Introduction to Optimization

Lecture 4: Continuous Optimization

October 9, 2015

TC2 - Optimisation

Université Paris-Saclay

Anne Auger

INRIA Saclay – Ile-de-France

Dimo Brockhoff

INRIA Lille – Nord Europe

Course Overview

Date		Topic
Fri, 18.9.2015	DB	Introduction and Greedy Algorithms
Fri, 25.9.2015	DB	Dynamic programming and Branch and Bound
Fri, 2.10.2015	DB	Approximation Algorithms and Heuristics
Fri, 9.10.2015	AA	Introduction to Continuous Optimization
Fri, 16.10.2015	AA	End of Intro to Cont. Opt. + Gradient-Based Algorithms I
Fri, 30.10.2015	AA	Gradient-Based Algorithms II
Fri, 6.11.2015	AA	Stochastic Algorithms and Derivative-free Optimization
16 - 20.11.2015		Exam (exact date to be confirmed)

all classes + exam are from 14h till 17h15 (incl. a 15min break)
here in PUIO-D101/D103

Further Details on Remaining 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

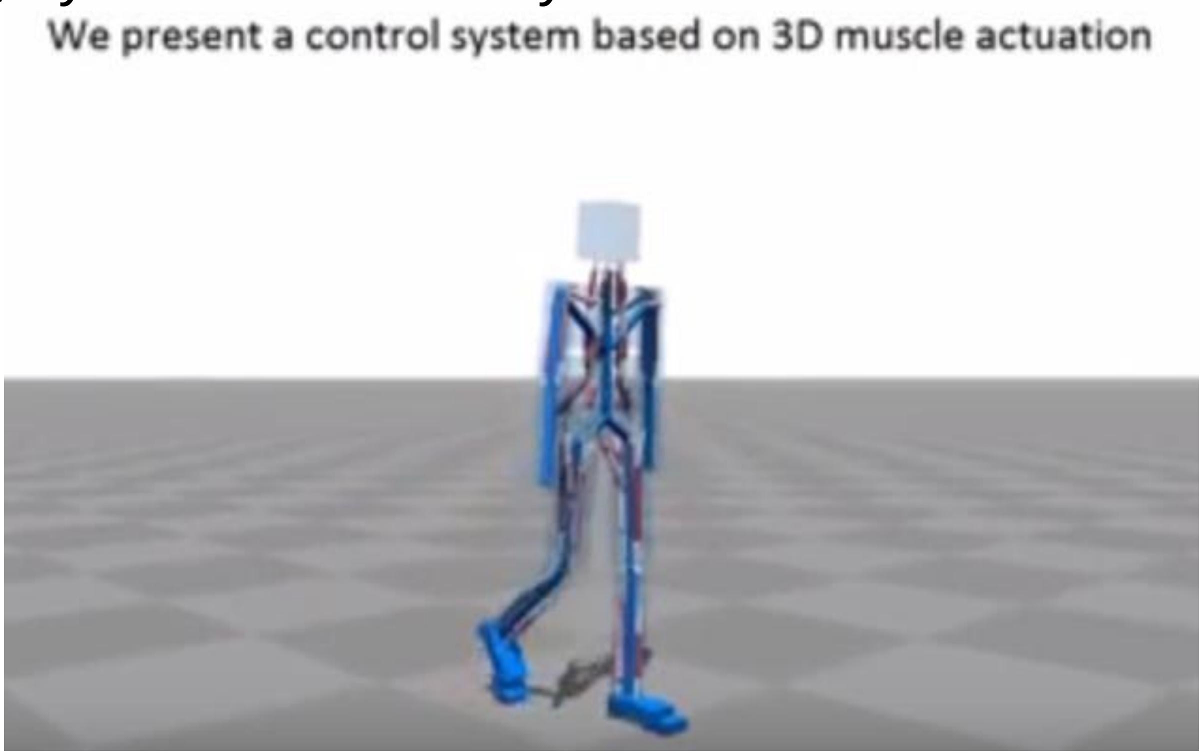
Learning in Optimization / Stochastic Optimization

