Non-Linear and High Dimensional Inference

Geometry and Statistics in Data Sciences
3-7 October 2022 - IHP, Paris

Monday

Sophie Langer
Overcoming the curse of dimensionality with deep neural networks.

Although the application of deep neural networks to real-world problems has become ubiquitous, the question of why they are so effective has not yet been satisfactorily answered. However, some progress has been made in establishing an underlying mathematical foundation. This talk surveys results on statistical risk bounds of deep neural networks. In particular, we focus on the question of when neural networks bypass the curse of dimensionality. Here we discuss results for vanilla feedforward and convolutional neural networks as well as regression and classification settings.

Adeline Fermanian
Scaling ResNets in the Large-depth Regime.

Deep ResNets are recognized for achieving state-of-the-art results in complex machine learning tasks. However, the remarkable performance of these architectures relies on a training procedure that needs to be carefully crafted to avoid vanishing or exploding gradients, particularly as the depth L increases. No consensus has been reached on how to mitigate this issue, although a widely discussed strategy consists in scaling the output of each layer by a factor α_L. We show in a probabilistic setting that with standard i.i.d. initializations, the only non-trivial dynamics is for $\alpha_L = 1/\sqrt{L}$ (other choices lead either to explosion or to identity mapping). This scaling factor corresponds in the continuous-time limit to a neural stochastic differential equation,

contrarily to a widespread interpretation that deep ResNets are discretizations of neural ordinary differential equations. By contrast, in the latter regime, stability is obtained with specific correlated initializations and $\alpha_L = 1/L$. Our analysis suggests a strong interplay between scaling and regularity of the weights as a function of the layer index. Finally, in a series of experiments, we exhibit a continuous range of regimes driven by these two parameters, which jointly impact performance before and after training.

Mikhail Belkin
Neural networks, wide and deep, singular kernels and Bayes optimality.

Wide and deep neural networks are used in many important practical settings. In this talk, I will discuss some aspects of width and depth related to optimization and generalization. I will first discuss what happens when neural networks become infinitely wide, giving a general result for the transition to linearity (i.e., showing that neural networks become linear functions of parameters) for a broad class of wide neural networks corresponding to directed graphs. I will then proceed to the question of depth, showing equivalence between infinitely wide and deep fully connected networks trained with gradient descent and Nadaraya-Watson predictors based on certain singular kernels. Using this connection we show that for certain activation functions these wide and deep networks are (asymptotically) optimal for classification but, interestingly, never for regression. (Based on joint work with Chaoyue Liu, Adit Radhakrishnan, Caroline Uhler and Libin Zhu.)

Tuesday

Clément Berenfeld
Understanding the geometry of high-dimensional data through the reach.

In high-dimensional statistics, and more particularly in manifold learning, the reach is an ubiquitous regularity parameter that encompasses the well-behavior of the support of the underlying probability measure. Enforcing a reach constraint is, in most geometric inference tasks, a necessity, which raises the question of the estimability of this parameter.We will try to understand how the reach relates to many other important geometric invariants and propose and estimation strategy that relies on estimating the intrinsic metric of the data. (Joint work with Eddie Aamari and Clément Levrard.)

Wolfgang Polonik
Topologically penalized regression on manifolds.

We study a regression problem on a compact manifold. In order to take advantage of the underlying geometry and topology of the data, we propose to perform the regression task on the basis of eigenfunctions of the Laplace-Beltrami operator of the manifold that are regularized with topological penalties. We will discuss the approach and the penalties, provide some supporting theory and illustrate the performance of the methodology on some data sets, illustrating the relevance of our approach in the case where the target function is "topologically smooth". This is joint work with O. Hacquard, K. Balasubramanian, G. Blanchard and C. Levrard.

John Harlim
Leveraging the RBF operator estimation for manifold learning.

I will discuss the radial-basis function pointwise and weak formulations for approximating Laplacians on functions and vector fields based on randomly sampled point cloud data, whose spectral properties are relevant to manifold learning. For the pointwise formulation, I will demonstrate the importance of the novel local tangent estimation that accounts for the curvature, which crucially improves the quality of the operator estimation. I will report the spectral theoretical convergence results of

these formulations and their strengths/weaknesses in practice. Supporting numerical examples, involving the spectral estimation of the Laplace-Beltrami operator and various vector Laplacians such as the Bochner, Hodge, and Lichnerowicz Laplacians will be demonstrated with appropriate comparisons to the standard graph-based approaches.

