CSE 493s/599s
Lecture 1

Jamie Morgenstern

Course Staff - Instructors

Jamie Morgenstern
Professor in CSE

Rachel Hong
PhD student

Bernie Zhu
PhD student

Contact: cse493s-staff@cs.washington.edu

Website: https://courses.cs.washington.edu/courses/cse493s/25au/

- Logistics
- Course outline
- Uncertainty motivation
- Introduction to learning theory

- Logistics
- Course outline
- ML research examples
- Introduction to learning theory

Basics

- Lectures
  - JHN100
  - Tuesdays / Thursdays 11:30AM-12:50 PM
- Website: https://courses.cs.washington.edu/courses/cse493s/25au/
  - Announcements
  - EdStem
  - Materials (lecture slides, deadlines, OHs)

Communication Channels

- **Announcements, questions about class, homework help**
  - EdStem (https://edstem.org/)
  - “I think there is a typo in the homework?”
  - “What does this notation mean?”
  - “Is this an accurate description of how this works?
- **Personal concerns (**cse493s-staff@cs.washington.edu)
  - “Was in hospital…”, “Laptop was stolen…”
- **Office hours**
  - Regular office hours + homework OH + project OHs.
  - Check website for updates
- **Regrade requests**
  - Directly submit on Gradescope
- **Anonymous feedback** (https://feedback.cs.washington.edu/)
  - “Your real-world example X lacked nuance. I would like you to…”

Prerequisites

- Familiarity with:
  - Linear algebra
    - Linear dependence, rank, linear equations
  - Multivariate calculus
  - Probability and statistics
    - Distributions, densities, marginalization, moments
  - Algorithms
    - Basic data structures, complexity

- Contact us if you feel like you need additional review materials!

Course Registration

- All CSE course registration processes are managed centrally by CSE.
- Resources:

- https://www.cs.washington.edu/academics/ugrad/advising/
- https://www.cs.washington.edu/academics/phd/advising

Lectures

- Will be broadcast on Zoom (please let us know if this s not the case).
- Will be recorded and posted shortly after class.
  - Find Zoom links and videos in Canvas—>Zoom.
- But most lectures will be on the white board.
- In-person attendance is **highly** encouraged.

Grading

- Two homeworks
  - HW1: theory-oriented homework tentatively due week 3
  - HW2: a second homework, likely theory-based, due week 7 (?)
  - Collaboration okay. You must write, submit, and understand your answers.
  - Do not Google for answers or ask chatGPT to do it.
  - Submit to Gradescope. Regrade requests on Gradescope.

- For CSE493s students
  - 25% homework 1
  - 25% homework 2
  - 50% final project

- For CSE599s students
  - 20% homework 1
  - 20% homework 2
  - 10% reading assignment
  - 50% final project

Project

- 3 project milestones
  - Proposal: Thursday, October 17, 2025 (each team up to 4 people)
  - Version 1: Thursday November 6, 2025, 11:59 PM
  - Final version
    - 5 minute presentation in exam slot, Dec 5th Thursday
    - Final report due Tue Dec 9, 2025

Project

- The project for this course is designed to give you an opportunity (i) to engage with the current state of machine learning research and (ii) to contribute useful knowledge to this community. Your team will choose a direction from the list below and develop a project based on recent research:
  - Replication of recent work
  - Summarizing a line of theoretical work, for instance:
    - Summarize a line of work that aims to provide a theoretical explanation for an empirical phenomenon, e.g.:
      - i. Neural networks generalize despite being a large hypothesis class
      - ii. Learning of neural networks succeeds despite the loss function being non-convex
      - iii. Mode connectivity in the loss landscape of neural networks
      - iv. The “lottery ticket” hypothesis (related to initialization of neural networks)
    - Summarize a line of work that develop a new algorithm (e.g. extensions to SGD optimizers for large mini-batch sizes)
  - Original research, such as:
    - Proposing and evaluating a new idea on top of an existing code base
    - Your own research project, if relevant

- Logistics
- Course outline
- Motivating understanding uncertainty
- Introduction to learning theory

Course outline

- Two parts:
  - Theoretical foundations (9 lectures)
    - Guiding principle: **generalization** and **empirical risk minimization**
    - We study both **statistical** aspects (generalization) and **algorithmic** aspects (optimization)
  - Topics in uncertainty and dynamic systems
    - Goal: understand what probabilities mean, how to interpret predictions and uncertainty
    - What sorts of behavior should we expect in systems that change (and whose environments change)?

