q_0 x$ (why?).

What do these functions look like in small dimensions?

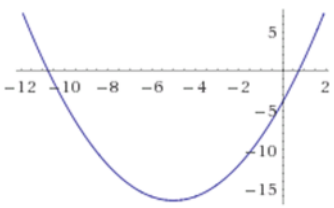
When $n = 1$, we have $f(x) = a\, x^2$ where $a \in \mathbb{R}$.

When $n = 2$, $Q = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, and $x^T Q x = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a x_1^2 + 2b\, x_1 x_2 + c x_2^2$.

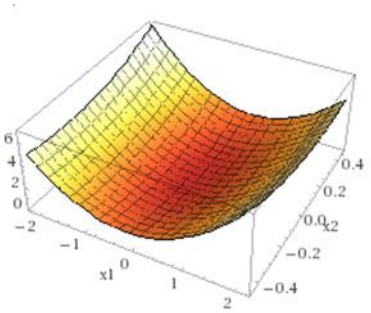
A *quadratic function* is a function that is the sum of a quadratic form and an affine function: $f(x) = x^T Q x + a^T x + b$.



Quadratic form $n = 1$



Quadratic function $n = 1$



Quadratic form $n = 2$

Eigenvalues and Eigenvectors

Definition

- Let A be an $n \times n$ square matrix. A scalar λ and a nonzero vector v satisfying the equation $Av = \lambda v$ are respectively called an eigenvalue and an eigenvector of A. In general, both λ and v may be complex.
- For λ to be an eigenvalue it is necessary and sufficient for the matrix $\lambda I - A$ to be singular, that is $\det(\lambda I - A) = 0$ (I here is the $n \times n$ identity matrix).
- We call the polynomial $\det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$ the characteristic polynomial of A.
- The fundamental theorem of algebra tells us that the characteristic polynomial must have n roots. These roots are the eigenvalues of A.
- Once an eigenvalue λ is computed, we can solve a linear system to $Av = \lambda v$ to obtain the eigenvectors.
- You should be comfortable with computing eigenvalues of 2×2 matrices.

Eigenvalues and eigenvectors of a symmetric matrix

M is a symmetric matrix if $M^T = M$.

- All eigenvalues of a symmetric matrix are real.

Proof. $Ax = \lambda x$ ①

$\Rightarrow x^* A^* = \lambda^* x^*$ $\xrightarrow{A = A^*}$ $x^* A = \lambda^* x^*$ ②

① $\Rightarrow x^* A x = \lambda x^* x$

② $\Rightarrow x^* A x = \lambda^* x^* x$

$\Rightarrow \lambda x^* x = \lambda^* x^* x \Rightarrow (\lambda - \lambda^*) x^* x = 0$ ($x^* x \neq 0$)

$\Rightarrow \lambda = \lambda^*$. □

"Conjugate": $\lambda = a + ib$ $\lambda^* = a - ib$

- Any real symmetric $n \times n$ matrix has a set of n real eigenvectors that are mutually orthogonal. (We did not prove this.)

Positive definite and Positive semidefinite matrices

Notation

A symmetric $n \times n$ matrix Q is said to be

- Positive semidefinite (psd) if $x^T Q x \geq 0$ for all $x \in \mathbb{R}^n$. $Q \succeq 0$
- Positive definite (pd) if $x^T Q x > 0$ for all $x \in \mathbb{R}^n, x \neq 0$. $Q \succ 0$
- Negative semidefinite if $-Q$ is positive semidefinite. $Q \preceq 0$
- Negative definite if $-Q$ is positive definite. $Q \prec 0$
- Indefinite if it is neither positive semidefinite nor negative semidefinite.

Note: The [CZ13] book uses the notation $Q \geq 0$ instead of $Q \succcurlyeq 0$ (and similarly for the other notions). We reserve the notation $Q \geq 0$ for matrices whose entries are nonengative numbers. The notation $Q \succcurlyeq 0$ is much more common in the literature for positive semidefiniteness.

Link with the eigenvalues of the matrix

- A symmetric matrix Q is postive semidefinite (resp. positive definite) if and only if all eigenvalues of Q are nonnegative (resp. positive).
- As a result, a symmetric matrix Q is negative semidefinite (resp. negative definite) if and only if the eigenvalues of Q are nonpositive (resp. negative).