Break

Marina Meila
Manifold Learning, Explanations and Eigenflows.

This talk will extend Manifold Learning in two directions.First, we ask if it is possible, in the case of scientific data where quantitative prior knowledge is abundant, to explain a data manifold by new coordinates, chosen from a set of scientifically meaningful functions? Second, we ask how popular Manifold Learning tools and their applications can be recreated in the space of vector fields and flows on a manifold.Central to this approach is the order 1-Laplacian of a manifold, Δ_1, whose eigen-decomposition into gradient, harmonic, and curl, known as the Helmholtz-Hodge Decomposition, provides a basis for all vector fields on a manifold. We present an estimator for Δ_1, and based on it we develop a variety of applications. Among them, visualization of the principal harmonic, gradient or curl flows on a manifold, smoothing and semi-supervised learning of vector fields, 1-Laplacian regularization. In topological data analysis, we describe the 1st-order analogue of spectral clustering, which amounts to prime manifold decomposition. Furthermore, from this decomposition a new algorithm for finding shortest independent loops follows. The algorithms are illustrated on a variety of real data sets. (Joint work with Yu-Chia Chen, Samson Koelle, Hanyu Zhang and Ioannis Kevrekidis.)

Franck Picard
A probabilistic Graph Coupling View of Dimension Reduction.

Dimension reduction is a standard task in machine learning, to reduce the complexity and represent the data at hand. Many (and more than many!) methods have been proposed for this purpose, among which the seminal principal component analysis (PCA), that approximates the data linearly with a reduced number of axes. In recent years, the field has witness the emergence of new non linear methods, like the

Stochastic Neighbor Embedding method (SNE) and the Uniform Manifold Approximation and Projection method (UMAP), that proposes very efficient low-dimensional representations of the observations. Though widely used, these approaches lack clear probabilistic foundations to enable a full understanding of their properties and limitations. A common feature of these techniques is to be based on a minimization of a cost between input and latent pairwise similarities, but the generative model is still missing. In this work we introduce a unifying statistical framework based on the coupling of hidden graphs using cross entropy. These graphs induce a Markov random field dependency structure among the observations in both input and latent spaces. We show that existing pairwise similarity dimension reduction methods can be retrieved from our framework with particular choices of priors for the graphs. Moreover this reveals that these methods suffer from a statistical deficiency that explains poor performances in conserving coarse-grain dependencies. Our model is leveraged and extended to address this issue while new links are drawn with Laplacian eigenmaps and PCA.

Alexander Cloninger
Learning on and near low-dimensional subsets of the Wasserstein Manifold.

Detecting differences and building classifiers between distributions $\{\mu_i\}_{i=1}^N$, given only finite samples, are important tasks in a number of scientific fields. Optimal transport (OT) has evolved as the most natural concept to measure the distance between distributions, and has gained significant importance in machine learning in recent years. There are some drawbacks to OT: computing OT can be slow, and because OT is a distance metric, it only yields a pairwise distance matrix between distributions rather than embedding those distributions into a vector space. If we make no assumptions on the family of distributions, these drawbacks are difficult to overcome. However, in the case that the measures are generated by push-forwards by elementary transformations, forming a low-dimensional submanifold of the Wasserstein manifold, we can deal with both of these issues on a theoretical and a computational level. In this talk, we'll show how to embed the space of distributions into a Hilbert space via linearized optimal transport (LOT), and how linear techniques can be used to classify different families of distributions generated by elementary transformations and perturbations. The proposed framework significantly reduces both the computational effort and the required training data in supervised settings. Similarly, we'll demonstrate the ability to learn a near isometric embedding of the low-dimensional submanifold. Finally, we'll provide non-asymptotic bounds on the error induced in both the supervised and unsupervised algorithms from finitely sampling the target

distributions and projecting the LOT Hilbert space into a finite dimensional subspace. We demonstrate the algorithms in pattern recognition tasks in imaging and provide some medical applications.

