Book of Abstracts

of the Brummer & Partners MathDataLab

Conference on the Mathematics of Complex Data

held at KTH Royal Institute of Technology
in Stockholm, June 13–16, 2022.

June 4, 2022

Index of talks by speaker

Plenary Talks

Metric Representations: Algorithms and Geometry

Anna Gilbert
Yale University

Given a set of distances amongst points, determining what metric representation is most "consistent" with the input distances or the metric that best captures the relevant geometric features of the data is a key step in many machine learning algorithms. In this talk, we discuss a number of variants of this problem, from convex optimization problems with metric constraints to sparse metric repair.

Morse-theoretic signal compression and reconstruction

Kathryn Hess
EPFL

In this lecture I will present work of three of my PhD students, Stefania Ebli, Celia Hacker, and Kelly Maggs, on cellular signal processing. In the usual paradigm, the signals on a simplicial or chain complex are processed using the combinatorial Laplacian and the resultant Hodge decomposition. On the other hand, discrete Morse theory has been widely used to speed up computations, by reducing the size of complexes while preserving their global topological properties. Ebli, Hacker, and Maggs have developed an approach to signal compression and reconstruction on chain complexes that leverages the tools of algebraic discrete Morse theory,, which provides a method to reduce and reconstruct a based chain complex together with a set of signals on its cells via deformation retracts, preserving as much as possible the global topological structure of both the complex and the signals. It turns out that any deformation retract of real degreewise finite-dimensional based chain complexes is equivalent to a Morse matching. Moreover, in the case of certain interesting Morse matchings, the reconstruction error is trivial, except on one specific component of the Hodge decomposition. Finally, the authors developed and implemented an algorithm to compute Morse matchings with minimal reconstruction error, of which I will show explicit examples.

Two problems in statistical estimation for high-dimensional dynamical systems

Mauro Maggioni
Johns Hopkins University

We discuss two problems at the intersection of dynamical systems and statistical estimation/inference/machine learning. In the first problem we consider systems of interacting agents or particles, which are commonly used for modeling across the sciences. Oftentimes the laws of interaction between the agents are quite simple, for example they depend only on pairwise interactions, and only on pairwise distance in each interaction. We consider the following inference problem for a system of interacting particles or agents: given only observed trajectories of the agents in the system, can we learn what the laws of interactions are? We would like to do this without assuming any particular form for the interaction laws, i.e. they might be "any" function of pairwise distances. We consider this problem both the mean-field limit (i.e. the number of particles going to infinity) and in the case of a finite number of agents, with an increasing number of observations, albeit in this talk we will mostly focus on the latter case. We cast this as an inverse problem, and present a solution in the simplest yet interesting case where the interaction is governed by an (unknown) function of pairwise distances. We discuss when this problem is well-posed, we construct estimators for the interaction kernels with provably good statistically and computational properties, and discuss extensions to second-order systems, more general interaction kernels, stochastic systems, and to the setting where the variables (e.g. pairwise distance) on which the interaction kernel depends are not known a priori. This is joint work with F. Lu, J.Miller, S. Tang and M. Zhong. The second problem we consider is that of estimating invariant manifolds, and effective equations on them, for fast-slow stochastic systems in high-dimensions. We introduce a nonlinear stochastic model reduction technique for high-dimensional stochastic dynamical systems that have a low-dimensional invariant effective manifold with slow dynamics, and high-dimensional, large fast modes. Given only access to a black box simulator from which short bursts of simulation can be obtained, we design an algorithm that outputs an estimate of the invariant manifold, a process of the effective stochastic dynamics on it, which has averaged out the fast modes, and a simulator thereof. This simulator is efficient in that it exploits of the low dimension of the invariant manifold, and takes time steps of size dependent on the regularity of the effective process, and therefore typically much larger than that of the original simulator, which had to resolve the fast modes. The algorithm and the estimation can be performed on-the-fly, leading to efficient exploration of the effective state space, without losing consistency with the underlying dynamics. This construction enables fast and efficient simulation of paths of the effective dynamics, together with estimation of crucial features and observables of such dynamics, including the stationary distribution, identification of metastable states, and residence times and transition rates between them. This is joint work with S. Yang and X.-F. Ye.

Efficient federated Bayesian sampling by Stochastic Averaging Langevin Dynamics

Eric Moulines
Ecole polytechnique

In this work, we develop new methods for Bayesian computation in a federated learning context (FL). While there are a variety of distributed MCMC algorithms, few have been developed to address the specific constraints of FL, such as data privacy, communication bottleneck, and statistical heterogeneity. To tackle these issues, we propose SALaD, a MCMC algorithm that combines the ideas of Stochastic Langevin Gradient Dynamics and Federated Averaging. In each round, each client executes SGLD to update its local parameter, which is sent to a central server. The central server then in turn sends the average of the local parameters to the clients. However, this method may suffer from the high variance of the stochastic gradients used by local SGLD and the heterogeneity of the data, which hinders and/or slows down convergence. To address these issues, we propose three alternatives based on a combination of control variates and bias reduction techniques for which theoretical improvements are derived. We illustrate our findings using several FL benchmarks for Bayesian inference.

Geometry of Dependency Equilibria

Bernd Sturmfels
MPI Leipzig

An n-person game is specified by n tensors of the same format. Its equilibria are points in that tensor space. Dependency equilibria satisfy linear constraints on conditional probabilities. These cut out the Spohn variety, named after the philosopher who introduced the concept. Nash equilibria are tensors of rank one. We discuss the real algebraic geometry of the Spohn variety and its payoff map, with emphasis on connections to oriented matroids and algebraic statistics.

