Machine Learning Algorithms for Matching Theories, Simulations, and Observations in Cosmology

DOE Award Number: DE-SC0011114

Final Project Report

PI Barnabas Poczos
Carnegie Mellon University
Machine Learning Department

Brief description of the project

The search for simple and elegant equations to explain and predict observable natural phenomena has been a driving force in scientific theory for several centuries. The great successes of Newton's laws, Maxwell's equations, the Schrodinger equation, and the theory of General Relativity have encouraged this perspective, and perhaps a beautiful equation comprising a “theory of everything” remains to be discovered. But now scientists are grappling with systems of ever-increasing complexity, fraught with nonlinearity, and with the quantities of interest often only indirectly observed. *Nature may resist a simple description, and the most important discoveries of the next century may be complex theories with countless variables and parameters*.

The era of big data opens up a promising new approach to scientific discovery in this setting. The predictions of modern theories, even if complex or nonlinear, can be examined through detailed and computationally-intensive simulations that take days or weeks to run and fill petabytes of disk space.
The upcoming generation of scientific instruments will provide petabytes of observations at unprecedented resolution and depth. The challenge is to compare data to simulations so as to test theories and identify those that best match the observations. The data are numerous and rich but are subject to noise and various systematic biases. The simulations are large and detailed but are so costly that the effective sample size is small. Moreover, the mapping from data to simulation to the parameters of interest is typically complicated and ill-conditioned.

Although these challenges, and the solutions we propose, apply across the sciences, we will ground our efforts by focusing on cosmology. Here, we are truly entering the era of big data in several senses. Ever larger experiments survey larger and larger cosmological volumes, resulting in enormous amounts of pixel data to process and extract the information of cosmological interest. These include large imaging surveys, which collect all the light in broad wavelength ranges over significant fractions of the sky, and spectroscopic surveys, which measure the light as a function of wavelength to map out 3D distances. These experiments are designed to shed light on the big unanswered questions in cosmology. Distilling the petabytes of data into observed quantities for comparison with theory is a daunting task. But these large experiments do not only result in large amounts of data to analyze: if we wish to compare to theories of physics, then we need increasingly extensive predictions for the observed quantities as well, which requires ever-larger cosmological simulations that take substantial -- and at times prohibitive -- computational resources to produce.

Our goal is to develop statistical and machine learning methods for using observed and simulated data to advance machine learning with applications cosmology. In particular, we focus our research efforts on the following tasks:

- The challenge for automated science is that it is computationally impossible and statistically dangerous to consider every possible model in order to find the best one. We will *develop **Bayesian Optimization based active-learning methods*** that accelerate both the execution of the simulations and the search for best-fitting parameters. The key idea is to make the simulations adaptive -- across resolutions, time, and parameters -- using the data to search as the simulation runs.

- Most ML methods operate on simple finite dimensional feature vectors. However, many cosmology and other science applications require ML methods that can operate on more complex objects as ***inputs or outputs such as functions, distributions, or sets and point clouds*** (Ntampaka et al., 2015a, 2015b). Our goal is to develop efficient ML methods for this problem and demonstrate their applicability in Cosmology, Astrophysics, and other science problems.

DOE Relevance. The goal of this project is to make fundamental contributions in machine learning, statistics, and cosmology. Our scientific focus on cosmology combines massive data, complex theories, and important open questions. We analyze existing simulation and observational data sets as well as new ones that become available during the project. This will produce scientific advances in machine learning and cosmology and demonstrate the value of our methods for use in other data and simulation-intensive sciences of importance to DOE.

Accomplishments

In this section we list our publications and accomplishments and provide a short description of them.

