CS 226r Fall 2024
Theory for Society: Algorithmic Fairness

MW 12:45-2:00pm, SEC LL2.229

Teaching Staff

Professor Cynthia Dwork
dwork@seas.harvard.edu
OH (tentative): Alternate Wednesdays (when there is no theory seminar) 4-5pm, and by appointment
Reading Discussions: Thursday 1/25 and thereafter on Mondays, 7:30-8:30pm via Zoom

David Brewster, head TF, dbrewster@g.harvard.edu
Sahil Kuchlous, sahilkuchlous@college.harvard.edu
Peihan Liu, peihanliu@fas.harvard.edu
Yanis Vandecasteele, yanis_vandecasteele@g.harvard.edu

Overview

As algorithms reach ever more deeply and broadly into our lives there is increasing interest that they be fair, despite a lack of consensus on the meaning of the term. The theory of algorithmic fairness is a still-new discipline exploring notions of fairness and their consequences: which fairness goals can be simultaneously achieved? How do various notions compose – are systems made up of parts that are fair in isolation also fair in toto? How can we move beyond fairness-as-correctness in the current, flawed, world, to fairness in a better world? The course will start with basics and move to highlights from a recent explosion of research showing broad applicability to problems in machine learning even when fairness is not a concern, as well as deep connections to notions in pseudorandomness and complexity theory.

Assignments and Deadlines

There will be 3-4 problem sets, a reading assignment, and a course project. The last problem set will be due during Week 10. You may have 8 late days. If you need more than 2 extra days on any given assignment, please notify a member of the teaching staff.

Problem Set Collaboration Policy

You may work on problem sets in groups of up to 3 people, but *you must write up your solutions independently. You must list your collaborators and any resources (papers, websites, LLMs) used.*

Policy on Language Models and Other Technologies

If you use any technology in a substantial way, you must include as an appendix to your written work a description of what you used, how you used it, and how it helped you to learn. A description of usage in preparation for a discussion of the readings must be submitted before the discussion begins.

Course Project

Projects can be done alone or in pairs, with groups of 3 permitted for ambitious projects. Course project proposal *ideas* are due March 18 (beginning of week 8). This is not graded. The final project proposal will be due March 25. A progress report is due April 8. There will be in-class presentations on April 17 and April 22. The project is due on May 1.

Reading Assignments

There is no book for this course, and lectures are based on papers. 1-2 groups, of up to 6 students each, will meet with the professor during each week to discuss some of the readings in detail. Learning to read research papers is an important part of this course, and contributions to this discussion will figure in the final grade. The set of students for each paper will be determined via a sign-up sheet. *Each student must sign up for at least one paper*.

Grading

The grade will be roughly 5% paper discussion, 65% homework, 30% final project.

Advice for Reading Papers

The reading assignments are important. This is a graduate course, and you should be able to read most of the material in the assigned papers. If you have trouble, prioritize (1) the definitions and (2) understanding the theorem statements. Read "around" the theorem and lemma statements; a good paper will often explain in informal terms, just before or just after the theorem statement, what the theorem is telling us.

Draft of Detailed Syllabus 2/21/2024

The syllabus will be updated during the semester. Check back regularly.

Week 1: January 22, 2024: Individual and Group Fairness

Primary Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012, January. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*
Additional Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness.
Additional Background on Linear Programming

Lecture 1, January 22:
Course Overview
Individual, group, and multi-group fairness
The meaning of individual probabilities

Outcome Indistinguishability as a new paradigm for machine learning
Complexity Roots

Lecture 2, January 24
What were the assumptions?
Paradigm from Cryptography, Catalog of evils
Statistical parity and Individual (metric) fairness
What should the metric capture?
The fairness LP
Two Challenges: Size of the universe of individuals & articulating a metric

Week 2: January 29, 2024

Primary Ilvento, Metric learning for individual fairness
Additional Bechavod, Jung, and Wu, Metric-Free Individual Fairness in Online Learning
Additional Background on PAC Learning

Lecture 3, January 29
Learning a metric from an arbiter

Lecture 4, January 31
(Agnostic) PAC Learning, ε-nets, ε-samples, VC dimension
Addressing the size of the universe
Ilvento's D'_r based on learning threshold functions
Generalizing correctness and fairness

Week 3: February 5, 2024

Primary Rothblum and Yona, Probably approximately metric-fair learning
Primary Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat]
Additional Mukherjee, Yurochin, Banajee, and Sun, Two Simple Ways to Learn Individual Fairness Metrics from Data (mlr.press)?