Logistic regression with little noise

Sara van de Geer
Seminar for Statistics, ETH Zürich

Theoretical results for the logistic regression model typically assume that the observed binary label has probabilities staying away from zero. We study a situation where this assumption is violated, which is the case where the the noise level is low. There may even be no noise at all, which is often the case in the literature on 1-bit compressed sensing. Consider a feature vector $\mathbf{x} \in \mathbb{R}^s$ and a vector of regression coefficients $\beta^* \in \mathcal{S}_2^{s-1}$ with $s \log n \ll n$ and let $\mathbf{y} \in \{\pm 1\}$ be the sign of $\mathbf{x}\beta^* + \sigma\zeta$ where the noise $\zeta \sim \mathcal{N}(0,1)$ is independent of $\mathbf{x}$ and $0 < \sigma \leq 1$. Then $1/\sigma$ is the signal-to-noise level, since only the ratio of signal strength and noise level is identified. We observe n i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^n$ of $(\mathbf{x}, \mathbf{y})$. With feature vector $x \in \mathbb{R}^s$ and label $y \in \{\pm 1\}$ the logistic loss function is $l_c(x, y) := \log(1 + \mathrm{e}^{-yxc})$. We examine the estimator $\hat{\gamma} := \arg\min_{c \in \mathbb{R}^s} \sum_{i=1}^n l_c(X_i, Y_i)$. Let $\hat{\beta} := \hat{\gamma}/\|\hat{\gamma}\|_2$. We show that for the case of Gaussian design, the rate of convergence for $\|\hat{\beta} - \beta^*\|_2$ is of order $\sqrt{\sigma s \log n/n} \vee s \log n/n$. Empirical risk minimization with $\{0,1\}$ loss $1\{\mathrm{sign}(xc) \neq y\}$ has the rate $(\sigma^2 s \log n/n)^{1/3} \vee s \log n/n$. Thus, with Gaussian design, logistic regression appears to have a faster rate. For the high-dimensional case, we discuss an approach using ℓ_0-regularization.

Scalable semidefinite programming

Joel Tropp
California Institute of Technology

Semidefinite programming (SDP) is a powerful framework from convex optimization that has striking potential for data science applications. This talk describes a provably correct randomized algorithm for solving large, weakly constrained SDP problems by economizing on the storage and arithmetic costs. Numerical evidence shows that the method is effective for a range of applications, including relaxations of MaxCut, abstract phase retrieval, and quadratic assignment problems. Running on a laptop equivalent, the algorithm can handle SDP instances where the matrix variable has over 10^14 entries. This talk will highlight the ideas behind the algorithm in a streamlined setting. The insights include a careful problem formulation, design of a bespoke optimization method, and use of randomized matrix computations. Joint work with Alp Yurtsever, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Based on arXiv 1912.02949 (Scalable SDP, SIMODS 2021) and other papers (SketchyCGM in AISTATS 2017, Nyström sketch in NeurIPS 2017).

Invited Talks

Estimating Gaussian mixtures using sparse polynomial moment systems

Carlos Améndola
MPI-MiS / TU Berlin

The method of moments is a statistical technique for density estimation that solves a system of moment equations to estimate the parameters of an unknown distribution. A fundamental question critical to understanding identifiability asks how many moment equations are needed to get finitely many solutions and how many solutions there are. Since the moments of a mixture of Gaussians are polynomial expressions in the means, variances and mixture weights, one can address this question from the perspective of algebraic geometry. With the help of tools from polyhedral geometry, we answer this fundamental question for several classes of Gaussian mixture models. Furthermore, these results allow us to present an algorithm that performs parameter recovery and density estimation, applicable even in the high dimensional case. Based on joint work with Julia Lindberg and Jose Rodriguez.

AI-generated hallucinations: Why do they happen when stable/accurate NNs exist?

Vegard Antun
University of Oslo

Artificial intelligence (AI) is currently entering many industries with full force. Yet, there is also overwhelming empirical evidence that much of modern AI is non-robust (unstable) and may produce non-trustworthy outputs. This is particularly true in high-risk applications such as medical imaging, where the issue of AI generated hallucinations now is causing serious concerns. In this talk, we present a comprehensive mathematical analysis explaining the many facets of AI generated hallucinations in imaging, their links to instabilities, but also how stable AI methods can hallucinate. Furthermore, we will show examples of basic well-conditioned problems in scientific computing where neural networks (NNs) with great approximation qualities are proven to exist, however, there does not exist any algorithm, even randomised, that can train (or compute) such a NN to even 1-digit of accuracy with a probability greater than 1/2. These results provide basic foundations for Smale's 18th problem ("What are the limits of AI?") and imply a potentially vast classification theory describing conditions under which (stable) NNs with a given accuracy can be computed by an algorithm. We begin this theory by initiating a unified theory for compressed sensing and deep learning, leading to sufficient conditions for the existence of algorithms that compute stable NNs in inverse problems.

Fully Stochastic Reconstruction Methods in Coupled Physics Imaging

Simon Arridge
University College London

Coupled Physics Imaging methods combine image contrast from one physical process with observations using a secondary process; several modalities in acousto-optical imaging follow this concept wherein optical contrast is observed with acoustic measurements. For the inverse problem both an optical and acoustic model need to be inverted. Classical methods that involve a non-linear optimisation approach can be combined with advances in stochastic subsamplings strategies that are in part inspired by machine learning applications. In such approaches the forward problem is considered deterministic and the stochasticity involves splitting of an objective function into sub functions that approach the fully sampled problem in an expectation sense. In this work we consider where the forward problem is also solved stochastically, by a Monte Carlo simulation of photon propagation. By adjusting the batch size in the forward and inverse problems together, we can achieve better performance than if subsampling is performed seperately. Joint work with S Powell, C Macdonald, N Hänninen, A Pulkkinen, T Tarvainen.

Modern Tensor Factorizations and Applications to Longitudinal 'Omics' Data Analysis

Haim Avron
Tel Aviv University

Many real-world data are inherently of multi-way structure. However, multidimensional data is often processed as two-dimensional arrays (matrices), thus, ignoring the inherent higher-order structure therein. Arguably, matrization of higher-order data is such a common practice due to the ubiquitousness and strong theoretical foundations of matrix algebra. In the talk, I will discuss recent developments in multilinear algebra that have established an Eckart-Young like best rank approximation for the tensor tubal singular value decomposition (tSVDM), providing theoretical justification for the superiority of tensor-based approximations over matrization for the first time. I will then further explain how we utilized the tSVDM to construct a principal component analysis (PCA) analog for 3rd order tensors which we refer to as the M-product based Tensor Component Analysis (TCAM). We derive optimality for the TCAM, namely, the maximization of variance and distortion minimization for the TCAM embedding. These theoretical guarantees put TCAM as a promising utility for multi-way data analysis tasks. We explore this utility in the context of analyzing high dimensional so-called "omics" data, which is collected in longitudinal biological experiments.

