Optimal Transport in Data Science
Poster Session Abstracts
Tuesday, May 8, 2023

A non-parametric generative model for conditional sampling

Ricardo Baptista, California Institute of Technology

Sampling conditional distributions is a fundamental task for Bayesian inference and density estimation. Generative models, such as normalizing flows and generative adversarial networks, characterize conditional distributions by learning a transformation that transports a simple reference (e.g., a standard Gaussian) to a target distribution. While these approaches can successfully describe many non-Gaussian problems, their performance is constrained by parametric bias and the reliability of (possibly adversarial) gradient-based optimizers to learn these transformations. This work proposes a non-parametric generative model, with naturally adaptive complexity, that iteratively maps samples between the reference and target distributions. Our formulation solves the optimal transport problem by minimizing a weighted cost function that yields block-triangular transport maps, thereby extending the approach in (Tabak and Trigila, 2014) to sampling conditionals. In this presentation, I will relate the approach to gradient flows on probability space and demonstrate the performance of the algorithm for parameter inference problems with nonlinear ODEs.

Variant of Paulsen's Problem in mathematical signal process: from probabilistic frames and optimal transport perspective

Dongwei Chen, Clemson University

The Paulsen problem is a basic problem in frame theory claiming that every $\epsilon$-nearly equal norm Parseval frame in $d$ dimension is within squared distance $O(\epsilon d^2)$ of an equal norm Parseval frame. A variant of Paulsen Problem is that the closest Parseval frame to a given frame is the canonical dual Parseval frame. In this work, we will focus on a similar variant of Paulsen Problem for probabilistic frame, which is a probability measure on $\mathbb{R}^d$ with an invertible second-moment matrix. We show that there exists a unique closest tight probabilistic frame with unit norm to a given probabilistic frame, where the distance is quantified by the 2-Wasserstein metric in optimal transport.

Identification of diverse trajectories and prediction of early differential gene expression in EMT by optimal-transport analysis of single-cell RNA sequencing

Yu-Chen Cheng, Dana-Farber Cancer Institute/Harvard University

Epithelial-mesenchymal transition (EMT) is a complex biological process involving multiple steps and changes in gene expression. Recent single-cell RNA sequencing studies have shown that EMT signaling pathways are activated sequentially along the trajectory from epithelial to mesenchymal features, resulting in cellular heterogeneity over time. However, the heterogeneity across divergent trajectories and cell fates is not yet fully understood. In this study, we used optimal-transport analysis to infer ancestor distributions of different cell fates and recover their most probable past trajectories. Our analysis identified three distinct temporal processes - failed EMT, partial EMT, and full EMT - each with unique cellular signatures of stemness, proliferation, and metabolism. By extending differential gene expression analysis from the end of EMT to all early time points, we identified a list of early differentially expressed genes that strongly predict the progression towards partial EMT with an increase in stemness. To validate our predictions, we found downregulation of EED and EZH2 genes in the early phase of the partial EMT trajectory, which is consistent with a recent CRISPR-associated knock-out screening study. We also found that fn1, KRT8, and POSTN were highly expressed in the very early phase, one day after TGF-beta treatment, of the partial EMT trajectory. The role of these genes in this phase had not been fully characterized previously, suggesting their potential as novel regulators of this process. Our study provides insights into the dynamic nature of EMT and offers a framework for identifying key regulators in the early phase of EMT. These findings have important implications for understanding the role of EMT in cancer progression and developing new cancer prevention strategies.

A constrained unbalanced optimal transport problem
Yuqing Dai, Duke University

The unbalanced optimal transport problem is a minimization problem of the total transport cost of transporting an initial mass to a target mass such that their values are not necessarily equal. Compared with the balanced optimal transport, unbalanced optimal transport allows the transporting mass to be changed, giving a smaller total transportation cost value. In this poster, I will introduce a modified dynamic formulation of Hellinger-Kantorovich problem with a non-negative constraint such that the transporting mass is non-decreasing during transportation. I will present some properties of this problem, its equivalent formulations, and some numerical experiments.

Aggregation Methods for Computing Steady States
Gabriel Earle, University of Massachusetts Amherst

In this work, we present a new estimate for the asymptotic rate of convergence of a multigrid method known as iterative aggregation-disaggregation (IAD) for computing the steady-state of Markov chains. This method is potentially useful in molecular dynamics and statistical physics, as it is well suited to systems which are both highly metastable and nonequilibrium. We show that IAD effectively accelerates the rate of sampling of such processes, with implications for its potential applications.