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

method strongly related to ML / new promising research area
interesting open questions

First Example of a Continuous Optimization Problem

Computer simulation teaches itself to walk upright (virtual robots (of different shapes) learning to walk, through stochastic optimization (CMA-ES)), by Utrecht University:



https://www.youtube.com/watch?v=yci5FuI1ovk

T. Geitjtenbeek, M. Van de Panne, F. Van der Stappen: "Flexible Muscle-Based Locomotion for Bipedal Creatures", SIGGRAPH Asia, 2013.

Continuous Optimization

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

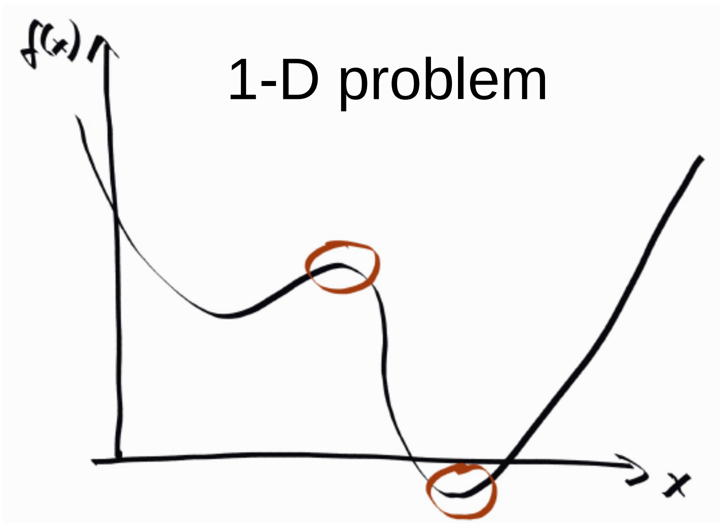
  $x_1 \in \mathbb{R}$

  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

Unconstrained vs. Constrained Optimization

Unconstrained optimization

$$\inf \{f(x) \mid x \in \mathbb{R}^n\}$$

Constrained optimization

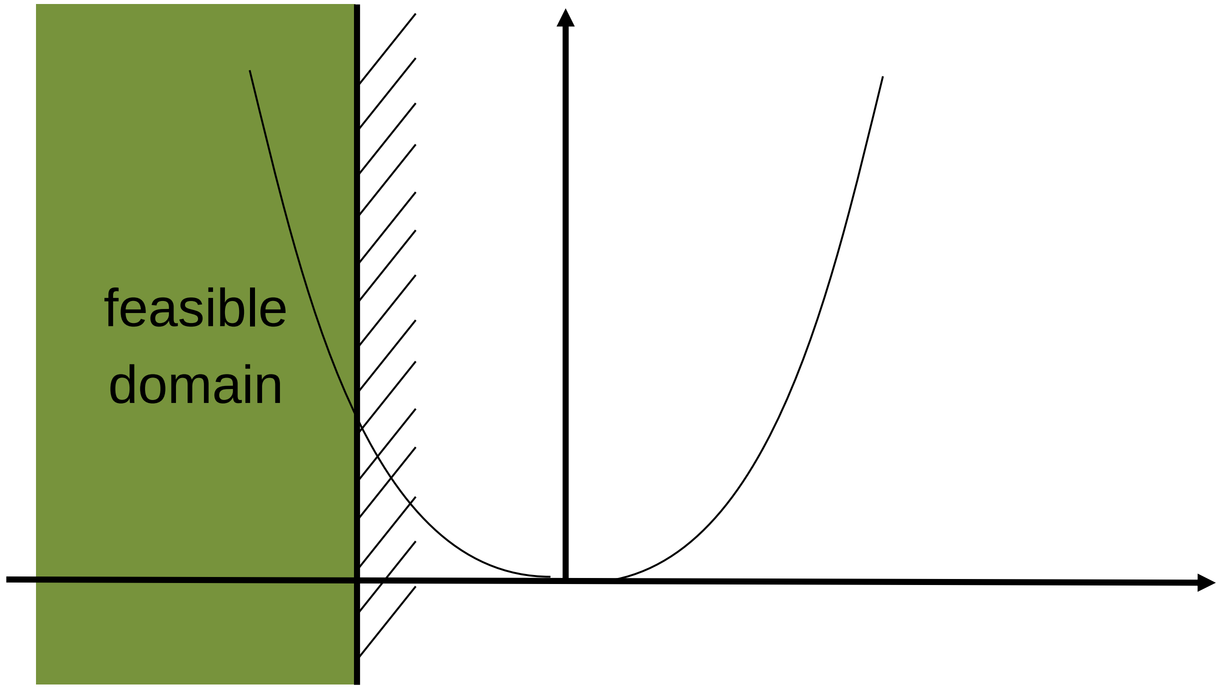
- Equality constraints: $\inf \{f(x) \mid x \in \mathbb{R}^n, g_k(x) = 0, 1 \leq k \leq p\}$

