Mathematical Statistics in the Information Age

Book of Abstracts

Wednesday 17.09.

9:30 - 10:15 Pragya Sur (Harvard) - Data Integration: Challenges and Opportunities for Interpolation Learning under Distribution Shifts

Abstract: Min-norm interpolators naturally emerge as implicit regularized limits of modern machine learning algorithms. Recently, their out-of-distribution risk has been studied in settings where test samples are unavailable during training. However, many applications have access to limited amounts of test data during training. The properties of min-norm interpolation in this setting remain poorly understood. In this talk, I will present a characterization of the risk of pooled min-L2-norm interpolation under covariate and concept shifts. I will demonstrate that the pooled interpolator captures both early fusion and a form of intermediate fusion. Our results yield several important implications. For instance, under concept shift, adding data always degrades prediction performance when the signal-to-noise ratio is low. Conversely, for higher signal-to-noise ratios, transfer learning provides benefits as long as the shift-to-signal ratio remains below a threshold that we characterize. Additionally, our findings reveal that under covariate shift, when the source sample size is small relative to the dimension, heterogeneity between domains actually improves the risk. Time permitting, I will introduce a novel anisotropic local law that facilitates these characterizations and holds independent interest for both random matrix theory and the study of other heterogeneous data problems. This is based on joint work with Kenny Gu, Yanke Song, and Sohom Bhattacharya.

10:15 - 11:00 Guillaume Lecué (ESSEC Business School) - On the feature space decomposition for ridge estimators and minimum ell q-norm interpolant estimators.

Abstract: In this talk I will present a methodology called the feature space decomposition and apply it to the statistical analysis of ridge estimators and minimum l_q-norm interpolant estimators. We establish high-probability non-asymptotic upper bounds for the excess risk of ridge estimators, minimum l_q-norm interpolating estimators in linear regression for all q larger than 1, and minimum l_q-norm interpolating classifiers in linear classification. We obtain sufficient conditions for benign overfitting behavior on the latter interpolant estimators. Our results rely on a features space decomposition where the self-regularization properties of minimum norm interpolant estimator is highlighted. Technically, we circumvent the convex min-max theorem, instead employing tools from Geometric Aspects of Functional Analysis - including the Dvoretzky-Milman theorem, Gluskin's theorem, and lower bounds on Gaussian mean widths of random polytopes. This provides a geometric perspective on benign overfitting, and crucially, our techniques remain valid beyond the Gaussian case. Consequently, we obtain the benign overfitting results with high probability when the design vector is not necessarily Gaussian. We particularly emphasize that feature space decomposition may potentially refine the uniform convergence approach, suggesting its promise as a new fundamental methodology in mathematical statistics. Based on joint works with George Gavrilopoulos and Zong Shang.

11:30 - 12:15 Boaz Nadler (Weizmann Institute) - Robustness of OLS to sample removals: Theoretical analysis and implications

Abstract: For learned models to be trustworthy, it is essential to verify their robustness to perturbations in the training data. Recent studies have found that for various datasets, learned models may change significantly with the removal of even less than one percent of the training samples. This led several authors to propose a more stringent form of robustness, denoted robustness auditing, along with methods to estimate it. Instead of the classical notion of uncertainty via confidence intervals and bootstrap methods, robustness auditing assesses the stability to the removal of any subset of k samples from the training set. In this talk, I'll present a theoretical study of this form of robustness for ordinary least squares (OLS), under the following two settings: a general miss-specified model and the standard Gaussian linear model. Given n i.i.d. training samples, we derive non-asymptotic concentration bounds as follows: (i) for the general miss-specified model under mild regularity conditions, we prove that with high probability, OLS is robust to removal of any k$\ll$n samples; (ii) for the Gaussian linear model, we derive sharper results. We show that OLS can withstand removal of a much larger number of samples, and still remain robust and achieve the same error rate as OLS applied to the full dataset. In contrast, if k is proportional to n, then OLS is provably non-robust. Finally, we revisit prior analyses that found several econometric datasets to be highly non-robust to sample removals. While this appears to contradict our theoretical results, we demonstrate that the sensitivity is due to heavy-tailed responses and is substantially mitigated by classical robust methods, such as linear regression with a Huber loss.

