Research Summary

From July 2018-July 2020, I have had four papers published/accepted in well-regarded peer-reviewed journals and another four papers published/accepted in well-regarded peer-reviewed conferences proceedings. Broadly speaking, the contributions of these papers are as follows:

- Reduction of computational complexity for measuring pairwise statistical distances (earth mover’s distance, maximum mean discrepancy) between K distributions from $O(K^2 N^2)$ to $O(K N^2)$
- Developed theory of convergence for data adaptive global and localized statistic to detect differences between distributions, and demonstrated importance of such statistic for generative models.
- Applications of pairwise statistical distances between distributions to public health, medical imaging, and unsupervised flow cytometry clustering.
- Creation of a generative model neural network that maintains the topological and geometric structure of the original data.
- Introduction of greedy optimization to the quadrature problem on graphs (estimating the mean of a function from limited samples), proofs of convergence, and state of the art results for choosing samples in certain graph neural net applications.
- Discovery of the non-trivial geometric structure of Laplacian eigenfunctions on certain graphs and discretized manifolds
- A novel theoretic result bounding the probability of misclassification of a layer-wise quantized neural network in terms of the Lipschitz constant of each layer of the network.

Below, I summarize my results in these areas:

1) Detecting similarities and differences between distributions is a classic question that remains important and difficult problem in statistics and data science, especially in high dimensions. Recent work on kernel Maximum Mean Discrepancy (MMD) by Gretton and others have led to a number of advances, but MMD suffers from known deficiencies when data is concentrated near low dimensional structure, when different regions of data are at different scales, and when being applied to large scale tests as each test has $O(N^2)$ memory and computation requirements.

In the paper [A-I-16], we provide an anisotropic version of kernel MMD that accounts for all of these previous issues. In particular, when the distributions are locally low-dimensional, the proposed test can be made more powerful to distinguish certain alternatives by incorporating local covariance matrices and constructing an anisotropic kernel. The kernel matrix is also asymmetric; it computes the affinity between n data points and a set of reference points, where the number of reference points can be drastically smaller than n. This same set of reference points can be used for each pairwise test to compare K distributions, which significantly reduces the computation and memory requirements. The consistency of the test is proved, under mild assumptions of the kernel, and a finite-sample lower bound of the testing power is obtained. We also demonstrate applications to flow cytometry and diffusion MRI data, providing a novel unsupervised cluster analysis of large patient databases.

In the paper [A-I-17], we introduce a theoretical analysis of the local statistic (witness function) that is generated in kernel MMD, which builds on the work of [A-I-16]. This provides a method of detecting, with a high degree of certainty, where each distribution dominates. We also introduce a non-positive kernel for such tests and prove optimal and local approximation results in a supremum norm in a probabilistic sense. Together with a permutation test developed with the same kernel, we prove that the kernel converges to the true witness function in classification problems. This approach can be used to modify pretrained algorithms, such as neural networks or nonlinear dimension reduction techniques, to identify in-class vs out-of-class regions for the purposes of generative models, classification uncertainty, or finding robust centroids.

In the paper [A-I-15], we take a slightly different perspective by measuring pairwise distances using data adaptive, fast optimal transport. We use a co-clustering diffusion metric to learn the
underlying distribution of people, and **introduce an approximate earth mover’s distance algorithm using this data adaptive transportation cost.** As with anisotropic MMD, this approximate EMD is done through embedding each distribution into a low-dimensional space and computing the pairwise distances in that lower dimensional space, which leads to **significant computational cost reduction when comparing a large number of distributions.** The application we consider in this paper is a large scale high dimensional survey of people living in the US, and the question of how similar or different are the various counties in which these people live. This provides a mathematical tool for **public health experts to examine temporal changes in the similarities and differences between counties.**