Optimization for statistical estimators (applications to quantum fidelity estimation)

Stephen Becker
University of Colorado Boulder, Dept. of Applied Mathematics

Statistics and optimization are closely linked, and it is common to use optimization to compute statistical *estimates*, e.g., maximum likelihood estimation. However, optimization can also be used to compute *estimators* which are the rules used to compute estimates. This usage is not novel but is less known. We review a program of minimax risk estimators starting with work of Donoho (Annals of Statistics '94) and generalized by Juditsky and Nemirovski (Annals of Statistics '09). Following the exposition, we show how the framework can be applied to make rigorous direct estimates of the fidelity of a quantum state without needing to reconstruct an estimate of the state itself (which would be computationally expensive and ill-posed). Joint work with Akshay Seshadri, Martin Ringbauer, Rainer Blatt, Thomas Monz, Stephen Becker; https://arxiv.org/abs/2112.07925 and https://arxiv.org/abs/2112.07947

Line Multiview Varieties

Paul Breiding
University of Osnabrück / MPI MiS Leipzig

The mathematical abstraction of a pinhole camera is a projective linear map given by a 3×4 matrix. Suppose that we have a line in three-dimensional projective space and take m images of this line under using m different pinhole cameras. This produces an arrangement of m lines in two-space, called a line correspondence. I will present line correspondences for pinhole cameras from the point of view of algebraic geometry. We define the line multiview variety as the (complex) Zariski closure of the set of all line correspondences with m fixed cameras. We prove that in the case of generic camera matrices it is characterized by a natural determinantal variety and we provide a complete description for any camera arrangement. We investigate basic properties of this variety such as dimension, smoothness and multidegree. This is a joint work with Felix Rydell, Elima Shehu and Angelica Torres.

Does the Data Induce Capacity Control in Deep Learning?

Pratik Chaudhari
University of Pennsylvania

Accepted statistical wisdom suggests that the larger the model class, the more likely it is to overfit the training data. And yet, deep networks generalize extremely well. The larger the deep network, the better its accuracy on new data. This talk seeks to shed light upon this apparent paradox. We will argue that deep networks are successful because of a characteristic structure in the space of learning tasks. The input correlation matrix for typical tasks has a peculiar ("sloppy") eigenspectrum where, in addition to a few large eigenvalues (salient features), there are a large number of small eigenvalues that are distributed uniformly over exponentially large ranges. This structure in the input data is strongly mirrored in the representation learned by the network. A number of quantities such as the Hessian, the Fisher Information Matrix, as well as others activation correlations and Jacobians, are also sloppy. Even if the model class for deep networks is very large, there is an exponentially small subset of models (in the number of data) that fit such sloppy tasks. This talk will demonstrate the first analytical non-vacuous generalization bound for deep networks that does not use compression. We will also discuss an application of these concepts that develops new algorithms for semi-supervised learning.

The Kick-Kac teleportation algorithm: boost your favorite Markov Chain Monte Carlo using Kac formula

Alain Durmus
ENS PS

In this work, we propose to target a given probability measure π by combining two Markov kernels with different invariant probability measures. In its basic form, the mechanism consists in picking up the current position and moving it according to a π-invariant Markov kernel as soon as the proposed move does not fall into a predefined region. If this is the case, then we resort to the last position in this region and move it according to another auxiliary Markov kernel before starting another excursion outside the region with the first kernel. These state dependent interactions allow to combine smoothly different dynamics that can be taylored to each region while the resulting process still targets the probability measure π thanks to an argument based on the Kac formula. Under weak conditions, we obtain the Law of Large numbers starting from any point of the state space, as a byproduct of the same property for the different implied kernels. Geometric ergodicity and Central Limit theorem are also established. Generalisations where the indicator function on the region target is replaced by an arbitrary acceptance probability are also given and allow to consider any Metropolis Hastings algorithm as a particular case of this general framework. Numerical examples, including mixture of Gaussian distributions are also provided and discussed. This is joint work with Randal Douc, Aurélien Enfroy and Jimmy Olsen

Proximal algorithms with matrix spectral functions

Hamza Fawzi
Department of Applied Mathematics and Theoretical Physics, University of Cambridge

A spectral function on matrices is a permutation-invariant function of its singular values. Such functions arise in many optimization problems (e.g., logdet for symmetric matrices, or the nuclear norm). Computing the proximal operator of a convex spectral function, a fundamental task in optimization algorithms, requires a full singular value decomposition which can be very costly. In this talk we consider the problem of *approximating* the proximal operator of a convex spectral function. We prove that an ε-approximate SVD can be used to approximate the proximal operator to within ε; our bounds are remarkably simple and do not depend on any additional or unknown constants. We illustrate our result with the Jacobi method, an old and surprisingly simple iterative method for eigenvalue decomposition. By using warm starts and early termination, we show that one can approximately compute the proximal step at each iteration of the optimization algorithm with a significant speed-up compared to standard baselines, and without compromising convergence of the optimization algorithm.

Some Mathematical Approaches to Explainable AI

Emily J. King
Colorado State University

Explainable AI (XAI), also known as Interpretable AI, or Explainable Machine Learning (XML), is currently a topic of interest. There are certain fields in particular, like medicine and atmospheric science, where the researchers would prefer not to use black box methods. In this talk, results from two mathematical approaches to XAI will be presented. The first class of approaches involve using mathematical tools to better understand neural networks. The second class of approaches involve using mathematical methods, some "old fashioned", to perform certain tasks in an interpretable way.

The Geometry of Linear Convolutional Networks

Kathlén Kohn
KTH

We discuss linear convolutional neural networks (LCNs) and their critical points. We observe that the function space (i.e., the set of functions represented by LCNs) can be identified with polynomials that admit certain factorizations, and we use this perspective to describe the impact of the network's architecture on the geometry of the function space. For instance, for LCNs with one-dimensional convolutions having stride one and arbitrary filter sizes, we provide a full description of the boundary of the function space. We further study the optimization of an objective function over such LCNs: We characterize the relations between critical points in function space and in parameter space and show that there do exist spurious critical points. We compute an upper bound on the number of critical points in function space using Euclidean distance degrees and describe dynamical invariants for gradient descent. This talk is based on joint work with Thomas Merkh, Guido Montúfar, and Matthew Trager.