14:00 - 14:45 Georg Köstenberger (University of Vienna) - Sharp oracle inequalities and universality of the AIC and FPE

Abstract: In two landmark papers, Akaike introduced the AIC and FPE, demonstrating their significant usefulness for prediction. In subsequent seminal works, Shibata developed a notion of asymptotic efficiency and showed that both AIC and FPE are optimal, setting the stage for decades-long developments and research in this area and beyond. Conceptually, the theory of efficiency is universal in the sense that it (formally) only relies on second-order properties of the underlying process, but, so far, almost all (efficiency) results require the much stronger assumption of a linear process with independent innovations. In this work, we establish sharp oracle inequalities subject only to a very general notion of weak dependence, establishing a universal property of the AIC and FPE. A direct corollary of our inequalities is asymptotic efficiency of these criteria. Our framework contains many prominent dynamical systems such as random walks on the regular group, functionals of iterated random systems, functionals of (augmented) Garch models of any order, functionals of (Banach space valued) linear processes, possibly infinite memory Markov chains, dynamical systems arising from SDEs, and many more.

15:15 - 16:00 Nina Dörnemann (Aarhus University) - Monitoring for a phase transition in a time series of Wigner matrices

Abstract: We develop methodology and theory for the detection of a phase transition in a time-series of high-dimensional random matrices. In the model we study, at each time point $t = 1, 2, \ldots$, we observe a deformed Wigner matrix $\mathbf{M}_t$, where the unobservable deformation

represents a latent signal. This signal is detectable only in the supercritical regime, and our objective is to detect the transition to this regime in real time, as new matrix-valued observations arrive. Our approach is based on a partial sum process of extremal eigenvalues of $\mathbf{M}_t$, and its theoretical analysis combines state-of-the-art tools from random-matrix theory and Gaussian approximations. The resulting detector is self-normalized, which ensures appropriate scaling for convergence and a pivotal limit, without any additional parameter estimation. Simulations show excellent performance for varying dimensions. Applications to pollution monitoring and social interactions in primates illustrate the usefulness of our approach. This talk is based on a joint work with P. Kokoszka (Colorado State University), Tim Kutta (Aarhus University) and Sunmin Lee (Colorado State University).

16:00 - 16:45 Wei Biao Wu (University of Chicago) - Concentration bounds for statistical learning for time dependent data

Abstract: Classical statistical learning theory primarily concerns independent data. In comparison, it has been much less investigated for time dependent data, which are commonly encountered in economics, engineering, finance, geography, physics and other fields. In this talk, we focus on concentration inequalities for suprema of empirical processes which plays a fundamental role in the statistical learning theory. We derive a Gaussian approximation and an upper bound for the tail probability of the suprema under conditions on the size of the function class, the sample size, temporal dependence and the moment conditions of the underlying time series. Due to the dependence and heavy-tailness, our tail probability bound is substantially different from those classical exponential bounds obtained under the independence assumption in that it involves an extra polynomial decaying term. We allow both short- and long-range dependent processes, where the long-range dependence case has never been previously explored. We showed our tail probability inequality is sharp up to a multiplicative constant. These bounds work as theoretical guarantees for statistical learning applications under dependence.

Thursday 18.09.

9:30 - 10:15 Kengo Kato (Cornell) - Inference with Gromov-Wasserstein distances

Abstract: The Gromov-Wasserstein (GW) distance enables comparing metric measure spaces based solely on their internal structure, making it invariant to isomorphic transformations. This property is particularly useful for comparing datasets that naturally admit isomorphic representations, such as unlabelled graphs or objects embedded in space. However, apart from the recently derived empirical convergence rates for the quadratic GW problem, a statistical theory for valid estimation and inference remains largely obscure. Pushing the frontier of statistical GW further, this work derives the first limit laws for the empirical GW distance across several settings of interest: (i) discrete, (ii) semi-discrete, and (iii) general distributions under moment constraints under the entropically regularized GW distance. The derivations rely on a novel stability analysis of the GW functional in the marginal distributions. The limit laws then follow by an adaptation of the functional delta method. As asymptotic normality fails to hold in most cases, we establish the consistency of an efficient estimation procedure for the limiting law in the discrete case, bypassing the need for computationally intensive

resampling methods. We apply these findings to testing whether collections of unlabelled graphs are generated from distributions that are isomorphic to each other.