Here is the easier direction of the proof (the other direction is also straightforward; see [CZ13]):

$Q \succeq 0 \Rightarrow$ eigenvalues ≥ 0

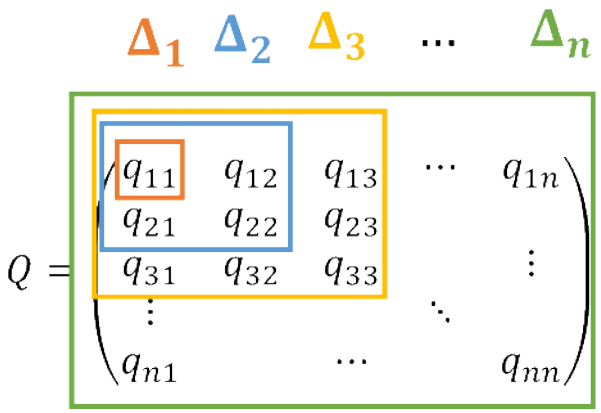
Indeed if $\lambda < 0$ and $Ax = \lambda x$, then $x^T A x = \underbrace{\lambda}_{<0} \underbrace{x^T x}_{\geq 0} < 0 \Rightarrow Q \not\succeq 0$.

Positive definite and positive semidefinite matrices (cont'd)

Sylvester's criterion

Sylvester's criterion provides another approach to testing positive definiteness or positive semidefiniteness of a matrix.

- A symmetric matrix Q is *positive definite* if and only if $\det(\Delta_1), \det(\Delta_2), \ldots, \det(\Delta_n)$ are *positive,* where $\Delta_1, \Delta_2, \ldots, \Delta_n$ are submatrices defined as in the drawing below. These determinants are called the *leading principal minors* of the matrix Q.
- There are always n leading principal minors.



- A symmetric matrix Q is *positive semidefinite* if and only if $\det(\Gamma_1), \det(\Gamma_2), \ldots, \det(\Gamma_{2^n-1})$ are *nonnegative,* where $\Gamma_1, \ldots, \Gamma_{2^n-1}$ are submatrices obtained by choosing a subset of the rows and the same subset of the columns from the matrix Q. The scalars $\det(\Gamma_1), \det(\Gamma_2), \ldots, \det(\Gamma_{2^n-1})$ are called the *principal minors* of Q.

2x2:

$$Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad \left[Q \succ 0 \Leftrightarrow \begin{array}{l} a > 0 \\ ac - b^2 > 0 \\ \uparrow \det Q \end{array} \right], \quad \left[Q \succeq 0 \Leftrightarrow \begin{array}{l} a \geq 0,\ c \geq 0 \\ ac - b^2 \geq 0 \end{array} \right]$$

3x3:

$$Q = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}, \quad \left[Q \succ 0 \Leftrightarrow \begin{array}{l} a > 0 \\ ad - b^2 > 0 \\ \det Q > 0 \end{array} \right], \quad \left[Q \succeq 0 \Leftrightarrow \begin{array}{l} a \geq 0,\ d \geq 0,\ f \geq 0 \\ ad - b^2 \geq 0,\ af - c^2 \geq 0,\ df - e^2 \geq 0 \\ \det Q \geq 0 \end{array} \right]$$

Gradients, Jacobians, and Hessians

Partial derivatives

Recall that the partial derivative of a function $f: \mathbb{R}^n \to \mathbb{R}$ with respect to a variable x_i is given by

$$\frac{\partial f}{\partial x_i} = \lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha},$$

where e_i is the i-th standard basis vector in $\mathbb{R}^n$; i.e., the i-th column of the $n \times n$ identity matrix.

The Jacobian matrix

For a function $f: \mathbb{R}^n \to \mathbb{R}^m$ given as $f(x) = \left(f_1(x), \ldots, f_m(x)\right)^T$, the *Jabocian matrix* is the $m \times n$ matrix of first partial derivatives:

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}$$

(The notation of the CZ book is $D_f(x)$)

The *first order approximation* of f near a point x_0 is obtained using the Jacobian matirx: $A(x) = f(x_0) + J_f(x_0)^T(x - x_0)$. Note that this is an affine function of x.

The gradient vector

The gradient of a real-valued function $f: \mathbb{R}^n \to \mathbb{R}$ is denoted by $\nabla f(x)$ and is given by

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} = J_f(x)^T.$$

This is a very important vector in optimization. As we will see later, at every point, the gradient vector points in a direction where the function grows most rapidly.

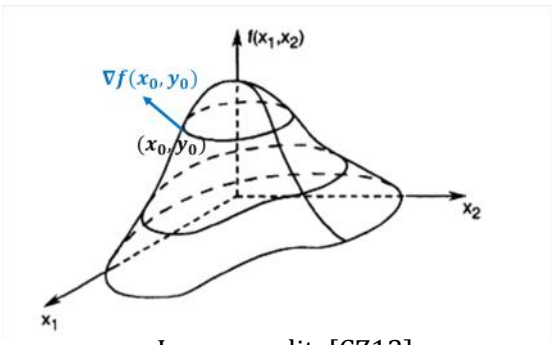
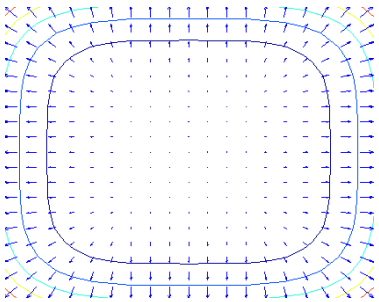