Lecture 5, February 5
Addressing the size of the universe (given metric information)
Generalizing correctness and fairness

Lecture 6, February 7
Wrap up Rademacher Complexity
Fairness and language models 1: Bolukbasi et al., begin Mukherjee et al

Week 4: February 12, 2024

Primary Dwork and Ilvento, Fairness Under Composition, ITCS 2019
Primary Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), pp.153-163, Sections 1 and 2.
Primary Functions: Kleinberg, J., Mullainathan, S. and Raghavan, M., 2016. Inherent trade-offs in the fair determination of risk scores, Sections 1 and 2
Primary Neil and Winship, Methodological challenges and opportunities in testing for racial discrimination in policing
Primary Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms | NEJM
Primary The JAMA editorial (June 6, 2022) on race-free estimation of kidney function
Additional article

Lecture 7, February 12
Wrap up Mukeherjee et al.
Fairness Under Composition
Group Fairness: Examples and Impossibility Results
Classifiers: Chouldechova,
Scoring Functions: Kleinberg, Mullainathan, and Raghavan

Lecture 8, February 14
Group Fairness: Benchmarking, Auditing, and Correcting
Neil and Winship, Methodological challenges and opportunities in testing for racial discrimination in policing.
Race Correction in Medicine Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms | NEJM
The JAMA editorial (June 6, 2022) on race-free estimation of kidney function and article referenced therein

Week 5: February 19, 2024

Interlude: Some complexity theory
Primary Tulsiani, Trevisan, and Vadhan Regularity, Boosting, and Efficiently Simulating Every High-Entropy Distribution Lemma, FOCS 2009, Sections 1, 3, 4, and 5
Additional Maciej Skorski. A cryptographic view of regularity lemmas: Simpler unified proofs and refined bounds. In Theory and Applications of Models of Computation, page 586–599, 2017

No class February 19 (Presidents' Day)

Lecture 9, February 21
Szemeredi's strong regularity theorem – statement and application
TTV: the regularity lemma and proof of Frieze-Kannan Regularity Theorem
Sketch of TTV Regularity ⇒ Impagliazzo's Hard-Core Lemma (Section 5 of TTV)

Week 6: February 26, 2024

Individual Probabilities and Multigroup Fairness
Primary Dawid, P., 2017. On Individual Risk. *Synthese*, *194*(9), pp.3445-3474. ArXiv version (2014), Sections 1-4
Primary Dwork, C., Kim, M.P., Reingold, O., Rothblum, G.N. and Yona, G., 2021, June. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (pp. 1095-1108), Sections
Primary Hebert-Johnson, Kim, Reingold, and Rothblum, Calibration for the (Computationally-Identifiable) Masses (ArXiv version). In *International Conference on Machine Learning* (pp. 1939-1948). PMLR, Sections 1-3.1 and 4

Lecture 10, February 26
Predictors and Individual Probabilities
Outcome Indistinguishability
Equivalence of no-access OI and TTV regularity

Lecture 11, February 28
Multigroup Fairness, Multicalibration (partition version), and relationship with weak agnostic learning

Week 7: March 4, 2024

Primary: Dwork, Lee, Lin, Tankala From Pseudorandomness to Multi-Group Fairness and Back, COLT 2023, Sections 1-4 and 6 and relevant proofs in the Appendix.
Additional: Yi-Hsiu Chen, Kai-Min Chung, and Jyun-Jie Liao. On the complexity of simulating auxiliary input. In EUROCRYPT: Annual International Conference on the Theory and Applications of Cryptographic Techniques, volume 3, pages 371–390, 2018.