- Logistics
- Course outline
- Motivating understanding uncertainty
- Introduction to learning theory

Questions we will ask

- When can we make forecasts that "look like" real probabilities in various ways, and how is this useful?
- How can we quantify how sure we are about specific predictions?
- When do these guarantees continue to hold when we zoom in on particular subsets of the data?
- When can we safely make decisions downstream of our predictions by treating probabilistic predictions as if they were correct? i.e. when can we "trust" predictions?
- What are minimal assumptions we need to make about data generating processes in order to get these guarantees?

Slides adapted from Aaron Roth

How can we make sense of probabilities?

- "What is the probability that if I flip a fair coin 16 times I get exactly 9 heads?"
  - We have a mathematical model that maps well onto reality; we can compute this in closed form.
  - We can also conduct the experiment repeatedly and empirically estimate.

Slides adapted from Aaron Roth

How can we make sense of probabilities?

- **“What is the probability that Kamala Harris wins the 2024 US Presidential election?”**
  - **If we posit a probabilistic model of the universe, this is perhaps philosophically coherent, but it is not a repeatable event; we can’t get empirical estimates.**

Slides adapted from Aaron Roth

How can we make sense of probabilities?

- **"ChatGPT says that the leading cause of death globally in children under five years old is Diarrheal diseases. What is the probability that it is correct?"**
  - **This is either correct or not, and (with some research) is knowable. But it would still be useful to have "confidence scores" for LLMs…**

Question	Answer
The limbic system plays an important role in	regulating emotion.
According to aristotle, happiness is:	activity of the soul in accordance with virtue.
When we speak of time dilation, we mean that time	is related to space.
What is the biggest cause of death in children under five years old (as of 2017)?	Diarrheal diseases

Slides adapted from Aaron Roth

Individual Probabilities Pervade ML

- In the practice of ML and statistics we frequently refer to individual probabilities:
  - "The probability that it will rain tomorrow" (weather forecasting)
  - "The probability that Alice will die in the next 12 months" (life insurance)
  - "The probability that Bob will be arrested for a violent crime 18 months after release on parole" (recidivism prediction)
  - "The probability that Carol will develop breast cancer before the age of 50" (predictive medicine)
  - …

Slides adapted from Aaron Roth

The measurement problem

- Individual probabilities refer to a model of the world in which there is some distribution D over feature/outcome pairs $(x, y) \in X \times \{0,1\}$.
  - Within the model, an individual probability of an outcome for an individual with observable features x is: $p(x) = \Pr_D[y = 1 \mid x]$,
  - Model is consistent with determined outcomes (e.g. $p(x) \in \{0,1\}$)
- The Basic Problem:
  - We observe each individual at most once, and so cannot measure individual probabilities.
  - All we can measure are averages over sufficiently large sets $S \subseteq X$:
  - $p(S) = \Pr_D[y = 1 \mid x \in S]$
    - (The more data we have, the smaller the sets S we can estimate $p(S)$ for).
- Given this, what are we to make of individual probabilities?

Slides adapted from Aaron Roth

Can't even distinguish a randomized from a deterministic world.

- A toy weather prediction problem: $x \in X$ a rich feature space (never see the same x twice), and $y \in \{0,1\}$.
  - Two models:
    - For every x, $\Pr[y = 1 \mid x] = \frac{1}{2}$
    - For a uniformly random subset $S \subseteq X$,
      $\Pr[y = 1 \mid x \in S] = 1, \Pr[y = 1 \mid x \notin S] = 0$
  - These two models are observationally equivalent.
  - So "learning individual probabilities accurately" is (unfortunately) not an achievable goal without making strong assumptions.

Slides adapted from Aaron Roth

Calibration: Level-0 Consistency

- What is the probability of heads at each period t?
- HTTHHTHHTTHTHTHHTT
- $p_t = \frac{1}{2}$ seems reasonable for all t. (The empirical frequency of heads conditional on prediction $p_t = \frac{1}{2}$ is $\frac{1}{2}$...
- Calibration: For all values v, the empirical frequency of heads conditional on prediction $p_t = v$ is v.

