COMP 677: Seminar in Learning Theory

Lecture 1

Maryam Aliakbarpour

Fall 2023
Today’s lecture

• Introduction
• Class format
• Policies
• Introduction to the topic
Introduction

Instructor: Maryam Aliakbarpour

Email: maryama@rice.edu

Office hour: By appointment (email me)

Lectures: Wednesdays 4-5pm, Duncan Hall 1075

Website: https://maryamaliakbarpour.com/courses/23F/seminar.html + Canvas

Your turn!
Class objectives

Studying fundamental problems in learning theory from a new perspective:

• Computational aspects: limited time or memory
• Societal aspects: privacy and fairness

We will return to this!

Practicing research soft skills:

• How to approach a problem
• How to review / write a paper
• Presenting technical material
Class Prerequisites

- solid understanding of mathematical proofs
- basic algorithms, and probability
- A graduate level course in algorithms or machine learning is recommended.
Class format

• In each class, we focus on one paper.

• Before class:
  • Reading assignment: read the paper
  • Provide a review on canvas

• Presentation:
  • A student presents the paper (45 min presentation)

• Questions / Discussion
Class format

- A list of suggested papers: Syllabus

- You may also pick papers that are not listed but are relevant to the topic of the class.

- Pick two* papers.

- Fill out this form by this Monday: https://forms.gle/Qu3duqfyc1QoY5Dp9

- First presenter? (By Friday)
Class format: presentation

A 45-minute long presentation:

• Introduction: What and why?
• Related work
• Problem definition
• Solution
• Technical part*
Class format: presentation

Practice your talk! (many times)

(Optional) Meet with me on Friday or Monday before your presentation.

• Set an appointment (maryama@rice.edu)
Class format: reading assignment

Read the paper before class, and be present.

Think of it as a mini-review.

Canvas assignment:

• Summary of the paper.

• Your opinion: Strengths / Limitations. Next steps?
Class format: class project

Only if you register for 3-credit

Two options:

• Survey of results
• Research project
Policies

Read Syllabus

• An inclusive environment
• Rice Honor Code
• Disability Resource Center
• Wellbeing and Mental Health
• Title IX Responsible Employee Notification
Our topic
Our daily activities produce vast amounts of data.
Our daily activities produce vast amounts of data.

How can we extract meaningful information?
Statistical inference

Data:
samples from $D$

$x_1, x_2, \ldots, x_m$

Algorithm

Information about $D$

Image from: https://tilics.dmi.unibas.ch/the-turing-machine
Statistical inference

Data: samples from $D$

$x_1, x_2, \ldots, x_m$

What do you want to know?

Estimation:
Estimate parameters of distribution
e.g. mean, variance

Testing:
Test distribution $D$ has a specific property
e.g. uniformity, unimodal

Learning:
Learn distribution $D$ in a class
e.g. Gaussians

Classification:
Learn a classifier from labeled data
e.g. learning half-spaces
New goal: understanding the relationship between all of these aspects

Use as few data points as possible

Sample complexity
# data points

Accuracy
Dependencies on the error parameter

Computational aspect:
Memory / Time

Societal aspect:
Privacy / Fairness
Statistical inference

Data: samples from \( D \)
\[ x_1, x_2, \ldots, x_m \]

Algorithm with limited memory
limited time
private
fair

Information about \( D \)

Image from: https://tilics.dmi.unibas.ch/the-turing-machine
This talk

Part I: Inference with privacy

Part II: Inference with limited memory
Sensitive data requires privacy preserving algorithms.
Privacy

- Learn about community, but not individuals
- Anonymization \(\neq\) not-identifiable
- Global information leaks information about individuals!

Example: Average net worth of patients in oncology
Differential privacy

- Mathematical formulation
- Not ambiguous
- Irrefutable claims
- Extensive use in practice:
  - Apple, Google, US census
Differential privacy (central)

Dataset → Processing via trusted server → Output
Differential privacy

Output should not depend on a single data point.

Dataset

Bob

Dataset

Alice

Output stays similar.
Differential privacy

$\epsilon$-differentially private algorithm $A$:

- Any possible output $Y$
- Two neighboring datasets $X, X'$ s.t. they differ in one sample

$$\Pr[A(X) = Y] \leq e^\epsilon \Pr[A(X') = Y]$$

[Diur and Nissim’03, Dwork, McSherry, Nissim, and Smith’06, Dwork’06]
Laplace Mechanism

For two neighboring datasets $X, X'$ such that $|X - X'| = 1$, the sensitivity of $f$ is:

$$\Delta f \triangleq \max_{X, X'} |f(X) - f(X')|$$

Can make $f$ a $\xi$-differentially private function by adding Laplace noise to it.

Function $f(X)$  

$\xrightarrow{\text{Laplace noise}}$  

$\tilde{f}(X)$

$+ \text{Lap}(\Delta f / \xi)$
This talk

Part I: Inference with privacy

Part II: Inference with limited memory
Why limited memory?