The Convex Geometry of Matrix Completion Revisited

Felix Krahmer
Technical University of Munich

Low-rank matrix recovery from structured measurements has been a topic of intense study in the last decades. An instance of this problem that is of particular interest due to its relevance for many applications such as recommender systems is the matrix completion problem, where the observations consist of randomly selected matrix entries. An important benchmark method to solve this and related problems is to minimize the nuclear norm, a convex proxy for the rank. A common approach to establish recovery guarantees for this convex program relies on the construction of a so-called approximate dual certificate. However, this approach provides only limited insight in various respects. In particular, the best-known bounds for the reconstruction error under adversarial noise involve seemingly spurious dimensional factors. In this talk, we analyze the problem from a geometric perspective and show that these dimension factors in the noise bounds are not an artefact of the proof, but cannot be avoided in the framework commonly applied for the analysis, which aims to establish a linear scaling of the error in terms of the noise level. At the same time, we establish that these factors only arise for very small noise levels, and if one accepts a square root scaling, the constants can be chosen independently of the dimension and depending only mildly on the rank. This is joint work with Yulia Kostina (TUM) and Dominik Stöger (KU Eichstätt/Ingolstadt).

Log-concave density estimation in undirected graphical models

Kaie Kubjas
Aalto University

Given i.i.d. samples, we study the maximum likelihood estimation problem over probability densities that factor according to an undirected graph G such that the factors are log-concave. We show that the maximum likelihood estimate exists and is unique with probability 1 if the sample size is at least the size of a maximal clique in a chordal cover of G. The optimal solution to this problem is a product of exponentials of tent functions, one for each clique. We use this observation to implement an algorithm for log-concave density estimation in undirected graphical models. This talk is based on joint work with Olga Kuznetsova, Elina Robeva, Pardis Semnani and Luca Sodomaco.

Topologies of Random Geometric Complexes on Riemannian Manifolds in the Thermodynamic Limit

Antonio Lerario
SISSA

I will discuss the topology of random geometric complexes built over random points sampled on Riemannian manifolds, in the so-called "thermodynamic" regime, proving universal limit laws for their homotopy types. Joint work with A. Auffinger and E. Lundberg.

Deep learning for PDE-based inverse problems

Peter Maass
University Bremen, Gemany

In this talk we analyse and compare different deep learning approaches such as Deep Ritz, Fourier Neural Operators, DeepOnet, TorchPhysics and others for a hierarchy of PDE-based problems: 1) forward solvers of single PDEs, 2) parametric studies of PDEs, 3) parameter identification for PDEs (inverse problems). The last topic typically is not covered by the original concepts and requires novel approaches. The main part of the talk is on numerical experiments with those different schemes. We also present preliminary results for the theoretical research in progress and include an application in collaboration with Volkswagen. This is joint work with Derick Nganyu Tanyu and Jianfeng Ning.

Random matrix theory and modern machine learning

Michael W. Mahoney
ICSI, LBNL, and Department of Statistics, UC Berkeley

Random matrix theory (RMT) has a long history and has proven useful in many areas, most recently modern machine learning (ML). Some of the most interesting and promising uses of RMT in ML go beyond a direct application of popular Wigner and Marcenko-Pastur ideas to linear models in a high-dimensional regime, and instead must account for the peculiar properties of modern ML systems. These properties include: that the data dimension is not vanishingly small compared to the number of data points; that there are "layers" with intermediate features that can be large, comparable to, or smaller in size than the number of data points; that there are strongly non-linear models; and that, compared to parameter vectors or matrices, we are in general more interested in (scalar) functionals of those things (e.g., regression error and classification accuracy). Here, we provide an overview of recent developments in RMT that go beyond traditional ideas and that are well-suited for the analysis of modern ML models. Central to our approach is understanding different behaviors of linear models in the high-dimensional regime compared to their low-dimensional counterparts (where infinite norm and operator norm of matrices must not be considered "equivalent"), and how these ideas extend to non-linear models.

Adaptive Algorithms

Yura Malitsky
Linköping University, Sweden

In optimization, convergence rates and complexity are essential notions to characterize the performance of algorithms. In most cases, however, such results hold for an arbitrary function class with generic properties. On the other hand, each time we minimize a particular function which is not necessarily the worst case function. Adaptive algorithms aim to harness the further structure of optimization problems to go beyond such worst-case analysis. In this talk, I will discuss some examples of such algorithms and show how the step sizes in these algorithms give rise to compelling practical performance for applications such as neural network training or min-max games.

Directed Gaussian graphical models with toric vanishing ideal

Pratik Misra
KTH Royal Institute of Technology

Directed Gaussian graphical models are statistical models that use a directed acyclic graph (DAG) to represent the conditional independence structures between a set of jointly random variables. The study of generators of the vanishing ideal of a DAG is an important problem for constraint based inference for inferring the structure of the underlying graph from data. In this talk, I will make an attempt to characterize the DAGs whose vanishing ideals are toric. In particular, I will give some combinatorial criteria to construct such DAGs from smaller DAGs which have toric vanishing ideals. In the end, for DAGs having toric vanishing ideal, I will present some results about the generating sets of those ideals.

A stochastic control approach to quasi-stationary distributions

Pierre Nyquist
KTH Royal Institute of Technology

Quasi-stationary distributions (QSDs) are a core concept within applied and computational probability. For example, they are at the heart of the study of population processes, and for systems exhibiting metastability, QSDs determine important quantities such as mean exit times and exit points from metastable states. More recently QSDs have appeared as central objects in the study and design of efficient Monte Carlo methods. In this talk, I will introduce a new approach for studying QSDs based on ergodic stochastic control problems. In the setting of diffusions on a bounded domain, I will describe the link between QSDs and such control problems, along with how the associated Hamilton-Jacobi-Bellman equations can be used to characterise important properties of the QSD. Time permitting, I will also mention how this connection can be used to construct efficient numerical schemes, and understand and explain non-uniqueness of QSDs in unbounded domains. This is based on joint work with A. Budhiraja, P. Dupuis and G.-J. Wu.

