Program YES Workshop 2022

Optimal Transport, Statistics, Machine Learning and moving in between

Monday September 5

10.30 – 11.20	Welcome and Registration
11.20 – 11.30	Opening
11.30 – 12.30	**Tutorial: Yoav Zemel**
12.30 – 14.00	Lunch
14.00 – 15.00	**Tutorial: Yoav Zemel**
15.00 – 15.30	Break
15.30 – 16.30	**Tutorial: Jonathan Niles-Weed**
16.30 – 16.50	Break
16.50 - 17.10	Jorge Justiniano
17.10 - 17.30	Hongwei Wen
17.30 – 19.00	Welcome reception

Tuesday September 6

09.00 – 10.00	**Tutorial: Jonathan Niles-Weed**
10.00 – 10.20	Break
10.20 – 11.20	**Tutorial: Marco Cuturi**
11.20 – 11.40	Break
11.40 – 12.40	**Tutorial: Marco Cuturi**
12.40 – 14.00	Lunch
14.00 – 16.00	**Discussion and Practical Tutorials**
16.00 – 16.30	Break
16.30 – 16.50	Guanyu Jin
16:50 – 17:10	Christopher B. Scarvelis

Wednesday September 7

09.00 – 10.00	**Tutorial: Yoav Zemel**
10.00 – 10.20	Break
10.20 – 11.20	**Tutorial: Jonathan Niles-Weed**
11.20 – 11.40	Break
11.40 – 12.40	**Tutorial: Marco Cuturi**
12.40 – 14.00	Lunch
14.00 – 16.00	**Discussion and Practical Tutorials**
16.00 – 16.30	Break
16.30 – 16.50	Flor de María
16:50 – 17:10	Petr Zamolodtchikov

Thursday September 8

09.00 – 10.00	**Invited: Anna Korba**
10.00 – 11.00	**Invited: Chin-Wei Huang**
11.00 – 11.30	Break
11.30 – 12.30	**Invited: Axel Munk**
12.30 – 14.00	Lunch
14.00 – 14.30	Hongjian Shi
14.30 – 15.00	Jia-Jie Zhu
15.00 – 15.30	Break
15.30 – 16.30	**Invited: Valentin de Bortoli**
16:30 – 17:00	Valentina Masarotto
17:00 – 17:30	Olga Mula
19:00 - …	Conference dinner

Friday September 9

09:00 – 10:00	**Invited: Alexandra Suvorikova**
10.00 – 10.30	Break
10.30 – 11.00	Arshak Minasyan
11.00 – 12.00	**Invited: Alberto González-Sanz**
12.00 – 14.00	Closing and Lunch

Tutorials

Jonathan Niles-Weed (NYU)

ESTIMATING WASSERSTEIN DISTANCES

Arguably the most basic statistical problem in optimal transport is the question of estimating Wasserstein distances between distributions on the basis of independent samples. We will explore several different perspectives on this question and develop the main techniques used in proving finite-sample statistical bounds. In the first lecture, we will focus on the "Wasserstein Law of Large Numbers" and give several proofs of the convergence rate of empirical measures in Wasserstein distance, based on the "transportation method" and the "empirical process method." In the second lecture, we will focus on improvements to the rates discussed in Lecture 1, and give refined versions of the transportation and empirical process methods suited to measures and couplings with additional structure. In the third lecture, we will consider alternate versions of the Wasserstein distance popular in machine learning, and discuss their benefits and drawbacks from the perspective of estimation rates.

Marco Cuturi (CREST - ENSAE, Apple ML Research)

HOW TO COMPUTE OPTIMAL TRANSPORT

In this short course I will present optimal transport from a computational perspective, motivating the relevant of OT for applications by starting from the canonical optimal matching problem. I will focus more particularly on two fruitful approaches to approximate OT: either by solving a regularized formulation of the discrete OT problem, or by attacking the more ambitious continuous OT problem between two densities using a neural network parameterization of OT maps. For the first approach, I will introduce the Sinkhorn algorithm and its differential properties. For the second part, I will prove reminders on the c-concavity and prove the Brenier theorem, to justify the application of input convex neural networks as tools to solve OT.

