Abstracts

Andrea Agazzi
Università di Pisa, Italy

Wide neural networks for learning dynamical systems: a mean-field theory approach

In this talk, I will build on groundbreaking results on wide, feedforward neural networks in the supervised learning setting to discuss the performance analogous models when learning dynamical systems. More specifically, I will discuss how, under an appropriate scaling of parameters at initialization, the training dynamics of these models converge towards a hydrodynamic, so-called “mean-field”, limit. This will be done first for feedforward neural networks in the reinforcement learning framework and then, coming back to the "original" supervised learning setting, for recurrent neural network architectures trained with gradient descent.

Back to Table of Contents

Pierre Alquier
ESSEC Business School, Singapore

Robust estimation and regression with MMD

Maximum likelihood estimation (MLE) enjoys strong optimality properties for statistical estimation, under strong assumptions. However, when these assumptions are not satisfied, MLE can be extremely unreliable. In this talk, we will explore alternative estimators based on the minimization of well chosen distances. In particular, we will see that the Maximum Mean Discrepancy (MMD, based on suitable kernels) leads to estimation procedures that are consistent without any assumption on the model nor on the data-generating process. This leads to strong robustness properties in practice, and this method was already used in complex models with promising results: estimation of SDE coefficients, ccopulas, data compression, generative models in AI...

In the second part of this talk, I will discuss the extension of this method to the estimation of conditional distributions, which allows to use MMD-estimators in various regression models. On the contrary to mean embeddings, very technical conditions are required for the existence of a conditional mean embedding that allows defining an estimator. In most papers, these conditions are often assumed, but rarely checked. It turns out that, in most generalized linear regression models, we proved that these conditions can be met, at the cost of more restrictions on the kernel choice.

This is based on joint works with: Badr-Eddine Chérief-Abdellatif (CNRS, Paris), Mathieu Gerber (University of Bristol), Daniele Durante (Bocconi University), Sirio Legramanti (University of Bergamo), Jean-David Fermanian (ENSAE Paris), Alexis Derumigny (TU Delft), Geoffrey Wolfer (RIKEN-AIP, Tokyo).

Back to Table of Contents

Jason Altschuler
University of Pennsylvania, USA

Acceleration by Stepsize Hedging

Can we accelerate the convergence of gradient descent without changing the algorithm — just by optimizing stepsizes? Surprisingly, we show that the answer is yes. Our proposed Silver Stepsize Schedule optimizes strongly convex functions in $k^{\log_p 2} = k^{0.7864}$ iterations, where $p=1+\sqrt{2}$ is the silver ratio and k is the condition number. This is intermediate between the textbook unaccelerated rate k and the accelerated rate $\sqrt{k}$ due to Nesterov in 1983. The non-strongly convex setting is conceptually identical and leads to an analogously accelerated rate $\eps^{-\log_p 2} = \eps^{-0.7864}$. We conjecture and provide partial evidence that these rates are optimal among all possible stepsize schedules.

The Silver Stepsize Schedule is an explicit non-monotonic fractal. Why should such stepsizes help? The core intuition is "hedging" between individually suboptimal strategies — short steps and long steps — since bad cases for the former are good cases for the latter, and vice versa. Properly combining these stepsizes yields faster convergence due to the misalignment of worst-case functions. This talk is based on a line of work with Pablo Parrilo that originates from my 2018 Master's Thesis — which established for the first time that judiciously chosen stepsizes can enable accelerated convex optimization. Prior to this thesis, the only such result was for the special case of quadratics, due to Young in 1953.

Back to Table of Contents

Gerard Ben Arous
New York University, USA

Distinguished Visitor Lecture Series

Dynamical spectral transition for optimization in very high dimensions

---

In recent work with Reza Gheissari (Northwestern), Aukosh Jagannath (Waterloo) we gave a general context for the existence of projected "effective dynamics" of SGD in very high dimensions, for "summary statistics" in much smaller dimensions. These effective dynamics (and, in particular, their so-called 'critical regime") define a dynamical system in finite dimensions which may be quite complex, and rules the performance of the learning algorithm.

The next step is to understand how the system finds these "summary statistics". This is done in the last work with the same authors and with Jiaoyang Huang (Wharton, U-Penn). This is based on a dynamical spectral transition of Random Matrix Theory: along the trajectory of the optimization path, the Gram matrix or the Hessian matrix develop outliers which carry these effective dynamics.

I will naturally first come back to the Random Matrix Tools needed here (the behaviour of the edge of the spectrum and the BBP transition). And then illustrate the use of this point of view on a few central examples of ML: multilayer neural nets for classification (of Gaussian mixtures), and the XOR task.

Ng Kong Beng Public Lecture Series

Beating the odds: Learning or hallucinating?
What is the science of data doing today?

The science of data, from its beginnings in classical probability in the 17th century, to statistics and to machine learning today, has been a constant driver of human progress and a tool of enquiry for the human mind. It has constantly aimed at helping us get information, observing and estimating the world around us; that is, learning. The recent explosion of uses of Artificial Intelligence is obviously a major step to quench the thirst and the need for learning.

But what about the recent fear of "hallucinations" in so-called generative AI? Is that a natural part of this long path of discovery -- is this to be expected, or feared as a new and fatal twist?

This talk will be geared to a general audience with an interest in science, and will try to introduce this debate lightly, from the point of view of a mathematician.

Back to Table of Contents

Arnab Bhattacharyya
National University of Singapore, Singapore

Learning bounded-degree polytrees with samples

We establish finite-sample guarantees for efficient proper learning of bounded-degree polytrees, a rich class of high-dimensional probability distributions and a subclass of Bayesian networks, a widely-studied type of graphical models. Very recently, Bhattacharyya-Gayen-Price-Vinodchandran (STOC '21) obtained finite-sample guarantees for recovering tree-structured Bayesian networks, i.e., 1-polytrees. We considerably extend their results by providing an efficient algorithm which learns d-polytrees in polynomial time and sample complexity when the in-degree d is constant, provided that the underlying undirected graph (skeleton) is known. We complement our algorithm with an information-theoretic lower bound, showing that the dependence of our sample complexity is nearly tight in both the dimension and target accuracy parameters.