1. M. Ravanbakhsh, J. Schneider, and B. Póczos. "**Deep Learning with Sets and Point Clouds**". International Conference on Learning Representations (ICLR) – workshop track. Toulon, France, 2017. (Ravanbakhsh et al 2017a)

2. S. Ravanbakhsh, J. Schneider, and B. Póczos. "**Equivariance Through Parameter-Sharing**". International Conference on Machine Learning (ICML). Sydney, Australia, 2017. (25% acceptance rate). (Ravanbakhsh et al 2017b)

3. M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. Smola. "**Deep Sets**". Proceedings of the Neural Information Processing Systems (NIPS). Long Beach , CA, 2017. (Accepted for oral presentation, 1.23% acceptance rate). (Zaheer et al 2017)

   In the above listed three papers we study the problem of designing models for machine learning tasks defined on sets / point clouds. In contrast to traditional approach of operating on fixed dimensional vectors, we consider objective functions defined on sets that are invariant to permutations. Such problems are widespread, ranging from estimation of population statistics (poczos13aistats), to anomaly detection in piezometer data of embankment dams (Jung15Exploration), to cosmology (Ntampaka et al. 2015a, Ntampaka et al. 2015b, Ravanbakhsh 2016). Our main theorem characterizes the permutation invariant functions and provides a family of functions to which any permutation invariant objective function must belong. This family of functions has a special structure which enables us to design a deep network architecture that can operate on sets and which can be deployed on a variety of scenarios including both unsupervised and supervised learning tasks. We also derive the necessary and sufficient conditions for permutation equivariance in deep models. We demonstrate the applicability of our method on population statistic estimation, point cloud classification, set expansion, set-outlier detection, and semi-supervised learning with clustering side-information outlier detection.

4. S. Ravanbakhsh, F. Lanusse, R. Mandelbaum, J. Schneider, and Póczos. "**Enabling Dark Energy Science with Deep Generative Models of Galaxy Images**". Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). San Francisco, CA, 2017. (24.6% acceptance rate). (Ravanbakhsh et al 2017c)

   Understanding the nature of dark energy, the mysterious force driving the accelerated expansion of the Universe, is a major challenge of modern cosmology. The next generation of cosmological surveys, specifically designed to address this issue, rely on accurate measurements of the apparent shapes of distant galaxies. However, shape measurement methods suffer from various unavoidable biases and therefore will rely on a precise calibration to meet the accuracy requirements of the science analysis. This calibration process remains an open challenge as it requires large sets of high quality galaxy images. To this end, we study the application of deep conditional generative models in generating realistic galaxy images. In particular we consider variations on conditional variational autoencoder and introduce a new adversarial objective for training of conditional generative networks. Our results suggest a reliable alternative to the acquisition of expensive high quality observations for generating the calibration data needed by the next generation of cosmological surveys.

5. K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos. "**Asynchronous Parallel Bayesian Optimisation via Thompson Sampling**". AutoML workshop, ICML 2017, 2017. (Kandasamy et al 2017)

   We design and analyse variations of the classical Thompson sampling (TS) procedure for Bayesian optimisation (BO) in settings where function evaluations are expensive, but can be performed in parallel. Our theoretical analysis shows that a direct application of the sequential Thompson sampling algorithm in either synchronous or asynchronous parallel settings yields a surprisingly powerful result: making n evaluations distributed among M workers is essentially equivalent to performing n evaluations in sequence. Further, by modeling the time taken to complete a function evaluation, we show that, under a time constraint, asynchronously parallel TS achieves asymptotically lower regret than both the synchronous and sequential versions. These results are complemented by an experimental analysis, showing that asynchronous TS outperforms a suite of existing parallel BO algorithms in simulations and in a hyper-parameter tuning application in convolutional neural networks. In addition to these, the proposed procedure is conceptually and computationally much simpler than existing work for parallel BO.

6. F. Lanusse, Q. Ma, N. Li, T. Collett, C. Li, S. Ravanbakhsh, R. Mandelbaum, and B. Póczos. **"CMU DeepLens: Deep Learning For Automatic Image-based Galaxy-Galaxy Strong Lens Finding"**. Monthly Notices of the Royal Astronomical Society Main Journal (MNRAS), 2017. (Lanusse et al 2017)