- Inequality constraints: $\inf \{f(x) \mid x \in \mathbb{R}^n, g_k(x) \leq 0, 1 \leq k \leq p\}$

where always $g_k\colon \mathbb{R}^n \to \mathbb{R}$

Example of a Constraint

$$\min_{x \in \mathbb{R}} f(x) = x^2 \text{ such that } x \leq -1$$



Analytical Functions

Example: 1-D

$$f_1(x) = a(x - x_0)^2 + b$$

where $x, x_0, b \in \mathbb{R}, a \in \mathbb{R}$

Generalization:

convex quadratic function

$$f_2(x) = (x - x_0)^T A \, (x - x_0) + b$$

where $x, x_0, b \in \mathbb{R}^n, A \in \mathbb{R}^{\{\mathrm{n} \times n\}}$

and A symmetric positive definite (SPD)

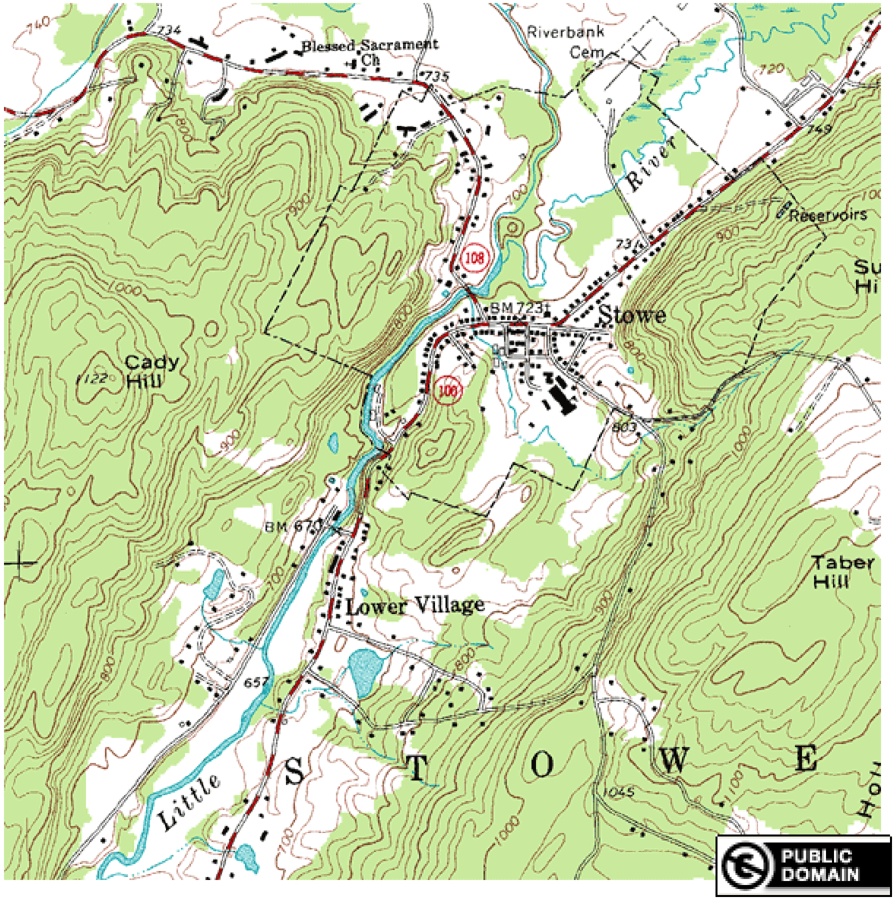
Exercise:
What is the minimum of $f_2(x)$?

