Advanced Optimization
Lecture 2: Continuous Optimization I

October 20, 2021

CentraleSupélec / ESSEC Business School

dimo.brockhoff@inria.fr

Dimo Brockhoff

Inria Saclay – Ile-de-France

Course Overview

		Topic
Wed, 13.10.2021	PM	Introduction, examples of problems, problem types
Wed, 20.10.2021	PM	Continuous (unconstrained) optimization: convexity, gradients, Hessian, … [technical test Evalmee]
Wed, 27.10.2021	PM	Continous optimization II: gradient descent, Newton direction, quasi-Newton (BFGS) [1st mini-exam] Linear programming: duality, maxflow/mincut, simplex algo
Wed, 03.11.2021	PM	Constrained optimization: Lagrangian, optimality conditions
Wed, 10.11.2021	PM	Gradient-based and derivative-free stochastic algorithms: SGD and CMA-ES
Wed, 17.11.2021	PM	Other blackbox optimizers: Nelder-Mead, Bayesian optimization [2nd mini-exam]
Wed, 24.11.2021	PM	Benchmarking solvers: runtime distributions, performance profiles
Tue, 30.11.2021	23:59	Deadline open source project (PDF sent by email)
Wed, 01.12.2021	PM	Discrete optimization: branch and bound, branch and cut, k-means clustering
Wed, **15.**12.2021	PM	Exam

Details about the Group Project Grading

What is **not** graded?

- Whether the contribution to the project is actually accepted into the production code/master branch/…
- Your contributions when writing issues, interacting with others developers or users etc.
  - Just because I cannot check everything
  - But please interact with people (and also mention this in the report if you feel it is relevant; in this case it is graded ☺)

What **is** graded?

  - Report: readability, structure, clearness, …
  - Contribution itself: difficulty, amount/scale, …

I will try to be as fair as possible by grading all groups relatively to each other

Organization of the Groups

https://docs.google.com/spreadsheets/d/1WV8yfI1T0rYqtdoPYzOu7ORVx9qKC1kwvOSFE6MVaX4/edit?usp=sharing

back to lecture

Details on Continuous Optimization Lectures

Introduction to Continuous Optimization

- examples and 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
  - linear programming, dual problem
  - Lagrangian, optimality conditions

Gradient-based Algorithms

- stochastic gradient
- quasi-Newton method (BFGS)

Learning in Optimization / Optimization in Machine Learning

- Stochastic gradient descent (SGD) + Adam
- CMA-ES (adaptive algorithms / Information Geometry)
- Other derivative-free algorithms: Nelder-Mead, Bayesian opt.

- Optimize f: $\begin{cases} \Omega \subset \mathbb{R}^n \to \mathbb{R} \\ x = (x_1, \ldots, x_n) \to f(x_1, \ldots, x_n) \end{cases}$

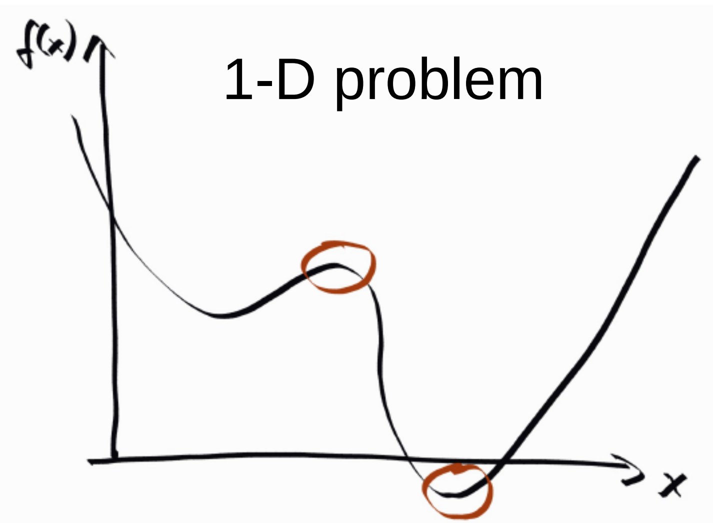
  $\in \mathbb{R}$

  unconstrained optimization

- Search space is continuous, i.e. composed of real vectors $x \in \mathbb{R}^n$

- $n = \begin{cases} \text{dimension of the problem} \\ \text{dimension of the search space } \mathbb{R}^n \text{ (as vector space)} \end{cases}$

