Sydney Workshop on Mathematics of Data Science
04–06, December 2024, University of Sydney, Sydney, Australia

Program

Venue for all sessions: Law Annex Seminar Room 340, Law Building Annex
A campus map is available here `https://maps.sydney.edu.au`
Campus Wi-Fi: eduroam, or "UniSydney-Guest" (Username: sydneymds98u3b7ef   Password: 518152)

Day 1, December 04, 2024, Wednesday

- 08:00–08:50, Registration

- 08:50–09:00, Welcome Speech, by Dingxuan Zhou

Session I, Chair: Georg Gottwald

- 09:00–09:30, **Masashi Sugiyama**, Training Error and Bayes Error in Deep Learning (See abstract, p. 13)

- 09:30–10:00, **Clara Grazian**, Invertibility of Deep Biometric Representations (See abstract, p. 7)

- 10:00–10:30, **Ha Quang Minh**, Infinite-dimensional statistical distances for functional data analysis (See abstract, p. 10)

- 10:30–11:00, **Coffee Break and Group Photo**

Session II, Chair: Clara Grazian

- 11:00–11:30, **Nicole Mücke**, Recent Trends in Learning Operators (See abstract, p. 11)

- 11:30–12:00, **Georg Gottwald**, A localized Schrödinger bridge sampler for generative modelling (See abstract, p. 7)

- 12:00–12:30, **Quoc Le Gia**, Evolution of time-fractional stochastic hyperbolic diffusion equations on the unit sphere (See abstract, p. 9)

- 12:30–13:40, **Lunch Break**

Session III, Chair: Caroline Wormell

- 13:40–14:10, **Gerlind Plonka**, MOCCA: A Fast Algorithm for Parallel MRI Reconstruction Using Model Based Coil Calibration (See abstract, p. 12)

- 14:10–14:40, **Rachel Wang**, Adaptive conformal classification with noisy labels (See abstract, p. 14)

- 14:40–15:10, **Tongliang Liu**, Revealing causal information from data (See abstract, p. 10)

- 15:10–15:40, **Junbin Gao**, Ground Metric Learning for Optimal Transport with Riemmian Optimization (See abstract, p. 6)

- 15:40–16:00, **Coffee Break**

Session IV, Chair: Rachel Wang

- 16:00–16:30, **Caroline Wormell**, Dictionary error of trigonometric least squares approximation with non-uniform distribution (See abstract, p. 15)

- 16:30–17:00, **Tiangang Cui**, Tensor-Train Methods for Sequential State and Parameter Estimation in State-Space Models (See abstract, p. 4)

- 17:00–17:30, **Ian Gallagher**, Stable Dynamic Network Embeddings (See abstract, p. 6)

- 17:30–18:00, **Lei Shi**, Solving PDEs on Spheres with Physics-Informed Convolutional Neural Networks (See abstract, p. 13)

Day 2, December 05, 2024, Thursday

Session I, Chair: Tongliang Liu

- 09:00–09:30, **Tianbao Yang**, On Discriminative Probabilistic Modeling for Self-Supervised Representation Learning (See abstract, p. 16)

- 09:30–10:00, **Simon Foucart**, Worst-case learning under multifidelity models (See abstract, p. 5)

- 10:00–10:30, **Flora Salim**, Data-Efficient Learning for Imperfect Spatio-Temporal Data: Leveraging SSLs and LLMs (See abstract, p. 12)

- 10:30–10:50, **Coffee Break and Registration**

Session II, Chair: Xin Guo

- 10:50–11:20, **Lindon Roberts**, An adaptively inexact first-order method for bilevel optimization with application to hyperparameter learning (See abstract, p. 12)

- 11:20–11:50, **Linh Nghiem**, Random effects model-based sufficient dimension reduction for independent clustered data (See abstract, p. 11)

- 11:50–12:20, **Zheng-Chu Guo**, Learning theory of spectral algorithms under covariate shift (See abstract, p. 8)

- 12:20–13:30, **Lunch Break**

Session III, Chair: Qiuzhuang Sun

- 13:30–14:00, **Ata Kaban**, Compression and heterogeneity in PAC-learning (See abstract, p. 9)

- 14:00–14:30, **Andreas Christmann**, On Algorithmic Stability and Robustness of Bootstrap SGD (See abstract, p. 4)

- 14:30–15:00, **Hien Nguyen**, Lp Approximation Rates for Location-Scale Mixture Densities and Implications to Adaptive Least-Squares Estimation (See abstract, p. 11)

- 15:00–15:30, **Caleb Ju**, Strongly-polynomial time and validation analysis of policy gradient methods (See abstract, p. 8)

- 15:30–15:50, **Coffee Break**

Session IV, Chair: Linh Nghiem

- 15:50–16:20, **Markus Holzleitner**, On Polynomial Functional Regression (See abstract, p. 8)

- 16:20–16:50, **Qiuzhuang Sun**, Optimal stopping under imperfect condition monitoring for non-Markovian systems (See abstract, p. 14)

- 16:50–17:20, **Junhong Lin**, On Convergence of AdaGrad with Relaxed Assumptions (See abstract, p. 9)

- 17:20–17:50, **Daohong Xiang**, Online Outcome Weighted Learning (See abstract, p. 15)

Day 3, December 06, 2024, Friday

Session I, Chair: Dingxuan Zhou

- 09:00–09:30, **Bin Yu**, Adaptive Wavelet Distillation towards Interpretable Deep Learning (See abstract, p. 16)

- 09:30–10:00, **Hrushikesh Mhaskar**, Local approximation of operators (See abstract, p. 10)

- 10:00–10:30, **Mingming Gong**, Domain generalization via content factors isolation: a two-level latent variable modeling approach (See abstract, p. 6)

- 10:30–10:50, **Coffee Break and Registration**

Session II, Chair: Bin Yu

- 10:50–11:20, **Wei Cai**, SOC-MartNet: A Martingale Neural Network for the Hamilton-Jacobi-Bellman Equation without Explicit $\inf_u H$ in Stochastic Optimal Controls (See abstract, p. 4)