Joint work with Clément Canonne, Davin Choo, and Joy Yang

Back to Table of Contents

Xavier Bresson
National University of Singapore, Singapore

Graph Transformers and Developments

Graph Neural Networks (GNNs) have shown great potential in the field of graph representation learning. Standard GNNs define a local message-passing mechanism which propagates information over the whole graph domain by stacking multiple layers. This paradigm suffers from two major limitations, over-squashing and poor long-range dependencies, that can be solved using global attention but significantly increases the computational cost to quadratic complexity. In this work, we propose an alternative approach to overcome these structural limitations by leveraging the ViT/MLP-Mixer architectures introduced in computer vision. We introduce a new class of GNNs, called Graph MLP-Mixer/ViT, that holds three key properties. First, they capture long-range dependency as demonstrated on the long-range LRGB datasets and mitigate the over-squashing issue on the TreeNeighbour dataset. Second, they offer memory and speed efficiency, surpassing related techniques. Third, they show high expressivity in terms of graph isomorphism as they can distinguish at least 3-WL isomorphic graphs. As a result, this novel architecture provides significantly better results over standard message-passing GNNs for molecular datasets.

Back to Table of Contents

Caroline Chaux
CNRS@CREATE, Singapore

Formulation and resolution of inverse problems in signal and image processing - From classical methods to hybrid AI

In this talk, we will be interested in inverse problems arising in the signal and image processing field.

Solving such problems imply in a fist time to formalise the direct problem by understanding the physics behind and in a second time, to solve the associated inverse problem, through a variational formulation, that is, solving an optimization problem. Such issues are encountered in many areas such as biology, medical imaging, chemistry, audio signal processing, ... for which, different tasks have to be tackled such as deconvolution, restoration, unmixing, missing data reconstruction, ...

Classical optimization-based approaches consist in, once the optimization problem has been formulated, proposing iterative procedures (e.g. proximal algorithms) converging to a solution of the considered inverse problem. More recently, unrolled or unfolded neural networks have been proposed. They combine optimization and learning, constitute interpretable networks and integrate information about the direct model. We will study and describe such networks for the resolution of two inverse problems: image deconvolution and robust PCA.

Collaborations: this work has been done in collaboration with Vincent Tan, Emmanuel Soubiès, Pascal Nguyen and Elisabeth Tan.

Back to Table of Contents

Sinho Chewi
Massachusetts Institute of Technology, USA

Mini Course
Optimal transport and high-dimensional probability

Optimal transport, which began with the work of Gaspard Monge in the eighteenth century, has developed into a rich mathematical theory with applications to geometry, PDEs, physics, high-dimensional probability, and statistics and machine learning. In this minicourse, we will introduce the theory and its applications to topics such as concentration inequalities, gradient flows, and sampling.

Back to Table of Contents

Alexandre d'Aspremont
École Normale Supérieure, France

Approximation Bounds for Sparse Programs

We show that sparsity-constrained optimization problems over low dimensional spaces tend to have a small duality gap. We use the Shapley-Folkman theorem to derive both data-driven bounds on the duality gap, and an efficient primalization procedure to recover feasible points satisfying these bounds. These error bounds are proportional to the rate of growth of the objective with the target cardinality k, which means in particular that the relaxation is nearly tight as soon as k is large enough so that only uninformative features are added.

This is joint work with Armin Askari and Laurent El Ghaoui.

Back to Table of Contents

Zhou Fan
Yale University, USA

Gradient flows for empirical Bayes in high-dimensional linear models

Empirical Bayes provides a powerful approach to learning and adapting to latent structure in data. Theory and algorithms for empirical Bayes have a rich literature for sequence models, but are less understood in settings where latent variables and data interact through more complex designs.

In this work, we study empirical Bayes estimation of an i.i.d.\ prior in Bayesian linear models, via the nonparametric maximum likelihood estimator (NPMLE). We introduce and study a system of gradient flow equations for optimizing the marginal log-likelihood, jointly over the prior and posterior measures in its Gibbs variational representation using a smoothed reparametrization of the regression coefficients. A diffusion-based implementation yields a Langevin dynamics MCEM algorithm, where the prior law evolves continuously over time to optimize a sequence-model log-likelihood defined by the coordinates of the current Langevin iterate.

We show consistency of the NPMLE as $n,p \to \infty$ under mild conditions, including settings of random sub-Gaussian designs when $n \asymp p$. In high noise, we prove a uniform log-Sobolev inequality for the mixing of Langevin dynamics, for possibly misspecified priors and non-log-concave posteriors. We then establish polynomial-time convergence of the joint gradient flow to a near-NPMLE if the marginal negative log-likelihood is convex in a sub-level set of the initialization.

Joint work with Leying Guan, Yandi Shen, and Yihong Wu.

Back to Table of Contents

Borjan Geshkovski
Massachusetts Institute of Technology, USA

A mathematical perspective on Transformers

This talk will report on several results, insights and perspectives Cyril Letrouit, Yury Polyanskiy, Philippe Rigollet and I have found regarding Transformers. We model Transformers as interacting particle systems (each particle representing a token), with a non-linear coupling called self-attention. When considering pure self-attention Transformers, we show that trained representations cluster in long time to different geometric configurations determined by spectral properties of the model weights. We also cover Transformers with layer-normalisation, which amounts to considering the interacting particle system on the sphere. On high-dimensional spheres, we prove that all randomly initialized particles converge to a single cluster. The result is made more precise by describing the precise phase transition between the clustering and non-clustering regimes. The appearance of metastability, and ideas for the low-dimensional regime, will be discussed.

Back to Table of Contents

Jeremy Heng
ESSEC Business School, Singapore

Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modelling

Progressively applying Gaussian noise transforms complex data distributions to approximately Gaussian. Reversing this dynamic defines a generative model. When the forward noising process is given by a Stochastic Differential Equation (SDE), Song et al. (2021) demonstrate how the time inhomogeneous drift of the associated reverse-time SDE may be estimated using score-matching. A limitation of this approach is that the forward-time SDE must be run for a sufficiently long time for the final distribution to be approximately Gaussian. In contrast, solving the Schrödinger Bridge problem (SB), i.e. an entropy-regularized optimal transport problem on path spaces, yields diffusions which generate samples from the data distribution in finite time. We present Diffusion SB (DSB), an original approximation of the Iterative Proportional Fitting (IPF) procedure to solve the SB problem, and provide theoretical analysis along with generative modeling experiments. The first DSB iteration recovers the methodology proposed by Song et al. (2021), with the flexibility of using shorter time intervals, as subsequent DSB iterations reduce the discrepancy between the final-time marginal of the forward (resp. backward) SDE with respect to the prior (resp. data) distribution. Beyond generative modeling, DSB offers a widely applicable computational optimal transport tool as the continuous state-space analogue of the popular Sinkhorn algorithm (Cuturi, 2013).