Continuation of exercise:
What are the level sets of f_2?

Reminder: level sets of a function

$$L_c = \{x \in \mathbb{R}^n \mid f(x) = c\}$$

(similar to topography lines / level sets on a map)



Levels Sets of Convex Quadratic Functions

Continuation of exercise:
What are the level sets of f_2?

- Probably too complicated in general, thus an example here
- Consider $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}, b = 0, n = 2$
  a) Compute $f_2(x)$.
  b) Plot the level sets of $f_2(x)$.
  c) More generally, for $n = 2$, if A is SPD with eigenvalues $\lambda_1 = 9$ and $\lambda_2 = 1$, what are the level sets of $f_2(x)$?

Data Fitting – Data Calibration

Objective

- Given a sequence of data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, N$, find a model "$y = f(\boldsymbol{x})$" that explains the data

  experimental measurements in biology, chemistry, ...

- In general, choice of a parametric model or family of functions $(f_\theta)_{\theta \in \mathbb{R}^n}$

  use of expertise for choosing model or simple models only affordable (linear, quadratic)

- Try to find the parameter $\theta \in \mathbb{R}^n$ fitting best to the data

Fitting best to the data

Minimize the quadratic error:

$$\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{N} |f_\theta(\boldsymbol{x}_i) - y_i|^2$$

Supervised Learning:

Predict $y \in \mathcal{Y}$ from $\boldsymbol{x} \in \mathcal{X}$, given a set of observations (examples) $\{y_i, \boldsymbol{x}_i\}_{i=1,\ldots,N}$

(Simple) Linear regression

Given a set of data: $\{y_i, \underbrace{x_i^1, \ldots, x_i^p}_{\boldsymbol{x}_i^T}\}_{i=1\ldots N}$

$$\min_{\boldsymbol{w}\in\mathbb{R}^p, \beta\in\mathbb{R}} \underbrace{\sum_{i=1}^{N} |\, \boldsymbol{w}^T \boldsymbol{x}_i + \beta - y_i \big|^2}_{||\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}} - \mathbf{y}||^2}$$

$\widetilde{\boldsymbol{X}} \in \mathbb{R}^{N\times(p+1)}$, $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{p+1}$

same as data fitting with linear model, i.e. $f_{(\boldsymbol{w},\beta)}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \beta$, $\theta \in \mathbb{R}^{p+1}$