   Galaxy-scale strong gravitational lensing is not only a valuable probe of the dark matter distribution of massive galaxies, but can also provide valuable cosmological constraints, either by studying the population of strong lenses or by measuring time delays in lensed quasars. Due to the rarity of galaxy-scale strongly lensed systems, fast and reliable automated lens finding methods will be essential in the era of large surveys such as LSST, Euclid, and WFIRST. To tackle this challenge, we introduce CMU DeepLens, a new fully automated galaxy-galaxy lens finding method based on Deep Learning. This supervised machine learning approach does not require any tuning after the training step which only requires realistic image simulations of strongly lensed systems. We train and validate our model on a set of 20,000 LSST-like mock observations including a range of lensed systems of various sizes and signal-to-noise ratios (S/N). We find on our simulated data set that for a rejection rate of non-lenses of 99%, a completeness of 90% can be achieved for lenses with Einstein radii larger than 1.4" and S/N larger than 20 on individual g-band LSST exposures. Finally, we emphasize the importance of realistically complex simulations for training such machine learning methods by demonstrating that the performance of models of significantly different complexities cannot be distinguished on simpler simulations.

7. K. Kandasamy, J. Schneider, and B. Póczos. "**Query Efficient Posterior Estimation in Scientific Experiments via Bayesian Active Learning**". Artificial Intelligence Journal, 2016 (Kandasamy et al 2016a)

   A common problem in disciplines of applied Statistics research such as Astrostatistics is of estimating the posterior distribution of relevant parameters. Typically, the likelihoods for such models are computed via expensive experiments such as cosmological simulations of the universe. An urgent challenge in these research domains is to develop methods that can estimate the posterior with few likelihood evaluations. In this paper, we study active posterior estimation in a Bayesian setting when the likelihood is expensive to evaluate. Existing techniques for posterior estimation are based on generating samples representative of the posterior. Such methods do not consider efficiency in terms of likelihood evaluations. In order to be query efficient we treat posterior estimation in an active regression framework. We propose two myopic query strategies to choose where to evaluate the likelihood and implement them using Gaussian processes. Via experiments on a series of synthetic and real examples we demonstrate that our approach is significantly more query efficient than existing techniques and other heuristics for posterior estimation.

8. Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. "**Learning Theory for Distribution Regression"**. Journal of Machine Learning Research (JMLR), 2016. (Szabo et al 2016)

   We focus on the distribution regression problem: regressing to vector-valued outputs from probability measures. Many important machine learning and statistical tasks fit into this framework, including multi-instance learning and point estimation problems without analytical solution (such as hyperparameter or entropy estimation). Despite the large number of available heuristics in the literature, the inherent two-stage sampled nature of the problem makes the theoretical analysis quite challenging, since in practice only samples from sampled distributions are observable, and the estimates have to rely on similarities computed between sets of points. To the best of our knowledge, the only existing technique with consistency guarantees for distribution regression requires kernel density estimation as an intermediate step (which often performs poorly in practice), and the domain of the distributions to be compact Euclidean. In this paper, we study a simple, analytically computable, ridge regression-based alternative to distribution regression, where we embed the distributions to a reproducing kernel Hilbert space, and learn the regressor from the embeddings to the outputs. Our main contribution is to prove that this scheme is consistent in the two-stage sampled setup under mild conditions (on separable topological domains enriched with kernels): we present an exact computational-statistical efficiency trade-off analysis showing that our estimator is able to match the one-stage sampled minimax optimal rate [Caponnetto and De Vito, 2007; Steinwart et al., 2009]. This result answers a 17-year-old open question, establishing the consistency of the classical set kernel [Haussler, 1999; Gaertner et. al, 2002] in regression. We also cover consistency for more recent kernels on distributions, including those due to [Christmann and Steinwart, 2010].

9. K. Kandasamy, G. Dasarathy, B. Póczos, and J. Schneider. "**The Multi-fidelity Multi-armed Bandit**". Proceedings of the Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016. (23% acceptance rate). (Kandasamy et al 2016b)

   We study a variant of the classical stochastic K-armed bandit where observing the outcome of each arm is expensive, but cheap approximations to this outcome are available. For example, in online advertising the performance of an ad can be approximated by displaying it for shorter time periods or to narrower audiences. We formalise this task as a multi-fidelity bandit, where, at each time step, the forecaster may choose to play an arm at any one of M fidelities. The highest fidelity (desired outcome) expends cost λ(m). The mth fidelity (an approximation) expends λ(m)<λ(M) and returns a biased estimate of the highest fidelity. We develop MF-UCB, a novel upper confidence bound procedure for this setting and prove that it naturally adapts to the sequence of available approximations and costs thus attaining better regret than naive strategies which ignore the approximations. For instance, in the above online advertising example, MF-UCB would use the lower fidelities to quickly eliminate suboptimal ads and reserve the larger expensive experiments on a small set of promising candidates. We complement this result with a lower bound and show that MF-UCB is nearly optimal under certain conditions.

