Workshop on the Mathematical Foundations of Privacy, Fairness, and Adaptive Data Analysis

Israel Data Science and AI Initiative 3rd Annual Conference

Schedule

Tuesday, 4 June 2024		
Time	**Speaker**	**Title**
10:00-10:40	Shay Moran	What Is The Sample Complexity of Differentially Private Learning?
10:40-10:55	Hilla Schefler	A General 'Private Learning Implies Online Learning' Theorem
10:55-11:10	Bogdan Chornomaz	Topological barrier to replicable learning
11:10-11:35	**Coffee Break**	
11:35-12:15	Yishay Mansour	A Theory of Interpretable Approximations
12:15-12:30	Tal Herman	Verifying The Unseen: Interactive Proofs for Distribution Properties
12:30-12:45	Tomer Shoham	Differential privacy and hypothesis testing - from parametric testing to ECDFs
12:45-13:00	Moshe Shenfeld	Understanding Generalization via a Bayes Factor
13:00-14:20	**Lunch**	
14:20-15:00	Edith Cohen	Hot PATE: Private Aggregation of Distributions for Diverse Tasks
15:00-15:20	Adi Haviv	Not Every Image is Worth a Thousand Words: Quantifying Originality in Stable Diffusion
15:20-16:00	Ran Canetti	Co-Designing the Pilot Release of Israel's National Registry of Live Births: Reconciling Privacy with Accuracy and Usability

Abstracts

Shay Moran : What Is The Sample Complexity of Differentially Private Learning?

How much data is needed for differentially private learning?

How much more data does private learning require compared to learning without privacy constraints?

We will survey some of the recent progress towards answering these questions in the distribution-free PAC model, including the Littlestone-dimension-based *qualitative* characterization and the relationship with online learning.

If time allows, we will also discuss this question in more general (distribution- and data-dependent) learning models.

Hilla Schefler: A General 'Private Learning Implies Online Learning' Theorem

In this talk we will present a recent result regarding the link between differentially private (DP) and online learning. Alon, Livni, Malliaris, and Moran (2019) showed that for binary concept classes, DP learnability of a given class implies that it has a finite Littlestone dimension (equivalently, that it is online learnable). Their proof relies on a model theoretic result by Hodges (1997), which demonstrates that any binary concept class with a large Littlestone dimension contains a large subclass of thresholds. In a followup work, Jung, Kim, and Tewari (2020) extended this proof to multiclass PAC learning with a bounded number of labels. Unfortunately, Hodges's result does not apply in other natural settings such as multiclass PAC learning with an unbounded label space, and PAC learning of partial concept classes.

This naturally raises the question whether DP learnability continues to imply online learnability in more general scenarios such as partial concept classes and general multiclass setting. In this work, we give a positive answer to these questions showing that for general classification tasks, DP learnability implies online learnability. Our proof reasons directly about Littlestone trees, without relying on thresholds.

This talk is based on joint work with Simone Fioravanti, Steve Hanneke, Shay Moran, and Iska Tsubari.

Bogdan Chornomaz: Topological barrier to replicable learning

Replicable/stable learning is intrinsically connected to the topological properties of a certain topological space, arising from the underlying concept class. We use this connection to derive an obstruction to replicability using a local version of the famous Borsuk-Ulam theorem.

Yishay Mansour: A Theory of Interpretable Approximations

Can a deep neural network be approximated by a small decision tree based on simple features? This question and its variants are behind the growing demand for machine learning models that are *interpretable* by humans. In this work we study such questions by introducing "interpretable approximations", a notion that captures the idea of approximating a target concept c by a small aggregation of concepts from some base class H.

In particular, we consider the approximation of a binary concept c by decision trees based on a simple class H (e.g., of bounded VC dimension), and use the tree depth as a measure of complexity.

Our primary contribution is the following remarkable trichotomy. For any given pair of $\mathcal{H}$ and c, exactly one of these cases holds:

(i) c cannot be approximated by $\mathcal{H}$ with arbitrary accuracy;

(ii) c, can be approximated by $\mathcal{H}$ with arbitrary accuracy, but there exists no distribution-free upper bound on the complexity of the approximations; or

(iii) there exists a constant κ that depends only on $\mathcal{H}$ and c such that, for *any* data distribution and any desired accuracy level, c can be approximated by H with a complexity not exceeding κ.

This taxonomy stands in stark contrast to the landscape of supervised classification, which offers a complex array of distribution-free and universally learnable scenarios. Our theory shows that, in the case of interpretable approximations, even a slightly nontrivial a-priori guarantee on the complexity of approximations implies approximations with constant (distribution-free and accuracy-free) complexity.

Our results also include extensions of the trichotomy to classes $\mathcal{H}$ of unbounded VC dimension and characterizations of interpretability based on the algebras generated by $\mathcal{H}$.

This is a joint work with Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Shay Moran and Maximilian Thiessen.

Tal Herman: Verifying The Unseen: Interactive Proofs for Distribution Properties

Learning about an unknown distribution using i.i.d. samples is a basic challenge in many scientific disciplines. There are many tasks we could consider: from estimating statistical quantities, such as the distribution's support size or Shannon entropy, to supervised or unsupervised learning. In many cases, these tasks require drawing a large number of samples from the distribution and significant computational resources.

