Workshop on Functional Inference and Machine Intelligence 2019

28th March – 29th March 2019

Auditorium, The Institute of Statistical Mathematics.

Program

28 March (Thu)

10:00-10:05	Opening
10:05-11:05	Dino Sejdinovic (University of Oxford)
11:15-12:15	Yen-Chi Chen (University of Washington)
12:15-13:45	Lunch Break
13:45-14:45	Taiji Suzuki (The University of Tokyo)
14:55-15:55	Bharath Sriperumbudur (Pennsylvania State University)
16:05-17:05	Masaaki Imaizumi (The Institute of Statistical Mathematics)

29 March (Fri)

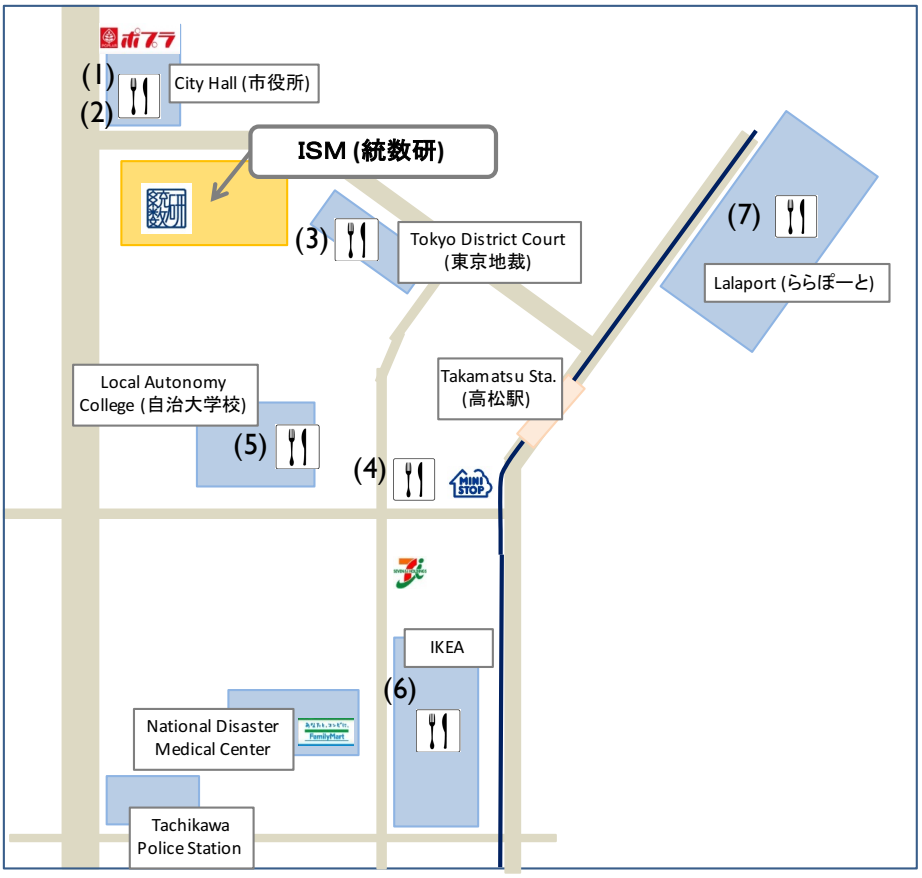
09:30-10:30	Arthur Gretton (University College London)
10:40-11:40	Heishiro Kanagawa (University College London)
11:40–13:10	Lunch Break
13:10–14:10	Motonobu Kanagawa (University of Tuebingen)
14:20-15:20	Akifumi Okuno (Kyoto University / RIKEN AIP)
15:30-16:20	Kenji Fukumizu (The Institute of Statistical Mathematics)
16:20-16:30	Closing

Organizers

Masaaki Imaizumi (The Institute of Statistical Mathematics)
Kenji Fukumizu (The Institute of Statistical Mathematics)

This workshop is supported by
Research Center for Statistical Machine Learning, ISM.

