Algorithmic Foundations of Data Science[*]

Spring 2022, ETH Zurich

Contents

1 Current course website **2**

2 Synopsis **2**

3 General information **2**

3.1 Important: Optimization for Data Science 2018–2021 2

3.2 Exam information 3

4 Lectures **3**

5 Weekly exercise sheets **4**

6 Programming exercises **4**

7 Background material **4**

8 Special assignments **4**

9 Past exams **5**

10 Reference texts **5**

[*] home — versions: web / pdf / gitlab

1 Current course website

This is the website of the iteration of the course that took place in spring 2022. You can find the website of the current iteration «here».

2 Synopsis

This course provides theoretical foundations for the design and mathematical analysis of efficient algorithms that can solve fundamental tasks relevant to data science. We consider statistical models for such tasks and aim to design efficient (polynomial-time) algorithms that achieve the strongest possible rigorous (statistical) guarantees under these models. We also incorporate adversarial (worst-case) components into these models as a way to reason about the robustness of algorithms.

3 General information

In order to access the course material on this website, use your «nethz account» (same login as for mailbox at ETH Zurich).

	time	room
lecture	Th 10-12	CAB G 51
lecture	Fr 12-13	HG D 3.2
exercise class	Fr 13-15	HG E 22
office hours	Wed 17-18	ML H 41.1

If you are a student enrolled in this course, please have a close look at the course logistics as well as the official information in the «course catalogue of ETH Zurich».

3.1 Important: Optimization for Data Science 2018–2021

This course was created after a reorganization of the course “Optimization for Data Science” (ODS). A significant portion of the material for this course has previously been taught as part of ODS. Consequently, *it is not possible to earn credit points for both this course and ODS as offered in 2018–2021*. This restriction does not apply to ODS offered in 2022 or afterwards and you can earn credit points for both courses in this case.

3.2 Exam information

The exam takes four hours but is designed so that it can be solved in approximately two hours. The additional time is supposed to take away time pressure. No additional material is allowed at the exam, we will hand out a cheat sheet containing facts you can use without proof. Further, you can use without proof all results from lecture, the exercises, and the special assignments. The exam is designed to test if you understood the core of the arguments and results seen in lecture, the exercises, and the special assignments and can reproduce them. It is not designed to test how well you can memorize complicated proofs, etc. As an example, knowing properties of the singular value decomposition (best low-rank approximation, poly-time computable, etc.) would count as "understanding the core of the arguments and results seen in lecture".

4 Lectures

You can access «video recordings of the lectures».[1] In addition, we provide handwritten notes that underly the individual lectures.

The following collection of typeset notes serve as a companion to the lectures. These notes cover approximately the same material as the lectures but sometimes contain additional discussions or more general expositions. For the final examination, we expect that you have read and understood the material in these notes.

- Linear regression and ordinary least squares
- Minimax lower bounds and Bayesian linear regression
- Sparse linear regression and lasso
- Huber loss and oblivious outliers
- Low-rank models and statistical guarantees of principal component analysis
- Matrix completion and Bernstein tail bounds for matrices
- Community detection in networks and Grothendieck's inequality
- Non-negative matrix factorization and topic models
- Tensor decomposition: algorithms and applications
- Clustering, Gaussian mixture models, and sum-of-squares
- Robust mean estimation

[1] Due to a technical issue, the lectures in the first week as well as on April 1, May 6, and the first hour of June 2 have not been recorded. Unfortunately, for some early lectures, only an audio recording is available.

5 Weekly exercise sheets

Weekly exercise sheets will be released each Friday afternoon. The deadline is 11:59 the following Friday. We strongly encourage you to try to solve them and to hand in your (partial) solutions.

- Exercise sheet 0
- Exercise sheet 1
- Exercise sheet 2
- Exercise sheet 3
- Exercise sheet 4
- Exercise sheet 5
- Exercise sheet 6
- Exercise sheet 7
- Exercise sheet 8
- Exercise sheet 9
- Exercise sheet 10
- Exercise sheet 11
- Exercise sheet 12
- Exercise sheet 13

6 Programming exercises

- Ridge regression
- Lasso and sparse linear regression
- Principal component analysis
- Matrix completion

7 Background material

- Notation and background
- Central limit theorems and the Gaussian distribution
- Concentration bounds
- Convex optimization

8 Special assignments

In Week 6, specifically, on April 1, and Week 12, on May 20, we will assign two graded homeworks, called special assignments, as compulsory continuous performance assessments, accounting together for 30% of the final grade (15%

for each graded homework). For more information, see the section on grading in the course logistics.

- Special assignment 1
- Special assignment 2

9 Past exams

Below you find exams of the course "Optimization of Data Science" which previously contained a large part of the material of this course. However, since this course also contained additional material, we indicate which exercises are relevant for the current course.

- 2018 - solution - relevant exericses: Assignments 5 and 6
- 2019 - solution - relevant exericses: Assignments 4,5, and 6
- 2020 - solution - relevant exericses: Assignments 4,5, and 6
- 2021 - solution - relevant exericses: Assignments 4 and 5

10 Reference texts

While this course does not adhere to a particular textbook, the following references cover (different) parts of the course:

- Martin Wainwright, «High-Dimensional Statistics: A Non-Asymptotic Viewpoint».
- Ankur Moitra, «Algorithmic Aspects of Machine Learning»
- Philippe Rigollet and Jan-Christian Hütter, «High Dimensional Statistics: Lecture Notes»
- Roman Vershynin, «High-Dimensional Probability: An Introduction with Applications in Data Science»