Slides adapted from Aaron Roth

Multicalibration: More Consistency Constraints

- What is the probability of heads at each period t?
- HTHTHTHTHTHTHTHTHTHTHTHT
- $p_t = \frac{1}{2}$ is still calibrated but seems less reasonable.
- Why? Not calibrated on the subsequences of even periods and odd periods.
- Multicalibration: Simultaneous calibration subject to some set of (not necessarily disjoint) conditioning events.

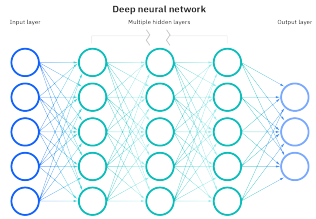
Slides adapted from Aaron Roth

Beyond Means

- **Calibration is about predicting averages, but can also ask for collections of consistency constraints for other losses/distributional quantities.**
- **E.g. quantiles are useful for uncertainty quantification.**
- **General philosophy: If predictions are indistinguishable from real probabilities with respect to a certain class of tests, then they are as good as real probabilities for any application that only interacts with them using tests from that class.**

Slides adapted from Aaron Roth

Prediction Sets



- **We have many black box methods that can make predictions.**
- **Which predictions should we feel confident in?**
- **Need a way of quantifying uncertainty. One way: prediction sets.**

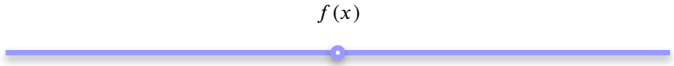
Image from Angelopoulos, Bates, Malik, Jordan ICLR 2021

Goal: The prediction set should contain the true label with probability (e.g.) 95%.

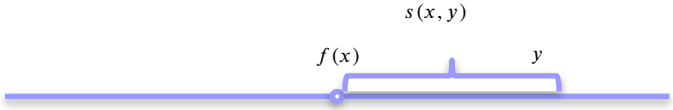
Slides adapted from Aaron Roth

Conformal Prediction [Shafer, Vovk]

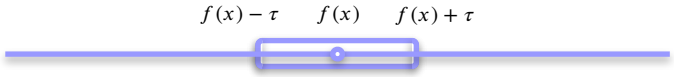
- Simple, elegant method to affix prediction sets to black box models.
- Pick an arbitrary model $f: X \to Y$ for making point predictions.
  - e.g. a regression model $f: X \to \mathbb{R}$.



- Pick a non-conformity score $s: X \times Y \to R$ that measures how different a label $\hat{y}$ is from the predicted label $f(x)$.
  - e.g. $s\left(x, \hat{y}\right) = |f(x) - \hat{y}|$



- On a holdout set $D_H \sim \mathcal{D}^n$, find the smallest τ such that $\Pr_{(x,y) \in D_H}\left[s\left(x, y\right) \leq \tau\right] \geq 0.95$
- On new examples x, produce the prediction set:
- $P(x) = \{\hat{y} : s\left(x, \hat{y}\right) \leq \tau\}$



- Promise: If $\left(x, y\right) \sim D$, then $\Pr\left[y \in P(x)\right] \geq 0.95$
- (A marginal guarantee: Probability is over the randomness of x, y, D_h...)

Slides adapted from Aaron Roth

What's wrong with marginal guarantees?

Given your features x, our model predicts your blood oxygen level in 24 hours will be $f(x)$

How sure are you?

I have a 95% prediction interval that your blood oxygen level will be in $|\ell(x), u(x)|$

Hmmm…

Slides adapted from Aaron Roth

What's wrong with marginal guarantees?

Ideally

$$\Pr_y\left[y \in [\ell(x), u(x)] \mid x\right] = 0.95$$

Randomness is entirely over the unrealized/unmeasured randomness of the world, conditional on all of your observable attributes.

Conformal Prediction

Marginal Coverage: $\Pr_{(x,y)}[\,|f(x) - y| \leq \tau] \geq 1 - \delta$

Randomness is averaging over people.

Slides adapted from Aaron Roth

Marginal Guarantees.

$|\ell(x), u(x)|$ is a 95% marginal prediction interval.

But I'm part of a demographic group representing less than 5% of the population…

For disjoint groups, could just calibrate separately on [Romano, Foygel-Barber, Sabatti, Candes '20]

But…

Slides adapted from Aaron Roth

Marginal Guarantees.

What about for people like me?

For African Americans under the age of 50 the 95% prediction interval is [a,b]

For women with a family history of diabetes the 95% prediction interval is [c,d]

For people with egg allergies and no history of smoking, the 95% prediction interval is [e, f].

