Introduction to Optimization

Lecture 3: Continuous Optimization I

September 30, 2016

TC2 - Optimisation

Université Paris-Saclay

Anne Auger

Inria Saclay – Ile-de-France

Dimo Brockhoff

Inria Saclay – Ile-de-France

Course Overview

1	Fri, 16.9.2016	Introduction to Optimization
	Wed, 21.9.2016	groups defined via wiki
	Thu, 22.9.2016	everybody went (actively!) through the Getting Started part of github.com/numbbo/coco
2	Fri, 23.9.2016	Lecture: Benchmarking; final adjustments of groups everybody can run and postprocess the example experiment (~1h for final questions/help during the lecture)
3	Fri, 30.9.2016	Today's lecture: Continuous Optimization I
4	Fri, 7.10.2016	Lecture: Continuous Optimization II
	Mon, 10.10.2016	deadline for intermediate wiki report: what has been done and what remains to be done?
5	Fri, 14.10.2016	Lecture
6	Tue, 18.10.2016	Lecture
	Tue, 18.10.2016	deadline for submitting data sets
	Fri, 21.10.2016	deadline for paper submission
		vacation
7	Fri, 4.11.2016	Final lecture
	7.-11.11.2016	oral presentations (individual time slots)
	14 - 18.11.2016	Exam (exact date to be confirmed)

All deadlines:
23:59pm Paris time

Details on Continuous Optimization Lectures

Introduction to Continuous Optimization

- examples (from ML / black-box problems)
- typical difficulties in optimization

Mathematical Tools to Characterize Optima

- reminders about differentiability, gradient, Hessian matrix
- unconstraint optimization
  - first and second order conditions
  - convexity
- constraint optimization

Gradient-based Algorithms

- quasi-Newton method (BFGS)
- DFO trust-region method

Learning in Optimization / Stochastic Optimization

- CMA-ES (adaptive algorithms / Information Geometry)
- PhD thesis possible on this topic

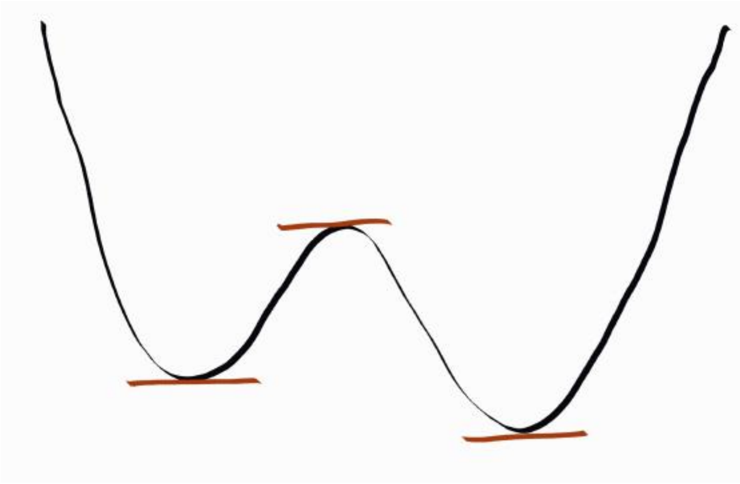
method strongly related to ML / new promising research area

interesting open questions

Mathematical Characterization of Optima

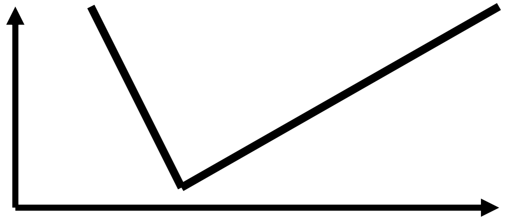
Objective: Derive general characterization of optima

Example: if $f: \mathbb{R} \to \mathbb{R}$ differentiable,
$f'(x) = 0$ at optimal points



- generalization to $f: \mathbb{R}^n \to \mathbb{R}$?
- generalization to constrained problems?

Remark: notion of optimum independent of notion of derivability



optima of such function can be easily approached by certain type of methods

A Few Reminders...