How about *verifying* the results of such an analysis? Can an untrusted analyst provide a proof of correctness for the results? We are interested in proofs that can be verified using fewer resources (samples and running time) than what is needed for performing the task on one's own.

The talk will shortly survey a line of recent works on constructing interactive proof systems for verifying properties of unknown distributions.

Joint work with Guy Rothblum.

Tomer Shoham: Differential privacy and hypothesis testing - from parametric testing to ECDFs

Hypothesis testing plays a central role in statistical inference and is used in many settings where privacy concerns are paramount. When noise is introduced to ensure differential privacy, analysts can adopt various strategies. One intuitive approach is to retain the original model and adjust either the confidence level, the test structure, or the statistic (query). However, this approach often leads to suboptimal results due to the sensitivity of the statistic. A more effective strategy involves rethinking how to test the hypothesis in light of the added perturbation noise and sampling error.

In my talk, I will highlight key aspects of hypothesis testing and differential privacy based on my past publications and current work. Key takeaways include the importance of adjusting queries to enhance statistical inference ("don't ask for what you need"), the potential for non-parametric statistics to succeed where parametric statistics fail, and the crucial role of the cumulative distribution function (CDF) under the appropriate semi-metric.

Moshe Shenfeld: Understanding Generalization via a Bayes Factor

Learning (as machine learning practitioners call it) or estimating (as statisticians would say) is the act of modeling population characteristics from properties observed in a dataset. Success hinges on avoiding overfitting and instead achieving a small generalization gap, meaning the model's fit to the data extends to similar performance on a broader population.

Using stability notions initially developed in the context of differential privacy and adaptive data analysis, we demonstrate that the algorithmic stability (and by extension, the generalization gap) equals the covariance between a model's loss with respect to the data and a Bayes Factor term reflecting the sample-specific information encoded by the learned model. This observation yields new, straightforward proofs of various information-based techniques for ensuring generalization.

Based on joint work with Katrina Ligett.

Edith Cohen: Hot PATE: Private Aggregation of Distributions for Diverse Tasks

The Private Aggregation of Teacher Ensembles (PATE) framework is a versatile approach to privacy-preserving machine learning. In PATE, teacher models that are not privacy-preserving are trained on distinct portions of sensitive data. Privacy-preserving knowledge transfer to a student model is then facilitated by privately aggregating teachers' predictions on new examples. Employing PATE with large language models presents both challenges and opportunities. These models excel in open ended diverse (aka hot) tasks with multiple valid responses. Moreover, the knowledge of models is often encapsulated in the response distribution itself and preserving this diversity facilitates fluid knowledge transfer from teachers to student. In all prior designs, however, higher diversity resulted in lower teacher agreement and what appeared to be an inherent tradeoff between diversity and privacy.

We present a nuanced model for preserving diversity and propose hot PATE that allows the transfer of high diversity with no privacy penalty. We demonstrate empirically the benefits of hot PATE for in-context learning via prompts and potential to unleash more of the capabilities of generative models.

Adi Haviv: Not Every Image is Worth a Thousand Words: Quantifying Originality in Stable Diffusion

This work addresses the challenge of quantifying originality in text-to-image (T2I) generative diffusion models, with a focus on copyright originality. We begin by evaluating T2I models' ability to innovate and generalize through controlled experiments, revealing that stable diffusion models can effectively recreate unseen elements with sufficiently diverse training data. Then, our key insight is that concepts and combinations of image elements the model is familiar with, and saw more during training, are more concisly represented in the model's latent space. Hence, we propose a method that leverages textual inversion to measure the originality of an image based on the number of tokens required for its reconstruction by the model.

Our approach is inspired by legal definitions of originality and aims to assess whether a model can produce original content without relying on specific prompts or having the training data of the model. We demonstrate our method using both a pre-trained stable diffusion model and a synthetic dataset, showing a correlation between the number of tokens and image originality. This work contributes to the understanding of originality in generative models and has implications for copyright infringement cases.

Ran Canetti: Co-Designing the Pilot Release of Israel's National Registry of Live Births: Reconciling Privacy with Accuracy and Usability

On February 2024, Israel's Ministry of Health released microdata of live births in Israel in 2014. The dataset is based on Israel's National Registry of Live Births and offers substantial value in multiple areas, such as scientific research and policy-making. At the same time, the data was processed so as to protect the privacy of 2014's mothers and newborns. The release was co-designed by the authors together with stakeholders from both inside and outside the Ministry of Health. This paper presents the methodology used to obtain that release. It also describes the considerations involved in choosing the methodology and the process followed.

We used differential privacy as our formal measure of the privacy loss incurred by the released dataset. More concretely, we prove that the released dataset is differentially private with privacy loss budget of epsilon = 9.98. We extensively used the private selection algorithm of Liu and Talwar (STOC 2019) to bundle together multiple steps such as data transformation, model generation algorithm, hyperparameter selection, and evaluation. The model generation algorithm selected was PrivBayes (Zhang et al., SIGMOD 2014). The evaluation was based on a list of acceptance criteria, which were also disclosed only approximately so as to provide an overall differential privacy guarantee. We also discuss concrete challenges and barriers that appear relevant to the next steps of this pilot project, as well as to future differentially private releases.

Joint work with Shlomi Hod.