Back to Table of Contents

Masaaki Imaizumi
The University of Tokyo, Japan

Statistical Analysis on Generalization Ability of In-Context Learning

In this study, we analyze sample complexity in in-context learning, a type of meta-learning. In-context learning is a framework that consists of an identical learner capable of handling multiple tasks and has attracted strong attention in recent artificial intelligence technologies. As an approach to understanding this learning framework, several studies have raised a hypothesis that the learner learns an algorithm itself. In this study, we study this hypothesis of algorithmic learning through statistical sample complexity analysis. Specifically, we evaluate the generalization ability of in-context learning using task selection and prompt length, as well as the complexity of the mapping on an empirical distribution. Through these quantitative assessments, we try to gain a better understanding of in-context learning.

Back to Table of Contents

Aukosh Jagannath
University of Waterloo, Canada

Spectral alignment for high-dimensional SGD

Over the last decade, a body of rich predictions has been made about the spectra of empirical Hessian and information matrices over the course of training (via SGD) in overparametrized networks. I'll present a recent work, in collaboration with G. Ben Arous (NYU Courant), R.Ghessari (Northwestern U.), and J. Huang (U. Penn), where we rigorously establish some of these predictions. We prove that in two canonical classification tasks for multi-class high-dimensional mixtures and either 1 or 2-layer neural networks, the SGD trajectory rapidly aligns with emerging low-rank outlier eigenspaces of the Hessian and gradient matrices. Moreover, in multi-layer settings this alignment occurs per layer, with the final layer's outlier eigenspace evolving over the course of training and exhibiting rank deficiency when the SGD converges to sub-optimal classifiers.

Back to Table of Contents

Kengo Kato
Cornell University, USA

Semidiscrete optimal transport maps: stability, limit theorems, and asymptotic efficiency

---

We study statistical inference for the optimal transport (OT) map (also known as the Brenier map) from a known absolutely continuous reference distribution onto an unknown finitely discrete target distribution. We derive limit distributions for the integral and linear functionals of the empirical OT map, together with their moment convergence. The former has a non-Gaussian limit, whose explicit density is derived, while the latter attains asymptotic normality. For both cases, we also establish consistency of the nonparametric bootstrap. The derivation of our limit theorems relies on new stability estimates of functionals of the OT map with respect to the dual potential vector, which may be of independent interest. We also discuss applications of our limit theorems to the construction of confidence sets for the OT map and inference for a maximum tail correlation. Finally, we discuss asymptotic efficiency of the empirical OT map in an infinite dimensional setting.

Back to Table of Contents

Anya Katsevich
Massachusetts Institute of Technology, USA

(Skew) Gaussian surrogates for high-dimensional posteriors: from tighter bounds to tighter approximations

Computing integrals against a high-dimensional posterior is the major computational bottleneck in Bayesian inference. A popular technique to reduce this computational burden is to use the Laplace approximation, a Gaussian distribution, in place of the true posterior. Despite its widespread use, the Laplace approximation's accuracy in high dimensions is not well understood. The body of existing results does not form a cohesive theory, leaving open important questions e.g. on the dimension dependence of the approximation rate. We address many of these questions through the unified framework of a new, leading order asymptotic decomposition of high-dimensional Laplace integrals. In particular, we (1) determine the tight dimension dependence of the approximation error, leading to the tightest known Bernstein von Mises result on the asymptotic normality of the posterior, and (2) derive a simple correction to this Gaussian distribution to obtain a higher-order accurate approximation to the posterior.

Back to Table of Contents

Yuehaw Khoo
University of Chicago, USA

Randomized tensor-network algorithms for random data in high-dimensions

Tensor-network ansatz has long been employed to solve the high-dimensional Schrödinger equation, demonstrating linear complexity scaling with respect to dimensionality. Recently, this ansatz has found applications in various machine learning scenarios, including supervised learning and generative modeling, where the data originates from a random process. In this talk, we present a new perspective on randomized linear algebra, showcasing its usage in estimating a density as a tensor-network from i.i.d. samples of a distribution, without the curse of dimensionality, and without the use of optimization techniques. Moreover, we illustrate how this concept can combine the strengths of particle and tensor-network methods for solving high-dimensional PDEs, resulting in enhanced flexibility for both approaches.

Back to Table of Contents

Tam Le
The Institute of Statistical Mathematics, Japan

Local Structures for Large-Scale Optimal Transport

Optimal transport (OT) theory provides a set of powerful tools to compare measures. OT has a wide range of applications, e.g., computer vision, natural language processing and machine learning. However, OT has a high computational complexity (i.e., super-cubic) which hinders its applications in large-scale settings. One of popular approaches is to exploit local structure of supports in measures, e.g., one-dimensional structure in sliced-Wasserstein (SW). The SW projects supports into a random one-dimensional space and relies on the closed-form solution of the univariate OT. However, SW suffers a curse of dimensionality since using one-dimensional projection limits its ability to capture topological structures of measures in high-dimensional settings. In this talk, I will show that we can leverage more general structures such as tree and graph over supports to alleviate the curse of dimensionality in SW and scale up OT and its variant problems, especially for large-scale applications.

Back to Table of Contents

Qianxiao Li
National University of Singapore, Singapore

Approximation Theory of Deep Learning for Sequence Modelling

In this talk, we present some recent results on the approximation theory of deep learning architectures for sequence modelling. In particular, we formulate a basic mathematical framework, under which different popular architectures such as recurrent neural networks, dilated convolutional networks (e.g. WaveNet), encoder-decoder structures, and most recently - transformers - can be rigorously compared. These analyses reveal some interesting connections between approximation, memory, sparsity/low-rank, graphical structures that may guide the practical selection and design of these network architectures.

Back to Table of Contents

Qin Li
University of Wisconsin-Madison, USA

The perfect diffusion model does not generate

The diffusion model has emerged as a highly successful strategy in generative modeling. A beautiful set of theory based on the application of Girsanov theorem shows that a model with well-learned score function can generate samples from a distribution that approximates the ground truth, achieving the task of ``generation.'' It is tempting to draw the conclusion from here that the success of generation is equivalent to that of learning.