Lunch Map



Pack lunches （at the lobby lounge） 11：00～13：00

Rin rin:	450~500yen each	
Haiji (curry)	500yen each	(a light dessert offered on Wednesdays)
Zuikyo (Chinese)	500yen each	
Hello Lunch	300~400yen each	(miso-soup included)

Restaurant

(1) City Hall dining room, 3rd floor. (立川市役所食堂(3F)) 300~500yen

(2) Café Harmony at the City Hall, 1st floor. (カフェハーモニー, 立川市役所 1F) 600yen

(3) Pole Light, Tokyo District Court dining room, 1st basement floor. (食堂ポールライト, 東京地裁 B1F)

(4) Chinese Restaurant Zuikyo, near MINISTOP. (中華料理店 瑞京),

(5) Local Autonomy College dining room. (自治大学校食堂)

(6) IKEA restaurant and café, 2nd floor. (IKEA レストラン 2F)

(7) Restaurants in Lalaport (shopping mall). (ららぽーと レストラン街)

Convenience Stores

Popula (City Hall 1F) / MINISTOP / 7-Eleven / Family Mart

Abstracts (28th)

Dino Sejdinovic (University of Oxford)

Hyperparameter Learning via Distributional Transfer

Bayesian optimisation is a popular technique for hyperparameter learning but typically requires initial 'exploration' even in cases where potentially similar prior tasks have been solved. We propose to transfer information across tasks using kernel embeddings of distributions of training datasets used in those tasks. The resulting method has a faster convergence compared to existing baselines, in some cases requiring only a few evaluations of the target objective.
Joint work with Leon Law, Peilin Zhao, Junzhou Huang.

Yen-Chi Chen (University of Washington)

Analyzing GPS data using density ranking

A common approach for analyzing a point cloud is based on estimating the underlying probability density function. However, in complex datasets such as GPS data, the underlying distribution function is singular so the usual density function no longer exist. To analyze this type of data, we introduce a statistical model for GPS data in the form of a mixture model with different dimensions. To derive a meaningful surrogate of the probability density, we propose a quantity called density ranking. Density ranking is a quantity representing the intensity of observations around a given point that can be defined in a singular measure. We then show that one can consistently estimate the density ranking using a kernel density estimator even in a singular distribution such as the GPS data. We apply density ranking to GPS datasets to analyze activity spaces of individuals.

Taiji Suzuki (The University of Tokyo)

Generalization error of deep learning with connection to sparse estimation in function space

In this talk, we will talk about the learning ability of deep ReLU-neural network, in particular, in connection to sparsity in a functional space. The superior learning ability of deep models is essentially due to its ability to construct bases in an adaptive way to each target function. For this purpose, the non-convexity of the model is quite important. This point shares several similarities to the sparse learning methods such as L0 regularization and low rank matrix estimation. In this talk, we will show that the non-convexity of the model

gives superior performance with connection to sparse estimation. In particular, we consider the Besov space and mixed-smooth Besov space, and show deep learning can outperform any linear estimators on near non-convex functional spaces. Moreover, we will show that the low rank property of the internal layers can give better generalization error. Consequently, a compression based generalization error bound is obtained even for non-compressed networks.

Bharath Sriperumbudur (Pennsylvania State University)

On Distribution Regression

In this work, we continue our investigation on distribution regression problem: regressing to vector-valued outputs from probability measures where the probability measures are not fully observed but only through finite samples, each of size m drawn from each of the probability measures. In (Zoltan et al., JMLR 2016), we developed a ridge regressor based on kernel mean embeddings and showed that it achieves minimax rates as long as $m=N^h log(N)$ where N is the size of the training set (i.e., the number of probability measures) and $1¥le h¥le 2$ with h depending on the smoothness of the true regressor. In this work, we construct a ridge regressor based on a certain U-statistics estimate of the covariance operator and show that minimax rates are attained while requiring m to behave sublinearly in N.

Masaaki Imaizumi (The Institute of Statistical Mathematics)

Statistical Analysis for Generative Adversarial Networks

We investigate statistical generalization properties of generative adversarial networks (GANs) and show that GANs can achieve minimax optimality with some settings. GANs estimate probability measures from observations using a notion of a generator and a discriminator. While many empirical results show that the estimators by GANs outperform other estimators, clarifying theoretical properties of GANs needs several simplifications of settings of GANs. This work studies a statistical generalization property of GANs with a general setting, and show that a convergence rate of a generalization error of GANs can attain minimax optimality. To obtain the optimality, we provide a smoothing operation for an empirical distribution as a precondition. Also, we provide a new approximation and convergence analysis of generators and discriminators. Based on the theory, we also provide a guideline for selecting deep network architecture for GANs.