- 11:20–11:50, **Susan Wei**, Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition (See abstract, p. 15)

- 11:50–12:20, **Tian-Yi Zhou**, Optimal Classification-based Anomaly Detection with Neural Networks: Theory and Practice in Cybersecurity (See abstract, p. 16)

- 12:20–13:30, **Lunch Break**

Session III, Chair: Yiming Ying

- 13:30–14:00, **Han Feng**, Feature Sparsity in CNNs and a Driven Learning Strategy (See abstract, p. 5)

- 14:00–14:30, **Jun Fan**, Functional neural network on infinite-dimensional data (See abstract, p. 5)

- 14:30–15:00, **Xin Guo**, Learning Green's functions from data (See abstract, p. 7)

Titles and Abstracts of the Talks

SOC-MartNet: A Martingale Neural Network for the Hamilton-Jacobi-Bellman Equation without Explicit $\inf_u H$ in Stochastic Optimal Controls

Wei Cai
Southern Methodist University, United States
cai@mail.smu.edu

In this talk, we present a martingale-based neural network, SOC-MartNet, for solving high-dimensional Hamilton-Jacobi-Bellman (HJB) equations where no explicit expression is needed for the infimum of the Hamiltonian, $\inf u \in UH(t, x, u, z, p)$, and stochastic optimal control problems (SOCPs) with controls on both drift and volatility. We reformulate the HJB equations for the value function by training two neural networks, one for the value function and one for the optimal control with the help of two stochastic processes- a Hamiltonian process and a cost process. The control and value networks are trained such that the associated Hamiltonian process is minimized to satisfy the minimum principle of a feedback SOCP, and the cost process becomes a martingale, thus, ensuring the value function network as the solution to the corresponding HJB equation. Moreover, to enforce the martingale property for the cost process, we employ an adversarial network and construct a loss function characterizing the projection property of the conditional expectation condition of the martingale. Numerical results show that the proposed SOC-MartNet is effective and efficient for solving HJB-type equations and SOCPs with a dimension up to 2000 in a small number of epochs (less than 20) or iteration steps (less than 2000) of training.

On Algorithmic Stability and Robustness of Bootstrap SGD

Andreas Christmann
University of Bayreuth, Germany
andreas.christmann@uni-bayreuth.de

In this talk some methods to use the empirical bootstrap approach for stochastic gradient descent (SGD) to minimize the empirical risk over a separable Hilbert space are investigated from the view point of algorithmic stability and statistical robustness. Two types of approaches are based on averages and are investigated from a theoretical point of view. Another type of bootstrap SGD is proposed to demonstrate that it is possible to construct purely distribution-free pointwise confidence intervals and distribution-free pointwise tolerance intervals of the median curve using bootstrap SGD.

Tensor-Train Methods for Sequential State and Parameter Estimation in State-Space Models

Tiangang Cui
University of Sydney, Australia
tiangang.cui@sydney.edu.au

Numerous real-world applications require the estimation, forecasting, and control of dynamic systems using incomplete and indirect observations. These problems can be formulated as state-space models, where the challenge lies in learning the model states and parameters from observed data. We present new tensor-based sequential Bayesian learning methods that jointly estimate parameters and states. Our methods provide manageable error analysis and potentially mitigate the particle degeneracy encountered in many particle-based approaches. Besides offering new insights into algorithmic design, our methods naturally incorporate conditional transports, enabling filtering, smoothing, and parameter estimation within a unified framework.

Functional neural network on infinite-dimensional data

Jun Fan

Hong Kong Baptist University, Hong Kong, China

junfan@hkbu.edu.hk

Neural networks have proven their versatility in approximating continuous functions, but their capabilities extend far beyond. In this talk, we delve into the realm of functional neural networks, which offer a promising approach for approximating nonlinear smooth functionals. By investigating the convergence rates of approximation and generalization errors under different regularity conditions, we gain insights into the theoretical properties of these networks under the empirical risk minimization framework. This analysis contributes to a deeper understanding of functional neural networks and opens up new possibilities for their effective application in domains such as functional data analysis and scientific machine learning.

Feature Sparsity in CNNs and a Driven Learning Strategy

Han Feng

City University of Hong Kong, Hong Kong, China

hanfeng@cityu.edu.hk

In this talk, we explore the intersection of sparse coding theory and convolutional neural networks (CNNs) to enhance our understanding of feature extraction capabilities in deep learning models. We begin by introducing a framework for deep sparse coding models, where we establish the conditions necessary for the uniqueness and stability of these models. This framework provides a robust method for learning data representations through sparsity. We then transition to a rigorous analysis of the capability of CNNs to approximate sparse features, offering a theoretical foundation for their effectiveness in feature extraction and representation learning.

Building on these theoretical insights, we propose a feature sparsity learning strategy tailored specifically for CNNs. We conduct extensive numerical experiments to validate the performance and efficacy of our proposed strategy. The results demonstrate significant improvements in feature representation and image processing tasks.

Worst-case learning under multifidelity models

Simon Foucart

Texas A&M University, United States

foucart@tamu.edu

This talk showcases the speaker's recent results in the field of Optimal Recovery, viewed as a trustworthy Learning Theory focusing on the worst case. At the core of several results is a scenario, resolved in the global and the local settings, where the model set is the intersection of two hyperellipsoids. This has implications in optimal recovery from deterministically inaccurate data and in optimal recovery under a multifidelity-inspired model, emphasized here. The theory becomes even richer when considering the optimal estimation of linear functionals.

Stable Dynamic Network Embeddings

Ian Gallagher
University of Melbourne, Australia
ian.gallagher@unimelb.edu.au

Network embeddings are low-dimensional representations of nodes in a network which capture the underlying structure, for example, grouping people online with similar political beliefs or capturing the shape of a computer network. In dynamic networks, evolving node behaviour is represented as paths moving through the embedding space. These can describe sudden anomalous node behaviour, communities merging or splitting, or the entire network changing structure. However, this is only possible if the dynamic network embedding is stable in the sense that similarly behaving nodes over time are represented in a consistent way, a property that many existing techniques are lacking.