However, we aim to sound a cautionary note – perfection in learning, as evidenced by a straightforward proof and a very simple simulation, actually leads to memorization and shying the system away from true generative capabilities.

This is light hearted talk. Please join me in the discussion on the nuances between learning and generation in the context of diffusion model.

Back to Table of Contents

Cheng Mao
Georgia Institute of Technology, USA

Information-Theoretic Thresholds for Planted Dense Cycles

We study a random graph model for small-world networks which are ubiquitous in social and biological sciences. In this model, a dense cycle of expected bandwidth $n \tau$, representing the hidden one-dimensional geometry of vertices, is planted in an ambient random graph on n vertices. For both detection and recovery of the planted dense cycle, we characterize the information-theoretic thresholds in terms of n, τ, and an edge-wise signal-to-noise ratio λ. The information-theoretic thresholds differ from the computational thresholds established in an earlier companion work for low-degree polynomial algorithms, thereby justifying the existence of statistical-to-computational gaps for this problem.

The talk is based on joint work with Alex Wein and Shenduo Zhang.

Back to Table of Contents

Robert McCann
University of Toronto, Canada

A geometric approach to apriori estimates for optimal transport maps

A key inequality which underpins the regularity theory of optimal transport for costs satisfying the Ma-Trudinger-Wang condition is the Pogorelov second derivative bound. This translates to an apriori interior C^1 estimate for smooth optimal maps.

Here we give a new derivation of this estimate which relies in part on Kim, McCann and Warren's observation that the graph of an optimal map becomes a volume maximizing spacelike submanifold when the product of the source and target domains is endowed with a suitable pseudo-Riemannian geometry that combines both the marginal densities and the cost.

Back to Table of Contents

Govind Menon
Brown University, USA

Mini Course
The geometry of the deep linear network

The deep linear network (DLN) is a phenomenological model for deep learning introduced by Arora, Cohen and Hazan. These two lectures will provide an introduction to the surprising geometric structure of this model and its interplay with training dynamics.

Back to Table of Contents

Marco Mondelli
Institute of Science and Technology Austria, Austria

From Spectral Estimators to Approximate Message Passing… And Back

In a generalized linear model (GLM), the goal is to estimate a d-dimensional signal x from an n-dimensional observation of the form f(Ax, w), where A is a design matrix and w is a noise vector. Well-known examples of GLMs include linear regression, phase retrieval, 1-bit compressed sensing, and logistic regression. We focus on the high-dimensional setting in which both the number of measurements n and the signal dimension d diverge, with their ratio tending to a fixed constant. Spectral methods provide a popular solution to obtain an initial estimate, and they are also commonly used as a 'warm start' for other algorithms. In particular, the spectral estimator is the principal eigenvector of a data-dependent matrix, whose spectrum exhibits a phase transition.

In the talk, I will start by (i) discussing the emergence of this phase transition for an i.i.d. Gaussian design A, and (ii) combining spectral methods with Approximate Message Passing (AMP) algorithms, thus solving a key problem related to their initialization. I will then focus on two instances of GLMs that capture the heterogeneous and structured nature of practical data models: (i) a mixed GLM with multiple signals to recover, and (ii) a GLM with a correlated design matrix. To study spectral estimators in these challenging settings, the plan is to go back to Approximate Message Passing: I will demonstrate that the AMP framework not only gives Bayes-optimal algorithms, but it also unveils phase transitions in the spectrum of random matrices, thus leading to a precise asymptotic characterisation of spectral estimators.

Based on a series of joint works with Hong Chang Ji, Andrea Montanari, Ramji Venkataramanan, and Yihan Zhang.

Back to Table of Contents

Jaouad Mourtada
ENSAE/CREST, France

Finite-sample performance of the maximum likelihood estimator in logistic regression

The logistic model is a classical linear model to describe the probabilistic dependence of binary responses to multivariate features. We consider the predictive performance of the maximum likelihood estimator (MLE) for logistic regression, assessed in terms of the logistic loss of its probabilistic forecasts. We consider two questions: first, that of existence of the MLE (which occurs when the data is not linearly separated), and second that of its accuracy when it exists. These properties depend on both the dimension of covariates and on the signal strength.

In the case of Gaussian covariates and a well-specified logistic model, we obtain sharp non-asymptotic guarantees for the existence and excess prediction error of the MLE. This complements asymptotic results of Sur and Candès, and refines non-asymptotic upper bounds of Ostrovskii and Bach and Chinot, Lecué and Lerasle. It also complements independent recent results by Kuchelmeister and van de Geer. We then extend these results in two directions: first, to non-Gaussian covariates satisfying a certain regularity condition, and second to the case of a misspecified logistic model.

Back to Table of Contents

Somabha Mukherjee
National University of Singapore, Singapore

Least Squares Estimation of a Multivariate Quasiconvex Regression Function

Nonparametric least squares estimation of a multivariate function based on the economic axiom of quasiconvexity is fundamentally different from least-squares estimation under the classical shape constraints of monotonicity and convexity, because unlike the latter two shape constraint problems, the least squares constraint space for the former problem is not convex. In this talk, I will show how to construct a quasiconvex function estimate through a mixed integer quadratic optimization technique, and discuss about the consistency and finite sample risk bounds of the proposed estimate. Towards the end, I will also illustrate the performance of this method on two real life datasets.

Back to Table of Contents

Soumendu Sundar Mukherjee
Indian Statistical Institute, Kolkata, India

Learning under latent group sparsity via heat flow dynamics on networks

In this talk, we will consider the problem of variable selection in high-dimensional regression under latent group sparsity. We will present a new penalty that automatically selects variables in groups without being explicitly told what those groups are. This will be done by incorporating into the penalty a suitable Laplacian matrix (containing group information) in the form of a heat flow. At equilibrium, the proposed penalty coincides with the classical group lasso penalty. We will present some numerical and theoretical results on the performance of the proposed penalty. This is based on joint work with Subhroshekhar Ghosh.