- $(E, ||\ \ ||)$ will be a K-general vector space endowed with a norm $||\ \ ||$ and a corpus K.
- If not familiar with this notion, think about $E = \mathbb{R}^n$, $\boldsymbol{x} \in \mathbb{R}^n$, $K = \mathbb{R}$, and $||\boldsymbol{x}|| = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$

Linear Mapping:

- $u: E \to E$ is a linear mapping if $u(\lambda x + \mu y) = \lambda u(x) + \mu u(y)$ for all $\lambda, \mu \in K$ and for all $x, y \in E$

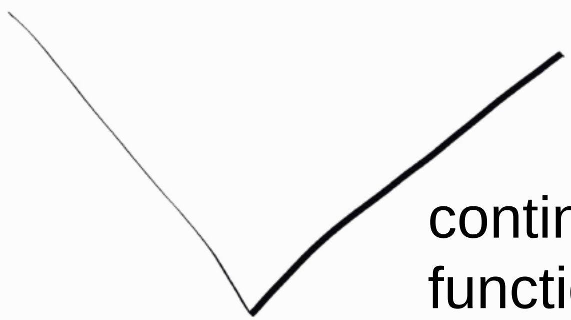
Exercise:

Let $E = \mathbb{R}^n$, $K = \mathbb{R}$ and $A \in \mathbb{R}^{n\times \mathrm{n}}$ be a matrix.
Show that $x \longmapsto Ax$ is a linear mapping.

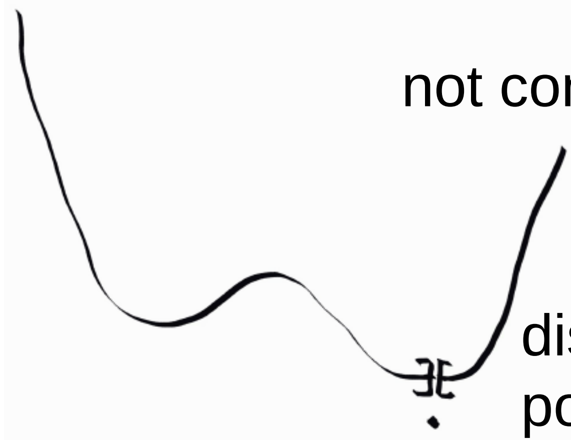
Continuity of a Function

$f: (E, ||\quad||) \longrightarrow (E, ||\quad||)$ is continuous in $x \in E$ if

$\forall \epsilon > 0, \exists \eta > 0$ such that $\forall y$: $||x - y|| \leq \eta; ||f(x) - f(y)|| \leq \epsilon$



continuous function



not continuous

discontinuity point

Scalar Product

$\langle\ ,\ \rangle: E \times E \to \mathbb{R}$ is a scalar product if it is

- a bilinear application
- symmetric (i.e. $\langle x, y \rangle = \langle y, x \rangle$
- positive (i.e. $\forall x \in E$: $\langle x, x \rangle \geq 0$)
- definite (i.e. $\langle x, x \rangle = 0 \Longrightarrow x = 0$)

Given a scalar product $\langle\ ,\ \rangle$, $||x|| = \sqrt{\langle x, x \rangle}$ is a norm.

(home exercise)

Example in $\mathbb{R}^n$: $\langle x, y \rangle = x^T y$

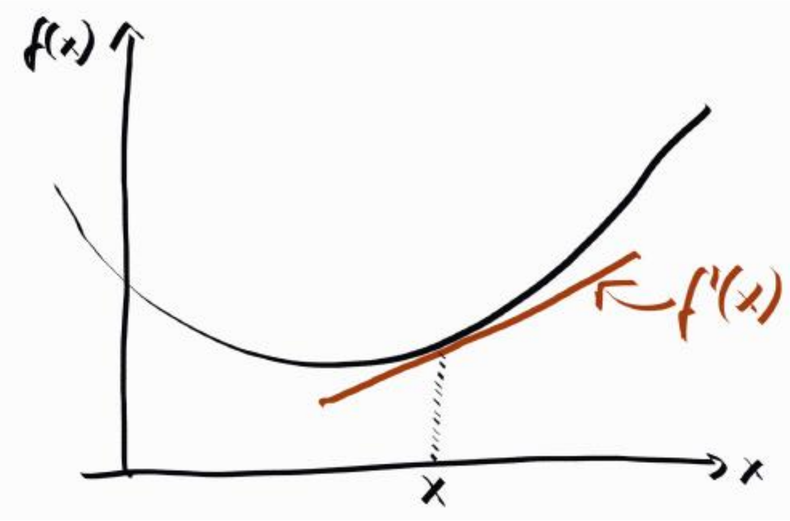
Reminder: Derivability in 1D (n=1)

$f: \mathbb{R} \to \mathbb{R}$ is derivable in $x \in \mathbb{R}$ if

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \text{ exists, } h \in \mathbb{R}$$

Notation:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



The derivative corresponds to the slope of the tangent in x.