Yoav Zemel (University of Cambridge)

INTRODUCTION TO OPTIMAL TRANSPORT FOR STATISTICIANS

Optimal transport (or Wasserstein) distances quantify the difference between probability distributions by measuring the minimal effort required to reconfigure one distribution in order to recover the other. They have become a popular tool in a wide range of applications, where objects with complex geometric structure are to be compared. This course will give a brief introduction to some of the properties of Wasserstein distances that make them a versatile tool for statisticians. The first lecture will survey important properties of optimal transport, and the second will continue in this theme, with emphasis on the barycenter (Fréchet mean) in Wasserstein-2 space. Emphasis will be given to intuitive or rigorous arguments as to why these properties hold. The final lecture will describe a resampling scheme, based on the empirical measure (see Jonathan's talks), for computation of discrete Wasserstein-Fréchet means.

Invited Speakers

Valentin De Bortoli (CNRS ENS Ulm, Paris)

GENERATIVE MODELING AND SCHRÖDINGER BRIDGES

Abstract. Generative modeling is the task of drawing new samples from an underlying distribution known only via an empirical measure. There exists a myriad of models to tackle this problem with applications in image and speech processing, medical imaging, forecasting and protein modeling to cite a few. Among these methods score-based generative models (or diffusion models) are a new powerful class of generative models that exhibit remarkable empirical performance. They consist of a "noising" stage, whereby a diffusion is used to gradually add Gaussian noise to data, and a generative model, which entails a "denoising" process defined by approximating the time-reversal of the diffusion. A well-known limitation of diffusion models is that the forward-time stochastic process must be run for a sufficiently long time for the final distribution to be approximately Gaussian. In contrast, solving the Schrödinger Bridge problem, i.e. an entropy-regularized optimal transport problem on path spaces, yields diffusions which generate samples from the data distribution in finite time. I will present Diffusion Schrödinger Bridge, an original approximation of the Iterative Proportional Fitting procedure to solve the Schrödinger Bridge problem.

Alberto González-Sanz (Uni. Toulouse)

NONPARAMETRIC MULTIPLE-OUTPUT CENTER-OUTWARD QUANTILE REGRESSION

Abstract. This talk addresses the problem of non-parametric multiple-output quantile regression based on the novel concept of measure-transportation-based multivariate center-outward quantiles introduced in Chernozhukov et al. (2017) and Hallin et al. (2021). The conditional quantile regions and contours are obtained via the conditional center-outward quantile mapping. A new cyclically monotone interpolation, with non-necessarily constant weights, is proposed to define their empirical versions. This method is completely non-parametric and produces interpretable empirical regions/contours which converge in probability to their population counterparts. The presentation includes some real and simulated examples, demonstrating the power of the method as a data-analytic tool, and its ability to catch heteroskedasticity and non-linear trends.

Chin-Wei Huang (Microsoft Research)

CONVEX POTENTIAL FLOWS, DIFFUSION MODELS, AND TRANSITION PATHS

Abstract. Optimal transport provides an elegant mathematical framework for characterizing the displacement of probability distributions which has found a multitude of applications in generative modeling of static data and dynamical systems. In the first part of the talk, we will see how OT can be used to motivate the parameterization of generative models such as normalizing flows, which opens a new door for designing expressive invertible architectures for modelling arbitrary data distributions. In the second part of the talk, we will consider dynamical systems defined by stochastic differential equations known as diffusion models, and how they relate to the Schrodinger bridge problem and entropy-regularized OT. At the end, I will draw connections to some of the problems that we are solving in the space of chemical kinetics and highlight a few recent works in this direction.