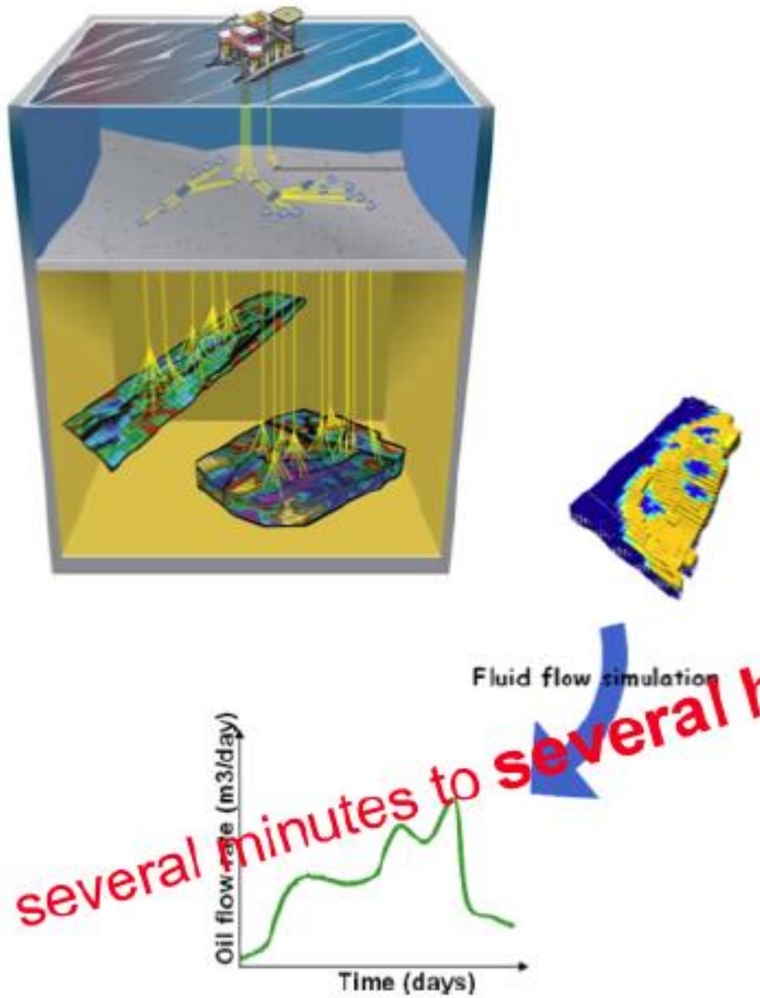
Regression

- *Data:* N observations $\{y_i, x_i\} \in \mathbb{R} \times \mathcal{X}$
- $\Phi(x_i) \in \mathbb{R}^p$ features of x_i
- prediction as a linear function of the feature $\hat{y} = \langle \theta, \Phi(x) \rangle$
- *empirical risk minimization:* find $\hat{\theta}$ solution of

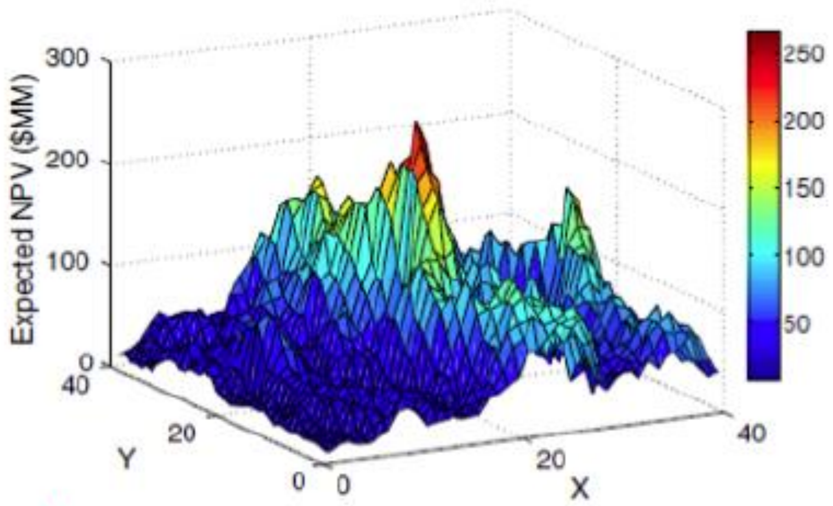
$$\min_{\theta \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} I(y_i, \langle \theta, \Phi(x_i) \rangle)$$

where I is a loss function [example: quadratic loss $I(y, \hat{y}) = 1/2(y - \hat{y})^2$]

Well Placement Problem



several minutes to **several hours** !!