What does this mean for me?

Slides adapted from Aaron Roth

Multicalibrated quantile estimation can give:

The prediction sets $P(x, f) = \{\hat{y} : s\left(x, \hat{y}\right) \leq f(x)\}$ satisfy group conditional coverage with respect to a set of groups G if for every $g \in G$:

$$\Pr_{(x,y)}\left[y \in P(x, f) \,\middle|\, x \in g\right] = 0.95$$

Slides adapted from Aaron Roth

Distributional Assumptions

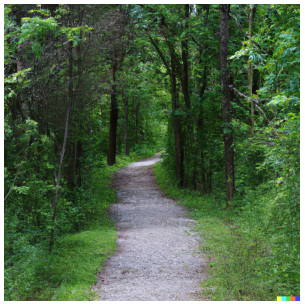
- Standard Assumption for Conformal Prediction:
  - Exchangeability (e.g. iid data): The future must look like the past.
- But this is often violated:

Distributional Assumptions

- **We will see how much we can do with no assumptions on the data generating process: data arrives sequentially, selected by an adversary.**
- **It will turn out we can do quite a lot --- close connections with game theory and the minimax theorem.**

Slides adapted from Aaron Roth

A (Rough) Roadmap

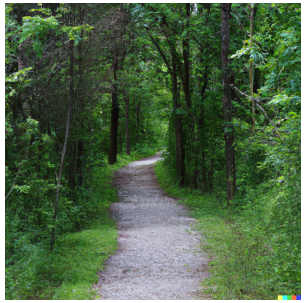


- **Introduction to classical learning theory, and the guarantees it provides**
- **Simple Marginal Guarantees**
  - **Marginal Mean Consistency improves squared error**
  - **Marginal Quantile Consistency improves pinball loss**
  - **Marginal Quantile Consistency is already useful for conformal prediction.**

From marginal guarantees to calibration
Mean calibration: Marginal mean consistency conditional on mean prediction.
Quantile calibration: Marginal quantile consistency conditional on quantile prediction
Why calibration?
…

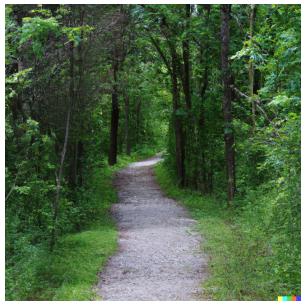
Slides adapted from Aaron Roth

A (Rough) Roadmap



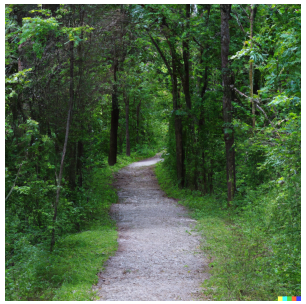
Adding more conditioning events.
Group Conditional Mean Calibration
(Gradient Descent on Squared Loss)
Group Conditional Quantile Calibration
(Gradient Descent on Pinball Loss)
Decision Calibration and (Swap) Regret

A (Rough) Roadmap



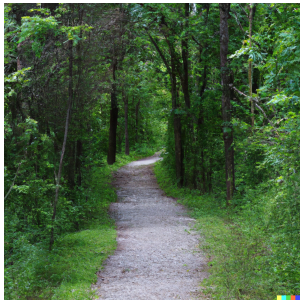
Interlude! Zero Sum Games and the Minimax Theorem
Intro to Zero Sum Games
Proving the Minimax Theorem via Sequential Decision Making
Computing Minimax Strategies via Sequential Decision Making
Application of the Minimax Theorem: Passing any empirical test in an adversarial environment.
Existential proof of sequential calibration algorithms with guarantees against adversaries --- and much more.

A (Rough) Roadmap



Multiobjective Sequential Learning:
A General Game Theoretic Framework for Multiobjective Learning
Efficiently Controlling Regret Across Multiple Subsequences
Efficiently Making (Multi)calibrated Predictions Against an Adversary

A (Rough) Roadmap



Some Applications:
Predicting for many loss functions (“Omniprediction”)
Accuracy under distribution shift (“Universal Adaptability”)
Ensembling and Agreement
Resolving the “Reference Class Problem at Scale”
Generalizations of Aumann’s Agreement Theorem
Boosting for Squared Error Regression
More?

- Logistics
- Course outline
- Motivating uncertainty as a topic
- Introduction to learning theory