10. K. Kandasamy, G. Dasarathy, J. Oliva, J. Schneider, and B. Póczos. "**Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations**". Proceedings of the Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016. (23% acceptance rate). (Kandasamy et al 2016c)

    In many scientific and engineering applications, we are tasked with the optimisation of an expensive to evaluate black box function f. Traditional methods for this problem assume just the availability of this single function. However, in many cases, cheap approximations to f may be obtainable. For example, the expensive real world behaviour of a robot can be approximated by a cheap computer simulation. We can use these approximations to eliminate low function value regions cheaply and use the expensive evaluations of f in a small but promising region and speedily identify the optimum. We formalise this task as a multi-fidelity bandit problem where the target function and its approximations are sampled from a Gaussian process. We develop MF-GP-UCB, a novel method based on upper confidence bound techniques. In our theoretical analysis we demonstrate that it exhibits precisely the above behaviour, and achieves better regret than strategies which ignore multi-fidelity information. MF-GP-UCB outperforms such naive strategies and other multi-fidelity methods on several synthetic and real experiments.

11. A. Tallavajhula, A. Kelly, and B. Póczos. **"Nonparametric Distribution Regression Applied to Sensor Modeling"**. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Korea. (Accepted for oral presentation). (Tallavajhula et al 2016)

    Sensor models, which specify the distribution of sensor observations, are a widely used and integral part of robotics algorithms. Observation distributions are commonly approximated by parametric models, which are limited in their expressiveness, and may require careful design to suit an application. In this paper, we propose nonparametric distribution regression as a procedure to model sensors. It is a data-driven procedure to predict distributions that makes few assumptions. We apply the procedure to model raw distributions from real sensors and demonstrate its utility to a mobile robot state estimation task. We show that nonparametric distribution regression adapts to characteristics in the training data, leading to realistic predictions. The same procedure competes favorably with baseline parametric models across applications. The results also help develop intuition for different sensor modeling situations. Our procedure is useful when distributions are inherently noisy and sufficient data is available.

12. S. Ravanbakhsh, J. Oliva, S. Fromenteau, L. Price, S. Ho, J. Schneider, and B. Póczos. **"Estimating Cosmological Parameters from the Dark Matter Distribution"**. International Conference on Machine Learning (ICML). NYC, NY, 2016. (24% acceptance rate). (Ravanbakhsh et al 2016)

    A grand challenge of the 21st century cosmology is to accurately estimate the cosmological parameters of our Universe. A major approach in estimating the cosmological parameters is to use the large scale matter distribution of the Universe. Galaxy surveys provide the means to map out cosmic large-scale structure in three dimensions. Information about galaxy locations is typically summarized in a "single" function of scale, such as the galaxy correlation function or powerspectrum. We show that it is possible to estimate these cosmological parameters directly from the distribution of matter. This paper presents the application of deep 3D convolutional networks to volumetric representation of dark-matter simulations as well as the results obtained using a recently proposed distribution regression framework, showing that machine learning techniques are comparable to, and can sometimes outperform, maximum-likelihood point estimates using "cosmological models". This opens the way to estimating the parameters of our Universe with higher accuracy.

