Programme

Monday, 9th September

- 9:00am–9:30am **Registration**
- 9:30am–10.15am
  - Ismael Castillo (Sorbonne Université): Uniform estimation of a class of random graph functionals
- 10:20am–11.05am
  - Olivier Collier (Université Paris-Nanterre): Estimators of general functionals in the sparse Gaussian mean model
- **Refreshment Break**
- 11.40am–12.25pm
  - Mathew Penrose (University of Bath): Limit theory for entropy and other estimators based on multidimensional spacings
- **Lunch**
- 2pm–2.45pm
  - Tom Berrett (University of Cambridge): Efficient functional estimation and the super-oracle phenomenon
- **Refreshment Break**
- 4.10pm–4.55pm
  - Varun Jog (University of Wisconsin-Madison): Bridging the inequality gap
- **Reception**, 5pm
- **Workshop Dinner** 7.30pm, at St John's College, for invited speakers and organisers.

Tuesday, 10th September

- 9.30am–10.15am
  - Po-Ling Loh (University of Wisconsin-Madison): Teaching and learning in uncertainty
- 10:20am–11.05am

  - Ramji Venkataramanan (University of Cambridge): Stronger risk lower bounds for high-dimensional estimation

- **Refreshment Break**
- 11.40am–12.25pm
  - Chao Gao (University of Chicago): Optimal estimation of variance in nonparametric regression with random design
- **Lunch**
- 2pm–2.45pm
  - Alon Orlitsky (University of California San Diego): Profile maximum likelihood: optimal, universal, plug-and-play, functional estimation
- 2.50pm–3.35pm
  - Barnabás Póczos (Carnegie Mellon University): Distribution regression and nonparametric density estimation under adversarial losses
- **Refreshment Break**

Wednesday, 11th September

- 9:00am–9.45am
  - Judith Rousseau (University of Oxford): On Bayesian inference in a family of sparse graphs and multigraphs
- 9:50am–10.35am
  - Yury Polyanskiy (Massachusetts Institute of Technology): Smoothed empirical measures and entropy estimation
- **Refreshment Break**
- 11.10am–11.55am
  - Zhiyi Zhang (University of North Carolina at Charlotte): Unfolding entropy
- 12:00pm–12:45pm
  - Peter Grünwald (CWI and Leiden University): Safe Testing: attacking the reproducibility crisis via the joint information projection

Abstracts

Ismael Castillo (Sorbonne Université)
Monday, 9.30am–10.15am

We consider estimation of certain functionals of a random graph generated by a stochastic block model (SBM). The number of classes is fixed or grows with the number of vertices. Minimax lower and upper bounds of estimation along specific submodels are derived. The results are nonasymptotic and imply that uniform estimation of a single connectivity parameter is much slower than the expected asymptotic pointwise rate. Specifically, the uniform quadratic rate does not scale as the number of edges, but only as the number of vertices. The lower bounds are local around any possible SBM. An analogous result is derived for functionals of a class of smooth graphons. This is joint work with Peter Orbanz (Columbia).

Olivier Collier (Université Paris-Nanterre)
Monday, 10.20am–11.05pm

We observe a sparse mean vector through Gaussian noise and we want to estimate general functionals of this vector in the minimax sense. Adapting the method that has been successfully applied to the linear and quadratic functionals, ie the sums of the coefficients and of the squared coefficients, we obtain new results for the sums of the other powers of the coefficients, then for general additive functionals under very broad assumptions. However, our new estimators crucially depend on the knowledge of the noise level and the noise distribution, which can lead to larger rates of estimation. This is a joint work with Latitia Comminges, Mohamed Ndaoud and Alexandre Tsybakov.

Mathew Penrose (University of Bath)
Monday, 11.40am–12.25pm

Consider an empirical point process governed by a probability density function in d-space or in a sumbmanifold thereof. We discuss the limit theory (laws of large numbers and central limit theorems) for certain statistics based on nearest-neighbour distances within the sample, which is one way of defining spacings for a multivariate sample. As well as entropy estimators based on these spacings, we intend to discuss estimators of dimension and of divergence. The limit theorems are based on the theory of stabilizing functionals. Most of the results discussed are joint work with Yuliy Barishnikov and Joe Yukich.

Tom Berrett (University of Cambridge)
Monday, 2pm–2.45pm

We consider the estimation of two-sample density functionals

$$T(f,g) = \int f(x)\phi(f(x), g(x), x)dx,$$

based on independent d-dimensional random vectors $X_1, \ldots, X_m$ with density f and $Y_1, \ldots, Y_n$ with density g. The interest in such functionals arises from many applications: for instance, many divergences such as the KL divergence, total variation and Hellinger distances are of this form.