2) Generative models of data have well known issues of posterior mismatch and mode collapse during training, in which the generated distribution fails to match the moments of the training data. To avoid this, the goal of a decoder/generator should be to approximate a homeomorphism between the data distribution and the sampling space, which requires choice of a suitable prior that matches the data topology. However, this is not the case for most generative models on a variety of manifold supported data sets. The paper [A-IV-10], inspired by manifold learning literature, **constructs the first generative neural network that automatically infers the data topology while preserving the scalability of deep models.** This creates a generative model that guarantees that small steps in the latent space result in small steps in the data space, and **provides approximation theoretic results for the dimension dependence of our proposed method.** [A-IV-10] also demonstrates **state-of-the-art generative performance** on a number of image and video **datasets known to be difficult for traditional generative models.**

3) The paper [A-IV-8] constructs a greedy algorithm for subsampling a point cloud or graph in order to find a small number of well-separated points. The algorithm also provides dynamically chosen weights so that, given function values at these small number of points, one can well approximate the mean of the function over the entire data set. [A-IV-8] bounds this quadrature error in terms of the points and weights, and **provides the first rate of convergence results** for the algorithmically chosen points and weights. [A-IV-8] also introduces a modification that can incorporate a non-uniform cost of sampling for the points, and demonstrates experimentally **state-of-the-art active selection of points for graph neural network semi-supervised learning.** These points are crucial for a number of applications, and are being incorporated into [A-IV-10] and [A-I-16] to construct more memory efficient models.

4) Eigenfunctions of Laplacians on graphs and manifolds have been of significant importance in applied mathematics and data science, as they retain geometric information about the domain. The “dual” geometry of Laplacian eigenfunctions, which describes the relationship between these eigenfunctions, is well understood on the torus and on the reals. The dual geometry is of tremendous role in various fields of pure and applied mathematics. However, on all other structures, not much is known about the relationship between these eigenfunctions other than the relative gap between their associated eigenvalues.

The paper [A-I-14] introduces a notion of similarity between eigenfunctions that allows one to reconstruct that geometry in very general settings. Our measure is based on a global average of local correlations, and can be computed in $O(N^2)$ time for each pair of eigenfunctions. This notion recovers all classical notions of duality but is equally applicable to other (rough) geometries and graphs. We provide many numerical examples in different continuous and discrete settings illustrate the result.

The importance of this result is two-fold. First, having a computational method of capturing the dual geometry allows one to build **anisotropic spectral filters,** and opens a rich new class of spectral graph theoretic data analysis. Second, discovering the dual geometric structure of the domain allows for **discovery of interesting structures in the eigenspace that are either not obvious or even previously unknown.** This includes the discovery that low-frequency eigenfunctions of the unnormalized Laplacian on random graphs have smaller 1-norm than the
vast majority, a phenomenon not present in the normalized Laplacian and a previously unknown phenomenon of the spectrum.

5) While neural networks have led to a number of advancements in machine learning, they are difficult to implement on low-memory chips due to the large number of parameters and matrix multiplication. Because of this, there is a lot of interest in compression of networks to reduce the number of non-zero parameters. Similarly, finding the “important” parameters through compression can lead to better understanding of the relevant features for classification.

In the paper [A-IV-9], we introduce a novel method of quantizing networks in parallel by quantizing each layer independently while maintaining minimal accuracy loss. This is demonstrated experimentally, and we also provide deterministic bounds on the misclassification error of the resulting quantized network. The bounds come from an argument that each layer is a Lipschitz function with controllable Lipschitz constant, and we use a classification margin argument to bound the total error.

In the paper [A-IV-7], we take a different perspective of quantization and seek to build a validation metric to measure the quality of the compressed network, specifically in the context of generative models where there is not a clear measure of error. We use permutation testing combined with MMD (with various underlying metrics of similarity) to create a hypothesis test of whether the generated images are statistically from the same distribution as the testing data. Using the permutation test provides a bar of acceptance for early stopping of training/compression in order to avoid common generative model issues, such as mode collapse. We also demonstrate that several state-of-the-art generative models still statistically deviate from the distribution of the data on which they are trained.

Additional: An additional product of major research, [B-IV-1], is a patent for methods of analysis of vastly subsampled NMR relaxometry data. Collecting partial relaxometry data results in significant speed up of testing time, and the advancements in the patent result in equivalent levels of reconstruction error using this partial data as compared to the full data. The patent mathematically pertains to inverting a subsampled 2D Fredholm integral of the first kind.
Bibliography