In this talk, I will describe how any static network embedding technique can be used to produce stable dynamic network embeddings. This theory allows for new dynamic embedding algorithms using skip-grams and graph neural networks with applications in dynamic node classification, conformal prediction, and anomaly detection.

Ground Metric Learning for Optimal Transport with Riemmian Optimization

Junbin Gao
University of Sydney, Australia
junbin.gao@sydney.edu.au

Optimal transport (OT) theory has garnered significant attention in machine learning and signal processing applications due to its ability to define distances between probability distributions of source and target data points. A key factor impacting OT-based distances is the ground metric of the embedding space where these data points reside. In this study, we introduce a method to learn an optimal latent ground metric, parameterized by a symmetric positive definite matrix. Leveraging the rich Riemannian geometry of symmetric positive definite matrices, we jointly learn the OT distance alongside the ground metric. Empirical results demonstrate the effectiveness of the learned metric in enhancing OT-based domain adaptation.

Domain generalization via content factors isolation: a two-level latent variable modeling approach

Mingming Gong
University of Melbourne, Australia
mingming.gong@unimelb.edu.au

The purpose of domain generalization is to develop models that exhibit a higher degree of generality, meaning they perform better when evaluated on data coming from previously unseen distributions. Models obtained via traditional methods often cannot distinguish between label-specific and domain-related features in the latent space. To confront this difficulty, we propose formulating a novel data generation process using a latent variable model and postulating a partition of the latent space into content and style parts while allowing for statistical dependency to exist between them. In this model, the distribution of content factors associated with observations belonging to the same class depends on only the label corresponding to that class. In contrast, the distribution of style factors has an additional dependency on the domain variable. We derive constraints that suffice to recover the collection of content factors block-wise and the collection of style factors component-wise while guaranteeing the isolation of content factors. Our simulations with dependent latent variables produce results consistent with our theory, and real-world experiments show that our method outperforms the competitors.

A localized Schrödinger bridge sampler for generative modelling

Georg Gottwald

University of Sydney, Australia

`georg.gottwald@sydney.edu.au`

We consider the generative problem of sampling from an unknown distribution for which only a sufficiently large number of training samples are available. We show how combining Schrödinger bridges and Langevin dynamics provides an efficient framework for such problems, in particular for data which are concentrated on a compact sub-manifold. A key bottleneck of this approach though is the exponential dependence of the required training samples on the dimension d of the ambient state space. We propose a localization strategy which exploits conditional independence of conditional expectation values. Our localization replaces a single high-dimensional Schrödinger bridge problem by d low-dimensional Schrödinger bridge problems over the available training samples. In this context, a connection to multi-head self attention transformer architectures is established. The localized sampler is stable and geometric ergodic. The sampler also naturally extends to conditional sampling and to Bayesian inference. We demonstrate the performance of our proposed scheme through experiments on a Gaussian problem with increasing dimensions, on a temporal stochastic process, and on a stochastic subgrid-scale parametrization conditional sampling problem. This is joint work with Sebastian Reich.

Invertibility of Deep Biometric Representations

Clara Grazian

University of Sydney, Australia

`clara.grazian@sydney.edu.au`

Biometric recognition systems are pattern-recognition systems designed to identify users based on a vector of features, which may pertain to physiological or behavioral characteristics. These features are typically stored in databases. Despite the encryption algorithms employed to secure these databases, there is always a chance, however small, that an attacker could decrypt the data and gain access to the full set of features. Subsequently, the attacker could compare the acquired biometric information with templates corresponding to user identities, potentially leading to unauthorized identification.

This study focuses on leveraging convolutional neural networks (CNNs) to compress images in a manner that increases the difficulty of restoring them, thereby reducing the effectiveness of such attacks. We employ CNNs to generate deep representations (features) and utilize deep convolutional generative adversarial networks (DCGANs) to reconstruct the images from these features. Finally, the accuracy of the reconstructed images is analyzed using Bayesian beta regression and Random Forest to identify the CNN hyperparameters that have the greatest impact on reconstruction accuracy. Our findings indicate that for simpler datasets, such as black-and-white images of letters or digits (e.g., MNIST), the number of convolutional layers and the presence of pooling layers are the two most critical structural hyperparameters in securing image features. However, for more complex datasets, such as colored facial images (e.g., FaceScrub), activation functions and optimization algorithms become more significant in determining the security of biometric representations.

Learning Green's functions from data

Xin Guo

University of Queensland, Australia

`xin.guo@uq.edu.au`

We studied the problem of learning the Green's functions of partial differential equations from data, through reproducing kernel methods. With the help of a novel kernel design, we derived an algorithm of time complexity $O(m^3+m^2N)$ only, where N is the size of training sample, and m is the number of grid points. We demonstrated that the kernel we designed could safely replace general kernels though sometimes constrain the hypothesis spaces. Minimax lower bound and upper bound of learning rates were derived.

Learning theory of spectral algorithms under covariate shift

Zheng-Chu Guo
Zhejiang University, China
guozc@zju.edu.cn

In machine learning, it is commonly assumed that the training and test samples are drawn from the same underlying distribution. However, this assumption may not always hold true in practice. A scenario is delved into where the distribution of the input variables (also known as covariates) differs between the training and test phases. This situation is referred to as a covariate shift. To address the challenges posed by covariate shift, various techniques have been developed, such as importance weighting, domain adaptation, and reweighting methods. In this talk, we consider the weighted spectral algorithm within the context of covariate shift. Under mild conditions imposed on the weights, it is demonstrated that this algorithm achieves satisfactory convergence rates.