Back to Table of Contents

Praneeth Netrapalli
Google Research, India

Steering Deep Feature Learning with Backward Aligned Feature Updates

Deep learning succeeds by doing hierarchical feature learning, yet tuning Hyper-Parameters (HP) such as initialization scales, learning rates etc., only give indirect control over this behavior. In this paper, we propose the alignment between the feature updates and the backward pass as a key notion to predict, measure and control feature learning. On the one hand, we show that when alignment holds, the magnitude of feature updates after one SGD step is related to the magnitude of the forward and backward passes by a simple and general formula. This leads to techniques to automatically adjust HPs (initialization scales and learning rates) at initialization and throughout training to attain a desired feature learning behavior. On the other hand, we show that, at random initialization, this alignment is determined by the spectrum of a certain kernel, and that well-conditioned layer-to-layer Jacobians (aka dynamical isometry) implies alignment. Finally, we investigate ReLU MLPs and ResNets in the large width-then-depth limit. Combining hints from random matrix theory and numerical experiments, we show that (i) in MLP with iid initializations, alignment degenerates with depth, making it impossible to start training, and that (ii) in ResNets, the branch scale 1/\sqrt{depth} is the only one maintaining non-trivial alignment at infinite depth.

Joint work with Lenaic Chizat (EPFL).

Back to Table of Contents

Ariel Neufeld
Nanyang Technological University, Singapore

Deep Learning based algorithm for nonlinear PDEs in finance and gradient descent type algorithm for non-convex stochastic optimization problems with ReLU neural networks

In this talk, we first present a deep-learning based algorithm which can solve nonlinear parabolic PDEs in up to 10'000 dimensions with short run times, and apply it to price high-dimensional financial derivatives under default risk. Then, we discuss a general problem when training neural networks, namely that it typically involves non-convex stochastic optimization. To that end, we present TUSLA, a gradient descent type algorithm (or more precisely : stochastic gradient Langevin dynamics algorithm) for which we can prove that it can solve non-convex stochastic optimization problems involving ReLU neural networks.

This talk is based on joint works with C. Beck, S. Becker, P. Cheridito, A. Jentzen, and D.-Y. Lim, S. Sabanis, Y. Zhang, respectively.

Back to Table of Contents

Tan Minh Nguyen
National University of Singapore, Singapore

Transformers Meet Image Denoising: Mitigating Over-smoothing in Transformers via Regularized Nonlocal Functionals

Transformers have achieved remarkable success in a wide range of natural language processing and computer vision applications. However, the representation capacity of a deep transformer model is degraded due to the over-smoothing issue in which the token representations become identical when the model's depth grows. In this work, we show that self-attention layers in transformers minimize a functional which promotes smoothness, thereby causing token uniformity. We then propose a novel regularizer that penalizes the norm of the difference between the smooth output tokens from self-attention and the input tokens to preserve the fidelity of the tokens. Minimizing the resulting regularized energy functional, we derive the Neural Transformer with a Regularized Nonlocal Functional (NeuTRENO), a novel class of transformer models that can mitigate the over-smoothing issue. We empirically demonstrate the advantages of NeuTRENO over the baseline transformers and state-of-the-art methods in reducing the over-smoothing of token representations on various practical tasks, including object classification, image segmentation, and language modelling.

Back to Table of Contents

Jonathan Niles-Weed
New York University, USA

Optimal transport map estimation in general function spaces

We present a unified methodology for obtaining rates of estimation of optimal transport maps in general function spaces. Our assumptions are significantly weaker than those appearing in the literature: we require only that the source measure P satisfy a Poincare inequality and that the optimal map be the gradient of a smooth convex function that lies in a space whose metric entropy can be controlled. As a special case, we recover known estimation rates for Holder transport maps, but also obtain nearly sharp results in many settings not covered by prior work. For example, we provide the first statistical rates of estimation when P is the normal distribution, between log-smooth and strongly log-concave distributions, and when the transport map is given by an infinite-width shallow neural network.

Joint with Vincent Divol and Aram-Alexandre Pooladian.

Back to Table of Contents

Soumik Pal,
University of Washington, USA

Mirror gradient flows in the Wasserstein space

The Sinkhorn algorithm is a widely popular iterative algorithm to approximately compute an optimal transport coupling between two probability measures. However, much of its behaviour is still shrouded in mystery. We will talk about scaling limit of the iterates as it converges to an absolutely continuous curve on the Wasserstein space. This curve can be described as a Wasserstein counterpart of the Euclidean mirror gradient flow. An equivalent description of this flow is provided by the parabolic Monge-Ampere PDE. We will introduce this novel family of flows and talk about its properties including the hidden Hessian geometry that controls their rates to equilibrium.

Back to Table of Contents

Courtney Paquette
McGill University, Canada

Hitting the High-D(imensional) Notes: SGD learning dynamics

In this talk, I will present a framework, inspired by random matrix theory, for analyzing the dynamics of stochastic optimization algorithms (e.g., stochastic gradient descent (SGD) and momentum (SGD + M)) when both the number of samples and dimensions are large. Using this new framework, we show that the dynamics of optimization algorithms on generalized linear models and multi-index problems with random data become deterministic in the large sample and dimensional limit. In particular, the limiting dynamics for stochastic algorithms are governed by an ODE. From this model, we identify a stability measurement, the implicit conditioning ratio (ICR), which regulates the ability of SGD+M to accelerate the algorithm. When the batch size exceeds this ICR, SGD+M converges linearly at a rate of $O(1/\kappa)$, matching optimal full-batch momentum (in particular performing as well as a full-batch but with a fraction of the size). For batch sizes smaller than the ICR, in contrast, SGD+M has rates that scale like a multiple of the single batch SGD rate. We give explicit choices for the learning rate and momentum parameter in terms of the Hessian spectra that achieve this performance. Finally we show this model matches performances on real data sets.

This is joint with the talk by Elliot Paquette.

Back to Table of Contents

Elliot Paquette
McGill University, Canada

High-dimensional limits of streaming and multi-pass SGD on least squares

Traditional complexity analysis of SGD is formulated in terms of minimax optimality, wherein tight upper and lower bounds for complexity are given over a class of objective functions. Often these bounds are pessimistic, even for strongly convex quadratic objectives, and they often do not consider speedups that can occur in high-dimensional settings. We show here a different analysis, for both streaming and multi-pass (aka random shuffle) SGD, leveraging simplifications that occur in high-dimensional random settings. This leads to sharp complexity estimates for single problems, up to error terms that are small with dimension. We also give a sketch of the mathematical differences between the multi-pass and streaming analyses.