Diffusion models, neural nets, and stochastic calculus of variations

Maxim Raginsky
UIUC

The probabilistic approach to establishing universal approximation theorems for neural nets, pioneered by Andrew Barron, hinges on representing a function of interest as an expectation of a nonlinear activation function whose (finite-dimensional) parameters are drawn from some probability measure. One can then obtain a finite-size neural net by Monte Carlo sampling, and the resulting approximation error is governed by the second moment of this probability measure. In this talk, based on joint work with Tanya Veeravalli, I will discuss a natural generalization of this idea to the setting when the underlying parameters are infinite-dimensional, corresponding to a realization of a diffusion process on a finite time interval. Sharp upper and lower bounds on transition densities of diffusion processes were obtained by Shuenn-Jyi Sheu in terms of the solution of a certain deterministic optimal control problem. I will build on these results to show that the expressive power of diffusion models can be quantified in a manner similar to Barron's probabilistic approach.

ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally!

Peter Richtárik
KAUST, Saudi Arabia

We introduce ProxSkip—a surprisingly simple and provably efficient method for minimizing the sum of a smooth (f) and an expensive nonsmooth proximable (ψ) function. The canonical approach to solving such problems is via the proximal gradient descent (ProxGD) algorithm, which is based on the evaluation of the gradient of f and the prox operator of ψ in each iteration. In this work we are specifically interested in the regime in which the evaluation of prox is costly relative to the evaluation of the gradient, which is the case in many applications. ProxSkip allows for the expensive prox operator to be skipped in most iterations: while its iteration complexity is $O(\kappa \log 1/\epsilon)$, where κ is the condition number of f, the number of prox evaluations is $O(\sqrt{\kappa} \log 1/\epsilon)$ only. Our main motivation comes from federated learning, where evaluation of the gradient operator corresponds to taking a local GD step independently on all devices, and evaluation of prox corresponds to (expensive) communication in the form of gradient averaging. In this context, ProxSkip offers an effective acceleration of communication complexity. Unlike other local gradient-type methods, such as FedAvg, Scaffold, S-Local-GD and FedLin, whose theoretical communication complexity is worse than, or at best matching, that of vanilla GD in the heterogeneous data regime, we obtain a provable and large improvement without any heterogeneity-bounding assumptions.

Partially observing graphs – when can we infer underlying community structure?

Fiona Skerman
Uppsala University

Suppose edges in an underlying graph G appear independently with some probability in our observed graph G' – or alternately that we can query uniformly random edges. We describe how high a sampling probability we need to infer the modularity of the underlying graph. Modularity is a function on graphs which is used in algorithms for community detection. For a given graph G, each partition of the vertices has a modularity score, with higher values indicating that the partition better captures community structure in G. The (max) modularity $q^*(G)$ of the graph G is defined to be the maximum over all vertex partitions of the modularity score, and satisfies $0 \leq q^*(G) \leq 1$. In the seminar I will spend time on intuition for the behaviour of modularity, how it can be approximated and links to other graph parameters. Joint work with Colin McDiarmid.

Time Square sampling: stochastic approximation in the wild

Michael Snarski
D. E. Shaw Research

Times Square sampling (TSS) is a sampling algorithm that runs on the molecular dynamics supercomputer Anton 3, as well as on commodity hardware, and estimates free energy differences on-the-fly using stochastic approximation. In this talk, I'll briefly cover the drug discovery pipeline and explain how free energy differences play an important role in the search for drug-like molecules. I'll discuss the key modifications to the stochastic approximation procedure which allow TSS to retain good convergence properties for a broad range of sampling problems, as is necessitated by the complex and diverse nature of chemical systems studied in practice, and I will highlight some promising research directions.

Modeling Soft Interventions in Context-Specific Settings

Liam Solus
KTH Royal Institute of Technology

A well-known limitation of modeling causal systems via DAGs is their inability to encode context-specific information. Among the several proposed representations for context-specific causal information are the staged tree models, which are colored probability trees capable of expressing highly diverse context-specific information. The expressive power of staged trees comes at the cost of easy interpretability and the admittance of desirable properties useful in the development of causal discovery algorithms. Using the algebro-geometric properties of staged tree models we find that a subfamily of staged tree models admit an alternative representation as a sequence of DAGs and allow us to generalize the key properties driving causal discovery algorithms. These results generalize to context-specific causal models under both hard and soft interventions. We will discuss these results and their potential applications to causal discovery algorithms for context-specific models based on interventional and observational data.

Normalization Effects in Machine Learning and Deep Learning algorithms

Konstantinos Spiliopoulos
Boston University

Machine learning, and in particular neural network models, have revolutionized fields such as image, text, and speech recognition. Today, many important real-world applications in these areas are driven by neural networks. There are also growing applications in finance, engineering, robotics, and medicine. Despite their immense success in practice, there is limited mathematical understanding of neural networks. We characterize the performance of neural networks when trained with stochastic gradient descent as the number of hidden units N and gradient descent steps grow to infinity. In particular, we investigate the effect of different scaling schemes, which lead to different normalizations of the neural network, on the network's statistical output, closing the gap between the $1/\sqrt{N}$ and the mean-field $1/N$ normalization. We develop an asymptotic expansion for the neural network's statistical output pointwise with respect to the scaling parameter as the number of hidden units grows to infinity. Based on this expansion we demonstrate mathematically that to leading order in N there is no bias-variance trade off, in that both bias and variance (both explicitly characterized) decrease as the number of hidden units increases and time grows. In addition, we show that to leading order in N, the variance of the neural network's statistical output decays as the implied normalization by the scaling parameter approaches the mean field normalization. Numerical studies on the MNIST and CIFAR10 datasets show that test and train accuracy monotonically improve as the neural network's normalization gets closer to the mean field normalization. Time permitting, I will also discuss applications of these ideas to optimization and global convergence of pde-constrained models with neural network terms.