13. J. Oliva, A. Dubey, A. Wilson, B. Póczos, J. Schneider, and E. Xing. "**Bayesian Nonparametric Kernel-Learning**". International Conference on Artificial Intelligence and Statistics (AISTATS). Cadiz, Spain, 2016. (30% acceptance rate) (Oliva et al 2016)

    Kernel methods are ubiquitous tools in machine learning. However, there is often little reason for the common practice of selecting a kernel a priori. Even if a universal approximating kernel is selected, the quality of the finite sample estimator may be greatly affected by the choice of kernel. Furthermore, when directly applying kernel methods, one typically needs to compute a N ×N Gram matrix of pairwise kernel evaluations to work with a dataset of N instances. The computation of this Gram matrix precludes the direct application of kernel methods on large datasets, and makes kernel learning especially difficult. In this paper we introduce Bayesian nonparmetric kernel-learning (BaNK), a generic, data-driven framework for scalable learning of kernels. BaNK places a nonparametric prior on the spectral distribution of random frequencies allowing it to both learn kernels and scale to large datasets. We show that this framework can be used for large scale regression and classification tasks. Furthermore, we show that BaNK outperforms several other scalable approaches for kernel learning on a variety of real world datasets.

14. D. Sutherland, J. Oliva, B. Póczos, and J. Schneider. "**Linear-time Learning on Distributions with Approximate Kernel Embeddings**". 30th AAAI Conference on Artificial Intelligence (AAAI-16). Phoenix, AZ, 2016. (26% acceptance rate). (Sutherland et al 2016)

    Many interesting machine learning problems are best posed by considering instances that are distributions, or sample sets drawn from distributions. Previous work devoted to machine learning tasks with distributional inputs has done so through pairwise kernel evaluations between pdfs (or sample sets). While such an approach is fine for smaller datasets, the computation of an NxN Gram matrix is prohibitive in large datasets. Recent scalable estimators that work over pdfs have done so only with kernels that use Euclidean metrics, like the L2 distance. However, there are a myriad of other useful metrics available, such as total variation, Hellinger distance, and the Jensen-Shannon divergence. This work develops the first random features for pdfs whose dot product approximates kernels using these non-Euclidean metrics, allowing estimators using such kernels to scale to large datasets by working in a primal space, without computing large Gram matrices. We provide an analysis of the approximation error in using our proposed random features and show empirically the quality of our approximation both in estimating a Gram matrix and in solving learning tasks in real-world and synthetic data.

15. M. Ntampaka, H. Trac, D.J. Sutherland, N. Battaglia, B. Poczos, & J. Schneider, **"A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters"**, 2015, *Astrophysical Journal, 803, 2, http://iopscience.iop.org/article/10.1088/0004-637X/803/2/50/pdf,* (Ntampaka et al. 2015a)

16. M. Ntampaka, H. Trac, D.J. Sutherland, S. Fromenteau, B. Poczos, & J. Schneider, **Dynamical Mass Measurements of Contaminated Galaxy Clusters using Machine Learning**, The Astrophysical Journal*, http://arxiv.org/abs/1509.05409,* (Ntampaka et al. 2015b).,

    In the Ntampaka et al. (2015a) and Ntampaka et al. (2015b) papers we use Support Distribution Machines (SDM; Sutherland et al. 2012) to improve dynamical mass measurements of galaxy clusters. SDM is a new machine learning classification method developed by our team. Conventionally, a standard power-law scaling relation is used to infer cluster mass from line-of-sight velocity dispersion. When applied to a mock cluster catalog, the resulting fractional mass error distribution is very broad and has extended high-error tails. The problem with this method is that much useful information is thrown away. With SDM, we train and test on the entire distribution of galaxy velocities. We demonstrate that it can reduce the width of the error distribution by a factor of two and effectively eliminate the problematic high-error tails. Remarkable, we show that SDM applied to realistic clusters contaminated by interloper galaxies is better able to recover masses than even the scaling relation approach applied to idealistic clusters without interlopers. Decreasing cluster mass errors will improve measurements of the growth of structure and lead to tighter constraints on cosmological parameters.