Reminder: Derivability in 1D (n=1)

Taylor Formula (Order 1)

If f is derivable in x then

$$f(x+h) = f(x) + f'(x)h + o\big(||h||\big)$$

i.e. for h small enough, $h \longmapsto f(x+h)$ is approximated by $h \longmapsto f(x) + f'(h)$

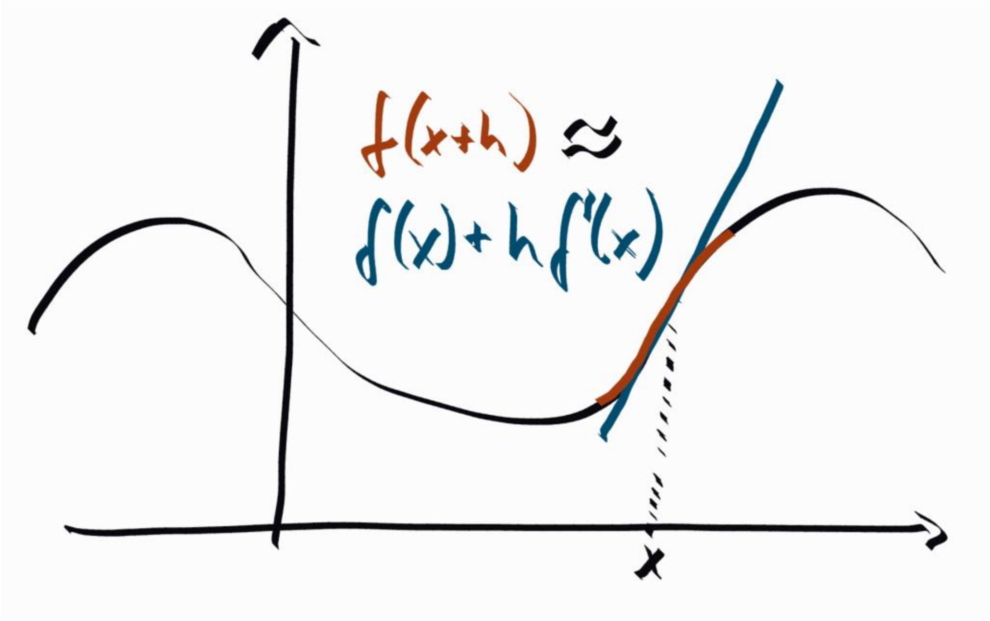
$h \longmapsto f(x) + f'(x)h$ is a linear approximation of $f(x+h)$

Exercise:
Why is it linear?

$h \longmapsto f(x) + f'(x)h$ is a first order approximation of $f(x+h)$

Reminder: Derivability in 1D (n=1)

Geometrically:



The notion of derivative of a function defined on $\mathbb{R}^n$ is generalized via this idea of a linear approximation of $f(x+h)$ for h small enough.

Differentiability: Generalization from 1D

Given a normed vector space $(E, ||.||)$ and complete (Banach space), consider $f: U \subset E \to \mathbb{R}$ with U open set of E.

- f is differentiable in $\boldsymbol{x} \in U$ if there exists a continuous linear mapping $Df(\boldsymbol{x})$ such that

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})(\boldsymbol{h}) + o(||\boldsymbol{h}||)$$

$Df(\boldsymbol{x})$ is the differential of f in $\boldsymbol{x}$

Exercise:

Consider $E = \mathbb{R}^n$ with the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$. Let $\boldsymbol{a} \in \mathbb{R}^n$, show that

$$f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle$$

is differentiable and compute its differential.