Anna Korba (ENSAE)

Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM

Abstract. Many problems in machine learning can be formulated as optimizing a convex functional over a space of measures. This paper studies the convergence of the mirror descent algorithm in this infinite-dimensional setting. Defining Bregman divergences through directional derivatives, we derive the convergence of the scheme for relatively smooth and strongly convex pairs of functionals. Applying our result to joint distributions and the Kullback–Leibler (KL) divergence, we show that Sinkhorn's primal iterations for entropic optimal transport in the continuous setting correspond to a mirror descent, and we obtain a new proof of its (sub)linear convergence. We also show that Expectation Maximization (EM) can always formally be written as a mirror descent, and, when optimizing on the latent distribution while fixing the mixtures, we derive sublinear rates of convergence.

Joint work with Pierre-Cyril Aubin-Frankowski and Flavien Léger.

Axel Munk (Uni. Göttingen)

Transport dependency: Optimal transport based dependency measures

Abstract. Finding meaningful ways to determine the dependency between two random variables ξ and ζ is a timeless statistical endeavor with vast practical relevance. In recent years, several concepts that aim to extend classical means (such as the Pearson correlation or rank-based coefficients like Spearman's ρ) to more general spaces have been introduced and popularized, a well-known example being the distance correlation. In this talk, we propose and study an alternative framework for measuring statistical dependency, the transport dependency $\tau \geq 0$ (TD), which relies on the notion of optimal transport and is applicable in general Polish spaces. It can be estimated via the corresponding empirical measure, is versatile and adaptable to various scenarios by proper choices of the cost function. It intrinsically respects metric properties of the ground spaces. Notably, statistical independence is characterized by $\tau = 0$, while large values of τ indicate highly regular relations between ξ and ζ. Based on sharp upper bounds, we exploit three distinct dependency coefficients with values in $[0, 1]$, each of which emphasizes different functional relations: These transport correlations attain the value 1 if and only if $\zeta = \varphi(\xi)$, where φ is a) a Lipschitz function, b) a measurable function, c) a multiple of an isometry.

Besides a conceptual discussion of transport dependency, we address numerical issues and its ability to adapt automatically to the potentially low intrinsic dimension of the ground space. Monte Carlo results suggest that TD is a robust quantity that efficiently discerns dependency structure from noise for data sets with complex internal metric geometry. The use of TD for inferential tasks is illustrated for independence testing on a data set of trees from cancer genetics.

This is joint work with Giacomo Nies and Thomas Staudt.

Alexandra Suvorikova (WIAS)

Robust k-means in metric spaces and spaces of probability measures

Abstract. In this work we investigate the theoretical properties of the robust k-means clustering under assumption of adversarial data corruption. We provide non-asymptotic rates for excess distortion under weak model assumptions on the moments of the distribution.

Contributed Speakers

Valentina Masarotto (Leiden University)

OPTIMAL TRANSPORT MEETS TONAL COARTICULATION

Abstract. All our ways of verbal interacton and communication comprise melodic components. Sometimes more than words, volume and intonation express feelings, imply jokes, differentiate affirmations from questions, and overall play an essential part in conveying a message. Such "melodic variations" assume an even more dramatic role in tonal languages, as, in this setting, varying a sound's pitch while pronouncing the same word conveys different lexical meanings. An interesting aspect of tonal languages is tonal coarticulation. In everyday speech, words (and tones) are never pronounced in isolation, but are always concatenated to one another. This implies that each tone is not stable, but is affected by the neighbouring ones, preceding and following the tone of interest.

