Modern Topics in Statistical Learning Theory

Spring 2023

Lecturer: Qi Lei
Email: ql518@nyu.edu
Office hour: Friday 4-5 pm, CDS 706 (60 5th Ave)
remark: Students only have unlimited access to CDS buildings during 7 am - 6 pm Monday - Friday.

Section leader: Yunzhen Feng
Email: yf2231@nyu.edu
Office hour: Thursday 3-4 pm (preferred), & Wednesday 2:30-3:30 pm, CDS 242 (60 5th Ave)
Structure: This is an in-person class. There is a lecture from 10 am - 11:40 am every Friday morning at Silver Center room 520. Zoom meeting is possible when instructor is away for conferences.

Course Description:

This course is a graduate-level topic course focusing on the theoretical grounding and statistical properties of the modern learning algorithms — with a focus on weakly supervised learning.

The intended topics to cover include: basics in machine learning, optimization and generalization bound, followed by the introduction and theoretical understanding surrounding meta-learning, self-supervised learning, and domain adaptation.

To benefit from this class, strong linear algebra, probability, and optimization background are required. Students should be familiar with basic machine learning and deep learning concepts.

The class consists of 3 units. In the first unit, we will cover the more standard theoretical analysis tools used in deep learning including stochastic gradient, uniform convergence theory and statistical learning theory.

After the first unit, the course will move on to specific topics (Unit 2: transfer learning and 3: self-supervised learning). Since some topics covered in this course are quite recent, some content will be based on recent papers instead of a textbook.

Resources for the class: Even though the content of this course is not based on a specific textbook, the following materials are good references for certain topics of the course.

- *Stanford CS 229M notes (http://web.stanford.edu/class/stats214/)*
- *CMSC 828W notes: Foundations of Deep Learning (http://www.cs.umd.edu/class/fall2022/cmsc828W/info.html)*
- *Wainwright Book (High dimensional statistics non-asymptotic (https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E)*
- *Vershynin book (https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf)*

Tentative schedule: We will follow the following the approximate schedule.

Unit 1: Deep Learning optimization and generalization

- Week 1: Basics in ML
- Week 2: Generalization bound: concentration inequality
- Week 3: Generalization bound: uniform convergence
- Week 4: Generalization bound: complexity measure
- Week 5: Theory of deep learning: (non-convex) optimization
- Week 6: Theory of deep learning: neural tangent kernel

- Week 7: Theory of deep learning: implicit/algorithmic regularization

Unit 2: Transfer learning

- Week 8: Meta learning
- Week 9: Domain adaptation
- Week 10: Domain generalization

Unit 2: Self-supervised learning

- Week 11: Data augmentation
- Week 12: Self-supervised learning
- Week 13: Transformer

Final presentation

- Week 14: Final presentation

Final Grade:
30% Homework, 10% Scribing, 30% Midterm, 30% Final project

1. **Homework** is given in the first half of the course on the fundamentals of machine learning. Homework is required to be LaTeXed (a good online platform for latex is overleaf) instead of hand-written. We allow for 2 weeks for each homework. You should always try to finish it within one week and use the second week as a buffer for any unpredictable events. We do not accept late submissions; we do not drop the lowest score as there are only 3 assignments.

2. **Scribing.** Please sign-up here before Feb 2nd. Each student is responsible for scribing: latex a certain topic. In general, around 2 students (depending on the class size) will be assigned on each topic. They should work together in editing the scribed notes, complete omitted details, and provide proper reference. A nice example is this Lecture notes 1 and its Tex file that you may find in this course.

3. **Midterm** will be take-home and consists of both written problems and coding problems.

4. **Final project.** In teams of at most three, you will do a kaggle competition using any tools we learned in class. (15% of score is based on the ranking on the competition.)
A written report is required. (15% of score is based on the report.) It is optional to give an oral presentation in the last week (with up to 10% bonus points).

Technology

- *NYU Brightspace* will be used to coordinate the course, class notes, and grades. Slides or written notes from the presenter will be made available. Scribed notes from the students will be uploaded on a timely manner.
- *Overleaf* (`https://www.overleaf.com/`) is a convenient online platform for writing and compiling tex files.
- *Colab.* (`https://colab.research.google.com/`) We will provide some programming modules on Colab to supplement the lecture.
- *Kaggle* (`https://www.kaggle.com/`) will be used for one data competition in the end of the class.

Diversity Statement

- As an instructor, I will strive to create a safe, respectful, and inclusive environment for all students regardless of their identity. I recognize and value diversity inside and outside of the classroom, and recognize that each student has a unique contribution to make and brings with them different strengths and weaknesses. I welcome your ideas for how to promote a better understanding and deeper learning in this class as a community. Please feel free to ask questions, to participate in discussions, and to suggest new approaches to the class content. Please also feel welcome to raise any issue you may have in class or outside of class, including reporting incidents of bias or discrimination, whether intentional or unintentional, either to me, to your advisor(s)/mentor(s), or by using the NYU Bias Response Line.