Lecture 12, March 4
1. Equivalence of (1) no-access OI and multiaccuracy; (2) multiaccuracy and regularity; (3) sample-access OI and multicalibration
2. No predictor smaller than the distinguishers fools all the distinguishers, in contrast to the situation with pseudo-random generators, pseudo-random functions.

Lecture 13, March 6 (Pranay Tankala)
The OI/Graph Regularity connections (from DLLT)

No Class Week of March 11, 2024 (Spring Break)

Week 8: March 18, 2024

Casacuberta, S., Dwork, C. and Vadhan, S., 2023. Complexity-Theoretic Implications of Multicalibration.

Lecture 14, March 18
Review/Catch Up; partition view of multicalibration (maybe omniprediction)

Lecture 15, March 20 (Silvia Casacuberta)
Complexity-Theoretic Implications of Multicalibration: new hardcore theorem and simple proof of an old hardcore theorem

Week 9: March 25, 2024

Rothblum and Yona, Multi-group agnostic PAC learnability

Dwork, Lee, Lin, Tankala From Pseudorandomness to Multi-Group Fairness and Back

Gopalan, P., Hu, L., Kim, M.P., Reingold, O. and Wieder, U., 2023. Loss Minimization Through the Lens Of Outcome Indistinguishability (ITCS 2023)

Lecture 16, March 25
Indistinguishability and Loss: Rothblum and Yona
Omniprediction (from DLLT)

Lecture 17, March 27
Loss OI

Week 10: April 1, 2024

Sergiu Hart's lecture (video)

Sandroni, A, The reproducible properties of correct forecasts | International Journal of Game Theory (springer.com)

Dwork et al., Learning from Outcomes: Evidence-Based Rankings

Lecture 18, April 1
Forecasting (online prediction)

Hart's proof of the min-max theorem video (start at minute 10) and the meaninglessness of online calibration

Online-to-Batch converstion

Lecture 19, April 3
Beyond calibration tests (Sandroni)
Fair Ranking

The schedule below has been modified due to the eclipse. Class on April 8 is canceled and a make-up class will be held on April 29.

Week 11: April 8, 2024

Heidari, H. and Kleinberg, J., Allocating Opportunities in a Dynamic Model of Intergenerational Mobility
Kohler-Haussman, What Does 'Race Neutral' Admissions Mean?

No class Monday, April 8

Lecture 20, April 10 (Wednesday)
Affirmative Action
Fairness Through Awareness
Affirmative Action via Fair Ranking
Heidari & Kleinberg
+ 15 minutes of discussion. Come to class prepared to discuss Kohler-Haussman's paper.

Week 12: April 15, 2024

Dwork, Reingold, and Rothblum, From the Real Towards the Ideal: Risk Prediction in a Better World, FORC 2023

Lecture 21, April 15 (Monday)
Better Worlds

Wednesday April 17 and Monday April 22

Project Presentations: ungraded brief (<10 minute) class presentations describing your projects. If needed, we will also book some time in the evening or format as a poster session.

Wednesday, April 24, 2024: Flipped Class

We will discuss the presentations from Wrong at the Root and the Hu & Kohler-Hausmann paper What's Sex Got to Do with Fair Machine Learning?

Readings for April 24:
Hu and Kohler-Hausmann, What's Sex Got to Do with Fair Machine Learning?

Videos from Wrong at the Root:
Watch Jay Kaufmann's talk here at 49:15
Watch Dorothy Roberts's talk here at 1:31:54
Watch Morris Hardt's talk here at 22:08
Watch at least one additional presentation from the Wrong at the Root workshop and come to class prepared to share something you learned.

Make-Up Class Monday, April 29, 2024

Lecture 23: (Un)Fairness in Networking

Okafor, C.O., 2020. Social networks as a mechanism for discrimination

Reading period begins April 25 (Th)
Reading period ends May 1