On Polynomial Functional Regression

Markus Holzleitner
University of Genoa, Italy
markus.holzleitner1@gmail.com

Functional Regression (FR) revolves around datasets composed of functions sampled from a population. Most FR research is rooted in a modified version of the functional linear model initially introduced by Ramsay and Dalzell in 1991. Recently, Yao and Müller (2010) discussed a more expansive form of polynomial functional regression, highlighting quadratic functional regression as a prominent case. Constructing FR models entails addressing a pivotal challenge: the combination of information both across and within observed functions, denoted as replication and regularization by Ramsay and Silverman (1997). In this presentation, we will unveil a comprehensive approach for analyzing regularized polynomial functional regression of arbitrary order, by formulating it as an inverse problem. We will explore the potential utilization of a technique developed recently in the realm of supervised learning. Additionally, we will delve into the application of multiple penalty regularization within the FR framework, showcasing its advantages, and we also present a theoretically grounded strategy for dealing with the associated parameters. Finally, we will touch upon the application of FR in stenosis detection. This is based on joint work with Sergei Pereverzyev (RICAM, Linz).

Strongly-polynomial time and validation analysis of policy gradient methods

Caleb Ju
Georgia Institute of Technology, United States
calebju4@gatech.edu

This paper proposes a novel termination criterion, termed the advantage gap function, for finite state and action Markov decision processes (MDP) and reinforcement learning (RL). By incorporating this advantage gap function into the design of step size rules and deriving a new linear rate of convergence that is independent of the stationary state distribution of the optimal policy, we demonstrate that policy gradient methods can solve MDPs in strongly-polynomial time. To the best of our knowledge, this is the first time that such strong convergence properties have been established for policy gradient methods. Moreover, in the stochastic setting, where only stochastic estimates of policy gradients are available, we show that the advantage gap function provides close approximations of the optimality gap for each individual state and exhibits a sublinear rate of convergence at every state. The advantage gap function can be easily estimated in the stochastic case, and when coupled with easily computable upper bounds on policy values, they provide a convenient way to validate the solutions generated by policy gradient methods. Therefore, our developments offer a principled and computable measure of optimality for RL, whereas current practice tends to rely on algorithm-to-algorithm or baselines comparisons with no certificate of optimality.

Compression and heterogeneity in PAC-learning

Ata Kaban

University of Birmingham, United Kingdom

a.kaban@cs.bham.ac.uk

With the advent of big data, large models, and small devices, the idea of model compression, or using approximate predictors, has been very natural and surprisingly successful in practice. The first part of the talk will analyse the role of approximability in PAC learning, both in the full precision and in the approximated learning settings. We find semi-supervised learning algorithms whose approximated version preserves the learning guarantees of their full precision version. We then highlight some natural examples of structure in the class of approximation-sensitivities that can eliminate the otherwise abundant requirement of additional unlabelled data, and at the same time shed more light onto what makes one problem instance easier to learn than another. The second part of the talk is motivated by the idea of aggregating approximate predictors. Such ensembles are often heterogeneous. We develop a general approach to measure the complexity of heterogeneous sets that takes advantage of low complexity components. We then leverage this to improve distance preservation guarantees in random projections, and to justify heterogeneous ensembles in PAC learning.

Evolution of time-fractional stochastic hyperbolic diffusion equations on the unit sphere

Quoc Le Gia

University of New South Wales, Australia

qlegia@unsw.edu.au

We studied the evolution of a two-stage stochastic model for spherical random fields using a time-fractional stochastic hyperbolic diffusion equation on the unit sphere. This equation incorporates a time-fractional derivative in the Caputo sense. The model has two stages: in the first, an isotropic Gaussian random field on the sphere serves as the initial condition for a homogeneous problem. In the second stage, it shifts to an inhomogeneous problem driven by time-delayed Brownian motion on the sphere. The solution is represented by an expansion in spherical harmonics, which we approximate by truncating at a certain degree. Truncation error analysis shows convergence rates influenced by the angular power spectra of the noise and initial conditions. Numerical simulations, inspired by the cosmic microwave background (CMB) data, illustrate the theoretical results. This is a joint work with Tareq Alodat from La Trobe University.

On Convergence of AdaGrad with Relaxed Assumptions

Junhong Lin

Zhejiang University, China

junhong@zju.edu.cn

In this study, we revisit the convergence of AdaGrad with momentum (covering AdaGrad as a special case) on non-convex smooth optimization problems. We consider a general noise model where the noise magnitude is controlled by the function value gap together with the gradient magnitude. This model encompasses a broad range of noises including bounded noise, sub-Gaussian noise, affine variance noise and the expected smoothness, and it has been shown to be more realistic in many practical applications. Our analysis yields a probabilistic convergence rate which, under the general noise, could reach at $\tilde{\mathcal{O}}(1/\sqrt{T})$. This rate does not rely on prior knowledge of problem-parameters and could accelerate to $\tilde{\mathcal{O}}(1/T)$ where T denotes the total number iterations, when the noise parameters related to the function value gap and noise level are sufficiently small. The convergence rate thus matches the lower rate for stochastic first-order methods over non-convex smooth landscape up to logarithm terms [Arjevani et al., 2023]. We further derive a convergence bound for AdaGrad with momentum, considering the generalized smoothness where the local smoothness is controlled by a first-order function of the gradient norm.

Revealing causal information from data

Tongliang Liu
University of Sydney, Australia
tongliang.liu@sydney.edu.au

Many tasks in sciences or engineering require the underlying causal information. Since it is typically expensive and time-consuming to conduct randomized experiments, there has been significant attention towards revealing causal relations through the analysis of purely observational data, commonly known as causal discovery. Over the past few years, with the rapid development of big data, causal discovery is facing great opportunities and challenges. In this talk, I will first introduce some classical causal discovery methods, including PC algorithm and LiNGAM, which has been successfully applied to the cases without latent variable. However, in complex systems, we typically fail to collect and measure all task-relevant variables. In the second part of the talk, I will focus on causal structure recovery in the presence of latent variables. In particular, I will briefly review some researches in this line and introduce our recent work, the latter requires less restrictive assumption and hence can handle more general cases.