Joint Work with Brian Van Koten (UMass Amherst)

Lipschitz regularized f-divergences flows and generative particles algorithm

Hyemin Gu, UMass Amherst

We constructed gradient flows which minimize Lipschitz regularized f-divergences which are written in variational formulation. Variational formulation enables to approximate a function of likelihood ratio dP/dQ between two empirical distributions obtained by samples. In case of KL-divergence, this function is the log likelihood ratio. We allow flexibility in choosing f depending on the probability distribution to learn, so that heavy-tailed distributions can be fitted using alpha divergences, instead of the KL divergence. On the other hand, Lipschitz regularization leads to the f-divergences bounded even between non-absolutely continuous distributions.
In terms of the transport equation of probability distributions in the Wasserstein space, the gradient flow evolves the empirical distribution in direction of the gradients of the function of likelihood ratio that are learned from data. This function is parametrized by neural networks, and its gradients give us the particle dynamics. Hence we transport the particles through the ODEs and generate more samples from the particles trajectory.
Moreover, in order to reduce the dimensions, we developed our particle transportation algorithm in latent spaces and applied to high dimensional problems such as image generation and gene expression data merging.

Linearized Wasserstein dimensionality reduction with approximation guarantees

Varun Khurana, University of California, San Diego

We introduce LOT Wassmap, a computationally feasible algorithm to uncover low-dimensional structures in the Wasserstein space. The algorithm is motivated by the observation that many datasets are naturally interpreted as probability measures rather than points in $\mathbb{R}^n$, and that finding low-dimensional descriptions of such datasets requires manifold learning algorithms in the Wasserstein space. Most available algorithms are based on computing the pairwise Wasserstein distance matrix, which can be computationally challenging for large datasets in high dimensions. Our algorithm leverages approximation schemes such as Sinkhorn distances and linearized optimal transport to speed-up computations, and in particular, avoids computing a pairwise distance matrix. We provide guarantees on the embedding quality under such approximations, including when explicit descriptions of the probability measures are not available and one must deal with finite samples instead. Experiments demonstrate that LOT Wassmap attains correct embeddings and that the quality improves with increased sample size. We also show how LOT Wassmap significantly reduces the computational cost when compared to algorithms that depend on pairwise distance computations.

Transport subspace models and invariance encoding

Shiying Li, University of North Carolina - Chapel Hill

Transport-based metrics and related embeddings have recently been used to model data classes where nonlinear structures or variations are present. We will describe several transport transforms and their mathematical properties related to convexification under various algebraic generative modeling assumptions, enabling efficient modeling of data classes as subspaces in the transform domain. Such modeling also gives rise to simple machine learning algorithms with the ability to incorporate meaningful invariances, which are robust to out-of-distribution samples (generalizability). We will show applications in time series classification and face recognition under varying illumination conditions.

This poster is based on joint work with Akram Aldroubi, Yan Zhuang, Hasnat Rubaiyat, Gustavo Rohde, M Shifat Rabbi, and Xuwang Yin.

Multispecies Optimal Transport and its Linearization

Dorde Nikolic, University of California Santa Barbara

The discovery of linear optimal transport by Wang et al., in 2013 improved the computational efficiency of optimal transport algorithms for grayscale image classification. Our main goal is to classify special kinds of multicolor images, arising in collider events. We will introduce the basics of optimal transport theory, linear optimal transport and the multispecies distance. I will discuss similarities of the multispecies case with the Hellinger-Kantorovich distance, which was linearized in 2021 by Cai et al., via its Riemannian structure. This is a work in progress with Katy Craig and Nicolás García Trillos.

Entropic regularized Wasserstein distances between infinite-dimensional Gaussian measures and Gaussian processes

Minh Ha Quang, RIKEN

Optimal transport (OT) has been attracting much research attention in various fields, in particular machine learning and statistics.
It is well-known that the exact OT distances are generally computationally demanding and suffer from the curse of dimensionality.
One approach to alleviate these problems is via regularization. In this work, we present recent results on the entropic regularization of OT in the setting of Gaussian measures on Euclidean space and their generalization to the infinite-dimensional setting of Gaussian measures on Hilbert space and Gaussian processes. In these settings, the entropic regularized Wasserstein distances admit closed form expressions, which satisfy many favorable theoretical properties, especially in comparison with the exact distance. In particular, we show that the infinite-dimensional regularized distances can be consistently estimated from the finite-dimensional versions, with dimension-independent sample complexities. The methodology of reproducing kernel Hilbert spaces (RKHS) plays a crucial role in the theoretical analysis. The mathematical formulation will be illustrated with numerical experiments on Gaussian processes.