A large part of phonetic analysis is based on speech recording, and focusses on modelling fundamental frequency curves (F0 curves). The amplitude of F0 curves measures how high or low a speaker's voice sounds, and their intrinsic structure identifies them naturally as functional data. Modelling the effect of tonal coarticulation in the F0 curves is a complex problem, especially when tonal variation is not captured in the sound-curve shape, but it is of finer and higher order. Functional covariances are the functional objects that capture second-order (tonal) variation. Covariances play a fundamental role in functional data analysis but their statistical analysis is made harder by both their infinite-dimensionality and their intrinsic non-linearity. Working within such non-linearity constraints however, can give rise to very powerful statistical procedures. This talk will link to optimal transport theory and describe how working with covariances might answer some questions for the linguistic community.

Arshak Minasyan (ENSAE-CREST)

OPTIMAL EXACT RECOVERY OF THE MATCHING MAP OF UNKNOWN SIZE

Abstract. We consider the problem of finding the matching map of unknown size k^* between two sets of size n consisting of d-dimensional noisy observations of feature vectors. The main result shows that, in the high-dimensional setting, if the signal-to-noise ratio of the feature vectors is of order at least $d^{1/4}$ then it is possible to recover the true matching map exactly (making no errors) with high probability. We also prove the corresponding lower bound establishing the optimality of this rate. This rate is achieved using the estimated matching defined as the minimizer of the sum of squares of distances between matched pairs of points. Since the number of matching pairs is unknown we first estimate the parameter k^*. Then, we show that the resulting optimization problem can be formulated as a minimum-cost flow problem, and thus solved efficiently, with complexity $\widetilde{O}(\sqrt{k^*}\, n^2)$. Finally, we report the results of numerical experiments on both synthetic and real-world data that illustrate our theoretical results and provide further insight into the properties of the estimators and algorithms studied in this work. POINTGUARD0.GITHUB.IO

Olga Mula (TU/e)

SPARSE, ADAPTIVE INTERPOLATION OF MEASURES WITH WASSERSTEIN BARYCENTERS. APPLICATION TO MODEL ORDER REDUCTION

Abstract. We develop a general framework for sparse and adaptive interpolation of histograms/measures from the Wasserstein space of probability measures. The strategy relies on approximation of measures with Wasserstein barycenters. The optimal performance of the approach in terms of approximation error is characterized by a notion of best n-term barycentric approximation which we introduce in the talk. This best approximation is the minimizer of a highly non-convex, bi-level optimization problem, and we develop algorithmic strategies for practical numerical computation. We next leverage this approximation tool in order to build interpolation strategies to address structured prediction problems where the family of measures to approximate presents common general structural features. To illustrate the potential of the method, we focus on Model Order Reduction (MOR) of parametrized PDEs. The whole methodology is computationally feasible thanks to state of the art entropy regularized Sinkhorn algorithms from the field of numerical optimal transport.

Hongjian Shi (TU Munich)

On universally consistent and fully distribution-free rank tests of vector independence

Abstract. Rank correlations have found many innovative applications in the last decade. In particular, suitable rank correlations have been used for consistent and distribution-free tests of independence between pairs of random variables. However, the traditional concept of ranks relies on ordering data and is, thus, tied to univariate observations. As a result, it has long remained unclear how one may construct distribution-free yet consistent tests of independence between random vectors. In this talk, I will discuss how this problem can be addressed via a general framework for designing multivariate dependence measures and associated test statistics based on the recently introduced concept of center-outward ranks and signs, a multivariate generalization of traditional ranks. In this framework, we obtain new multivariate Hájek asymptotic representation results and use them for local power analyses that demonstrate the statistical efficiency of our tests.

Jia-Jie Zhu (WIAS)

Distributionally robust learning and optimization in the MMD geometry and beyond

Abstract. The last deep learning revolution focused on scaling up to large models, big data, and utilizing better stochastic optimization. In contrast, the principles of learning robustly under the so-called **distribution shift**, i.e., train-test distribution mismatch, is far from being clear. This distribution shift can be a consequence of causal confounding, unfairness due to data biases, and adversarial attacks.