Local approximation of operators

Hrushikesh Mhaskar
Claremont Graduate University, United States
Hrushikesh.Mhaskar@cgu.edu

Many applications, such as system identification, classification of time series, direct and inverse problems in partial differential equations, and uncertainty quantification lead to the question of approximation of a non-linear operator between metric spaces $\mathfrak{X}$ and $\mathfrak{Y}$. We study the problem of determining the degree of approximation of such operators on a compact subset $K_{\mathfrak{X}} \subset \mathfrak{X}$ using a finite amount of information. If $\mathcal{F} : K_{\mathfrak{X}} \to K_{\mathfrak{Y}}$, a well established strategy to approximate $\mathcal{F}(F)$ for some $F \in K_{\mathfrak{X}}$ is to encode F (respectively, $\mathcal{F}(F)$) in terms of a finite number d (respectively m) of real numbers. Together with appropriate reconstruction algorithms (decoders), the problem reduces to the approximation of m functions on a compact subset of a high dimensional Euclidean space $\mathbb{R}^d$, equivalently, the unit sphere $\mathbb{S}^d$ embedded in $\mathbb{R}^{d+1}$. The problem is challenging because d, m, as well as the complexity of the approximation on $\mathbb{S}^d$ are all large, and it is necessary to estimate the accuracy keeping track of the inter-dependence of all the approximations involved.

We describe constructive methods to do this efficiently; i.e., with the constants involved in the estimates on the approximation on $\mathbb{S}^d$ being $\mathcal{O}(d^{1/6})$. We study different smoothness classes for the operators, and also propose a method for approximation of $\mathcal{F}(F)$ using only information in a small neighborhood of F, resulting in an effective reduction in the number of parameters involved. To further mitigate the problem of large number of parameters, we propose prefabricated networks, resulting in a substantially smaller number of effective parameters.

Infinite-dimensional statistical distances for functional data analysis

Ha Quang Minh
RIKEN Centre for Advanced Intelligence, Japan
minh.haquang@riken.jp

In this talk, we present an overview of recent results on infinite-dimensional statistical distances between covariance operators and stochastic processes in the setting of functional data analysis. Our focus will be on distances and divergences arising from information geometry and optimal transport, including the Fisher-Rao distance and the Wasserstein distance and its entropic regularization. We discuss the challenges faced in dealing with the infinite-dimensional setting and show in particular that, by using regularization, the infinite-dimensional distances and related quantities can be consistently and efficiently estimated from their finite-dimensional counterparts. The resulting numerical algorithms will be illustrated by applications in functional data analysis.

Recent Trends in Learning Operators

Nicole Mücke
Technical University of Braunschweig, Germany
`nicole.muecke@tu-braunschweig.de`

We consider the problem of learning operators between Hilbert spaces from empirical observations, which we interpret as least squares regression in infinite dimensions with random design. We show that for linear operators, this goal can be reformulated as an inverse problem with the feature that its forward operator is generally non-compact. However, we prove that, in terms of spectral properties and regularization theory, this inverse problem is equivalent to the known compact inverse problem associated with scalar response regression.

Our framework allows for the elegant derivation of dimension-free rates for generic learning algorithms under Hölder-type source conditions. The proofs rely on a combination of techniques from kernel regression and recent results on concentration of measure for sub-exponential Hilbertian random variables. The obtained rates hold for a variety of practically relevant scenarios in functional regression as well as nonlinear regression with operator-valued kernels, and they match those of classical kernel regression with scalar response.

Finally, we review recent trends in learning non-linear operators.

Random effects model-based sufficient dimension reduction for independent clustered data

Linh Nghiem
University of Sydney, Australia
`linh.nghiem@sydney.edu.au`

Sufficient dimension reduction (SDR) is a popular class of regression methods which aim to find a small number of linear combinations of covariates that capture all the information of the responses i.e., a central subspace. The majority of current methods for SDR focus on the setting of independent observations, while the few techniques that have been developed for clustered data assume the linear transformation is identical across clusters. We introduce random effects SDR, where cluster-specific random effect central subspaces are assumed to follow a distribution on the Grassmann manifold, and the random effects distribution is characterized by a covariance matrix on a tangent space. We incorporate random effect SDR within model-based inverse regression frameworks that can handle mixed types of predictors (time-variant/time-invariant, continuous/binary). A two-stage algorithm is proposed to estimate the overall fixed effect central subspace, and predict the cluster-specific random effect central subspaces. We demonstrate the consistency of the proposed estimators, while simulation studies demonstrate the superior performance of the proposed approach compared to global and cluster-specific SDR approaches. Finally, we apply the method to study the longitudinal association between the life expectancy of women and socioeconomic variables across 117 countries from 1990-2015.

Lp Approximation Rates for Location-Scale Mixture Densities and Implications to Adaptive Least-Squares Estimation

Hien Nguyen
La Trobe University, Australia
`H.Nguyen5@latrobe.edu.au`

Approximation and estimation of probability densities on Euclidean spaces constitute an important class of statistical problems. Among the solutions are the use of classes of location-scale mixtures of some fixed density as a basis for conducting approximation. In previous works, we established some denseness results that provide qualitative approximation guarantees for such mixture classes for density approximation in Lp spaces. We iterate on these works by providing the approximation rates in Lp spaces when p is between 1 to infinity (exclusive), which depends only on the the dimension of the supporting space, p, and the smoothness of the target function. We then use these new results to provide novel estimation rates for adaptive least-squares mixture estimators.

MOCCA: A Fast Algorithm for Parallel MRI Reconstruction Using Model Based Coil Calibration

Gerlind Plonka
University of Gottingen, Germany
plonka@math.uni-goettingen.de

We propose a new fast algorithm for simultaneous recovery of the coil sensitivities and the magnetization image from incomplete Fourier measurements in parallel MRI. Our approach is based on a parameter model for the coil sensitivities using bivariate trigonometric polynomials of small degree. The derived MOCCA algorithm has low computational complexity of $\mathcal{O}(N_c N^2 \log N)$ for $N \times N$ images and N_c coils and achieves very good performance for incomplete MRI data. We present a complete mathematical analysis of the proposed reconstruction method. Further, we show that MOCCA achieves similarly good reconstruction results as ESPIRiT with a considerably smaller numerical effort which is due to the employed parameter model. Our numerical examples indicate that MOCCA can outperform several other reconstruction methods. These results haven been obtained in collaboration with Yannick Riebe, University of Goettingen.