- Size of working memory < size of data
- Facilitates communication and processing of distributed data
- Insightful: what summarizes the data
Memory restriction can affect learning drastically!

- [Raz, FOCS. 2016] Parity learning problem
- [Chien, Ligett, McGregor. ITCS 2010] Robust statistics and distribution testing
- [Diakonikolas, Gouleakis, Kane, Rao. COLT 2019] Distribution testing
- [Sharam, Sidford, Valiant. STOC 2019] Memory-Sample Tradeoffs for Linear Regression
- [Brown, Bun, Smith. COLT 2022] Memory lower bounds for sparse linear predictors

And many more…
Memory restriction can affect learning drastically!

[Raz’16]: Fast learning requires good memory!

Parity learning problem:
- Goal: find \( w \in \{0,1\}^n \)
- Samples: a random \( x \in \{0,1\}^n \) and \( w \cdot x \)

By Gaussian elimination

\( O(n^2) \) bits of memory

\( O(n) \) samples

[Raz’16]: Any algorithm using

\[ \leq \frac{n^2}{25} \] bits of memory

needs exponentially many samples
Example I: Private Hypothesis Testing

Joint work with Daniel Kane (UCSD), Ilias Diakonikolas (UW Madison), Ronitt Rubinfeld (MIT)
Hypothesis testing

Does $D$ have a particular property or not?

Distribution $D$ \Rightarrow x_1, x_2, \ldots, x_m$ \Rightarrow Algorithm \Rightarrow Accept or Reject
Hypothesis Testing

Clinical trials:
- Treatment efficacy

Applications
- Social sciences:
  - Correlation between gender and income
- E-commerce:
  - Efficacy of a new ad. strategy
- Technical property
  - Mixture of Gaussians
Sensitive data requires privacy preserving algorithms.
Goal:
Design testing algorithms:
- **Accurate**
- **Optimal** number of data points
- **Privacy** preserving

**Active area of research:** [Rogers, Roth, Smith, Thakkar’16], [Gaboardi, Lim, Rogers, Vadhan’16], [Cai, Daskalakis, Kamath’17], [A, Diakonikolas, Rubinfeld’18], [Acharya, Sun, Zhang’18]: [Bun, Kamath, Steinke, Wu’19], [Canonne, Kamath, McMillan, Smith, Ullman’19], [Canonne, Kamath, McMillan, Ullman, Zakynthinou’20], [Vepakomma, Amiri, Canonne, Raskar, Pentland’22]
Our problem:

Closeness testing:
Are two distributions equal?
Example: treatment efficacy

Closeness testing:
Are two distributions equal?

Pain level after treatment: 2, 10, 3, 1, 2, 9, 3, 1

Pain level in the control group: 6, 2, 7, 2, 3, 6, 2, 3
Example: treatment efficacy

Closeness testing:
Are two distributions equal?

Number of sold items per day: 2, 10, 3, 1, 2, 9, 3, 1

Number of sold items after price drop: 6, 2, 7, 2, 3, 6, 2, 3
Our problem: closeness testing

Distribution $p$ over $[n]$ \hspace{4cm} iid samples: $x_1, x_2, \ldots, x_s$

Distribution $q$ over $[n]$ \hspace{4cm} iid samples: $y_1, y_2, \ldots, y_s$

Tester

Output = \begin{align*}
\text{Accept} & \quad \text{if } q = p \\
\text{Reject} & \quad \text{if } p \text{ and } q \text{ are } \alpha\text{-far} \\
& \quad \text{in } \ell_1\text{-distance}
\end{align*}

[Batu, Fortnow, Rubinfeld, Smith, White'00]
Closeness Testing

- Testing k-histograms
- Independence $p = p_1 \times p_2$
- Mixture testing
- Uniformity $p = \text{uniform}
- Identity Known $p = q$
- Closeness Unequal sized sample sets
Closeness testing implies independence testing

\[(X, Y) \sim p.\]

**Question:** Are \(X\) and \(Y\) independent?

\[
\begin{align*}
X \text{ and } Y \text{ are independent} & \iff p = p_1 \times p_2 \\
X \text{ and } Y \text{ are far from being independent} & \iff |p - p_1 \times p_2|_1 \geq \Theta(\alpha)
\end{align*}
\]

[Batu, Fischer, Fortnow, Kumar, Rubinfeld, White’01]
Our results

- New flattening-based (FB) private tester for closeness testing

- Characterizing the non-private reductions that results in private testers automatically

- Private testers for other properties
  - Testing Property $P$
    - e.g., independence
  - Reduction
  - FB Closeness testing
  - Differentially private

[Non-private tester by [Diakonikolas, Kane’16]

[A, Diakonikolas, Kane, Rubinfeld NeurIPS19]
Our results

New flattening-based (FB) private tester

Why this tester?
- Exploits the underlying structure of distributions
- Only known optimal results for some problems

[A, Diakonikolas, Kane, Rubinfeld NeurIPS19]
Our result on closeness: privacy is almost free!

