Algorithmic Foundations of Data Science*

Spring 2024, ETH Zurich

Contents

*home — versions: web / pdf / gitlab

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 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	Th 10-12	CAB G 51
lecture	Fr 12-13	ML F 36
exercise class	Fr 14-16	ML F 36
office hours	Wed 16-17	HG D 3.3

Office hours start on Wednesday, February 28, i.e., there are no office hours during the first week of classes. 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, one of the exam exercises will be taken from the additional exercises we publish via the issue tracker each week. Refer to the logistics for more information.

3 Lectures

First, we provide handwritten notes that correspond closely to the individual lectures.

Second, 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.

- Linear regression and ordinary least squares
- Minimax lower bounds and Bayesian linear regression
- Compressed sensing
- 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
- Restricted isometry property of subsampled Fourier matrices

4 Weekly exercise sheets

Weekly exercise sheets will be released each Friday night. The deadline is 11:59 the following Friday. We strongly encourage you to try to solve them and to hand in your (partial) solutions. Refer to the course logistics for information on how to submit your solutions.

- 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 Additional exercises

In addition to the exercise sheets, we publish one exercise separately each week. To solve this, we envision the following procedure: You try to solve it on your own within the first week after publication, then, a "discussion phase" starts in which we post the same exercise, still without solutions, via the «issue tracker» where you solve it collaboratively with your peers. After one more week, we will publish its solutions. We encourage you to also post partial solutions and/or questions in the "discussion phase". As an incentive for you to solve these exercises, one of them will appear as is in the exam. This is supposed to make it easier for you, if you know the rest of the material well, all of the exam exercises can still be solved without having seen them before.

- Bonus exercise 1
- Bonus exercise 2
- Bonus exercise 3
- Bonus exercise 4
- Bonus exercise 5
- Bonus exercise 6
- Bonus exercise 7
- Bonus exercise 8
- Bonus exercise 9

- Bonus exercise 10
- Bonus exercise 11

6 Background material

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

7 Special assignments

We will assign two graded homework assignments, called special assignments, as compulsory continuous performance assessments, accounting together for 30% of the final grade (15% for each graded homework). We will release the first special assignment on Wednesday, April 10 and the second ~~on Friday, May 17~~ early in the week of May 20. You have roughly two weeks to solve each assignment (see the assignment itself for the precise deadline). For more information, see the section on grading in the course logistics.

- Special assignment 1
- Special assignment 2

8 Previous special assignments

- 2022, Special assignment 1
- 2022, Special assignment 2
- 2023, Special assignment 1
- 2023, 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 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

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»