References:
1) H a Quang Minh. Entropic regularization of Wasserstein distance between infinite-dimensional
Gaussian measures and Gaussian processes, Journal of Theoretical Probability, 2022, https://link.springer.com/article/10.1007/s10959-022-01165-1.
2) H a Quang Minh. Convergence and finite sample approximations of entropic regularized Wasserstein distances in Gaussian and RKHS settings, Analysis and Applications, 2022, https://www.worldscientific.com/doi/abs/10.1142/S0219530522500142

3) H a Quang Minh. Finite sample approximations of exact and entropic Wasserstein distances between covariance operators and Gaussian processes, SIAM/ASA Journal on Uncertainty Quanti cation, volume 10, number 1, pages 96-124, 2022, https://epubs.siam.org/doi/abs/10.1137/21M1410488

Wasserstein Graph Metric Computes Graph Laplacian Kernel and Geodesics

Michael Rawson, PNNL

We explore metrics and stabilities in labeled graph spaces. Graphs are used to describe and model many systems. Often, stable regions of this space can inform expected behavior of such systems. We use regularized Wasserstein (Wass) graph metrics to calculate stabilities, that is, unit balls. Wass gives a metric between labeled graphs. Wass unit balls encode the geometry of a graph. We show that degenerate dimensions correspond to clusters of vertices. This calculation is faster than known methods like graph traversal. It is also accurate in many cases where spectral methods or Fiedler vector clustering fails.

Error control in target measure diffusion maps and applications to transition path theory

Shashank Sule, University of Maryland, College Park

We prove strong pointwise consistency estimates for target measure diffusion map (TMD map), a recently proposed algorithm for approximating generators of non-degenerate Ito diffusions on high-dimensional pointclouds with arbitrary sampling densities. Our contributions include the computation of variance error and bias errors of the algorithm up to explicit formulae for the prefactors. We show that approximating the committor function--an important reaction coordinate in molecular dynamics--with a uniform sampling density causes several terms in the bias error prefactor to cancel. This enables us to use TMD map with a postprocessed sampling density for the numerical computation of the committor function in high dimensions. Our work thus justifies why TMD map is particularly well-suited for the committor problem and opens the door for transport based post-processing techniques for accuracy-boosting sampling of manifolds.

A mean-field games laboratory for generative modeling

Benjamin Zhang, University of Massachusetts Amherst

We demonstrate the versatility of mean-field games (MFGs) as a mathematical framework for explaining, enhancing, and designing generative models. There is a pervasive sense in the generative modeling community that the various flow and diffusion-based generative models have some foundational common structure and interrelationships. We establish connections between MFGs and major classes of flow and diffusion-based generative models including continuous-time normalizing flows, score-based models, and Wasserstein gradient flows. We derive these three classes of generative models through different choices of particle dynamics and cost functions. Furthermore, we study the mathematical structure and properties of each generative model by studying their associated MFG's optimality condition, which is a set of coupled nonlinear partial differential equations (PDEs). The theory of MFGs, therefore, enables

the study of generative models through the theory of nonlinear PDEs. Through this perspective, we investigate the well-posedness and structure of normalizing flows, unravel the mathematical structure of score-based generative modeling, and derive a mean-field game formulation of the Wasserstein gradient flow. From an algorithmic perspective, the optimality conditions of MFGs also allow us to introduce HJB regularizers for enhanced training a broader class of generative models. We present this framework as an MFG laboratory which serves as a platform for revealing new avenues of experimentation and invention of generative models. This laboratory will give rise to a multitude of well-posed generative modeling formulations, providing a consistent theoretical framework upon which numerical and algorithmic tools may be developed.

Efficient and Exact Multimarginal Optimal Transport with Pairwise Costs

Bohan Zhou, Dartmouth College

Optimal transport has profound and wide applications since its introduction in 1781 by Monge. Thanks to the Benamou-Brenier formulation, it provides a meaningful functional in the image science like image and shape registrations. However, exact computation through LP or PDE is in general not practical in large scale, while the popular entropy-regularized method introduces additional diffusion noise, deteriorating shapes and boundaries. Until the recent work [Jacobs and Leger, A Fast Approach to Optimal Transport: the back-and-forth method, Numerische Mathematik, 2020], solving OT in a both accurate and fast fashion finally becomes possible. Multi-marginal optimal transport is a natural extension from OT but has its own interest, and is in general more computationally expensive. The entropy method suffers from both diffusion noise and high dimensional computational issues. In this work with Matthew Parno, we extend from two marginals to multiple marginals, on a wide class of cost functions in the form of summed pairwise costs. This new method is fast and does not introduce diffusion. As a result, the new proposed method can be used in many fields those require accurate approach to MMOT.