This is based on joint works with Ben Adlam, Elizabeth Collins—Woodfin, Kiwon Lee, Courtney Paquette, Fabian Pedregosa and Jeffrey Pennington.

Back to Table of Contents

Vianney Perchet
ENSAE/CREST, France

On Preemption and Learning in Stochastic Scheduling

We study single-machine scheduling of jobs, each belonging to a job type that determines its duration distribution. We start by analyzing the scenario where the type characteristics are known and then move to two learning scenarios where the types are unknown: non-preemptive problems, where each started job must be completed before moving to another job; and preemptive problems, where job execution can be paused in the favor of moving to a different job. In both cases, we design algorithms that achieve sublinear excess cost, compared to the performance with known types, and prove lower bounds for the non-preemptive case. Notably, we demonstrate, both theoretically and through simulations, how preemptive algorithms can greatly outperform non-preemptive ones when the durations of different job types are far from one another, a phenomenon that does not occur when the type durations are known.

Back to Table of Contents

Andrej Risteski
Carnegie Mellon University, USA

Neural Networks for PDEs: Representational Power and Inductive Biases

Following breakthroughs in using deep learning in such diverse domains as computer vision and natural language processing, a burgeoning line of research leverages deep learning for scientific applications. Partial differential equations (PDEs) are a key primitive in many scientific applications, motivating a rapidly growing area of research in data-driven approaches to solving PDEs. The talk will survey several recent works on understanding PDEs for which neural networks constitute a good choice of a parametric family: in particular, in terms of representational strength, they circumvent "curse of dimensionality" style bounds. We will also show how theoretical insights can be used to elucidate and guide architectural design for neural operators.

Based on the works:
https://arxiv.org/abs/2103.02138
https://arxiv.org/abs/2210.12101
https://arxiv.org/abs/2312.00234

Back to Table of Contents

Mark Rudelson
University of Michigan, USA

Mini Course 1
How to check when a system of real quadratic equations has a solution

The existence and the number of solutions of a system of polynomial equations in n variables over an algebraically closed field is a classical topic in algebraic geometry. Much less is known about the existence of solutions of a system of polynomial equations over reals. Any such problem can be reduced to a system of quadratic equations by introducing auxiliary variables. Due to the generality of the problem, a computationally efficient algorithm for determining whether a real solution of a system of quadratic equations exists is believed to be impossible. We will discuss a simple sufficient condition for the existence of a solution which can be efficiently checked. While the problem and the condition are of algebraic nature, the approach lies entirely within the analysis/probability realm and relies on tools from Fourier analysis and concentration of measure.

Joint work with Alexander Barvinok.

Mini Course 2
Approximately Hadamard matrices and random frames

We will discuss a problem concerning random frames which arises in signal processing. A frame is an overcomplete set of vectors in the n-dimensional linear space which allows a robust decomposition of any vector in this space as a linear combination of these vectors. Random frames are used in signal processing as a means of encoding since the loss of a fraction of coordinates does not prevent the recovery. We will discuss a question when a random frame contains a copy of a nice (almost orthogonal) basis.

Despite the probabilistic nature of this problem it reduces to a completely deterministic question of existence of approximately Hadamard matrices. An n by n matrix with plus-minus 1 entries is called Hadamard if it acts on the space as a scaled isometry. Such matrices exist in some, but not in all dimensions. Nevertheless, we will construct plus-minus 1 matrices of every size which act as approximate scaled isometries. This construction will bring us back to probability as we will have to combine number-theoretic and probabilistic methods.

Joint work with Xiaoyu Dong.

Back to Table of Contents

Jonathan Scarlett,
National University of Singapore, Singapore

Recent Developments in Group Testing: Fundamental Limits and Algorithms

The group testing problem concerns discovering a small number of defective items within a large population by performing tests on pools of items. A test is positive if the pool contains at least one defective, and negative if it contains no defectives. This is a sparse inference problem with a combinatorial flavour, with applications in medical testing, biology, multi-access communication, database systems, and more. I will review recent advances in the mathematics of group testing, including both information-theoretic limits and performance bounds for practical algorithms, with an emphasis on the following defining features:

- Non-adaptive testing (all tests must be designed in advance) vs. adaptive testing (tests are designed sequentially based on previous outcomes)
- Noiseless testing (tests are perfectly reliable) vs. noisy tests (some test outcomes are corrupted)

Most of this talk is loosely based on a survey monograph available at https://arxiv.org/abs/1902.06002

Back to Table of Contents

Bodhisattva Sen
Columbia University, USA

A New Perspective On Denoising Based On Optimal Transport

In the standard formulation of the denoising problem, one is given a probabilistic model relating a latent variable and an observation Z, and the goal is to construct a map to recover the latent variable from Z. The posterior mean, a natural candidate for estimating the latent variable from observation, attains the minimum Bayes risk (under the squared error loss) but at the expense of over-shrinking the Z, and in general may fail to capture the geometric features of the prior distribution (e.g., low dimensionality, discreteness, sparsity, etc.). To rectify these drawbacks, we take a new perspective on this denoising problem that is inspired by optimal transport (OT) theory and use it to propose a new OT-based denoiser at the population level setting. We rigorously prove that, under general assumptions on the model, our OT-based denoiser is well-defined and unique, and is closely connected to solutions to a Monge OT problem.

We then prove that, under appropriate identifiability assumptions on the model, our OT-based denoiser can be recovered solely from information of the marginal distribution of Z and the posterior mean of the model, after solving a linear relaxation problem over a suitable space of couplings that is reminiscent of a standard multi-marginal OT (MOT) problem. In particular, thanks to Tweedie's formula, when the likelihood model is an exponential family of distributions, the OT-based denoiser can be recovered solely from the marginal distribution of Z. In general, our family of OT-like relaxations is of interest in its own right and for the denoising problem suggests alternative numerical methods inspired by the rich literature on computational OT.

Back to Table of Contents

Yair Shenfeld
Brown University, USA

Mini Course
Optimal transport and high-dimensional probability

Optimal transport, which began with the work of Gaspard Monge in the eighteenth century, has developed into a rich mathematical theory with applications to geometry, PDEs, physics, high-dimensional probability, and statistics and machine learning. In this minicourse, we will introduce the theory and its applications to topics such as concentration inequalities, gradient flows, and sampling.