10:15 - 11:00 Enno Mammen (University of Heidelberg) - Strong Approximations for Robbins-Monro Procedures

Abstract: The Robbins-Monro algorithm is a recursive, simulation-based stochastic procedure to approximate the zeros of a function that can be written as an expectation. It is known that under some technical assumptions, Gaussian limit theorems approximate the stochastic performance of the algorithm. Here, we are interested in strong approximations for Robbins-Monro procedures. The main tool for getting them are local limit theorems, that is, studying the convergence of the density of the algorithm. The analysis relies on a version of parametrix techniques for Markov chains converging to diffusions. The main difficulty that arises here is the fact that the drift is unbounded. The talk is based on joint work with Valentin Konakov, Moscow, and Lorick Huang, Toulouse.

11:30 - 12:15 Jason M. Klusowski (Princeton) - Statistical–Computational Trade-offs for Recursive Adaptive Partitioning Estimators

Abstract: Recursive adaptive partitioning estimators, such as decision trees and their ensembles, are effective for high-dimensional regression but typically rely on greedy training, which can become stuck at suboptimal solutions. We study this phenomenon in the estimation of sparse regression functions over binary features, showing that when the true function satisfies a structural property introduced by Abbe et al. (2022)—the Merged Staircase Property (MSP)—greedy training achieves low estimation error with only a logarithmic number of samples in the feature count. In contrast, without MSP, estimation becomes exponentially more difficult. Interestingly, this dichotomy between efficient and inefficient estimation resembles the behavior of two-layer neural networks trained with SGD in the mean-field regime. Meanwhile, ERM-trained recursive adaptive partitioning estimators achieve low estimation error with logarithmically many samples regardless of MSP, highlighting a fundamental statistical–computational trade-off for greedy training.

14:00 - 14:45 Dario Kieffer (University of Freiburg) - The Weak-Feature-Impact Effect on the NPMLE in Monotone Binary Regression

Abstract: The nonparametric maximum likelihood estimator (NPMLE) in monotone binary regression models is studied when the impact of the features on the labels is weak. Here, weakness is colloquially understood as "close to flatness" of the feature-label relationship $x \mapsto \mathbb{P}(Y = 1|X = x)$. Statistical literature provides limit distributions of the NPMLE for the two extremal cases: If the feature-label relation is strictly monotone and sufficiently smooth, then it converges at a nonparametric rate pointwise and in L^1 with scaled Chernoff-type and Gaussian limit distribution, respectively, and it converges at the parametric $\sqrt{n}$-rate if the underlying relation is flat. To explore the distributional transition of the NPMLE from the nonparametric to the parametric regime, we introduce a novel mathematical scenario. New restricted minimax lower bounds and matching pointwise and L^1-rates of convergence of the NPMLE in the weak-feature-impact scenario together with corresponding limit distributions

are derived. They are shown to exhibit an elbow and a phase transition respectively, solely characterized by the level of feature impact.

14:45 - 15:30 Yi Yu (University of Warwick) - Optimal Cox regression under federated differential privacy: coefficients and cumulative hazards

Abstract: We study two foundational problems in distributed survival analysis: estimating Cox regression coefficients and cumulative hazard functions, under federated differential privacy constraints, allowing for heterogeneous per-sever sample sizes and privacy budgets. To quantify the fundamental cost of privacy, we derive minimax lower bounds along with matching (up to poly-logarithmic factors) upper bounds. In particular, to estimate the cumulative hazard function, we design a private tree-based algorithm for nonparametric integral estimation. Our results reveal server-level phase transitions between the private and non-private rates, as well as the reduced estimation accuracy from imposing privacy constraints on distributed subsets of data.

To address scenarios with partially public information, we also consider a relaxed differential privacy framework and provide a corresponding minimax analysis. To our knowledge, this is the first treatment of partially public data in survival analysis, and it establishes a no-gain in accuracy phenomenon. Finally, we conduct extensive numerical experiments, with an accompanying R package FDPCox, validating our theoretical findings. These experiments also include a fully-interactive algorithm with tighter privacy composition, which demonstrates improved estimation accuracy.

https://arxiv.org/abs/2508.19640

Friday 19.09.