1-D problem

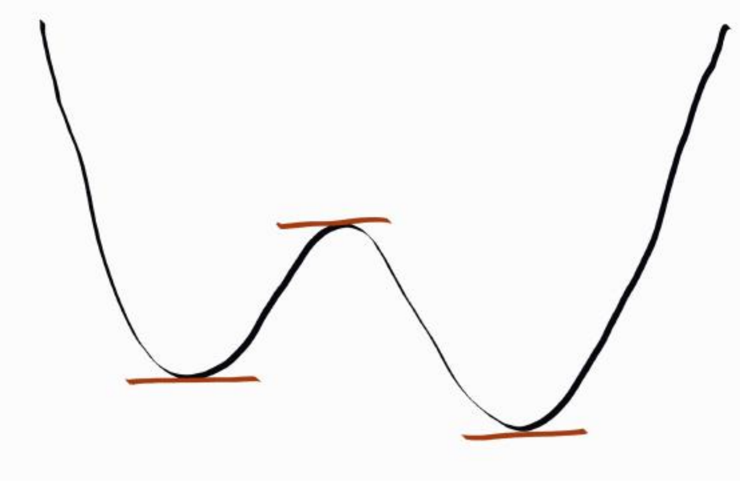


2-D level sets

Goal: Mathematical Characterization of Optima

Objective: Derive general characterization of optima

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



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

Reminder: Gradient Definition

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 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$

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

Exercise: Gradients

Exercise:

Compute the gradients of

a) $f(x) = x_1$ with $x \in \mathbb{R}^n$

b) $f(x) = a^T x$ with $\mathrm{a}, x \in \mathbb{R}^n$

c) $f(x) = x^T x \; (= ||\mathrm{x}||^2)$ with $x \in \mathbb{R}^n$

Link to the OneNote page with the solutions

Exercise: Gradients

Exercise:

Compute the gradients of

a) $f(x) = x_1$ with $x \in \mathbb{R}^n$
b) $f(x) = a^T x$ with $\mathrm{a}, x \in \mathbb{R}^n$
c) $f(x) = x^T x$ $(= ||\mathrm{x}||^2)$ with $x \in \mathbb{R}^n$

Some more examples:

- in $\mathbb{R}^n$, if $f(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$, then $\nabla f(\boldsymbol{x}) = (A + A^T)\boldsymbol{x}$
- in $\mathbb{R}$, $\nabla f(\boldsymbol{x}) = f'(\boldsymbol{x})$

Gradient: Geometrical Interpretation

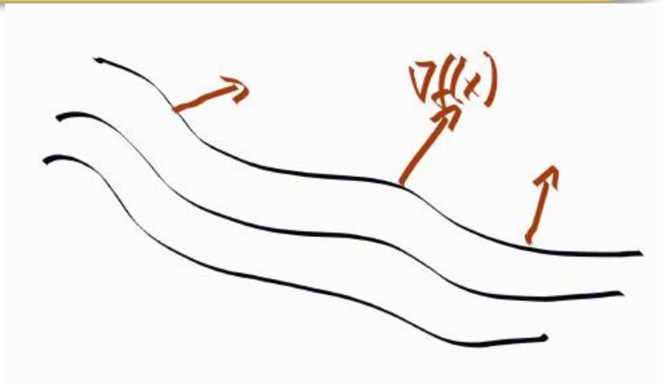
Exercise:

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 $\boldsymbol{x}_0 \in L_c \neq \emptyset$.

Compute the level sets for $f_1(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x}$ and $f_2(\boldsymbol{x}) = ||\boldsymbol{x}||^2$ and the gradient in a chosen point x_0 and observe that $\nabla f(\boldsymbol{x_0})$ is ***orthogonal*** to the level set in x_0.

If this seems too difficult, do it for two variables (and a concrete $\boldsymbol{a} \in \mathbb{R}^2$) and draw the level sets and the gradients.