In such cases, we use robustification strategy derived from *robust nonlinear optimization* and *distributionally robust optimization* for learning under distribution shift. Succinctly, we formulate the learning task as solving the **two-level distributionaly robust optimization**

$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_{Q} L(\theta, \xi)$$

where the plausible distribution shift is described by an ambiguity set $\mathcal{M}$. The distribution Q is then the **local worse-case distribution** we protect against. Previously, researchers have used the ϕ-divergences and Wasserstein distance as the underlying geometry for constructing ambiguity sets, e.g., $\mathcal{M} = \{Q : W_p(Q, \hat{P}_N) \leq \rho\}$, where W_p is the p-Wasserstein distances and $\hat{P}_N = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$. However, doing so places severely limitations on the learning models, e.g., one can only apply Wasserstein DRO to treat learning with linear models [2] or models with known Lipschitz constant. Those limitations make (exact) Wasserstein DRO unsuitable for modern machine learning.

To address those issues, we established a generalized kernel framework for DRO for **robust learning with nonlinear models – the Kernel DRO** [3], where we show the DRO problem can be exactly reformulated into a kernel learning problem. We prove a Kantorovich-type duality result for general DRO problem, which includes the special cases of DRO using the maximum mean discrepancy (MMD) and the integral probability metric family. This perspective unifies multiple existing robust and stochastic optimization methods. We show multiple applications of our frameworks to state-of-the-art topics such as adversarially robust deep learning [4] and chance constrained stochastic optimization and control [1]. We conclude with a discussion of the connection between DRO and topics such as Wasserstein gradient flow.

References

[1] Yassine Nemmour, Heiner Kremer, Bernhard Schölkopf, and Jia-Jie Zhu. Maximum Mean Discrepancy Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee. *arXiv:2204.11564 [cs, eess, math]*, April 2022. arXiv: 2204.11564.

[2] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. *arXiv:1509.09259 [math, stat]*, December 2015. arXiv: 1509.09259.

[3] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, March 2021. ISSN: 2640-3498.

[4] Jia-Jie Zhu, Christina Kouridi, Yassine Nemmour, and Bernhard Schölkopf. Adversarially Robust Kernel Smoothing. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4972–4994. PMLR, May 2022. ISSN: 2640-3498.

PhD Talks

Guanyu Jin (UvA)

Robust optimization for rank-dependent models with uncertain probabilities

Abstract. This paper studies robust optimization for a broad class of risk measures with uncertainty sets defined by ϕ-divergence measures. The risk measures are allowed to be non-linear in probabilities, are characterized by a Choquet integral induced by a probability weighting function, and include many well-known examples (such as CVaR, Mean Median Deviation, Gini-type). Optimization for this class of preference functionals is challenging due to their rank-dependent nature. We show that for general concave probability weighting functions, our robust optimization problem can be reformulated into a convex optimization problem with finitely many constraints. We prove that the conic representability of the reformulated robust counterpart is determined jointly by the conic representability of the probability weighting and ϕ-divergence functions and derive the explicit conic representations for a collection of canonical examples. While the number of constraints in general scales exponentially with the dimension of the state space, we provide four different approximation methods to circumvent this curse of dimensionality. They yield tight upper and lower bounds on the optimal value of the exact problem, and are formally shown to converge to the exact solution asymptotically. This is illustrated numerically in two examples given by a robust newsvendor problem and a robust portfolio choice problem. We also analyze non-concave probability weighting functions and show that the corresponding robust optimization problems are typically not equivalent to their concave envelope approximations, as opposed to other settings in the literature.