The estimators we consider can be expressed as weighted sums of preliminary estimators based on nearest neighbour distances. We provide conditions under which these estimators are efficient, in the sense of achieving the local asymptotic minimax lower bound, and under which they are asymptotically normal. Our results reveal an interesting phenomenon in which the natural 'oracle' estimator, requiring knowledge of f and g, can be outpeformed by our estimators. For some functionals of interest we show that the asymptotic limit of the ratio of the L_2 risks is strictly less than one, uniformly over suitable classes of densities. This is based on joint work with Richard Samworth and Ming Yuan.

Varun Jog (University of Wisconsin-Madison)
Monday, 4.10pm–4.55pm

Reconstructing probability distributions from projections is a fundamental problem in many scientific applications. Geometric and information theoretic inequalities provide important mathematical tools for understanding the behavior of such projections—in particular, for characterizing extremal distributions with respect to different lower-dimensional properties of interest. This talk will consist of two parts: First, we introduce new methods to bound the size of an unseen geometric object using information derived from its lower-dimensional projections. Second, we present a new information inequality that relates the entropy of a random variable to that of its lower-dimensional marginals. Both parts highlight the advantages of working with information inequalities instead of their equivalent geometric or functional formulations. This is joint work with Jing Hao (UW-Madison), Chandra Nair (CUHK), and Venkat Anantharam (UC Berkeley).

Po-Ling Loh (University of Wisconsin-Madison)
Tuesday, 9.30am–10.15am

We investigate a simple model for social learning with two characters: a teacher and a student. The teacher's goal is to teach the student the state of the world Theta. However, the teacher herself is not certain about Theta and needs to simultaneously learn it and teach it. We examine several natural strategies the teacher may employ to make the student learn as fast as possible when the state of the world is 0,1 and transmissions occur through a binary channel. Our primary technical contribution is analyzing the exact learning rates for these strategies by studying the large deviation properties of the sign of a transient random walk on the integer grid. We also discuss a Gaussian variant of this problem and contrast the conclusions reached in the binary vs. Gaussian settings. This is joint work with Varun Jog.

Ramji Venkataramanan (University of Cambridge)
Tuesday 10.20am–11.05am

In statistical inference problems, we wish to obtain lower bounds on the minimax risk, that is to bound the performance of any possible estimator. A standard technique to do this involves the use of Fano's inequality. However, recent work in an information-theoretic setting has shown that an argument based on binary hypothesis testing gives tighter converse results (error lower bounds) than Fano for channel coding problems. We adapt this technique to the statistical setting to obtain tighter lower bounds that can be easily computed and are asymptotically sharp. We illustrate our technique in three applications: density estimation, active learning of a binary classifier, and compressed sensing, obtaining tighter risk lower bounds in each case. (Joint work with Oliver Johnson, see doi:10.1214/18-EJS1419)

Chao Gao (University of Chicago)
Tuesday, 11.40am–12.25pm

We consider the heteroscedastic nonparametric regression model with random design. We derive the minimax rate of estimating the variance function. The result extends the fixed design rate derived in Wang et al. [2008] in a non-trivial manner, as indicated by the entanglement of the smoothness parameters of both mean function and variance function. An implication is that variance estimation is easier in the random design setting. In the special case of constant variance, we show that the minimax rate is n8/(4+1)n1 for variance estimation, which further implies the same rate for quadratic functional estimation and thus unifies the minimax rate under the nonparametric regression model with those under the density model and the white noise model.

Alon Orlitsky (University of California San Diego) Tuesday, 2.00pm–2.45pm

The profile of a sample is the multiset of number of times each symbol appears. For example, l, o, n, d, o, n has two letters appearing once and two appearing twice, hence its profile is {1,1,2,2}. We show that the distribution maximizing the probability of the observed profile yields simple, optimal, plug-in, estimates for a host of learning tasks, including estimating all additive Lipschitz distribution functionals, Renyi entropy, sorted probabilities, and identity testing. Based on work with Yi Hao and building on prior work with several former students in our group.

Barnabás Póczos (Carnegie Mellon University)
Tuesday, 2.50pm–3.35pm

In the first part of the talk we discuss new methods and applications for distribution regression. In this problem a response Y depends on a covariate P where P is a probability distribution. Typically, we do not observe P directly, but rather, we observe a sample from P. We discuss new theoretical results, open questions, and some applications in cosmology, high-energy physics, computer vision, and entropy estimation. In the second part of the talk we study minimax convergence rates of nonparametric density estimation under a large class of loss functions called "adversarial losses", which, besides classical L_p losses, includes maximum mean discrepancy, Wasserstein distance, and total variation distance. These losses are closely related to the losses encoded by discriminator networks in generative adversarial networks (GANs). In a general framework, we study how the choice of loss and the assumed smoothness of the underlying density together determine the minimax rate.