Gradient

If the norm $||.||$ comes from a scalar product, i.e. $||\boldsymbol{x}|| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ (the Banach space E is then called a Hilbert space), the gradient of f in $\boldsymbol{x}$ denoted $\nabla f(\boldsymbol{x})$ is defined as the element of E such that

$$Df(\boldsymbol{x})(\boldsymbol{h}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{h} \rangle$$

Riesz representation Theorem

Taylor formula – order one

Replacing the differential in the last slide by the above, we obtain the Taylor formula:

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{h} \rangle + o(||\boldsymbol{h}||)$$

Exercise: Gradients

Exercise:

Compute the gradient of the functions

- $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle$.
- $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(\boldsymbol{x}_n), \theta \rangle)^2$.

Gradient: Connection to Partial Derivatives

- In $(\mathbb{R}^n, ||\ \ ||_2)$ where $||\boldsymbol{x}||_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ is the Euclidean norm deriving from the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$

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

- Reminder: partial derivative in x_0

$$y \underset{f_i}{\rightarrow} f\left(x_0^1, \dots, x_0^{i-1}, y, x_0^{i+1}, \dots, x_0^n\right)$$

$$\frac{\partial \mathrm{f}}{\partial x_i}(x_0) = f_i{}'(x_0)$$

Gradient: More Examples

- if $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle, \nabla f(x) = \boldsymbol{a}$
- in $\mathbb{R}^n$, if $f(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$, then $\nabla f(\boldsymbol{x}) = (A + A^T)\boldsymbol{x}$
- particular case if $f(\boldsymbol{x}) = ||\boldsymbol{x}||^2$, then $\nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$
- in $\mathbb{R}$, $\nabla f(\boldsymbol{x}) = f'(\boldsymbol{x})$

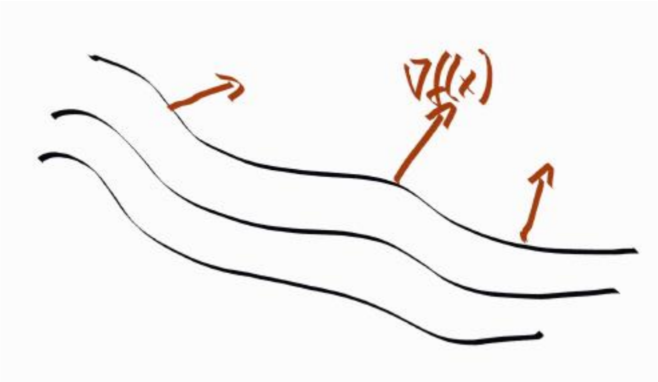
Gradient: Geometrical Interpretation

Exercise:

Let $L_c = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = c\}$ be again a level set of a function $f(\boldsymbol{x})$. Let $\boldsymbol{x}_0 \in L_c \neq \emptyset$.

Show for $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle$ and $f(\boldsymbol{x}) = ||\boldsymbol{x}||^2$ that $\nabla f(\boldsymbol{x_0})$ is ***orthogonal*** to the level sets in x_0.

More generally, the gradient of a differentiable function is orthogonal to its level sets.