Jorge Justiniano (Uni. Bonn)

Approximation of splines in Wasserstein spaces

Abstract. This paper investigates a time discrete variational model for splines in Wasserstein spaces to interpolate probability measures. Cubic splines in Euclidean space are known to minimize of the integrated squared acceleration subject to a set of interpolation constraints. As generalization on the manifold of probability measures the integral over the squared Riemannian acceleration is considered as a spline energy and adding the action functional a regularized spline energy is defined. Both energies are then discretized in time using local Wasserstein-2 distances and the generalized Wasserstein barycenter. The existence of time discrete regularized splines for given interpolation conditions is established and on the subspace of Gaussian distributions, explicit notions for the time discrete and time continuous splines are investigated. In particular, it is shown that discrete splines interpolating Gaussians are indeed families of Gaussians. The implementation is based on the entropy regulazation and the Sinkhorn algorithm. A variant of the iPALM method is applied for the minimization of the fully discrete functional. A variety of numerical examples demonstrate the robustness of the approach and show striking characteristics of the approach. As a particular application the spline interpolation for synthesized textures is presented.

Flor de María (CIMAT)

Variants of Wasserstein based kernels for distributional data

Abstract. In this work, we exploit the advantage of Kernel methods to extend classical statistical and machine learning techniques for distributional data (an observation is a distribution or a sample of a distribution, instead of a vector). We use the closed formula of the p-Wasserstein distance (with Euclidean ground distance) in one dimension to characterize the cases for which positive definite Wasserstein based kernels for univariate distributions can be defined. This characterization is used to extend the definition of the Sliced Wasserstein distance by replacing the squared 2-Wassserstein distance with other alternatives. Next, the variants of the Sliced Wasserstein distance are used to construct well-defined kernels between multivariate distributions. Other options of kernels for the multivariate case, based on bounds for the Wasserstein distance, are explored.

Joint work with Dr. Johan Van Horebeek

Christopher Basil Scarvelis (MIT)

Riemannian metric learning via optimal transport

Abstract. We introduce an optimal transport-based model for learning a metric tensor from cross-sectional samples of evolving probability measures on a common Riemannian manifold. We neurally parametrize the metric as a spatially-varying matrix field and efficiently optimize our model's objective using backpropagation. Using this learned metric, we can nonlinearly interpolate between probability measures and compute geodesics on the manifold. We show that metrics learned using our method improve the quality of trajectory inference on scRNA and bird migration data at the cost of little additional cross-sectional data.

Hongwei Wen (UT-EEMCS)

Random forest density estimation

Abstract. We propose a density estimation algorithm called *random forest density estimation (RFDE)* based on random trees where the split of cell is along the midpoint of the randomly chosen dimension. By combining the efficient random tree density estimation (RTDE) and the ensemble procedure, RFDE can alleviate the problems of boundary discontinuity suffered by partition-based density estimations. From the theoretical perspective, we first prove the fast convergence rates of RFDE if the density function lies in the Hölder space $C_{0,\alpha}$. Moreover, if the density function resides in the subspace $C_{1,\alpha}$, which contains smoother density functions, we for the first time manage to explain the benefits of ensemble learning in density estimation. To be specific, we show that the upper bound of the ensemble estimator RFDE turns out to be strictly smaller than the lower bound of its base estimator RTDE in terms of convergence rates. In the experiments, we verify the theoretical results, and show the promising performance of RFDE on both synthetic and real-world datasets. Moreover, we evaluate our RFDE through the problem of anomaly detection as a possible application.

Petr Zamolodtchikov (UT-EEMCS)

Local convergence rates of the least-squares estimator with applications to transfer learning

Abstract. Convergence properties of empirical risk minimizers can be conveniently expressed in terms of the associated population risk. To derive bounds for the performance of the estimator under covariate shift, however, pointwise convergence rates are required. Under weak assumptions on the design distribution, it is shown that least squares estimators (LSE) over 1-Lipschitz functions are also minimax rate optimal with respect to a weighted uniform norm, where the weighting accounts in a natural way for the non-uniformity of the design distribution. This moreover implies that although least squares is a global criterion, the LSE turns out to be locally adaptive. We develop a new indirect proof technique that establishes the local convergence behavior based on a carefully chosen local perturbation of the LSE. These local rates are then used to construct a rate-optimal estimator for transfer learning under covariate shift.