Judith Rousseau (University of Oxford)
Wednesday, 9.00am–9.45am

In the first part of this I will present some properties of the class of graphs based on exchangeable point processes: these include in particular the asymptotic expansions for the number of edges, nodes and triangles together with the degree distributions, identifying four regimes: (i) a dense regime, (ii) a sparse, almost dense regime, (iii) a sparse regime with power-law behaviour, and (iv) an almost extremely sparse regime. We also propose a class of models with this framework where one can separately control the local, latent structure and the global sparsity/power-law properties of the graph and we derive a central limit theorem for the number of nodes and edges in the graph.

In the second part I will recall the Bayesian approach to make inference in such graphs, both for simple and multigraphs proposed by Caron and Fox (2017). I will then study the asymptotic behaviour of the posterior distribution, both under well and mis-specified multigraph models.

Yury Polyanskiy (Massachusetts Institute of Technology)
Wednesday, 9.50am–10.35am

In this talk we discuss behavior of the empirical measure P_n corresponding to iid samples from a distribution P on a d-dimensional space. Let Q_n and Q denote the result of convolving P_n and P, respectively, with an isotropic standard Guassian kernel. We discuss convergence of the p-Wasserstein, KL and other distances between Q_n and Q. Curiously, for some distances (like 1-Wasserstein) we get parametric $1/\sqrt{n}$ speed of convergence regardless of dimension, whereas for some other distances (like 2-Wasserstein) the $1/\sqrt{n}$ rate can change to $\omega(1/\sqrt{n})$. We give an if and only if characterization in the class of subgaussian P for the parametric rate.

As an application, we show that differential entropy of Q_n converges to that of Q at parametric rate regardless of dimension. An estimator of differential entropy of Q, in turn, allows us to estimate the input-output mutual information in noisy neural networks.

Joint work with Ziv Goldfeld, Kristjan Greenewald and Jonathan Weed.

Zhiyi Zhang (University of North Carolina at Charlotte)
Wednesday, 11.10am–11.55pm

This talk includes three sections. In Section 1, an estimator of entropy is described and motivated in a perspective induced from Turings formula. The convergence rates and other asymptotic distributional properties of the estimator under different classes of underlying distributions are summarized. In Section 2, several fundamental questions are considered. What does entropy tell us? What do we really want to know through entropy, that is, what is the real underpinning object of interest behind entropy? Do we really want to estimate entropy, or are we creating a difficult mathematical problem without sufficient merit? The contemplation of these questions leads to the definition of a new notion: the entropic probability distribution, or the entropic distribution in short, which could serve as one of possibly many anchor points in the foundation of Statistics pertaining to information theory, or Information-Theoretical Statistics (for lack of a better term). Functionals that collectively characterize the entropic distribution are discussed. In Section 3, several results derived from the perspective of the entropic distribution are introduced, which may prove to be of some general interest in probability and statistics. The results discussed include 1) domains of attraction on countable alphabets, 2) diversity indices and their estimation, 3) generalization of mutual information, etc.

Peter Grünwald (CWI and Leiden University)
Wednesday, 12.00pm–12.45pm

In light of the 'replicability crisis' in the applied sciences, standard p-value based hypothesis testing has come under intense scrutiny. One of its many problems is this: if our test result is promising but nonconclusive (say, $p = 0.07$) we cannot simply decide to gather a few more data points. While this practice is ubiquitous in science, it invalidates p-values and error guarantees.

Here we propose an alternative hypothesis testing methodology based on supermartingales - it has both a gambling and a data compression interpretation. This method allows us to consider additional data and freely combine results from different tests, avoiding many other pitfalls of traditional testing as well. If the null hypothesis is simple (a singleton), it also has a Bayesian interpretation, and is very similar to classic Wald-style sequential testing. We work out the case of composite null hypotheses, which (unlike standard Bayes or sequential methods) allows us to formulate nonasymptotic versions of the most popular tests such as the t-test and the chi square tests that preserve error guarantees under optional stopping. The optimal safe tests for composite H0 are based on the 'joint information projection' of $(H1, H0)$, i.e. the distributions $\bar{P}_1$ and $\bar{P}_0$ that minimize the KL divergence $D(\bar{P}_1 \| \bar{P}_0)$ over the convex hull (set of Bayes mixtures) of H_1 and H_0. The surprising connection between information projection and gambling-based testing is proven via a minimax theorem with the logarithmic score replaced by a new proper scoring rule — this raises all kinds of interesting open questions in the intersection of information theory and statistics. Besides their theoretical interest the new tests work very well in practice: although for fixed n, more data are needed to achieve a desired power than classically, one can often stop early and ends up needing less data than classically. Joint Work with R. de Heide (CWI and Leiden Univ.) and W. Koolen (CWI).