17. Chen, Y., Ho, S., Brinkman, J., Freeman, P., Genovese, C., Schneider, D., Wasserman, L., **"Cosmic Web Reconstruction through density ridges: Catalog"**, arxiv:1509.06443, (Chen et al. 2015a)

    In Chen et al. (2015a) we construct a catalogue for filaments using a novel approach called subspace constrained mean shift (SCMS, Ozertem & Erdogmus 2011). SCMS is a gradient-based method that detects filaments through density ridges (smooth curves tracing high-density regions). A great advantage of SCMS is its uncertainty measure, which allows an evaluation of the errors for the detected filaments. To detect filaments, we use data from the Sloan Digital Sky Survey, which consist of three galaxy samples: the NYU main galaxy sample (MGS), the LOWZ sample and the CMASS sample. Each of the three dataset covers different redshift regions so that the combined sample allows detection of filaments up to z = 0.7. Our filament catalogue consists of a sequence of two-dimensional filament maps at different redshifts that provide several useful statistics on the evolution cosmic web. To construct the maps, we select spectroscopically confirmed galaxies within $0.050 < z < 0.700$ and partition them into 130 bins. For each bin, we ignore the redshift, treating the galaxy observations as a 2-D data and detect filaments using

SCMS. The filament catalogue consists of 130 individual 2-D filament maps, and each map comprises points on the detected filaments that describe the filamentary structures at a particular redshift. We also apply our filament catalogue to investigate galaxy luminosity and its relation with distance to filament. Using a volume-limited sample, we find strong evidence (6.1σ - 12.3σ) that galaxies close to filaments are generally brighter than those at significant distance from filaments.

18. Chen, Y., Ho, S., Mandelbaum, R., Bahcall, N., Brownstein J., Freeman, P., Genovese, C., Schneider, D., Wasserman, L., **"Detecting Effects of Filaments on Galaxy Properties in the Sloan Digital Sky Survey III"**, arxiv.1509:06376, (Chen et al. 2015b)

    In Chen et al. (2015b) we study the effects of filaments on galaxy properties in the Sloan Digital Sky Survey (SDSS) Data Release 12 using filaments from the "Cosmic Web Reconstruction" catalogue (Chen et al. 2015a), a publicly available filament catalogue for SDSS. Since filaments are tracers of medium-to-high density regions, we expect that galaxy properties associated with the environment are dependent on the distance to the nearest filament. Our analysis demonstrates a red galaxy or a high-mass galaxy tend to reside closer to filaments than a blue or low-mass galaxy. After adjusting the effect from stellar mass, on average, late-forming galaxies or large galaxies have a shorter distance to filaments than early-forming galaxies or small galaxies. For the Main galaxy sample, all signals are very significant (>5σ). For the LOWZ and CMASS samples, most of the signals are significant (with >3σ). The filament effects we observe persist until z = 0.7 (the edge of the CMASS sample). Comparing our results to those using the galaxy distances from redMaPPer galaxy clusters as a reference, we find a similar result between filaments and clusters. Our findings illustrate the strong correlation of galaxy properties with proximity to density ridges, strongly supporting the claim that density ridges are good tracers of filaments.

19. Chen, Y., Ho, S., Tenneti, A.,Mandelbaum, R., Freeman, P., Croft, R., DiMatteo, T., Genovese, C., Wasserman, L., **"Investigating Galaxy-Filament Alignments in Hydrodynamic Simulations using Density Ridges"**, 2015, MNRAS, 454, 3341C, (Chen et al. 2015c) http://adsabs.harvard.edu/abs/2015MNRAS.454.3341C:

    In Chen et al. (2015c), we study the filamentary structures and the galaxy alignment along filaments at redshift z=0.06 in the MassiveBlack-II simulation, a state-of-the-art, high-resolution hydrodynamical cosmological simulation which includes stellar and AGN feedback in a volume of (100 Mpc/h)^3. The filaments are constructed using the subspace constrained mean shift (SCMS; Ozertem & Erdogmus (2011) and Chen et al. (2015a)). First, we show that reconstructed filaments using galaxies and reconstructed filaments using dark matter particles are similar to each other; over 50% of the points on the galaxy filaments have a corresponding point on the dark matter filaments within distance 0.13 Mpc/h (and vice versa) and this distance is even smaller at high-density regions. Second, we observe the alignment of the major principal axis of a galaxy with respect to the orientation of its nearest filament and detect a 2.5 Mpc/h critical radius for filament's influence on the alignment when the subhalo mass of this galaxy is between 109M⨀/h and 1012M⨀/h. Moreover, we find the alignment signal to increase significantly with the subhalo mass. Third, when a galaxy is close to filaments (less than 0.25 Mpc/h), the galaxy alignment toward the nearest galaxy group depends on the galaxy subhalo mass. Finally, we find that galaxies close to filaments or groups tend to be rounder than those away from filaments or groups.