Onwunalu & Durlofsky (2010)

What Makes a Function Difficult to Solve?

- dimensionality

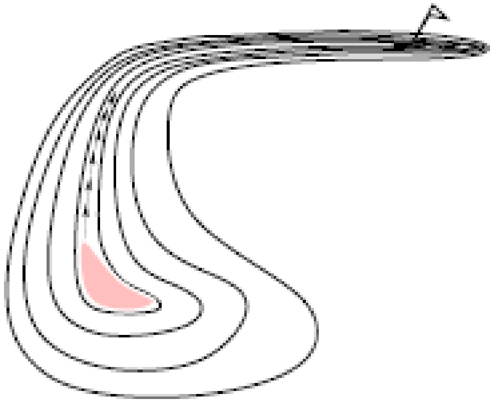
  (considerably) larger than three

- non-separability

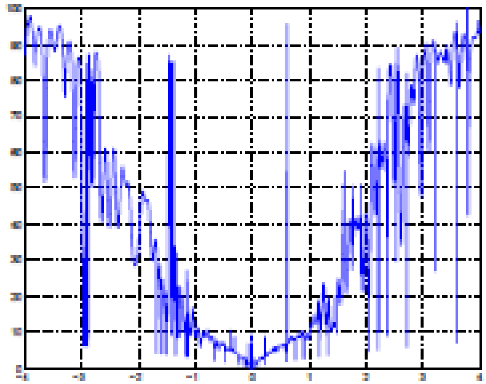
  dependencies between the objective variables

- ill-conditioning
- ruggedness

  non-smooth, discontinuous, multimodal, and/or noisy function



a narrow ridge



cut from 3D example, solvable with an evolution strategy

Curse of Dimensionality

- The term *Curse of dimensionality* (Richard Bellman) refers to problems caused by the rapid increase in volume associated with adding extra dimensions to a (mathematical) space.

- Example: Consider placing 100 points onto a real interval, say $[0,1]$. To get similar coverage, in terms of distance between adjacent points, of the 10-dimensional space $[0,1]^{10}$ would require $100^{10} = 10^{20}$ points. The original 100 points appear now as isolated points in a vast empty space.

- Consequently, a search policy (e.g. exhaustive search) that is valuable in small dimensions might be useless in moderate or large dimensional search spaces.

Separable Problems

Definition (Separable Problem)

A function f is separable if

$$\underset{(x_1,\dots,x_n)}{\operatorname{argmin}} f(x_1,\dots,x_n) = \left(\underset{x_1}{\operatorname{argmin}} f(x_1,\dots),\dots,\underset{x_n}{\operatorname{argmin}} f(\dots,x_n)\right)$$

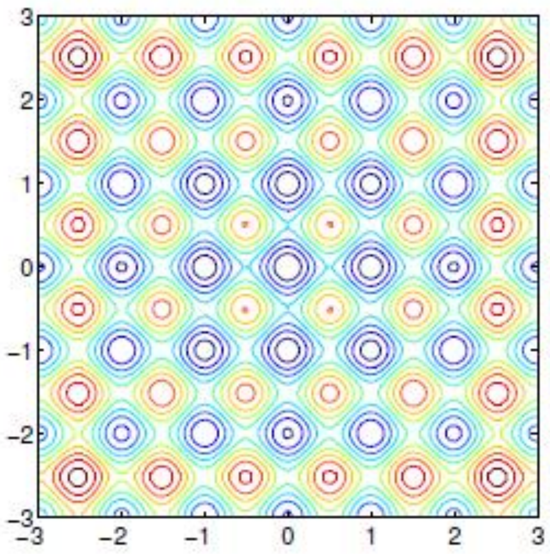
$\Longrightarrow$ it follows that f can be optimized in a sequence of n independent 1-D optimization processes

Example:

Additively decomposable functions

$$f(x_1,\dots,x_n) = \sum_{i=1}^{n} f_i(x_i)$$

Rastrigin function



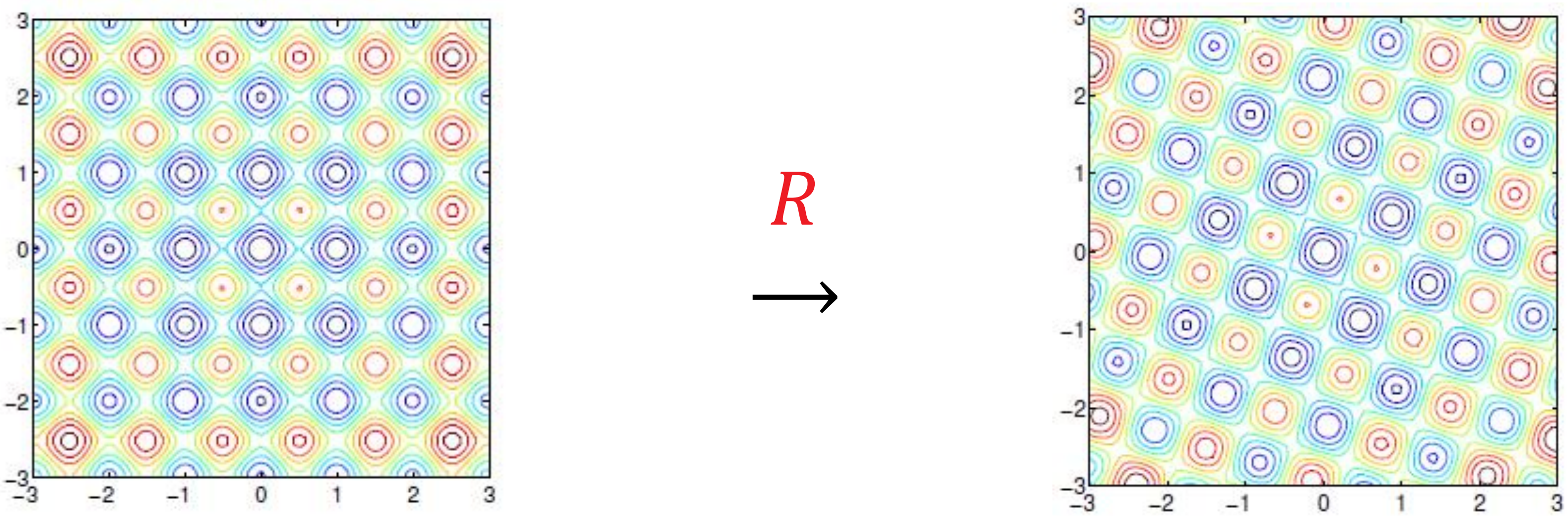
Non-Separable Problems

Building a non-separable problem from a separable one [1,2]

Rotating the coordinate system

- $f: \boldsymbol{x} \mapsto f(\boldsymbol{x})$ separable
- $f: \boldsymbol{x} \mapsto f(R\boldsymbol{x})$ non-separable

R rotation matrix



[1] N. Hansen, A. Ostermeier, A. Gawelczyk (1995). "On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation". Sixth ICGA, pp. 57-64, Morgan Kaufmann

