Optimization and Data Science

Lecture 1: Introduction

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Contents

Optimization

- What is Optimization?
  - Process or method to minimize or maximize a given quantity.
  - This quantity is expressed by a mathematical function that depends on some parameters.

Optimization

- What is Optimization?
  - Process or method to minimize or maximize a given quantity.
  - This quantity is expressed by a mathematical function that depends on some parameters.
- Why do we study these kind of problems?
  - Many practical applications in different disciplines (physics, chemistry, ..., economics etc.) can be expressed as optimization problems.
  - Human intention usually is not just to solve a problem, but to obtain the optimal solution.

Optimization

- What is Optimization?
  - Process or method to minimize or maximize a given quantity.
  - This quantity is expressed by a mathematical function that depends on some parameters.
- Why do we study these kind of problems?
  - Many practical applications in different disciplines (physics, chemistry, ..., economics etc.) can be expressed as optimization problems.
  - Human intention usually is not just to solve a problem, but to obtain the optimal solution.
- How does optimization work?
  - Optimization problems are formulated using mathematical functions.
  - Criteria for existence and uniqueness of optima are derived using mathematical techniques.
  - Solution algorithms are iterative and approximate.
  - $\leadsto$ Convergence speed and accuracy are issues.

Optimization

- What is Optimization?
    - Process or method to minimize or maximize a given quantity.
    - This quantity is expressed by a mathematical function that depends on some parameters.
- Why do we study these kind of problems?
    - Many practical applications in different disciplines (physics, chemistry, ..., economics etc.) can be expressed as optimization problems.
    - Human intention usually is not just to solve a problem, but to obtain the optimal solution.
- How does optimization work?
    - Optimization problems are formulated using mathematical functions.
    - Criteria for existence and uniqueness of optima are derived using mathematical techniques.
    - Solution algorithms are iterative and approximate.
    - $\rightsquigarrow$ Convergence speed and accuracy are issues.
- What if we know about optimization?
    - We can use these methods and algorithms for all kind of engineering and scientific problems.
    - They are useful tools for or elements of other methods.
    - We can improve these methods and adapt them for special problems.

Data Science

- What is Data Science?
  - Area of research that deals with data retrieval, representation, storing, processing, mining, getting information out of data
  - Broad topic, limits not completely clear or well-defined.
  - Big Data, Data Management, Data Analysis, Data Mining, Data Assimilation

Data Science

- What is Data Science?
  - Area of research that deals with data retrieval, representation, storing, processing, mining, getting information out of data
  - Broad topic, limits not completely clear or well-defined.
  - Big Data, Data Management, Data Analysis, Data Mining, Data Assimilation
- Why do we study these kind of problems?
  - Computer application generates huge amount of data
  - ... information therein can be exploited.

Data Science

- What is Data Science?
  - Area of research that deals with data retrieval, representation, storing, processing, mining, getting information out of data
  - Broad topic, limits not completely clear or well-defined.
  - Big Data, Data Management, Data Analysis, Data Mining, Data Assimilation
- Why do we study these kind of problems?
  - Computer application generates huge amount of data
  - ... information therein can be exploited.
- How does data science work?
  - Broad area $\rightsquigarrow$ broad range of technologies
  - Measurement, transfer, storage, representation, compression, reduction, clustering, obtain useful information out of data, comparing, optimization methods

Data Science

- What is Data Science?
  - Area of research that deals with data retrieval, representation, storing, processing, mining, getting information out of data
  - Broad topic, limits not completely clear or well-defined.
  - Big Data, Data Management, Data Analysis, Data Mining, Data Assimilation
- Why do we study these kind of problems?
  - Computer application generates huge amount of data
  - ... information therein can be exploited.
- How does data science work?
  - Broad area $\rightsquigarrow$ broad range of technologies
  - Measurement, transfer, storage, representation, compression, reduction, clustering, obtain useful information out of data, comparing, optimization methods
- What if we know about data science?
  - We know what is really meant when people talk about data science etc.
  - We know where and how which method can be appropriately applied.
  - We can transfer methods from other areas to data science.

Topic of the lecture: Optimization **and** Data Science

- What does the "and" mean?
- Does not mean we cover all of both topics (impossible).

- At first: We cover optimization methods

$\rightarrow$ ... to be used in any application where optimization is an issue.

$\rightarrow$ We use data science topics as examples.