20. Chen, Y., Ho, S,. Genovese, C., Wasserman, L., **"Optimal Ridge Detection Using Coverage Risk"**, Neural Information Processing Systems (NIPS) 2015, [Acceptance Rate for the

conference is 18%], (Chen et al. 2015d)

In Chen et al. (2015d), we introduce the concept of coverage risk as an error measure for density ridge estimation. The coverage risk generalizes the mean integrated square error to set estimation. We propose two risk estimators for the coverage risk and we show that we can select tuning parameters by minimizing the estimated risk. We study the rate of convergence for coverage risk and prove consistency of the risk estimators. We apply our method to three simulated datasets and to cosmology data. In all the examples, our proposed method successfully recovers the underlying density structure.

21. Chen, Y., Ho, S., Freeman, P., Genovese, C., Wasserman, L., **"Cosmic Web Reconstruction through Density Ridges: Method and Algorithm"** , arXiv:1501.05303, (Chen et al. 2015e) http://adsabs.harvard.edu/abs/2015MNRAS.454.1140C

    In Chen et al. (2015e), we demonstrate how one may apply the SCMS algorithm (Ozertem and Erdogmus (2011); Genovese et al. (2012)) to uncover filamentary structure in galaxy data. The detection and characterization of filamentary structures in the cosmic web allows cosmologists to constrain parameters that dictate the evolution of the Universe. While many filament estimators have been proposed, they generally lack estimates of uncertainty, reducing their inferential power. The SCMS algorithm is a gradient ascent method that models filaments as density ridges, one-dimensional smooth curves that trace high-density regions within the point cloud. We also demonstrate how augmenting the SCMS algorithm with bootstrap-based methods of uncertainty estimation allows one to place uncertainty bands around putative filaments. We apply the SCMS method to datasets sampled from the P3M N-body simulation, with galaxy number densities consistent with SDSS and WFIRST-AFTA and to LOWZ and CMASS data from the Baryon Oscillation Spectroscopic Survey (BOSS). To further assess the efficacy of SCMS, we compare the relative locations of BOSS filaments with galaxy clusters in the redMaPPer catalog, and find that redMaPPer clusters are significantly closer (with p-values <10−9) to SCMS-detected filaments than to randomly selected galaxies.

22. Oliva, J., Neiswanger, W., Poczos, B., Xing, E., Trac, H., Ho, S., & Schneider, J., **"Fast Function to Function Regression",** Artificial Intelligence & Statistics 2015 [Acceptance Rate for oral presentation= 26/441= 5.9%]. (Oliva et al, 2015)

    In Oliva et al. (2015) we analyze the problem of regression when both input covariates and output responses are functions from a nonparametric function class. Function to function regression (FFR) covers a large range of interesting applications including time-series prediction problems, and also more general tasks like studying a mapping between two separate types of distributions. However, previous nonparametric estimators for FFR type problems scale badly computationally with the number of input/output pairs in a data-set. Given the complexity of a mapping between general functions it may be necessary to consider large data-sets in order to achieve a low estimation risk. To address this issue, we develop a novel scalable nonparametric estimator, the Triple-Basis Estimator (3BE), which is capable of operating over datasets with many instances. To the best of our knowledge, the 3BE is the first nonparametric FFR estimator that can scale to massive datasets. We analyze the 3BE's risk and derive an upperbound rate. Furthermore, we show an improvement of several orders of magnitude in terms of prediction speed and a reduction in error over previous estimators in cosmological N-body simulations and various real-world data-sets.