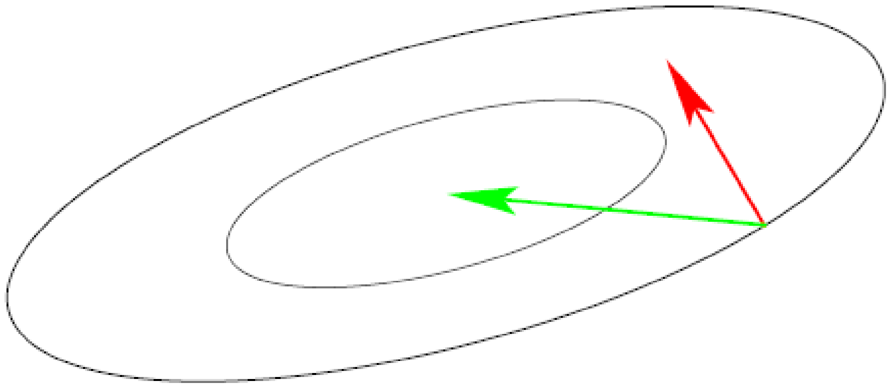
[2] R. Salomon (1996). "Reevaluating Genetic Algorithm Performance under Coordinate Rotation of Benchmark Functions; A survey of some theoretical and practical aspects of genetic algorithms." BioSystems, 39(3):263-278

Ill-Conditioned Problems: Curvature of Level Sets

Consider the convex-quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T H(\boldsymbol{x} - \boldsymbol{x}^*) = \frac{1}{2}\sum_i h_{i,i} x_i^2 + \frac{1}{2}\sum_{i,j} h_{i,j} x_i x_j$$

H is Hessian matrix of f and symmetric positive definite



gradient direction $-f'(x)^T$

Newton direction $-H^{-1} f'(x)^T$

Ill-conditioning means squeezed level sets (high curvature). Condition number equals nine here. Condition numbers up to 10^{10} are not unusual in real-world problems.

If $H \approx I$ (small condition number of H) first order information (e.g. the gradient) is sufficient. Otherwise second order information (estimation of H^{-1}) information necessary.

Different Notions of Optimum

Unconstrained case

- local vs. global
  - local minimum $\boldsymbol{x}^*$: ∃ a neighborhood V of $\boldsymbol{x}^*$ such that $\forall \boldsymbol{x} \in \mathrm{V}: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
  - global minimum: $\forall \boldsymbol{x} \in \Omega: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
- strict local minimum if the inequality is strict

Further Details on Remaining 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

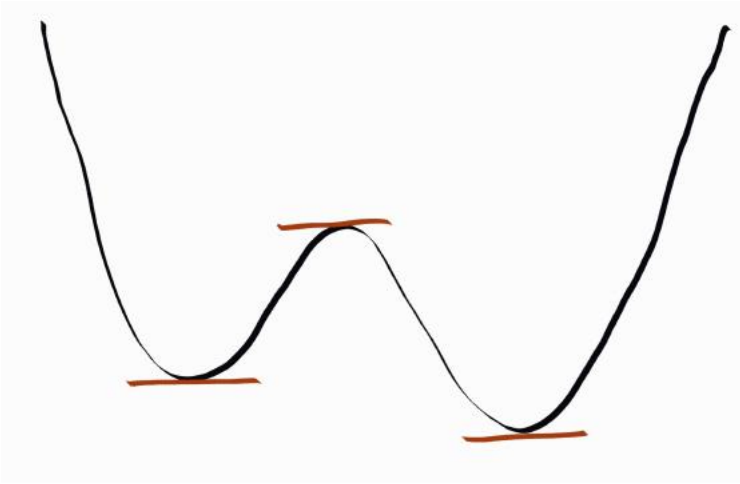
- CMA-ES (adaptive algorithms / Information Geometry)
- PhD thesis possible on this topic

method strongly related to ML / new promising research area
interesting open questions

Mathematical Characterization of Optima

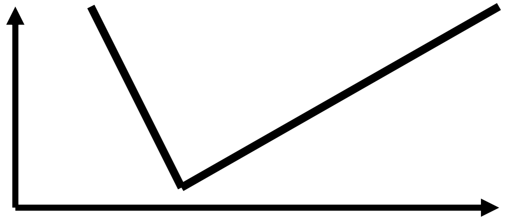
Objective: Derive general characterization of optima

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



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

- $(E, || \quad ||)$ will be a K-general vector space endowed with a norm $|| \quad ||$ 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.