- Moreover: We cover the parts of data science which are related to optimization, e.g.,

$\rightarrow$ data science methods are used to formulate or simplify an optimization problem

$\rightarrow$ data science problem is in fact an optimization problem

$\rightarrow$ data science methods use/need optimization methods.

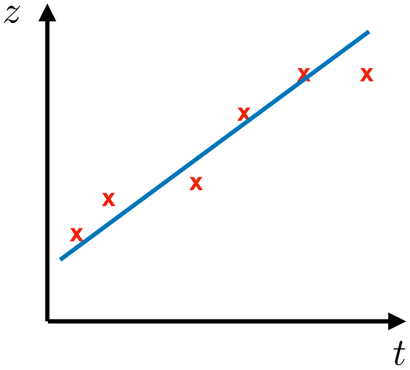
Overview of the course

- Relation data $\leftrightarrow$ optimization
- Methods for data analysis
- ... including statistics
- Methods for reduction of data complexity
- Formulation and classification of optimization problems
- Unconstrained optimization problems
  - Existence and uniqueness results
  - Optimality conditions
  - Solution algorithms
  - Algorithms applied in machine learning
- Optimization problems with additional constraints
  - (as above)

Example: Data-fitting (or: regression/reduced-order model)

- Given: data points

$$(t_k, z_k)_{k=1,\dots,m},\ t_k, z_k \in \mathbb{R}.$$



- Observation: approx. linear dependency
- Task: Detect parameters of this dependency
- Mathematical task: Find affine-linear function

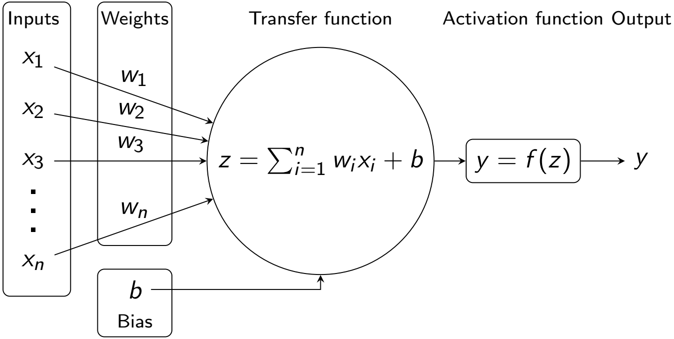
$$y(t) = at + b$$

that satisfies (at least approximately)

$$y(t_k) = at_k + b \approx z_k, \quad k = 1, \dots, m.$$

- Exact equality not possible for $m > 2$
- $\rightsquigarrow$ minimize distance between points and function (optimization problem)

Example: Training of artificial neural network



- Given: training data set with input $(x_{ik})_{i=1,\dots,n,k=1,\dots,m}$ and output $(y_k)_{k=1,\dots,m}$.
- adjust weights, bias to minimize misfit $f(\sum_i w_i x_{ik} + b) - y_k$ for training data set
- Reality: More than one layer $\rightsquigarrow$ concatenation of several transfer and activation functions

Contents

Example: Economic problem

- Boat rental wants to buy new boats
- two types of boats:
  - Standard (S) revenue/week: 600 EUR
  - Premium (P) revenue/week: 800 EUR
- total capacity: max. 350 boats
- additional requirement: max. 200 Premium boats
- but more P than S
- time needed per week: Standard 3 h, Premium: 4 h
- max. total working hours/week: 1400 h

Example: Sustainable Fishery (simplified)

- Temporal distribution of fish in one region of the ocean:

$$y_i: \text{ amount of fish at time } t_i, i = 1, \dots, n.$$

- Define quotas $q = (q_i)_{i=1}^n$ of the fish stock that is allowed to be fished for days/months/years in a given time interval.
- Aim: maximize profit of fishermen:

$$\max_q \sum_{i=1}^n q_i y_i$$

- sooner profit is worthier than later profit (parameter $\rho > 0$):

$$\max_q \sum_{i=1}^n q_i y_i e^{-\rho t_i}$$

Sustainable Fishery (2)

- Fishing effort depends on total amount of harvest (and thus on quota):
- Assumption: cost quadratic in q_i, parameter $\alpha > 0$:

$$\max_q \left(\sum_{i=1}^{n} q_i y_i e^{-\rho t_i} - \alpha \sum_{i=1}^{n} q_i^2 \right)$$

- Sustainable fishing: total stock should be always bigger than some lower limit

$\rightsquigarrow$ At the end of the time interval ($t = t_n$) there should be "some" fish remaining:

$\rightsquigarrow$ Reward if y_n "big" (with parameter $\beta > 0$):

$$\max_q \left(\sum_{i=1}^{n} q_i y_i e^{-\rho t_i} + \beta y_n - \alpha \sum_{i=1}^{n} q_i^2 \right)$$

Sustainable Fishery (3)

- Rewrite as minimization:

$$\max_q \left(\sum_{i=1}^{n} q_i y_i e^{-\rho t_i} + \beta y_n - \alpha \sum_{i=1}^{n} q_i^2 \right) \iff \min_q \left(-\sum_{i=1}^{n} q_i y_i e^{-\rho t_i} - \beta y_n + \alpha \sum_{i=1}^{n} q_i^2 \right).$$

- Relation of parameters

$$1 : \beta : \alpha$$

allows for different weights of the three aims

profit : sustainability : cost of fishing.

$\rightsquigarrow$ non-linear problem in q.

- In reality, this is even more complex (many fishermen, spatial distribution, fish stock depends on fishing $\rightsquigarrow$ dynamic problem...).

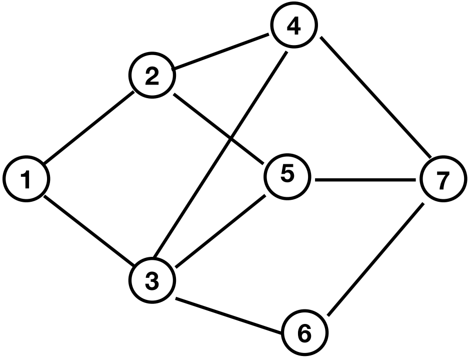
Example: Knapsack problem

- Task: Given
  - a number of items with needed space/weight and given value
  - a knapsack with given capacity
- select items such that the total value is maximized
- and maximum capacity is not violated.
- Example:

i	1	2	3	4	5	6	7	8
c_i	15	100	90	60	40	15	10	1
v_i	2	20	20	30	40	30	60	10

Example: Traveling salesman problem

- Task: find path in a graph with minimal length/cost
- and pass every node, but only once



Contents

General formulation of optimization problems

Most general form:

$$\min_{x \in X_{ad}} f(x)$$

Terminology:

- $f : X_{ad} \to \mathbb{R}$: **cost function, objective function**, usally real-valued
- min $\leftrightarrow$ max by replacement $f \leftrightarrow (-f)$
- x: **control** or **optimization parameters**
- $X_{ad} \subset X$: **admissible** or **feasible set**
- X: usually vector space or unbounded set
- $X_{ad} = X$ **unconstrained problem**
- $X_{ad} \neq X$ **constrained problem**: often X_{ad} is defined by functions:

$$X_{ad} := \{x \in X : g(x) \leq 0, h(x) = 0.\}$$

Contents

Classification of optimization problems

- w.r.t. the type of optimization parameters/variables:
  - $X = \{0,1\}^n, X = \mathbb{N}^n, X = \mathbb{Z}^n$: **integer** or **discrete** problems, also problems on graphs are of this form.
  - $X = \mathbb{R}^n$: **continuous** problems.
  - $X = X_1 \times X_2$ with X_1 discrete, X_2 continuous: **mixed-integer** problems.
- w.r.t. constraints:
  - $X_{ad} \neq X$: **constrained** problems
  - $X_{ad} = X$: **unconstrained** problems
- w.r.t. linearity:
  - f and g, h defining $X_{ad} := \{x \in X : g(x) \leq 0, h(x) = 0\}$ are linear: **linear** problem
  - f, g, or h nonlinear: **nonlinear** problem
  - special case: **least-squares** problems, if $f(x) = \sum_{i=1}^{m} F_i(x)^2$ **(non-)linear least-squares problem** if F_i (non-)linear
- for continuous problems:
  - **differentiability**: of f, g, h
  - **convexity**: of X_{ad} and f

Exercise:

- Classify the problems in the above examples.

Introduction: What is important?

- Data science is an emerging field in Computer Science, in science general and in society.
- It contains technical as well as algorithmic and mathematical topics and methods.
- In this lecture, we study basic methods of data science ...
- ... and the relation of data science to optimization.
- Many data science problems/methods have a close relation to (or are in fact) optimization problems/methods.
- Many optimization problems use data.
- Optimization problems themselves can be formulated in a mathematical (somehow standard) way.
- They can be devided into classes depending on their structure and the (data) type of parameters that are to be optimized.