Abstract (29th)

Arthur Gretton (University College London)

The Maximum Mean Discrepancy for Training Generative Adversarial Networks

Generative adversarial networks (GANs) use neural networks as generative models, creating realistic samples that mimic real-life reference samples (for instance, images of faces, bedrooms, and more). These networks require an adaptive critic function while training, to teach the networks how to move improve their samples to better match the reference data. I will describe a kernel divergence measure, the maximum mean discrepancy, which represents one such critic function. With gradient regularisation, the MMD is used to obtain current state-of-the art performance on challenging image generation tasks, including 160 × 160 CelebA and 64 × 64 ImageNet. In addition to adversarial network training, I'll discuss issues of gradient bias for GANs based on integral probability metrics, and mechanisms for benchmarking GAN performance.

Heishiro Kanagawa (University College London)

Informative Features for Model Comparison

Given two candidate models, and a set of target observations, we address the problem of measuring the relative goodness of fit of the two models. We propose two new statistical tests which are nonparametric, computationally efficient (runtime complexity is linear in the sample size), and interpretable. As a unique advantage, our tests can produce a set of examples (informative features) indicating the regions in the data domain where one model fits significantly better than the other. In a real-world problem of comparing GAN models, the test power of our new test matches that of the state-of-the-art test of relative goodness of fit, while being one order of magnitude faster.

This is a joint work with Wittawat Jitkrittum, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton.

Motonobu Kanagawa (University of Tuebingen)

Parameter Estimation and Model Selection for Simulator-based Statistical Models: A Kernel Approach

Computer simulation is nowadays a ubiquitous tool in various scientific fields, such as climate science, social science, economics and epidemiology. It is useful in studying time-evolving complex phenomena, for which models may be described by partial or ordinary differential equations, or by multi-agent systems. A key issue

regarding computer simulation is how to calibrate a simulation model to observed data, since there are in general parameters in the model that need to be tuned; the problem of parameter estimation must be addressed in practice. Further, typically there exist multiple candidate models for the phenomena of interest, so one also needs to perform model selection on the basis of observed data. These two issues, parameter estimation and model selection, are much more challenging than usual statistical problems; this is because the likelihood function, the conditional probability of observed data given parameters, is not available for a simulator-based statistical model in general, since the mapping from parameters to data is usually very complicated.

In this talk, I will describe our recent approach to parameter selection and model selection for simulator-based statistical models. This approach, termed Kernel Recursive ABC, recursively applies approximate Bayesian computations to the same observed data in an iterative way. It is realized as a combination of the kernel ABC and the kernel herding algorithm, and this results in robustness against the misspecification of parameter prior distributions. We discuss how the method works and why it is equipped with such robustness, and report experimental results showing the effectiveness of the proposed approach.

Akifumi Okuno (Kyoto University / RIKEN AIP)

Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability

Given a graph whose each node possesses a vector representation, graph embedding (GE) transforms these node vectors via a vector-valued neural network (NN) so that their similarities represent the weighted adjacency matrix. Whereas inner-product similarity (IPS) is often employed for NN-based GE, Mercer's theorem indicates that IPS is limited to approximating positive-definite (PD) similarities. To overcome this limitation of IPS, we propose shifted inner-product similarity (SIPS), which is IPS with added bias terms. Despite being a very simple extension of IPS, SIPS is novel: we theoretically prove that SIPS is capable of approximating not only PD but also conditionally PD (CPD) similarities, which includes many non-PD similarities such as cosine similarity, negative Poincare distance, or negative Wasserstein distance. As SIPS with sufficiently large neural networks learns a variety of similarities, SIPS alleviates the need for configuring the similarity function of GE. The approximation error rate is also evaluated, and experiments on two real-world datasets demonstrate that graph embedding using SIPS indeed outperforms existing methods.

This is a joint work with Geewook Kim and Hidetoshi Shimodaira.

Kenji Fukumizu (The Institute of Statistical Mathematics)

Flat local minima and saddle points in invariant structures of neural networks

While local minima of the loss function is an important issue in the learning of neural networks, the structure of the local minima in the landscape is not yet understood clearly. This talk will discuss a type of critical points given by the invariance for the permutations of units in a layer. Given a parameter point of a network with H0 units in a layer, we are able to embed it to a network with H (H > H0) units in the layer so that the embedded points can make an affine subset in the parameter space with the total input-output function unchanged. We can prove that, if the parameter of the smaller network is a stationary point of the loss function, all the embedded parameters in the larger network are also stationary points. Under some mild differentiability assumptions on the activation function, a sufficient condition for the stationary points to be local minima or saddle can be derived. I will also show some results on the network with ReLU units, which has further invariance caused by positive homogeneity.