Back to Table of Contents

Yong Sheng Soh
National University of Singapore, Singapore

Optimal Regularization for a Data Source

In optimization-based approaches to inverse problems and to statistical estimation, it is common to augment criteria that enforce data fidelity with a regularizer that promotes desired structural properties in the solution. The choice of a suitable regularizer is typically driven by a combination of prior domain information and computational considerations.

In this talk, we seek a systematic understanding of the power and the limitations of convex regularization by investigating the following questions: Given a distribution, what is the optimal regularizer for data drawn from the distribution? What properties of a data source govern whether the optimal regularizer is convex?

We will show that it suffices to parameterize the family of regularizers one considers with the collection of star bodies. Using ideas from dual Brunn-Minkowski theory as well as gamma-convergence from variational analysis, we will characterize these optimal regularizers and describe its behaviour with respect to the data source.

Back to Table of Contents

Vladimir Spokoiny
Weierstrass Institute for Applied Analysis and Stochastics, Germany

Inference for nonlinear inverse problems

Assume that a solution to a nonlinear inverse problem given e.g. by PDE is observed with noise. The target of analysis is typically a set of model parameters describing the corresponding forward operator and the corresponding denoised solution. The classical least squares approach faces several challenges and obstacles for theoretical study and numerically efficient implementation, especially if the parameter space is large and the observation noise is not negligible.

We propose a new approach that provides rather precise finite sample results about the accuracy of estimation and quantification of uncertainty and allows us to avoid any stability analysis of the inverse operator and advanced results from empirical processes theory. The approach is based on extending the parameter space by introducing a set of «observables» and careful treatment of the arising semiparametric problem.

Back to Table of Contents

Piyush Srivastava
Tata Institute of Fundamental Research, India

Sampling from convex bodies using multiscale decompositions

Sampling from convex bodies is a fundamental and widely-studied algorithmic primitive. We propose a new family of Markov chains based on lazily computed Whitney decompositions of convex bodies.

Aside from giving new algorithms for sampling from convex bodies, these new Markov chains also serve as a tool for mathematical analysis:
we use them to give the first polynomial-in-dimension mixing time bound for the often used coordinate hit-and-run chain when started from any interior point sufficiently far from the boundary of the body.

Joint work with Hariharan Narayanan (TIFR) and Amit Rajaraman (MIT).

Back to Table of Contents

Austin Stromme
Massachusetts Institute of Technology, USA

New statistical phenomena for entropic optimal transport

Optimal transport (OT) suffers from a well-known and severse statistical curse of dimensionality, obstructing its direct use in even moderate dimension. In practice, however, the OT problem is typically regularized with an entropic penalty term to afford the use of simpler and more scalable algorithms, forming entropic optimal transport (entropic OT). The ubiquity of entropic OT in practice, as well as the curse of dimensionality for un-regularized OT, motivates the statistical study of entropic OT. In this talk, we identify two novel statistical phenomena for entropic OT in the form of non-asymptotic bounds for various entropic OT quantities such as values, maps, and densities. Our first set of bounds are for high-dimensional settings, and give totally dimension-free rates of convergence, albeit with exponential dependence on the regularization parameter. And our second set of bounds identify a refined form of intrinsic dimension-dependence, which we call Minimum Intrinsic Dimension scaling (MID scaling), where the effective dimension is the minimum of the single-scale dimensions of the distributions. Our simple proof techniques are inspired by convex optimization, and notably avoid empirical process theory almost entirely.

Back to Table of Contents

Taiji Suzuki
The University of Tokyo, Japan

Convergence of mean field Langevin dynamics and its application to neural network optimization

The mean-field Langevin dynamics (MFLD) is a nonlinear generalization of the gradient Langevin dynamics (GLD) that minimizes an entropy regularized convex function defined on the space of probability distributions, and it naturally arises from the optimization of two-layer neural networks via (noisy) gradient descent. In this talk, I will present the convergence result of MFLD and explain how the convergence of MFLD is characterized by the log-Sobolev inequality of the so-called proximal Gibbs measure corresponding to the current solution. Moreover, I will provide a general framework to prove a uniform-in-time propagation of chaos for MFLD that takes into account the errors due to finite-particle approximation, time-discretization, and stochastic gradient approximation.

In the latter half, I will discuss the generalization error analysis of neural networks trained by MFLD. Addressing a binary classification problem, we have a general form of a test classification error bound that provides a fast learning rate based on a local Rademacher complexity analysis. By applying this general framework to the k-sparse parity problem, we demonstrate how the feature learning helps its sample complexity compared with the kernel methods.

Back to Table of Contents

Yanshuo Tan
National University of Singapore, Singapore

The Computational Curse of Big Data for Bayesian Additive Regression Trees: A Hitting Time Analysis

In this talk, we will be interested in inverse problems arising in the signal and image processing field.

Solving such problems imply in a fist time to formalise the direct problem by understanding the physics behind and in a second time, to solve the associated inverse problem, through a variational formulation, that is, solving an optimization problem. Such issues are encountered in many areas such as biology, medical imaging, chemistry, audio signal processing, ... for which, different tasks have to be tackled such as deconvolution, restoration, unmixing, missing data reconstruction, ...

Classical optimization-based approaches consist in, once the optimization problem has been formulated, proposing iterative procedures (e.g. proximal algorithms) converging to a solution of the considered inverse problem. More recently, unrolled or unfolded neural networks have been proposed. They combine optimization and learning, constitute interpretable networks and integrate information about the direct model. We will study and describe such networks for the resolution of two inverse problems: image deconvolution and robust PCA.

Collaborations: this work has been done in collaboration with Vincent Tan, Emmanuel Soubiès, Pascal Nguyen and Elisabeth Tan.

Back to Table of Contents

Vincent Y. F. Tan
National University of Singapore, Singapore

Multi-Armed Bandits with Abstention

We introduce a novel extension of the canonical multi-armed bandit problem that incorporates an additional strategic element: abstention. In this enhanced framework, the agent is not only tasked with selecting an arm at each time step, but also has the option to abstain from accepting the stochastic instantaneous reward before observing it. When opting for abstention, the agent either suffers a fixed regret or gains a guaranteed reward. Given this added layer of complexity, we ask whether we can develop efficient algorithms that are both asymptotically and minimax optimal. We answer this question affirmatively by designing and analyzing algorithms whose regrets meet their corresponding information-theoretic lower bounds. Our results offer valuable quantitative insights into the benefits of the abstention option, laying the groundwork for further exploration in other online decision-making problems with such an option. Numerical results further corroborate our theoretical findings.