An adaptively inexact first-order method for bilevel optimization with application to hyperparameter learning

Lindon Roberts
University of Sydney, Australia
lindon.roberts@sydney.edu.au

A common problem in data science is the determination of model hyperparameters. One approach for learning hyperparameters is to use bilevel optimisation, where the lower-level problem is the standard learning optimisation problem, and the upper-level problem is to learn the hyperparameters (e.g. by minimising validation error). In this setting, particularly for large-scale problems, neither exact function values nor exact gradients are attainable, necessitating methods that only rely on inexact evaluation of such quantities. I will present a new bilevel optimisation algorithm with adaptive inexactness suitable for hyperparameter learning. Numerical results on problems from imaging demonstrate its robustness and strong performance. This is joint work with Mohammad Sadegh Salehi, Matthias Ehrhardt (University of Bath) and Subhadip Mukherjee (IIT Kharagpur).

Data-Efficient Learning for Imperfect Spatio-Temporal Data: Leveraging SSLs and LLMs

Flora Salim
University of New South Wales, Australia
flora.salim@unsw.edu.au

Spatio-temporal data have been very instrumental in developing predictive models in different domains, including for energy, transport and mobility, retail, and public health management. However, modelling with spatiotemporal data is hampered by the limited availability and sparsity of training data, due to the lack of labels and/or semantic information on spatio-temporal data, and issues with the data, including noise and missing data. This talk explores recent advancements in self-supervised learning (SSL) and neural architectures for spatio-temporal data.

SSL, particularly contrastive learning, has shown promise in learning discriminative representations without labeled data, yet existing methods often struggle with multimodal data integration and computational inefficiencies. To address these challenges, we introduce several innovative models: a cross-modal SSL approach that can deal with missing sensor data, and a neural-ODE architecture that can handle sporadic time-series data. To address data sparsity and lack of semantic information in spatio-temporal data, we leverage LLMs to improve the generalizability, in zero-shot and fine-tuned settings. Examples will include the use of LLMs in dynamic contextual information embedding, mobility forecasting, next point-of-interest (POI) recommendation task, and energy load forecasting.

Solving PDEs on Spheres with Physics-Informed Convolutional Neural Networks

Lei Shi
Fudan University, China
`leishi@fudan.edu.cn`

Physics-informed neural networks (PINNs) have been demonstrated to be efficient in solving partial differential equations (PDEs) from a variety of experimental perspectives. Some recent studies have also proposed PINN algorithms for PDEs on surfaces, including spheres. However, theoretical understanding of the numerical performance of PINNs, especially PINNs on surfaces or manifolds, is still lacking. In this talk, we establish rigorous analysis of the physics-informed convolutional neural network (PICNN) for solving PDEs on the sphere. By using and improving the latest approximation results of deep convolutional neural networks and spherical harmonic analysis, we prove an upper bound for the approximation error with respect to the Sobolev norm. Subsequently, we integrate this with innovative localization complexity analysis to establish fast convergence rates for PICNN. Our theoretical results are also confirmed and supplemented by our experiments. In light of these findings, we explore potential strategies for circumventing the curse of dimensionality that arises when solving high-dimensional PDEs. This talk is based on the jointwork with Guanhang Lei, Chenyu Zeng, Prof. Zhen Lei and Prof. Ding-Xuan Zhou.

Training Error and Bayes Error in Deep Learning

Masashi Sugiyama
University of Tokyo, Japan
`sugi@k.u-tokyo.ac.jp`

When solving a classification problem with deep learning, it is not difficult to achieve perfect classification for training data . On the other hand, in many practical classification problems, the minimum achievable test error (i.e., the Bayes error) is not zero. In this talk, we discuss two topics regarding the training error and Bayes error in deep learning: Can we mitigate overfitting by avoiding too small training error and can we estimate the Bayes error accurately?

Optimal stopping under imperfect condition monitoring for non-Markovian systems

Qiuzhuang Sun

University of Sydney, Australia

qiuzhuang.sun@sydney.edu.au

Using mission abort as an example, we study the optimal stopping problem for non-Markovian systems. While most on-demand mission-critical systems are engineered to be reliable to support critical tasks, occasional failures may still occur during missions. To increase system survivability, aborting the mission before an imminent failure is a common practice. We consider optimal mission abort for a system whose deterioration follows a general three-state (normal, defective, failed) semi-Markov chain. The failure is assumed self-revealed, while the healthy and defective states have to be predicted from imperfect condition monitoring data. Due to the non-Markovian process dynamics, optimal mission abort for this partially observable system is an intractable stopping problem. For a tractable solution, we introduce a novel tool of Erlang mixtures to approximate non-exponential sojourn times in the semi-Markov chain. This allows us to approximate the original process by a surrogate continuous-time Markov chain whose optimal control policy can be solved through a partially observable Markov decision process (POMDP). We show that the POMDP optimal policies converge almost surely to the optimal abort decision rules when the Erlang rate parameter diverges. This implies that the expected cost by adopting the POMDP solution converges to the optimal expected cost. Next, we provide comprehensive structural results on the optimal policy of the surrogate POMDP. We develop a modified point-based value iteration algorithm based on the results to numerically solve the surrogate POMDP. We further consider mission abort in a multi-task setting where a system executes several tasks consecutively before a thorough inspection. Through a case study on an unmanned aerial vehicle, we demonstrate the capability of real-time implementation of our model, even when the condition-monitoring signals are generated with high frequency.

Adaptive conformal classification with noisy labels

Rachel Wang

University of Sydney, Australia

rachel.wang@sydney.edu.au

This talk presents recently developed conformal prediction methods for classification tasks that can automatically adapt to random label contamination in the calibration sample, leading to more informative prediction sets with stronger coverage guarantees compared to state-of-the-art approaches. This is made possible by a precise characterisation of the effective coverage inflation (or deflation) suffered by standard conformal inferences in the presence of label contamination, which is then made actionable through new calibration algorithms. Our solution is flexible and can leverage different modeling assumptions about the label contamination process, while requiring no knowledge of the underlying data distribution or of the inner workings of the machine-learning classifier. The advantages of the proposed methods are demonstrated through extensive simulations and an application to object classification with the CIFAR-10H image data set.

Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition

Susan Wei
University of Melbourne, Australia
susan.wei@unimelb.edu.au

We investigate **Bayesian phase transitions** in a Toy Model of Superposition (TMS). As described by [??] and [??], Bayesian learning can be viewed as a search for parameter neighborhoods that minimize free energy – a weighted sum of energy (representing loss) and entropy (representing complexity). For small sample sizes n, learning favors higher energy and lower entropy, while larger n transitions the preference toward lower energy and higher entropy. We derive the energy and entropy terms that govern these transitions in TMS. Empirically, we show that SGD training exhibits similar behavior, a phenomenon we refer to as **dynamical phase transitions**. This supports the conjecture that SGD follows a sequential learning mechanism, driven by a competition between loss and complexity.

Dictionary error of trigonometric least squares approximation with non-uniform distribution

Caroline Wormell
University of Sydney, Australia
caroline.wormell@sydney.edu.au

Functions are often estimated from data by linear least squares, particularly when constructing linear operators. This requires effective operator norm bounds on the error of these approximations, sometimes with control over errors in derivatives. Even in the finite-data limit, there are few results in this direction. I will discuss the case of trigonometric functions on a compact interval: when the data sampling measure is uniform, the problem reduces to Fourier series truncation, but in many use cases (e.g. Koopman operator approximation from time series data) the sampling density is non-uniform. Developing ideas from orthogonal polynomial theory, I will show that, up to a constant, least squares projection against a sufficiently smooth density is as accurate as Fourier truncation, i.e. close to the best possible in low dimensions.

Online Outcome Weighted Learning

Daohong Xiang
Zhejiang Normal University, China
daohongxiang@zjnu.cn

The pursuit of individualized treatment rules in precision medicine has generated significant interest due to its potential to optimize clinical outcomes for patients with diverse treatment responses. One approach that has gained attention is outcome weighted learning, which is tailored to estimate optimal individualized treatment rules by leveraging each patient's unique characteristics under a weighted classification framework. However, traditional offline learning algorithms, which process all available data at once, face limitations when applied to high-dimensional electronic health records data due to its sheer volume. Additionally, the dynamic nature of precision medicine requires that learning algorithms can effectively handle streaming data that arrives in a sequential manner. To overcome these challenges, we present a novel framework that combines outcome weighted learning with online gradient descent algorithms, aiming to enhance precision medicine practices. Our framework provides a comprehensive analysis of the learning theory associated with online outcome weighted learning algorithms, taking into account general classification loss functions. We establish the convergence of these algorithms for the first time, providing explicit convergence rates while assuming polynomially decaying step sizes, with (or without) a regularization term. Our findings present a non-trivial extension of online classification to online outcome weighted learning, contributing to the theoretical foundations of learning algorithms tailored for processing streaming input-output-reward type data.

This is a joint work with Yang Aoli and Fan Jun.

On Discriminative Probabilistic Modeling for Self-Supervised Representation Learning

Tianbao Yang

Texas A&M University, United States

`tianbao-yang@tamu.edu`

In this talk, I will present our recent work on improving self-supervised representation learning. I will present a discriminative probabilistic modeling framework and study its generalization error. I will further present a better method that aims to reduce the generalization error and finally show some experiments.

Adaptive Wavelet Distillation towards Interpretable Deep Learning

Bin Yu

University of California, Berkeley, United States

`binyu@berkeley.edu`

Sparse dictionary learning has a rich history and is known for producing wavelet-like filters when applied to natural image patches, analogous to the V1 primary visual cortex in the human brain. Wavelets, which function as localized Fourier Transforms, are widely interpretable across the physical sciences and beyond. In this talk, we introduce Adaptive Wavelet Distillation (AWD), a method that enhances interpretability of black-box deep learning models in applications such as cosmology and cellular biology, while simultaneously improving predictive performance.

Additionally, we present theoretical insights demonstrating that, under simple sparse dictionary models, gradient descent in autoencoder training converges to a point on a manifold of global minima, with the specific minimum influenced by batch size. Notably, we show that small batch sizes, as used in stochastic gradient descent (SGD), lead to a qualitatively different form of "feature selection", that is sparse.

Optimal Classification-based Anomaly Detection with Neural Networks: Theory and Practice in Cybersecurity

Tian-Yi Zhou

Georgia Institute of Technology, United States

`tzhou306@gatech.edu`

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behaviors. It has emerged as an important technique in many application areas, notably for network intrusion detection in cybersecurity. Existing deep learning approaches for AD produce sound empirical performance but lack theoretical guarantees. In the practice of network intrusion detection, collecting labeled network traffic data is expensive. Often, we are given only unlabeled data during training. In this work, we design a neural network based method to distinguish anomalies from normal samples in the presence of unlabeled data. We first model the unsupervised AD problem as a density level set estimation problem and then transform this level set estimation problem to a nonparametric classification problem. We proposed a neural network based method to solve this special classification problem. Our method achieves minimax optimal convergence rates on the excess risk trained on synthetic anomalies. We implement our method for various network intrusion detection tasks and witness competitive performance compared to other existing methods.