Theorem

There exists a $\epsilon$-private algorithm for testing closeness of two distributions $p$ and $q$ over domain of $[n]$ with error parameter $\alpha$ that uses

$$O\left(\frac{n^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{n}}{\alpha^2} + \frac{\sqrt{n}}{\alpha \sqrt{\epsilon}} + \frac{1}{\alpha^2 \epsilon}\right)$$

samples from $p$ and $q$. 

- Non-private cost
- Cost of privacy

[A, Diakonikolas, Kane, Rubinfeld’19]
Our results on other properties

- New $\epsilon$-DP tester for independence (domain = $[n] \times [m]$ when $m \leq n$)

\[
0\left(\frac{n^{2/3}}{m^{1/3}} \frac{1}{\alpha^{4/3}} + \sqrt{n} \frac{m}{\alpha^2} + \sqrt{n} \frac{m \log n}{(\alpha \epsilon)} + \frac{1}{(\alpha^2 \epsilon)}\right)
\]

  Non-private cost  
  Cost of privacy

- New $\epsilon$-DP tester for testing closeness with unequal sized samples

- Tighter result for closeness/uniformity/identity
Techniques
How? Simple approach

\[ \Omega \left( \frac{n}{\alpha^2} \right) \text{ samples} \]

Too much data!
Sub-linear?

An alternative way:

Statistic \( Z := \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i \)

\[
\begin{align*}
\text{if } p = q & \quad \text{Small } Z \\
\text{if } |p - q|_1 \geq \alpha & \quad \text{Large } Z
\end{align*}
\]
Sub-linear? Potential solution

Statistic: $Z := \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i$

Sample complexity = $\Omega \left( \frac{n \cdot \max(|p|_2, |q|_2)}{\alpha^2} \right) \propto \max \ell_2$-norm of $p$ and $q$

\[ p = q \implies p' = q' \]
\[ |p - q|_1 \geq \alpha \implies |p' - q'|_1 \geq \alpha \]

No change in $\ell_1$-distance!
How flattening reduces $\ell_2$-norm

On a new domain

Distribution $p$ → Detecting large elements → Distribution $p'$

How? Draw samples and see frequencies

$E[|p'|_2^2] < \frac{1}{|F|}$

Flattening Samples $F$: □ □ □ □ □ □ □ □ □ # bins = frequency in $F$ + 1

[Diakonikolas, Kane’16]
Testing closeness via flattening

Flattening samples

Flattening: Creates a mapping

Test samples from $p$ and $q$

Maps to samples from $p'$ and $q'$

Test $p'$ and $q'$

[Diakonikolas, Kane’16]
Flattening technique: strong, but sensitive...

Not easy to privatize

Hard to make it private!

Flattening samples:

Distribution $p'$

Very different $Z$
Noise make statistics similar

Find a more stable $Z$

Flattening samples:  \[ \hat{Z}_1 \]

Flattening samples:  \[ \hat{Z}_2 \]

Higher difference of $Z$’s

More noise
Noise make statistics similar

Flattening samples:

Find a more stable $Z$

+ noise

$\hat{Z}_1$

$\hat{Z}_2$

Similar

Higher difference of $Z$’s

More noise
Flattening samples

High sensitivity

Sample set $X$

Sample set $X'$

Flattening samples

Test samples

High sensitivity
Flattening samples

Test samples

Sample set $X$

Sample set $X'$

Not too high sensitivity
Our algorithm: derandomization

- Try all partitions for flattening and test samples
- Compute the mean of statistics

New statistic: \( \overline{Z} := E_\pi[Z] \)
Proof sketch: Why $\bar{Z}$ works

- Accuracy
- Privacy guarantee
- Efficiency: number of samples and time
Proof sketch: Why $\bar{Z}$ works

- **Accuracy**
- **Privacy guarantee**
- **Efficiency: number of samples and time**

- Not independent trials of the algorithms
- Flattening guarantees only worked in average
  - Requires a new analysis
Proof sketch: Why $\overline{Z}$ works

- Accuracy
- Privacy guarantee
- Efficiency: number of samples and time

- Analyze how $\overline{Z}$ changes after changing one sample
- Add noise to hide the change
- Does noise affect accuracy?
Proof sketch: Why $\bar{Z}$ works

- **Accuracy**
- **Privacy guarantee**
- **Efficiency: number of samples and time**

- Exponential time
- Alternative approach with linear time in sample size
Our result on closeness: privacy is almost free!

**Theorem** [A, Diakonikolas, Kane, Rubinfeld’19]

There exists a $\epsilon$-private algorithm for testing **closeness** of two distributions $p$ and $q$ over domain of $[n]$ with error parameter $\alpha$ that uses

$$O\left(\frac{n^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{n}}{\alpha^2} + \frac{\sqrt{n}}{\alpha \sqrt{\epsilon}} + \frac{1}{\alpha^2 \epsilon}\right)$$

samples from $p$ and $q$.  

- **Non-private cost**
- **Cost of privacy**