23. Garnett, R., Ho, S. & Schneider, J., **"Finding Galaxies in the Shadows of Quasars with Gaussian Processes",** International Conference on Machine Learning, 2015, [Acceptance Rate ~25%] (Garnett et al, 2015)

    In Garnett et al. (2015), we develop an automated technique for detecting damped Lyman-α absorbers (DLAs) along spectroscopic sightlines to quasi-stellar objects (QSOs or quasars). The detection of DLAs in large-scale spectroscopic surveys such as SDSS–III is critical to address outstanding cosmological questions, such as the nature of galaxy formation. We use nearly 50,000 QSO spectra to learn a tailored Gaussian process model for quasar emission spectra, which we apply to the DLA detection problem via Bayesian model selection. We demonstrate our method's effectiveness with a large-scale validation experiment on over 100,000 spectra, with excellent performance.

24. Kandasamy, K., Schneider, J., and Poczos, B., **"Bayesian Active Learning for Posterior Estimation"** *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015. (Buenos Aires, Argentina) [Distinguished Paper Award, 0.1% acceptance rate] (Kandasamy et al, 2015a)

    The Kandasamy et al. (2015a) paper studies active posterior estimation in a Bayesian setting when the likelihood is expensive to evaluate. Existing techniques for posterior estimation are based on generating samples representative of the posterior. Such methods do not consider efficiency in terms of likelihood evaluations. In order to be query efficient we treat posterior estimation in an active regression framework. We propose two myopic query strategies to choose where to evaluate the likelihood and implement them using Gaussian processes. Via experiments on a series of synthetic and real examples we demonstrate that our approach is significantly more query efficient than existing techniques and other heuristics for posterior estimation.

25. Kandasamy, K., Schneider, J., and Poczos, B. **"High Dimensional Bayesian Optimization and Bandits via Additive Models"**. International Conference on Machine Learning (ICML), 2015. (Lille, France) [26% acceptance rate], (Kandasamy et al, 2015b)

    Bayesian Optimisation (BO) is a technique used in optimizing a D-dimensional blackbox objective function which is typically expensive to evaluate. While there have been many successes for BO in low dimensions, scaling it to high dimensions has been notoriously difficult. Existing literature on the topic are under very restrictive settings. In this paper, we identify two key challenges in this endeavor. In the Kandasamy et al. (2015b) paper we tackle these challenges by assuming an additive structure for the function. This setting is substantially more expressive and contains a richer class of functions than previous work. We prove that, for additive functions the regret has only linear dependence on D even though the function depends on all D dimensions. We also demonstrate several other statistical and computational benefits in our framework. Via synthetic examples, a cosmology simulation, and a face detection problem we demonstrate that our method outperforms naive BO on additive functions and on several examples where the function is not additive.

Software artifacts

1. **Dragonfly** [https://github.com/dragonfly/dragonfly]
    This is an open source python library for scalable Bayesian optimisation. Bayesian optimisation is used for optimising black-box functions whose evaluations are usually expensive. Beyond vanilla

optimisation techniques, Dragonfly provides an array of tools to scale up Bayesian optimisation to expensive large-scale problems. These include features/functionality that are especially suited for high dimensional optimisation (optimising for a large number of variables), parallel evaluations in synchronous or asynchronous settings (conducting multiple evaluations in parallel), multi-fidelity optimisation (using cheap approximations to speed up the optimisation process), and multi-objective optimisation (optimising multiple functions simultaneously).

2. **FuncLearn** [https://github.com/junieroliva/funcLearn]
   This is a matlab package for performing machine learning tasks when inputs, and possibly outputs, are functions, distributions, or sets.

3. **SKL-Groups** [https://github.com/dougalsutherland/skl-groups]
   This is a package to perform machine learning on sets (or "groups") of features in Python. It extends the scikit-learn library with support for either transforming sets into feature vectors that can be operated on with standard scikit-learn constructs or obtaining pairwise similarity/etc matrices that can be turned into kernels for use in scikit-learn.

4. **Py-SDM** [https://github.com/dougalsutherland/py-sdm/]
   This is a Python implementation of nonparametric divergence estimators.

5. **CMU DeepLens** [https://github.com/McWilliamsCenter/CMUDeepLens]
   CMU DeepLens, or DeepLens for short, is a completely automated strong lens finder based on Deep Residual Networks, a state of the art Deep Learning architecture for image detection and classification tasks.