This is joint work with Junwen Yang and Tianyuan Jin.

Back to Table of Contents

Kim-Chuan Toh
National University of Singapore, Singapore

Convex Clustering: Theoretical Guarantee and Efficient Computations

Back to Table of Contents

Xin Tong
National University of Singapore, Singapore

Gradient flow for fairness in real and virtual worlds

Fairness is an important topic for modern day machine learning (ML). In the real world, social welfare applications often require the solution from ML to meet certain fairness constraints. In the virtual world applications like video game matching, fairness is essential for customer retention. These fairness requirements impose new and challenging problems for Bayesian ML. In particular, the target densities are defined implicitly as solution to different constrained density optimization problems. This makes classical sampling methods such as MCMC difficult to implement. We will discuss the constrained control gradient flow, which can be used to solve the aforementioned problems.

Back to Table of Contents

Eric Vanden-Eijnden
Courant Institute, New York University, USA

Stochastic Interpolants: A Unifying Framework for Flows and Diffusions

Stochastic interpolants are a class of generative models that unifies flow-based and diffusion-based methods and allow one to bridge any two arbitrary probability density functions exactly in finite time. These interpolants are built by combining data from the two prescribed densities with an additional latent variable that shapes the bridge in a flexible way. The time-dependent probability density function of the stochastic interpolant can be shown to satisfy a first-order transport equation as well as a family of forward and backward Fokker-Planck equations with tunable diffusion. Upon consideration of the time evolution of an individual sample, this viewpoint immediately leads to both deterministic and stochastic generative models based on probability flow equations or stochastic differential equations with an adjustable level of noise. The drift coefficients entering these models are time-dependent velocity fields characterized as the unique minimizers of simple quadratic objective functions, one of which is a new objective for the score of the interpolant density. Remarkably, minimization of these quadratic objectives leads to control of the likelihood for any of our generative models built upon stochastic dynamics. By contrast, generative models based upon a deterministic dynamics must, in addition, control the Fisher divergence between the target and the model. Connections with diffusion based models, other stochastic bridges will also be discussed, in particular to show that such models recover the Schrödinger bridge between the two target densities when explicitly optimizing over the interpolant.

Back to Table of Contents

Ke Wang
The Hong Kong University of Science and Technology, China

Random perturbation of low-rank matrices

The analysis of large matrices is a key aspect of high-dimensional data analysis, with computing the singular values and vectors of a matrix being a central task. However, real-world data is often disturbed by noise, which affects the essential spectral parameters of the matrix. While classical deterministic theorems can provide accurate estimates for the worst-case scenario, this talk will focus on the case when the perturbation is random. By assuming that the data matrix has a low rank, optimal subspace perturbation bounds can be achieved under mild assumptions.

This talk is based on joint works with Sean O'Rourke and Van Vu.

Back to Table of Contents

Wanjie Wang
National University of Singapore, Singapore

Network-Adjusted Covariates for Community Detection

Community detection is a crucial task in network analysis that can be significantly improved by incorporating subject-level information, i.e. covariates. Existing methods have shown the effectiveness of using covariates on the low-degree nodes, but rarely discuss the case where communities have significantly different density levels, i.e. multiscale networks.

In this work, we introduce a novel method that addresses this challenge by constructing network-adjusted covariates, which leverage the network connections and covariates with a node-specific weight to each node. This weight can be calculated without tuning parameters.
We present novel theoretical results on the strong consistency of our method under degree-corrected stochastic blockmodels with covariates, even in the presence of mis-specification and multiple sparse communities. Additionally, we establish a general lower bound for the community detection problem when both network and covariates are present, and it shows our method is optimal for connection intensity up to a constant factor.

Our method outperforms existing approaches in simulations and a LastFM app user network. We then compare our method with others on a statistics publication citation network where 30% of nodes are isolated, and our method produces reasonable and balanced results.

Back to Table of Contents

Ting-Kam Leonard Wong
University of Toronto, Canada

Bregman-Wasserstein divergence and a modified JKO scheme

Current applications of optimal transport are often based on the quadratic Wasserstein distance on Euclidean space. In a Riemannian setting the Riemannian distance, and hence the Wasserstein distance, are usually computationally intractable. We show that when the Riemannian metric is Hessian, the Bregman-Wasserstein divergence - the optimal transport cost with respect to a Bregman divergence - provides a useful alternative. After giving some of its properties, we show that it leads to a modified JKO scheme which converges to the associated Fokker-Planck equation. In fact, modified JKO schemes can be formulated for any cost whose Hessian agrees with the metric.

Based on joint works with Cale Rankin.

Back to Table of Contents

Denny Wu
University of Toronto, Canada

Feature Learning in Two-layer Neural Networks under Structured Data

Real-world learning problems are often high-dimensional but also exhibit certain low-dimensional structures. We study the performance of (i) kernel methods, (ii) neural networks optimized via gradient descent, when the low-dimensionality is encoded in two ways: 1. the target function is a single-index model defined by an unknown link function applied to a one-dimensional projection of the input; 2. the input features are drawn from a spiked covariance model which describes a low-dimensional signal (spike) "hidden" in high-dimensional noise (bulk). We characterize the interplay between structured data (the extent of input anisotropy, as well as the overlap between the input spike and the target direction) and the sample complexity of the learning algorithms, and show that both kernel ridge regression and neural network benefit from low-dimensional structure, but GD-trained neural network can adapt to such a structure more effectively due to feature learning.

Based on joint works with Jimmy Ba, Murat A. Erdogdu, Alireza Mousavi-Hosseini, Taiji Suzuki, and Zhichao Wang.

Back to Table of Contents

Jia-Jie Zhu
Weierstrass Institute for Applied Analysis and Stochastics, Germany

Approximation and Kernelization of Gradient Flow Geometry: Fisher-Rao and Wasserstein

Motivated by numerous machine learning applications of Wassersetin and Fisher-Rao gradient flows, we present an investigation of the approximation and kernelization of dissipation geometries. We apply our framework to a few concrete gradient systems of interest, including the Fisher-Rao, Wasserstein, and Wasserstein-Fisher-Rao type gradient flows, and uncover a few surprising connections between those gradient flows.

Joint work with Alexander Mielke.

Back to Table of Contents