Wednesday

Claudia Strauch
On high-dimensional Lévy-driven Ornstein–Uhlenbeck processes.

We investigate the problem of estimating the drift parameter of a high-dimensional Lévy-driven Ornstein–Uhlenbeck process under sparsity constraints. It is shown that both Lasso and Slope estimators achieve the minimax optimal rate of convergence (up to numerical constants), for tuning parameters chosen independently of the confidence level. The results are non-asymptotic and hold both in probability and conditional expectation with respect to an event resembling the restricted eigenvalue condition. (Based on joint work with Niklas Dexheimer.)

Botond Szabo
Linear methods for nonlinear inverse problems.

We consider recovering an unknown function f from a noisy observation of the solution u to a partial differential equation, where for the elliptic differential operator L, the map $L(u)$ can be written as a function of u and f, under Dirichlet boundary condition. A particular example is the time-independent Schrödinger equation. We transform this problem into the linear inverse problem of recovering $L(u)$, and show that Bayesian methods for this problem may yield optimal recovery rates not only for u, but also for f. The prior distribution may be placed on u or its elliptic operator. Adaptive priors are shown to yield adaptive contraction rates for f, thus eliminating the need to know the smoothness of this function. Known results on uncertainty quantification for the linear problem transfer to f as well. The results are illustrated by several numerical simulations. (Joint work with Geerten Koers and Aad van der Vaart.)

Judith Rousseau
Bayesian nonparametric estimation of a density living near an unknown manifold.

In high dimensions it is common to assume that the data have a lower dimensional structure. In this work we consider that the observations are iid and with a distribution whose support is concentrated near a lower dimensional manifold. Neither the

manifold nor the density is known. A typical example is for noisy observations on an unknown low dimensional manifold. We consider a family of Bayesian nonparametric density estimators based on location -scale Gaussian mixture priors and we study the asymptotic properties of the posterior distribution. Our work shows in particular that non conjuguate location -scale Gaussian mixture models can adapt to complex geometries and spatially varying regularity. This talk will also review the various aspects of mixtures of Gaussian for density estimation. (Joint work with Clément Berenfeld, Dauphine, and Paul Rosa, Oxford.)

Thursday

Denis Belometsny
Dimensionality reduction in reinforcement learning by randomisation.

In reinforcement learning an agent interacts with an environment, whose underlying mechanism is unknown, by sequentially taking actions, receiving rewards, and transitioning to the next state. With the goal of maximizing the expected sum of the collected rewards, the agent must carefully balance between exploring in order to gather more information about the environment and exploiting the current knowledge to collect the rewards. In this talk, we are interested in solving this exploration-exploitation dilemma by injecting noise into the agent's decision-making process in such a way that the dependence of the regret on the dimension of state and action spaces is minimised. We also review some recent approaches towards dimension reduction in RL.

Gilles Blanchard
Stein effect for estimating many vector means: a "blessing of dimensionality" phenomenon.

Consider the problem of joint estimation of the means for a large number of distributions in $\mathbb{R}^d$ using separate, independent data sets from each of them, sometimes also called "multi-task averaging" problem. We propose an improved estimator (compared to the naive empirical means of each data set) to exploit possible similarities between means, without any related information being known in advance. First, for each data set, similar or neighboring means are determined from the data by multiple testing. Then each naive estimator is shrunk towards the local average of its neighbors. We prove that this approach provides a reduction in mean squared error that can be significant when the (effective) dimensionality of the data is large, and when the unknown means exhibit structure such as clustering or concentration on a low-dimensional set. This is directly linked to the fact that the separation distance for testing is smaller than the estimation error in high dimension and generalizes the well-known James-Stein phenomenon. An application of this approach is the estimation ofmultiple kernel mean embeddings, which plays an important role in many modern applications. (Based on joint work with Hannah Marienwald and Jean-Baptiste Fermanian.)

Nicolas Verzelen
Optimal permutation estimation in crowd-sourcing problems.