Label-free Domain Adaptation and Multimodal Manifold Learning with Riemannian Geometry

Ronen Talmon
Viterbi Faculty of Electrical and Computer Engineering, Technion - IIT

Recently, Riemannian geometry has become a central ingredient in a broad range of data analysis and learning tools. Broadly, it facilitates features of complex high-dimensional data with known non-Euclidean geometry. In this talk, we will first consider the Riemannian geometry of symmetric positive-definite matrices and propose a method based on Procrustes analysis for label-free domain adaptation. We will present some theoretical guarantees and demonstrate the performance using simulations as well as real-measured data. While the majority of the talk will be focused on a particular geometry of symmetric positive-definite matrices, we will present generalizations to other geometries (e.g., hyperbolic spaces) and features (e.g., kernels and graph-Laplacians). Finally, we will introduce some intriguing spectral properties and demonstrate their application to multimodal manifold learning. Joint work with Or Yair, Ori Katz, Almog Lahav, Ya-Wei Lin, Roy Lederman, and Miri Ben-Chen.

On the inference of hidden Markov models with weakly informative observations

Matti Vihola
University of Jyväskylä

Particle Markov chain Monte Carlo methods allow for practical Bayesian inference with hidden Markov models (a.k.a. general nonlinear/non-Gaussian state space models), relying on efficient (conditional) particle filters. We discuss certain challenges of standard methods that arise with "weakly informative" observations, such as when inferring a time-discretised continuous-time path integral model. With suitable design choices and algorithmic modifications, efficient inference is possible also in this context.

Poster Presentations

A new compressed cover tree guarantees a near-linear complexity for k-nearest neighbors

Yury Elkin
University of Liverpool

This poster studies the classical problem of finding k nearest neighbors to m query points in a larger set of n reference points in any metric space. The well-known work by Beygelzimer, Kakade, and Langford in ICML 2006 introduced cover trees and claimed to guarantee a near-linear time complexity in the number n of reference points. Section 5.3 of Curtin's PhD (2015) pointed out that the proof of this result was wrong. The key step of the original proof attempted to show that the number of iterations can be estimated by multiplying the length of a longest root-to-leaf path in a cover tree by a constant factor. However, this estimate can miss many potential nodes in several branches of a cover tree, that should be considered during the search. The same erroneous argument was unfortunately repeated in several subsequent papers. This poster gives formal counterexamples to the time complexity estimates of the cover tree construction algorithm and the nearest neighbors search in the past work. We prove correct analogs of the past claims with slightly weaker bounds for any number $k > 0$ of neighbors. A new compressed cover tree guarantees a parameterized time complexity that is near-linear in the maximum size of both query and reference set.

Stable homological invariants from Wasserstein metrics

Andrea Guidolin
KTH Royal Institute of Technology

A main objective of topological data analysis is to provide invariants robust with respect to noise. In the context of persistent homology, metrics between persistence modules have been introduced with the purpose of formulating stability results. However, via a process called hierarchical stabilisation, metrics play a more active role in the definition of stable invariants. In this framework, a rich family of metrics between persistence modules is desirable to gain a high degree of flexibility in the analysis, which results in the possibility of searching for the most suited stable invariant for the task at hand. Metrics between persistence modules can be conveniently defined using the notion of noise systems, which indeed model a large family of metrics including the bottleneck distance. Wasserstein metrics are drawing increasing interest in persistence theory, as they tend to provide more refined information than the bottleneck distance, although still enjoying stability properties. Recent developments propose algebraic versions of Wasserstein metrics. Building on these contributions, we formulate p-Wasserstein metrics in the context of noise systems and use them to define invariants called stable ranks through hierarchical stabilisation. We provide an explicit and algorithmic way of computing stable ranks. Overall, our method is well-suited for statistical analysis and machine learning applications. In a supervised learning context for example it is possible to optimise the parameter p, ranging from the most discriminative metric for $p = 1$ to the bottleneck distance. We illustrate the behaviour and properties of p-Wasserstein based stable ranks on a real-world dataset.

Learning in block MDPs via efficient clustering

Yassir Jedra
KTH Royal Institute of Technology

This paper studies Reinforcement Learning tasks in episodic MDPs with rich observations(contexts) drawn from a limited number of hidden latent states, also referred to as clusters. As no prior information is available, we consider the case when uniformly random policy is used. We derive the information-theoretical lower bound on the error rate for the estimation of the latent state decoding function, and then present an algorithm (consisting of two parts), estimating this function and whose clustering error rate is close to that predicted by the lower bound. Finally, we combine this clustering algorithm to obtain sample complexity for reward-free exploration, which is much more efficient than doing the exploration without taking the block structures into account, especially as the sampling budget increases. More precisely, we identify three regimes depend- ing on the number of episodes T, the length of each episode H, and the number of contexts n: when $TH = O(n)$, we cannot gain from exploiting the latent structure (we do not have enough observations to estimate the cluster decoding function accurately); when $TH = \omega(n)$ and $TH = O(n \log n)$, exploiting the structure yields significant gains but the learning rate is mainly limited by the error made in the estimated cluster decoding function; when $TH = \omega(n \log(n))$, the clustering errors become negligible and leveraging the structure, we achieve a learning rate as high as if latent states were actually observed. In all results, we keep track of the aymptotics of the number of actions $|A|$ and the number of latent states $|S|$, leading to much more precise sample complexity guarantees.

A Network Model for Lattice Valued Data

Miguel Lopez
University of Pennsylvania

We construct a diffusion model over a network for lattice valued data. Using cellular sheaves valued in the category of lattices and Galois connections, we can define a suitable notion of a Laplacian, the Tarksi Laplacian, to model diffusive processes. Here we demonstrate an explicit model inspired by formal concept analysis, fuzzy sets and fuzzy Galois connections.

A kernel embedding of equivalence classes of causal models

Alex Markham
KTH Royal Institute of Technology

We introduce a distance covariance-based kernel designed to measure the similarity between the underlying nonlinear causal structures of different samples. We prove that the corresponding feature map is a statistically consistent estimator of nonlinear independence structure, rendering the kernel itself a statistical test for the hypothesis that sets of samples come from different generating causal structures. Even stronger, we prove that the kernel space is isometric to the space of causal ancestral graphs, so that distance between samples in the kernel space is guaranteed to correspond to distance between their generating causal structures. This kernel allows causal interpretability in a wide variety of machine learning methods, for example with applications in identifying causally heterogenous subpopulations, identifying different interventional settings, and causal dimensionality reduction/data visualization. Furthermore, it constitutes a novel connection between causality and basic machine learning methods, as well as suggesting interesting future work in graph kernel embeddings and algebraic statistics.

