Algorithmic Foundations of Data Science*

Spring 2025, ETH Zurich

Contents

1 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

*home — versions: web / pdf

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.

We will see that even for basic statistical estimation problems, the naive maximum-likelihood optimization problems are often NP-hard (in the worst case). Throughout the course, we will develop different strategies to circumvent these intractability barriers and to obtain polynomial-time estimators with close to optimal statistical guarantees.

2 General information

In order to access the course material on this website (e.g., lecture notes and weekly exercises), use your nethz account (same login as for mailbox at ETH Zurich).

	time	room
lecture	Tue 10-12	HG D 7.1
lecture	Tue 13-14	HG D 7.1
exercise class	Fr 12-14	CAB G 11
office hours	Mon 17-18	CHN D 29

There are no office hours during the first week of classes. The exercise session in the first week can be used to ask questions about the background material. 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».

2.1 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.

2.2 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". Similarly, we expect that you know the material in the section background material.

Also, two of the exam exercises will be taken from the weekly exercise sheets (one from the graded exercises and one from the ungraded exercises).

3 Lectures

We provide a collection of typeset notes that 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, unless explicitly excluded.

We also publish handwritten notes from a previous iteration of the course that in some cases are closer to what was discussed in lecture.

- 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
- Sum-of-squares for estimation problems
- Clustering and Gaussian mixture models via sum-of-squares
- Robust mean estimation

4 Weekly exercise sheets

Weekly exercise sheets will be released each Tuesday night. The deadline is Friday 14:00 the following week (before the exercise session starts) and you can submit your exercises via «Moodle». Unfortunately, we cannot promise you to give feedback on all exercises on a given sheet, but we will give feedback for at least one exercise per sheet. We will indicate for which one.

In Weeks 2 to 12 (with the exception of Week 9), one of the exercises will be graded and count toward your final grade, see course logistics We strongly encourage you to try to also solve the other exercises.

- 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

5 Special assignments

We will release two graded homework assignments, called special assignments, that count toward your final grade (see the section on grading in the course logistics for more information). We will release them (up to small modifications) in Week 6, Friday Mar 28, and Week 11, Friday May 16. You have roughly two weeks to solve each assignment (see the assignment itself for the precise deadline).

- Special assignment 1
- Special assignment 2

6 Background material

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

7 Past exams

Below you find exams from previous iterations. Some exams are from the course “Optimization of Data Science” which previously contained a large part of the material of this course. We indicate which exercises are relevant for the current course.

- 2018 - solution - relevant exercises: Assignments 5 and 6
- 2019 - solution - relevant exercises: Assignments 4,5, and 6
- 2020 - solution - relevant exercises: Assignments 4,5, and 6
- 2021 - solution - relevant exercises: Assignments 4 and 5
- 2022 - solution
- 2023 - solution
- 2024 - solution

8 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»