Motivated by crowd-sourcing applications, we consider a model where we have partial observations from a bivariate isotonic $n \times d$ matrix with an unknown permutation π^* acting on its rows. We consider the twin problems of recovering the permutation π^* and estimating the unknown matrix. We introduce a polynomial-time procedure achieving the minimax risk for these two problems, this for all possible values of n, d, and all possible sampling efforts. Along the way, we establish that, in some regimes, recovering the unknown permutation π^* is considerably simpler than estimating the matrix. (Based on joint work with Alexandra Carpentier, Potsdam, and Emmanuel Pilliat, Montpellier.)

Break

Claire Lacour
On the use of overfitting for estimator selection.

In this talk we consider theproblem of estimator selection. In the case of density estimation, we study a method called PCO, which is intermediate between Lepski's method and penalized empirical risk minimization. The key point is the comparison of all the estimators to the overfitted one. We provide some theoretical results which lead to some fully data-driven selection strategy. We will also show the numerical performance of the method. (Joint work with P. Massart, V. Rivoirard and S. Varet.)

Peter Bartlett
The Dynamics of Sharpness-Aware Minimization.

Optimization methodology has been observed to affect statistical performance in high-dimensional prediction problems, and there has been considerable effort devoted to understanding the behavior of optimization methods and the nature of solutions that they find. We consider Sharpness-Aware Minimization (SAM), a gradient-based optimization method that has exhibited performance improvements over gradient descent on image and language prediction problems using deep networks. We show that

when SAM is applied with a convex quadratic objective, for most random initializations it converges to oscillating between either side of the minimum in the direction with the largest curvature, and we provide bounds on the rate of convergence. In the non-quadratic case, we show that such oscillations encourage drift toward wider minima by effectively performing gradient descent, on a slower time scale, on the spectral norm of the Hessian. (Based on joint work with Olivier Bousquet and Phil Long)

Friday

Johannes Schmidt-Hieber
A statistical analysis of an image classification problem.

The availability of massive image databases resulted in the development of scalable machine learningmethods such as convolutional neural network (CNNs) filtering and processing these data. While the very recent theoretical work on CNNs focuses on standard nonparametric denoising problems, the variability in image classification datasets does, however, not originate from additive noise but from variation of the shape and other characteristics of the same object across different images. To address this problem, we consider a simple supervised classification problem for object detection on grayscale images. While from the function estimation point of view, every pixel is a variable and large images lead to high-dimensional function recovery tasks suffering from the curse of dimensionality, increasing the number of pixels in our imagedeformation model enhances the image resolution and makes the object classification problem easier. We propose and theoretically analyze two different procedures. The first method estimates the image deformation by support alignment. Under a minimal separation condition, it is shown that perfect classification is possible. The second method fits a CNN to the data. We derive a rate for the misclassification error depending on the sample size and the number of pixels. Both classifiers are empirically compared on images generated from the MNIST handwritten digit database. The obtained results corroborate the theoretical findings.(Joint work with Sophie Langer, Twente.)

Damien Garreau
What does LIME really see in images?

The performance of modern algorithms on certain computer vision tasks such as object recognition is now close to that of humans. This success was achieved at the price of complicated architectures depending on millions of parameters and it has become quite challenging to understand how particular predictions are made. Interpretability methods propose to give us this understanding. In this paper, we study LIME, perhaps one of the most popular. On the theoretical side, we show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which we give an explicit expression. We further this

study for elementary shape detectors and linear models. As a consequence of this analysis, we uncover a connection between LIME and integrated gradients, another explanation method. More precisely, the LIME explanations are similar to the sum of integrated gradients over the superpixels used in the preprocessing step of LIME.

Vasiliki Velona
Learning a partial correlation graph using only a few covariance queries.

In settings where the covariance matrix is too large to even store, we would like to learn the partial correlation graph with as few covariance queries as possible (in a partial correlation graph, an edge exists if the corresponding entry in the inverse covariance matrix is non-zero). In recent work with Gabor Lugosi, Jakub Truszkowski, and Piotr Zwiernik, we showed that it is possible to use only a quasi-linear number of queries if the inverse covariance matrix is sparse enough, in the sense that the partial correlation graph resembles a tree on a global scale. I will explain these results and discuss extensions and applications.