9:30 - 10:15 Stanislav Minsker (USC: Los Angeles) - Probabilistic Inequalities for Sums of Heavy-Tailed Random Matrices and Their Applications

Abstract: Matrix concentration inequalities have proven to be indispensable tools for the analysis of algorithms in high-dimensional statistics and machine learning. However, many of these results require the random matrices to have either bounded or light-tailed norms. On the other hand, classical Fuk-Nagaev and Rosenthal-type inequalities apply to sums of random variables with possibly heavy tails and provide useful deviation and moment bounds. In this talk, we will describe versions of these results that are applicable to random matrices. The key feature of our bounds is that they depend on the 'intrinsic' dimensional characteristics, such as the effective rank, rather than the dimension of the ambient space. We will illustrate the advantages of such results in several statistical applications, including new moment inequalities for the sample covariance operators and their eigenvectors. This talk is based on a joint work with Moritz Jirak, Martin Wahl and Yiqiu Shen.

10:15 - 11:00 Florentina Bunea (Cornell) - From softmax mixture ensembles to mixtures of experts, with applications to LLM output summarization

Abstract: Contemporary LLM models have billions of parameters, making them impossible to interpret or use directly in downstream analyses. Summarizing LLM output on the

basis of models that strike the balance between complexity and interpretability is therefore of immediate need. This talk will introduce a simple mixture-of-experts (MoE) model for contextually embedded corpora representation. Fitting and analyzing this MoE will be shown to reduce to the analysis of softmax mixture ensemble models, after an appropriate quantization of the corpus into p feature vectors embedded in R^L, for large L. Given a collection of discrete data samples, we identify each sample with a distribution in the probability simplex Δ_p, supported on p points in R^L. The softmax mixture ensemble model postulates that the distribution of each sample is a K-mixture of soft- max (multinomial logit) distributions, for $K \geq 2$, with softmax parameters common to the ensemble, and sample specific weights. Softmax mixture ensembles are of interest in their own right, beyond LLM applications, as they are instances of discrete choice models, widely used in econometrics, among other areas. Despite applicability and increasingly recognized potential, theory and methods for this model class are heavily under-developed. This talk will present solutions to open problems, with a focus on parameter estimation. We provide the first analysis of identifiability in softmax mixtures. Of note is that, for the ensemble model, we can exhibit testable identifiability conditions. We lay the theoretical foundations for parameter estimation in softmax mixtures, by providing the first theoretical analysis of the Expectation-Maximization (EM) algorithm in this model. We give a precise characterization of the size of the initialization neighborhood under which the mixture atoms can be estimated at parametric rates, in a number of iterations that is logarithmic in the sample size. We make use of a novel method-of-moments procedure to estimate the K-dimensional subspace in R^L spanned by the K mixture parameters. As a corollary, we show that EM with random start drawn from this estimated sub-space leads to optimal atom estimators in only exp K (relative to the typical, huge, exp L) draws. As an important feature of our analysis, we further show that the cross-entropy estimates of the mixture weights are exactly sparse, without need for extra regularization, and we also provide one- step corrected mixture weight estimates that are asymptotically normal, and thus amenable to statistical inference. The totality of these results provides a solution to the LLM representation problem. The estimated mixture atoms, that are common to the corpus, and the estimated mixture weights from each document, readily yield an estimated mixing measure that can serve as a sample-level (document) summary, while their average yields a corpus-level summary. These summaries can be used in downstream tasks involving LLM output evaluation and comparison, with the added advantage of parameter interpretability, typically lacking in existing summarization strategies. I will illustrate these theoretical and methodological results using a running data example that clarifies the net benefits of the MoE-type representation of an LLM embedded corpus, relative to a standard topic model-type representation of the same corpus, noting that, by definition, topic models cannot make use of contextual text embedding.

11:30 - 12:15 Harry Zhou (Yale) - From Score Estimation to Sampling

Abstract: Recent impressive advances in the algorithmic generation of high-fidelity images, audio, and video can be largely attributed to the success of score-based diffusion models. A crucial step in their implementation is score matching, which involves estimating the score function of the forward diffusion process from training data. In this work, we establish the rate-optimal estimation of the score function for smooth, compactly supported densities and explore its applications to estimation of density, transport, and optimal transport.