Image credit: [CZ13]

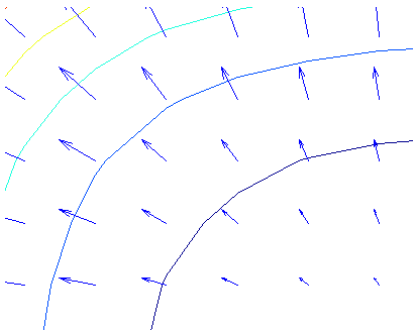
Level sets

For a scalar $\alpha \in \mathbb{R}$, the *α-level set* of a function $f: \mathbb{R}^n \to \mathbb{R}$ is defined as
$S_\alpha = \{x \in \mathbb{R}^n | \, f(x) = \alpha\}$,
and the *α-sublevel set* of f is given by
$\hat{S}_\alpha = \{x \in \mathbb{R}^n | \, f(x) \leq \alpha\}$.

Fact: At any point $x_0 \in \mathbb{R}^n$, the gradient vector $\nabla f(x_0)$ is orthogonal to the tangent to the level set going through x_0. See page 70 of [CZ14] for a proof.



Level sets and gradient vectors of a function.



Zooming in on the same picture to see orthogonality.

The Hessian matrix

For a function $f: \mathbb{R}^n \to \mathbb{R}$ that is twice differentiable, the Hessian matrix is the $n \times n$ matrix of second derivatives:

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

Remarks:

- If f is twice continuously differentiable, the Hessian matrix is always a symmetric matrix. This is because partial derivatives commute:
  $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.
- The [CZ13] book uses the notation $D^2 f$ for the Hessian matrix.
- Second derivatives carry information about the "curvature" of the function f.

Practical rules for differentiation

The sum rule

If $f: \mathbb{R}^n \to \mathbb{R}^m$ and $g: \mathbb{R}^n \to \mathbb{R}^m$ then $J_{f+g}(x) = J_f(x) + J_g(x)$.

The product rule

Let $f: \mathbb{R}^n \to \mathbb{R}^m$ and $g: \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Define the function $h: \mathbb{R}^n \to \mathbb{R}$ by $h(x) = f(x)^T g(x)$. Then h is also differentiable and

$J_h(x) = f(x)^T J_g(x) + g(x)^T J_f(x)$
and
$\nabla h(x) = J_h(x)^T$.

The chain rule

Let $f: \mathbb{R} \to \mathbb{R}^n$ and $g: \mathbb{R}^n \to \mathbb{R}$. We suppose that g is differentiable on an open set $D \subset \mathbb{R}^n$ and that $f: (a, b) \to D$ is diffentiable on (a,b). Then the composite function $h: (a, b) \to \mathbb{R}$ given by $h(t) = g(f(t))$ is differentiable on (a,b) and:

$$h'(t) = \nabla g\big(f(t)\big)^T \begin{pmatrix} f_1'(t) \\ \vdots \\ f_n'(t) \end{pmatrix}.$$

A special case that comes up a lot

Let x and y be two fixed vectors in $\mathbb{R}^n$ and let $g: \mathbb{R}^n \to \mathbb{R}$. Define a univarite function $h(t) = g(x + ty)$.
Then $h'(t) = y^T \nabla g(x + ty)$.

Gradients and Hessians of affine and quadratic functions

- If $f(x) = c^T x + b$, then $\nabla f(x) = c$ and $\nabla^2 f(x) = 0_{n \times n}$.
- If $f(x) = x^T Q x$ and Q is symmetric, then $\nabla f(x) = 2Qx$ and $\nabla^2 f(x) = 2Q$.

Taylor expansion

Little o and Big O notation
(Fall '15: I skipped the Big O notation and anything that uses it)

These notions are used to compare the growth rate of two functions near the origin.

Definition

Let $g: \mathbb{R}^n \to \mathbb{R}$ be a function that does not vanish in a neighborhood around the origin, except possibly at the origin . Let $f: \mathbb{R}^n \to \mathbb{R}^m$ be defined in a domain $\Omega \subset \mathbb{R}^n$ that includes the origin. Then we write:

- $f(x) = O\big(g(x)\big)$ (pronounced "f is big Oh of g") to mean that the quotient $\|f(x)\|/|g(x)|$ is bounded near 0; that is there exists $K > 0$ and $\delta > 0$ such that if $\|x\| < \delta, x \in \Omega$ then $\frac{\|f(x)\|}{|g(x)|} \leq K$.