Data Analysis with Tree-Shaped Topological Summaries

Matteo Pegoraro
Politecnico di Milano

Merge trees are tree-shaped representations of families of homology groups in dimension 0 obtained from filtrations of simplicial complexes. Such objects arise naturally in different scientific fields and fit very well into the framework of Topological Data Analysis. Compared to persistence diagrams (PDs), merge trees provide a finer representation of filtrations of topological spaces, making them an ideal alternative to PDs in several situations. Moreover, we show that this information can be enriched in many fruitful ways, obtaining new tools to analyze data and, lastly, we introduce a general way to endow such objects with a metric structure. We test this framework in several applications and case studies to showcase its effectiveness.

The volume of a tubular neighbourhood of the real Veronese variety

Andrea Rosana
SISSA

In this joint work with Alberto Cazzaniga and Antonio Lerario we addressed the following question: which is the probability for a symmetric tensor to be "close enough" to a decomposable one? It is well known that symmetric decomposable tensors can be parametrized by a Veronese variety. Therefore our problem can be reformulated as follows: which is the volume of a tubular neighbourhood of a Veronese variety? In our work we focused on the case of real norm-1 d-tensors, which correspond to the real spherical Veronese variety of degree d. We discuss the main properties of this variety and its isometries. We then introduce the notion of reach of a smooth submanifold of a Riemannian manifold and discuss the remarkable Weyl's tube formula, expressing the volume of tubular neighbourhoods of smooth submanifolds in the Euclidean or spherical space for radii smaller than the reach of the submanifold. In order to apply this powerful formula, we first compute the reach of the real spherical Veronese. The core of our work is then to show a correspondence between the second fundamental form of our variety and the Gaussian Orthogonal Ensemble (GOE): this indeed allows us to compute some integrals in the normal bundle of our variety by computing some expectations on the GOE space. Exploiting this result, we are finally able to give an explicit exact formula answering our guiding question.

Computing Complexity-aware Plans Using Kolmogorov Complexity

Elis Stefansson
KTH Royal Institute of Technology

In this paper, we introduce complexity-aware planning for finite-horizon deterministic finite automata with rewards as outputs, based on Kolmogorov complexity. Kolmogorov complexity is considered since it can detect computational regularities of deterministic optimal policies. We present a planning objective yielding an explicit trade-off between a policy's performance and complexity. It is proven that maximising this objective is non-trivial in the sense that dynamic programming is infeasible. We present two algorithms obtaining low-complexity policies, where the first algorithm obtains a low-complexity optimal policy, and the second algorithm finds a policy maximising performance while maintaining local (stage-wise) complexity constraints. We evaluate the algorithms on a simple navigation task for a mobile robot, where our algorithms yield low-complexity policies that concur with intuition.

Causal based domain generalization in deep learning

Xudong Sun
KTH Royal Institute of Technology

We address the task of domain generalization, where the goal is to train a predictive model such that it is able to generalize to a new, previously unseen domain. We choose a hierarchical generative approach within the framework of variational autoencoders and propose a domain-unsupervised algorithm that is able to generalize to new domains without domain supervision. We show that our method is able to learn representations that disentangle domain-specific information from class-label specific information even in complex settings where domain structure is not observed during training. Our interpretable method outperforms previously proposed generative algorithms for domain generalization as well as other non-generative state-of-the-art approaches in several hierarchical domain settings including sequential overlapped near continuous domain shift. It also achieves competitive performance on the standard domain generalization benchmark dataset PACS compared to state-of-the-art approaches which rely on observing domain-specific information during training, as well as another domain unsupervised method. Additionally, we proposed model selection purely based on Evidence Lower Bound (ELBO) and also proposed weak domain supervision where implicit domain information can be added into the algorithm.

Realisations of posets and tameness

Francesca Tombari
KTH Royal Institute of Technology

We introduce a construction called realisation which transforms posets into posets. Intuitively, realisations fill in gaps in a locally discrete posets. Realisations share several key features with upper semilattices which are essential in persistence. For example, we define local dimensions of points in a poset and show that these numbers for realisations behave in a similar way as they do for upper semilattices. Furthermore, similarly to upper semilattices, realisations have well behaved discrete approximations which are suitable for capturing homological properties of functors indexed by them. These discretisations are convenient and effective to describe tame functors. Homological properties of tame functors, particularly those indexed by realisations, are discussed, in particular, we show how to compute minimal free resolutions.

Improved analysis of randomized SVD for top-eigenvector approximation

Ruo-Chun Tzeng
KTH Royal Institute of Technology

Computing the top eigenvectors of a matrix is a problem of fundamental interest to various fields. While the majority of the literature has focused on analyzing the reconstruction error of low-rank matrices associated with the retrieved eigenvectors, in many applications one is interested in finding one vector with high Rayleigh quotient. In this paper we study the problem of approximating the top-eigenvector. Given a symmetric matrix A with largest eigenvalue λ_1, our goal is to find a vector $\hat{u}$ that approximates the leading eigenvector u_1 with high accuracy, as measured by the ratio $R(\hat{u}) = \lambda_1^{-1}\hat{u}TA\hat{u}/\hat{u}T\hat{u}$. We present a novel analysis of the randomized SVD algorithm of Halko et al. (2011b) and derive tight bounds in many cases of interest. Notably, this is the first work that provides non-trivial bounds for approximating the ratio $R(\hat{u})$ using randomized SVD with any number of iterations. Our theoretical analysis is complemented with a thorough experimental study that confirms the efficiency and accuracy of the method.

Fast Pure Exploration via Frank-Wolfe

Po-An Wang
KTH Royal Institute of Technology

We study the problem of active pure exploration with fixed confidence in generic stochastic bandit environments. The goal of the learner is to answer a query about the environment with a given level of certainty while minimizing her sampling budget. For this problem, instance-specific lower bounds on the expected sample complexity reveal the optimal proportions of arm draws an Oracle algorithm would apply. These proportions solve an optimization problem whose tractability strongly depends on the structural properties of the environment, but may be instrumental in the design of efficient learning algorithms. We devise Frank-Wolfe-based Sampling (FWS), a simple algorithm whose sample complexity matches the lower bounds for a wide class of pure exploration problems. The algorithm is computationally efficient as, to learn and track the optimal proportion of arm draws, it relies on a single iteration of Frank-Wolfe algorithm applied to the lower-bound optimization problem. We apply FWS to various pure exploration tasks, including best arm identification in unstructured, thresholded, linear, and Lipschitz bandits. Despite its simplicity, FWS is competitive compared to state-of-art algorithms.