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



Taylor Formula – Order One

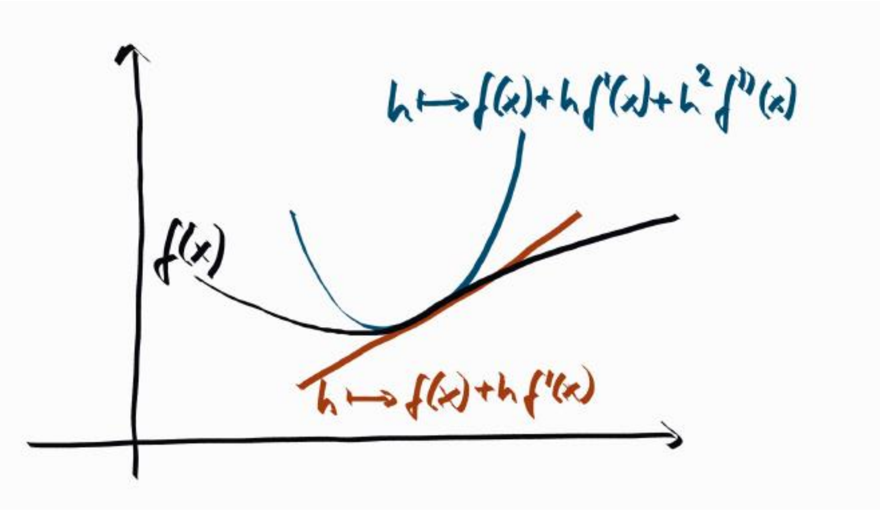
$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \left(\nabla f(\boldsymbol{x})\right)^T \boldsymbol{h} + o(||\boldsymbol{h}||)$$

Reminder: Second Order Derivability in 1D

- Let $f: \mathbb{R} \to \mathbb{R}$ be a differentiable function and let $f': x \to f'(x)$ be its derivative.
- If f' is differentiable 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 differentiable 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 $h \to f(x+h)$
- $h \to f(x) + hf'(x) + h^2 f''(x)$ 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.

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 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\, \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

Compute the Hessian matrix of f.

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

[Link to the OneNote page with the solutions]

Taylor Formula – Order Two

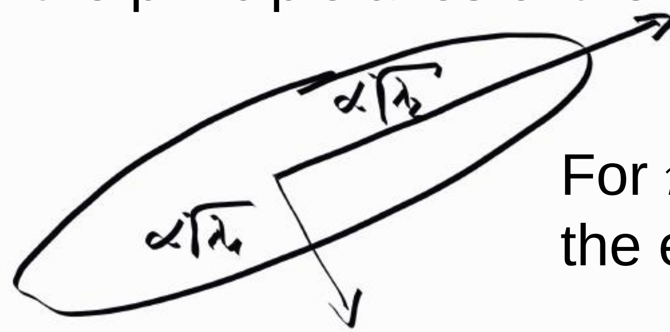
$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \left(\nabla f(\boldsymbol{x})\right)^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T\left(\nabla^2 f(\boldsymbol{x})\right)\boldsymbol{h} + o(||\boldsymbol{h}||^2)$$

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.

Organization of the Groups

https://docs.google.com/spreadsheets/d/1WV8yfI1T0rYqtdoPYzOu7ORVx9qKC1kwvOSFE6MVaX4/edit?usp=sharing

Test for Mini-Exam
in Evalmee

Gradient Direction Vs. Newton Direction

Gradient direction: $\nabla f(\boldsymbol{x})$

Newton direction: $\left(H(\boldsymbol{x})\right)^{-1} \cdot \nabla f(\boldsymbol{x})$

with $H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ being the Hessian at $\boldsymbol{x}$

Exercise:

Let again $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\, \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^2$, $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

❶ Plot the level sets of f(x).

❷ Compute the gradient and Newton direction of f in a point $x \in \mathbb{R}^n$ of your choice (which should not be on a coordinate axis) and plot them into the same plot with the level sets.

Link to the OneNote page with the solutions

Gradient Direction Vs. Newton Direction

Gradient direction: $\nabla f(\boldsymbol{x})$

Newton direction: $\left(H(\boldsymbol{x})\right)^{-1} \cdot \nabla f(\boldsymbol{x})$

with $H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ being the Hessian at $\boldsymbol{x}$

Exercise:

Let again $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\, \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^2$, $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

❶ Plot the level sets of f(x).
❷ Compute the gradient and Newton direction of f in a point $x \in \mathbb{R}^n$ of your choice (which should not be on a coordinate axis) and plot them into the same plot with the level sets.

- remind level sets: axis-parallel ellipsoids, axis-ratio=3
- remind gradient: $A\boldsymbol{x}$
- remind Hessian: A