- $f(x) = o\big(g(x)\big)$(pronounced "f is little oh of g") if $\lim_{x \to 0,\, x \in \Omega} \frac{\|f(x)\|}{|g(x)|} = 0$.
  Intuitively, this means that f goes to zero faster than g.

Examples

$f(x) = O\big(g(x)\big)$

- $x = O(x)$ as $\frac{|x|}{|x|} = 1 \; \forall \, x \in \mathbb{R}$

- $x^2 = O\left(\frac{1}{2}x\right)$ (can take $K = 1, \delta = \frac{1}{2}$.)

- $\cos(x) = O(1)$ (why?)

- $x \neq O(x^2)$ (why?)

- $\sin(x) = O(x)$ (why?)

Little o and Big O notation

Examples (cont'd)

$f(x) = o(g(x))$

- $x^2 = o(x)$
- $\begin{pmatrix} x^3 \\ 2x^2 + 3x^4 \end{pmatrix} = o(x)$
- $x^3 = o(x^2)$
- $x = o(1)$

Remarks

- We gave the definition of little o and big O for comparing growth rates around $x = 0$. One can give similar definitions around any other point. In particular, in many areas of computing, these notations are used to compare growth rates of functions at infinity; i.e. as $x \to \infty$.
- If $f(x) = o(g(x))$ then $f(x) = O(g(x))$ but the converse is not necessarily true.

• $f(x) = o(g(x)) \Rightarrow f(x) = O(g(x))$

Proof: Little o implies $\lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0$.

$\Rightarrow \forall \epsilon > 0, \exists \delta > 0$ s.t. $\|x\| < \delta \Rightarrow \frac{\|f(x)\|}{|g(x)|} < \epsilon$.

i.e., the K in the definition of big O can be anything.

• $f(x) = O(g(x)) \not\Rightarrow f(x) = o(g(x))$

Take $f(x) = x$, $g(x) = 2x$. $\begin{cases} x = O(2x) & (K = 2, \delta \text{ anything}) \\ \lim_{x \to 0} \frac{|x|}{|2x|} = \frac{1}{2} \neq 0. \end{cases}$

Taylor expansion

Taylor expansion in one variable

- The idea behind Taylor expansion is to approximate a function around a given point by functions that are "simpler"; in this case by polynomials. As we increase the order of the Taylor expansion, we increase the degree of this polynomial and we reduce the error in our approximation.
- The little o and big O notation that we just introduced nicely capture how our error of approximation scales around the point we are approximating.

Here are two theorems we can state for functions of a single variable:

Version 1

Assume that a function $f: \mathbb{R} \to \mathbb{R}$ is in C^m, i.e., m times continuously differentiable (meaning that $f, f', f'', \ldots, f^{(m)}$ all exist and are continuous). Consider a point $a \in \mathbb{R}$ around which we will Taylor expand and define $h = b - a$. Then,

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^m}{m!} f^{(m)}(a) + o(h^m).$$

Version 2

Assume that a function $f: \mathbb{R} \to \mathbb{R}$ is in C^{m+1}. Consider a point $a \in \mathbb{R}$ around which we will Taylor expand and define $h = b - a$. Then,

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^m}{m!} f^{(m)}(a) + O(h^{m+1}).$$

Extension to multiple variable functions

When $f: \mathbb{R}^n \to \mathbb{R}$, we will only care about first and second order Taylor expansions in this class. Here, the concepts of a gradient vector and a Hessian matrix need to come in to replace first and second derivatives. We state four different variations of the theorem below. The point we are approximating the function around is denoted by $a \in \mathbb{R}^n$.

First order

If f is C^1:

$$f(x) = f(a) + \nabla f(a)^T (x - a) + o(\|x - a\|)$$

If f is C^2:

$$f(x) = f(a) + \nabla f(a)^T (x - a) + O(\|x - a\|^2)$$

Second order

If f is C^2:

$$f(x) = f(a) + \nabla f(a)^T (x - a) + \frac{1}{2}(x - a)^T \nabla^2 f(a)(x - a) + o(\|x - a\|^2)$$

If f is C^3:

$$f(x) = f(a) + \nabla f(a)^T (x - a) + \frac{1}{2}(x - a)^T \nabla^2 f(a)(x - a) + O(\|x - a\|^3)$$

Notes:

The material here was a summary of the relevant parts of [CZ13] collected in one place for your convenience.
The relevant sections for this lecture are chapters 2,3,5 and more specifically sections:
2.1
3.1, 3.2, 3.4
5.2, 5.3, 5.4, 5.5, 5.6.

I filled in some more detail in class, with some examples and proofs given here and there. Your HW 1 will give you some practice with this material.

References:

- [CZ13] E.K.P. Chong and S.H. Zak. An Introduction to Optimization. Fourth edition. Wiley, 2013.