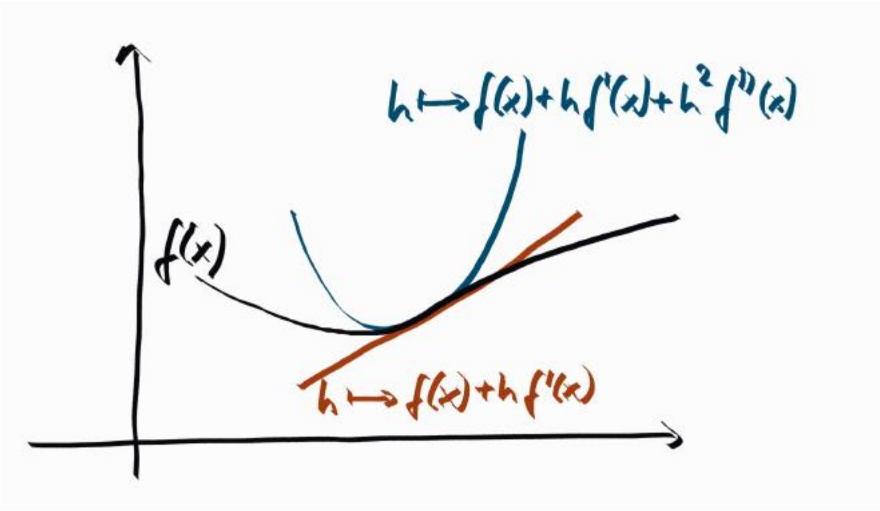


Reminder: Second Order Derivability in 1D

- Let $f: \mathbb{R} \to \mathbb{R}$ be a derivable function and let $f': x \to f'(x)$ be its derivative function.
- If f' is derivable in x, then we denote its derivative as $f''(x)$
- $f''(x)$ is called the *second order derivative* of f.

Taylor Formula: Second Order Derivative

- If $f: \mathbb{R} \to \mathbb{R}$ is two times derivable then
$$f(x+h) = f(x) + f'(x)h + f''(x)h^2 + o(||h||^2)$$
i.e. for h small enough, $h \to f(x) + hf'(x) + h^2 f''(x)$ approximates $h + f(x+h)$
- $h \to f(x) + hf'(x) + h^2 f''(x)$ is a quadratic approximation (or order 2) of f in a neighborhood of x



- The second derivative of $f: \mathbb{R} \to \mathbb{R}$ generalizes naturally to larger dimension.

Second Order Differentiability

- (first order) differential: gives a linear local approximation
- second order differential: gives a quadratic local approximation

Definition: second order differentiability

$f: U \subset E \to \mathbb{R}$ is differentiable at the second order in $\boldsymbol{x} \in U$ if it is differentiable in a neighborhood of $\boldsymbol{x}$ and if $u \longmapsto Df(u)$ is differentiable in $\boldsymbol{x}$

Second Order Differentiability (Cont.)

Another Definition:

$f: U \subset E \to \mathbb{R}$ is differentiable at the second order in $\boldsymbol{x} \in U$ iff there exists a continuous linear application $Df(\boldsymbol{x})$ and a bilinear symmetric continuous application $D^2 f(\boldsymbol{x})$ such that

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})(\boldsymbol{h}) + \frac{1}{2} D^2 f(\boldsymbol{x})(\boldsymbol{h},\boldsymbol{h}) + o(||\boldsymbol{h}||^2)$$

In a Hilbert space $(E, \langle \quad \rangle)$

$$D^2 f(\boldsymbol{x})(\boldsymbol{h},\boldsymbol{h}) = \langle \nabla^2 f(\boldsymbol{x})(\boldsymbol{h}), \boldsymbol{h} \rangle$$

where $\nabla^2 f(\boldsymbol{x})$: $E \to E$ is a symmetric continuous operator.

Hessian Matrix

In $(\mathbb{R}^n, \langle x, y \rangle = x^T y)$, $\nabla^2 f(x)$ is represented by a symmetric matrix called the Hessian matrix. It can be computed as

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

Exercise on Hessian Matrix

Exercise:

Let $f(\boldsymbol{x}) = \boldsymbol{x}^T A\, \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$.

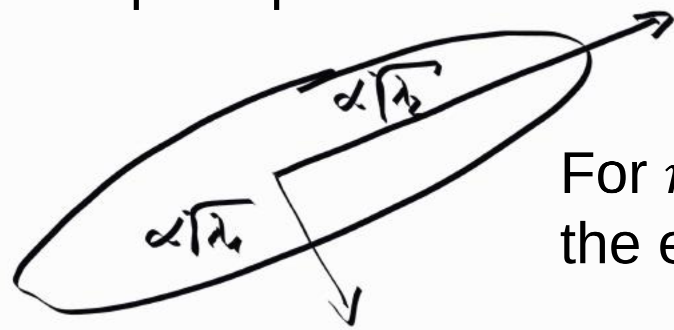
Compute the Hessian matrix of f.

If it is too complex, consider $f: \begin{cases} \mathbb{R}^2 \to \mathbb{R} \\ \boldsymbol{x} \to \boldsymbol{x}^T A\, \boldsymbol{x} \end{cases}$ with $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$

Back to Ill-Conditioned Problems

We have seen that for a convex quadratic function
$f(x) = \frac{1}{2}(x - x_0)^T A(x - x_0) + b$ of $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, A SPD, $b \in \mathbb{R}^n$:

1) The level sets are ellipsoids. The eigenvalues of A determine the lengths of the principle axes of the ellipsoid.