Path Signatures of Graph Laplacians

Ka Man (Ambrose) Yim
University of Oxford

The graph Laplacian is a fundamental tool in network science and graph signal processing. In particular, its eigenvalues and eigenvectors are well known to encode fundamental structural properties of the underlying graph. However, there are intrinsic difficulties in leveraging the eigendecomposition of the Laplacian to compare different graphs: namely, the eigenvector dimensions and number of eigenvectors can differ from one graph to another. We address this problem by regarding the eigendecomposition as a 'time series' of eigenvectors ordered by their eigenvalues. We then extract features from this time series using path signatures, a powerful method for analysing sequential data. Using these features to represent the underlying graphs, we are able to achieve comparable performance with state of the art graph neural networks on benchmark graph classification tasks.

Additional abstracts - will be included regularly in the booklet soon.

Recent developments in the use of stable ranks for data analysis

Martina Scolamiero
KTH Royal Institute of Technology

In the hierarchical stabilisation framework, topological features maps are defined based on metrics to compare algebraic representations of data. In this talk I will highlight the variety of metrics that can be constructed in an axiomatic way, via so called Noise Systems. In particular I will explain how also Wasserstein metrics can be interpreted as noise systems based metrics. The focus will then be on one invariant obtained through hierarchical stabilisation, the Stable Rank, which the TDA group at KTH has been studying for several years. Stable ranks are suitable for statistics and Machine Learning and can be used for classification tasks through the stable rank kernel. The use of stable ranks on real data will be illustrated through a project on microglia morphology description, in which the TDA group at KTH has been collaborating with S. Siegert's group, K. Hess and L. Kanari.

Invariants for multiparameter persistence

René Corbet
KTH Royal Institute of Technology

In topological data analysis, multiparameter persistent homology is the generalization of ordinary persistent homology to multiple parameters. These parameters may for instance capture scale, curvature, density, or properties arising directly from applications. While there are many use cases for a refined data analysis by taking several parameters into account at once, the theory of multiparameter persistence is provably much more complicated and computational challenges are substantially bigger than in ordinary persistence. A main difficulty of the computational pipeline of multiparameter persistence is the incompleteness of invariants, such as the Hilbert function, the rank invariant, and minimal presentations. We showcase a new invariant that in particular stabilizes the birth values of topological features in the multiparameter setting by reasonably clustering the minimal generators of its homology. It is the multiparameter generalization of stable rank, as presented in the talk of Martina Scolamiero in the same session. Like many constructions in multiparameter persistence, the computation of our invariant is NP-hard in general, but we present fast computational workarounds for important subclasses that can be used for practical data analysis. This is joint work with Wojciech Chachólski and Anna-Laura Sattelberger.

Learned operator correction in inverse problems

Andreas Hauptmann
University of Oulu

Iterative model-based reconstruction approaches for high-dimensional problems with non-trivial forward operators can be highly time consuming. Thus, it is desirable to employ model reduction techniques to speed-up reconstructions in variational approaches as well as to enable training of learned model-based techniques. Nevertheless, reduced or approximate models can lead to a degradation of reconstruction quality and need to be accounted for. For this purpose, we discuss in this talk the possibility of learning a nonlinear data-driven explicit model correction for inverse problems and whether such a model correction can be used within a variational framework to obtain regularized reconstructions.

Additional abstracts for the poster session - will be included regularly in the booklet soon.

Convergence of Stein Variational Gradient Descent under a Weaker Smoothness Condition

Avetik Karagulyan
KAUST

Stein Variational Gradient Descent (SVGD) is an important alternative to the Langevin-type algorithms for sampling from probability distributions of the form $\pi(x) \propto \exp(-V(x))$. In the existing theory of Langevin-type algorithms and SVGD, the potential function V is often assumed to be L-smooth. However, this restrictive condition excludes a large class of potential functions such as polynomials of degree greater than 2. Our paper studies the convergence of the SVGD algorithm for distributions with (L_0, L_1)-smooth potentials. This relaxed smoothness assumption was introduced by Zhang et al. for the analysis of gradient clipping algorithms. With the help of trajectory-independent auxiliary conditions, we provide a descent lemma establishing that the algorithm decreases the KL divergence at each iteration and prove a complexity bound for SVGD in the population limit in terms of the Stein Fisher information.

Geometry of the n-torus entropic trust region packing algorithm

Miloslav Torda
University of Liverpool

Stochastic relaxation is a well-known approach to solve problems in machine learning and artificial intelligence in cases of complicated optimization landscapes. Inspired by the information geometric optimization framework, we construct a non-Euclidean trust region as a variant of the natural gradient learning with adaptive selection quantile fitness rewriting, the entropic trust region, to solve the problem of densest packings of closed subsets of the n-dimensional Euclidean space restricted to a Crystallographic Symmetry Group (CSG). Since CSGs induce a toroidal topology on the configuration space, the entropic trust region search is performed on a statistical manifold of extended multivariate von Mises probability distributions, a parametric family of probability measures defined on an n-dimensional torus. Using the connection with the generalized proximal method we examine the geometry of the n-torus entropic trust region packing algorithm and provide a characterization of the algorithm via local dual geodesic flows which in fact maximize stochastic dependence among elements of the extended multivariate von Mises distributed random vector, thus providing a relationship between evolutionary computing, simulated annealing method and recurrent neural computing as instances of more general graphical interaction models.

Efficient learning of hidden state space models of unknown order

Othmane Mazhar
KTH Royal Institute of Technology

The aim of this study is to address two related estimation problems arising in the setup of hidden state space systems when the dimension of the hidden state is unknown. Namely, the estimation of any finite number of the system's Markov parameters and the estimation of a minimal realization for the system, both from the partial observation of a single trajectory. For both problems, we provide statistical guarantees in the form of various estimation error upper bounds, rank recovery conditions, and sample complexity estimates.