List of Participants

Shayan Azizi
University of New South Wales, Australia, `s.azizi@unsw.edu.au`

Stephen Baker
University of South Australia, Australia, `boobaker59@gmail.com`

Wei Cai (See abstract, p. 4)
Southern Methodist University, United States, `cai@mail.smu.edu`

Andreas Christmann (See abstract, p. 4)
University of Bayreuth, Germany, `andreas.christmann@uni-bayreuth.de`

Tiangang Cui (See abstract, p. 4)
University of Sydney, Australia, `tiangang.cui@sydney.edu.au`

Declan Curran
University of New South Wales, Australia, `d.curran@unsw.edu.au`

Edward Dong
University of New South Wales, Australia, `edward.dong1@unsw.edu.au`

Jun Fan (See abstract, p. 5)
Hong Kong Baptist University, Hong Kong, China, `junfan@hkbu.edu.hk`

Han Feng (See abstract, p. 5)
City University of Hong Kong, Hong Kong, China, `hanfeng@cityu.edu.hk`

Simon Foucart (See abstract, p. 5)
Texas A&M University, United States, `foucart@tamu.edu`

Ian Gallagher (See abstract, p. 6)
University of Melbourne, Australia, `ian.gallagher@unimelb.edu.au`

Junbin Gao (See abstract, p. 6)
University of Sydney, Australia, `junbin.gao@sydney.edu.au`

Mingming Gong (See abstract, p. 6)
University of Melbourne, Australia, `mingming.gong@unimelb.edu.au`

Georg Gottwald (See abstract, p. 7)
University of Sydney, Australia, `georg.gottwald@sydney.edu.au`

Clara Grazian (See abstract, p. 7)
University of Sydney, Australia, `clara.grazian@sydney.edu.au`

Xin Guo (See abstract, p. 7)
University of Queensland, Australia, `xin.guo@uq.edu.au`

Zheng-Chu Guo (See abstract, p. 8)
Zhejiang University, China, `guozc@zju.edu.cn`

Markus Holzleitner (See abstract, p. 8)
University of Genoa, Italy, `markus.holzleitner1@gmail.com`

Devin Hua
University of New South Wales, Australia, `devin.hua@unsw.edu.au`

Mahdi Jafari
University of New South Wales, Australia, `mahdi.jafari@unsw.edu.au`

Hakiim Jamaluddin
University of New South Wales, Australia, `a.jamaluddin@unsw.edu.au`

Caleb Ju (See abstract, p. 8)
Georgia Institute of Technology, United States, `calebju4@gatech.edu`

Ata Kaban (See abstract, p. 9)
University of Birmingham, United Kingdom, `a.kaban@cs.bham.ac.uk`

Quoc Le Gia (See abstract, p. 9)
University of New South Wales, Australia, `qlegia@unsw.edu.au`

Junhong Lin (See abstract, p. 9)
Zhejiang University, China, `junhong@zju.edu.cn`

Peilin Liu
University of Sydney, Australia, `P.Liu@maths.usyd.edu.au`

Tongliang Liu (See abstract, p. 10)
University of Sydney, Australia, `tongliang.liu@sydney.edu.au`

Pinak Mandal
University of Sydney, Australia, `pinak.mandal@sydney.edu.au`

Hrushikesh Mhaskar (See abstract, p. 10)
Claremont Graduate University, United States, `Hrushikesh.Mhaskar@cgu.edu`

Ha Quang Minh (See abstract, p. 10)
RIKEN Centre for Advanced Intelligence, Japan, `minh.haquang@riken.jp`

Nicole Mücke (See abstract, p. 11)
Technical University of Braunschweig, Germany, `nicole.muecke@tu-braunschweig.de`

Linh Nghiem (See abstract, p. 11)
University of Sydney, Australia, `linh.nghiem@sydney.edu.au`

Hien Nguyen (See abstract, p. 11)
La Trobe University, Australia, `H.Nguyen5@latrobe.edu.au`

Gerlind Plonka (See abstract, p. 12)
University of Gottingen, Germany, `plonka@math.uni-goettingen.de`

Alun Pope
Analytical Insight Pty Ltd & Monash University, Australia, `alunpope@gmail.com;Alun.Pope@monash.edu`

Arian Prabowo
University of New South Wales, Australia, `arian.prabowo@unsw.edu.au`

Hamada Rizk
Osaka University, Japan, `hamada_rizk@ist.osaka-u.ac.jp`

Lindon Roberts (See abstract, p. 12)
University of Sydney, Australia, `lindon.roberts@sydney.edu.au`

Hira Saleem
University of New South Wales, Australia, `h.saleem@unsw.edu.au`

Flora Salim (See abstract, p. 12)
University of New South Wales, Australia, `flora.salim@unsw.edu.au`

Lei Shi (See abstract, p. 13)
Fudan University, China, `leishi@fudan.edu.cn`

Masashi Sugiyama (See abstract, p. 13)
University of Tokyo, Japan, `sugi@k.u-tokyo.ac.jp`

Qiuzhuang Sun (See abstract, p. 14)
University of Sydney, Australia, `qiuzhuang.sun@sydney.edu.au`

Kejia Tang
University of Sydney, Australia, `K.Tang@maths.usyd.edu.au`

Maria Vivien Visaya
University of Johannesburg, South Africa, `mvvisaya@uj.ac.za`

Rachel Wang (See abstract, p. 14)
University of Sydney, Australia, `rachel.wang@sydney.edu.au`

Susan Wei (See abstract, p. 15)

University of Melbourne, Australia, `susan.wei@unimelb.edu.au`

Caroline Wormell (See abstract, p. 15)
University of Sydney, Australia, `caroline.wormell@sydney.edu.au`

Charles Wyatt
University of New South Wales, Australia, `charles.wyatt@student.unsw.edu.au`

Daohong Xiang (See abstract, p. 15)
Zhejiang Normal University, China, `daohongxiang@zjnu.cn`

Tianbao Yang (See abstract, p. 16)
Texas A&M University, United States, `tianbao-yang@tamu.edu`

Yiming Ying
University of Sydney, Australia, `yiming.ying@sydney.edu.au`

Bin Yu (See abstract, p. 16)
University of California, Berkeley, United States, `binyu@berkeley.edu`

Dingxuan Zhou
University of Sydney, Australia, `dingxuan.zhou@sydney.edu.au`

Junyu Zhou
University of Sydney, Australia, `junyu.zhou@sydney.edu.au`

Tian-Yi Zhou (See abstract, p. 16)
Georgia Institute of Technology, United States, `tzhou306@gatech.edu`