For $n = 2$, let λ_1, λ_2 be the eigenvalues of A.

2) The Hessian matrix of f equals to A.

Ill-conditioned convex quadratic problems are problems with large ratio between largest and smallest eigenvalue of A which means large ratio between longest and shortest axis of ellipsoid.

This corresponds to having an ill-conditioned Hessian matrix.

Optimality Conditions: First Order Necessary Cond.

For 1-dimensional optimization problems $\boldsymbol{f}$: $\mathbb{R} \to \mathbb{R}$

Assume f is derivable

- $\boldsymbol{x}^*$ is a local extremum $\Longrightarrow f'(\boldsymbol{x}^*) = 0$

not a sufficient condition: consider $f(\boldsymbol{x}) = \boldsymbol{x}^3$

proof via Taylor formula: $f(\boldsymbol{x}^* + \boldsymbol{h}) = f(\boldsymbol{x}^*) + f'(\boldsymbol{x}^*)h + o(||\boldsymbol{h}||)$

- points $\boldsymbol{y}$ such that $f'(\boldsymbol{y}) = 0$ are called critical or stationary points

Generalization to $\boldsymbol{n}$-dimensional functions

If $f: U \subset \mathbb{R}^n \longmapsto \mathbb{R}$ is differentiable

- necessary condition: If $\boldsymbol{x}^*$ is a local extremum of f, then $Df(\boldsymbol{x}^*) = 0$ and hence $\nabla f(\boldsymbol{x}^*) = 0$

proof via Taylor formula

Second Order Necessary and Sufficient Opt. Cond.

If f is twice continuously differentiable

- Necessary condition: if $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive semi-definite

proof via Taylor formula at order 2

- Sufficient condition: if $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, then $\boldsymbol{x}^*$ is a strict local minimum

Proof for sufficient condition:

- Let $\lambda > 0$ be the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x}^*)$, using a second order Taylor expansion, we have for all $\boldsymbol{h}$:
- $f(\boldsymbol{x}^* + \boldsymbol{h}) - f(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{h} + o(||\boldsymbol{h}||^2)$

$$> \frac{\lambda}{2} ||\boldsymbol{h}||^2 + \mathrm{o}(||\boldsymbol{h}||^2) = \left(\frac{\lambda}{2} + \frac{o(||\boldsymbol{h}||^2)}{||\boldsymbol{h}||^2} \right) ||\boldsymbol{h}||^2$$

Convex Functions

Let U be a convex open of $\mathbb{R}^n$ and $f: U \to \mathbb{R}$. The function f is said to be convex if for all $\boldsymbol{x}, \boldsymbol{y} \in U$ and for all $t \in [0,1]$

$$f\big((1-t)\boldsymbol{x} + t\boldsymbol{y}\big) \leq (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y})$$

Theorem

If f is differentiable, then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y}$

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq Df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

if $n = 1$, the curve is on top of the tangent

If f is twice continuously differentiable, then f is convex if and only if $D^2 f$ satisfies $\forall \boldsymbol{x} \in U, \boldsymbol{h} \in \mathbb{R}^n: D^2 f(\boldsymbol{x})(\boldsymbol{h}, \boldsymbol{h}) \geq 0$ (or $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for all x)

Convex Functions: Why Convexity?

Examples:

- $f(\boldsymbol{x}) = \langle a, \boldsymbol{x} \rangle + b$
- $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{x}, A\boldsymbol{x} \rangle + \langle a, \boldsymbol{x} \rangle + b$, A positive definite symmetric

the opposite of the entropy function: $f(\boldsymbol{x}) = -\sum_{i=1}^{n} \boldsymbol{x}_i \ln(\boldsymbol{x}_i)$ (the entropy function is then concave)

Exercise:

Let $f: U \to \mathbb{R}$ be a convex and differentiable function on a convex open U.
Show that if $Df(\boldsymbol{x}^*) = 0$, then $\boldsymbol{x}^*$ is a global minimum of f

Why convexity? local minima are also global under convexity assumption.