TRIPODS+X:RES: Investigations at the Interface of Data Science and Neuroscience

Project Description

1. Introduction

2. Distributed Processing (Objective 1)

Recent work in machine learning has studied the effect of communication constraints and parallelization in distributed estimation. There is a close parallel in vision, where any given input is sensed by multiple parts of the retina and an accurate percept needs to be constructed. We will consider different models for distributed processing, motivated by learning and perception in both lower-level organisms (visual processing in fruit flies) and higher-level cognition (visual cognition in humans).

Neuroscience Background. Animals use visual cues to guide many behaviors, from navigation to foraging and courtship. The perception of these visual cues is an inference problem (**?**). In this problem, the animal obtains light intensity information from an array of photoreceptors focused on different points in space. The animal must combine these light intensity signals in a way that allows it to infer and respond the true state of the world, across many different parallel dimensions of inference. For instance, one dimension of inference might be the global motion of the visual scene, while another might be the existence or non-existence of a predator in the scene. The neuronal circuits in the visual system perform this inference task, at both low levels (is there an edge at this location and angle?) and at high levels (is that object a predator?) (**?**). The operational processing of many visual neurons and circuits have been studied in depth, but it is frequently unknown how these operational descriptions relate to the inferences that guide behavior. In particular, these inferences require integrating distributed retinal information over space and over time, but we do not know how this integration relates to the statistics of the natural world, to channels of information flow within the circuits, or to noise or incomplete information about the world.

The fruit fly *Drosophila* has several advantages for studying the distributed visual processing that guides perception and behavior. First, there is a powerful genetic toolbox in fruit flies that allows researchers to genetically define, manipulate, and monitor specific classes of neurons (**?**). Those manipulations also allow specific neurons to be causally connected to behaviors. Second, the fly has a wealth of robust visual behaviors, including regulation of turning and speed, escape responses, and courtship behaviors (**???**). Third, the field has identified neuron types that appear to be making exactly the inferences described above: neurons sensitive to local motion direction and speed (**?**); neurons sensitive to wide-field motion that corresponds to rotational self-motion of the fly about various axes (**?**); neurons sensitive to looming (approaching) dark dots (**??**); and neurons sensitive to moving small dots (**?**). In each of these cases, we can silence the neurons and observe behavioral deficits. We can also record neural activity in these individual neurons and measure their response properties with well-controlled visual stimuli (**?**). Thus, these neuron classes act as handles for understanding how visual inferences are made, and how neurons extract specific visual features from a spatiotemporally distributed set of inputs.

Computation and Inference Background. Classic statistical theory studies the difficulty of estimation under various models, and attempts to find the optimal estimation procedures. Such studies usually assume

that all of the collected data are available to construct the estimators. Recent research has begun to study the problem of statistical estimation with data residing at multiple machines. Estimation in distributed settings is becoming common in modern data analysis tasks, as the data can be collected or stored at different locations. In order to obtain an estimate of some statistical functional, information needs to be gathered and aggregated from the multiple locations to form the final estimate. However, the communication between machines may be limited. In such a setting, it is important to understand how the statistical risk of estimation degrades as the communication budget becomes more limited.

The so-called CEO problem, first studied in the electrical engineering community in the context of rate-distortion-theory theory, treats a similar distributed estimation problem (**??**). More recently, several studies have focused on more specific statistical tasks and models, including mean estimation, regression, principal eigenspace estimation, and discrete density estimation (**????????**). Most of this existing research focuses on parametric and discrete models, where the parameter of interest has finite dimension. In a nonparametric setting, the effective dimension of the problem typically grows with the sample size, and these results no longer apply. Other results have been obtained on these problems in the normal means model of nonparametric estimation. The normal means model arises naturally when representing an estimator in terms of an orthogonal basis (**??**). One result gives a sharp constrained minimax analysis of nonparametric regression under quantization constraints (**?**); another characterizes lower bounds and achievability for distributed nonparametric regression (**?**).

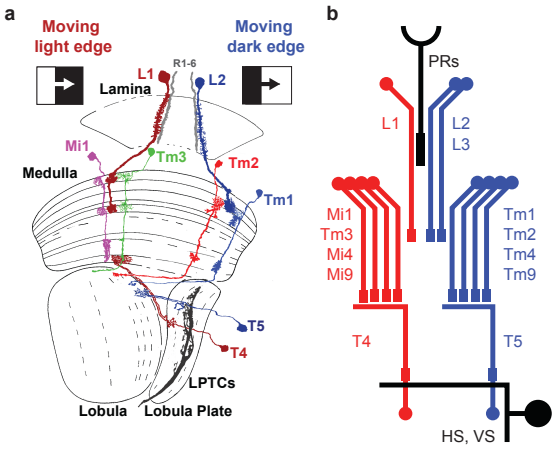

Figure 1: Motion circuits in the fly. (a) Light is detected at the retina (at the top of this diagram), and information is processed moving down through different neuropils. Each of the highlighted neurons is required for motion detection. Motion detection is split into one circuit that detects light edges moving over dark backgrounds and another that detects dark edges over light backgrounds. (b) Cartoon of neurons known to be involved in motion detection in the light edge (red) and dark edge (blue) pathways. There are four T4s and four T5s, selective for each of the four cardinal directions across the retina. This circuit repeats at all points in space.

Proposed Research A: Parallel channels for local motion detection. The fly's eye is arranged in a hexagonal lattice of repeated circuit motifs, with each column of circuitry representing one retinotopic point in visual space. Each eye consists of an array of roughly 800 of these pixels, which together cover approximately one half of visual space. Two classes of local motion detection cells exist in every column: there are T4 cells, which detect light edges moving across dark backgrounds, T5 cells, which detect dark edges moving across light backgrounds. There are 4 of each class, one for each cardinal direction. Thus, there are 8 parallel channels at each point in space representing motion in two dimensions. Why is the system organized in this way? How are naturalistic motion signals distributed across the 8 channels, and what encoding or decoding advantages does this serve? Under what conditions are the channels redundant? How would an optimal observer partition signals among these parallel channels? Could a data driven approach predict or give insight into this encoding scheme? One approach to begin studying these questions will be to adapt the classical framework of sparse coding (**?**) in a way that represents the neurobiology of the fly's visual system.

Proposed Research B: Detecting motion flow fields. When an animal moves or rotates in the world, its self-motion generates flow fields across its retina. These flow fields are often used as feedback to control orientation or speed. In *Drosophila*, some neurons downstream of the local motion detectors have large

receptive fields that integrate motion signals across the retina. They appear to be selective for specific flow fields, which correspond to the flow fields created by the rotation of the fly about different axes. These neural signals have been proposed to be linear filters, matching their weighting for local motion to specific optical flow fields ("matched filters"). However, it is not clear whether a linear weighting of local motion estimates represents a best estimate of each flow field, or whether more complex dendritic computations could improve encoding. In particular, it is not clear how these neurons might optimally integrate motion signals in the presence of occlusions or differential velocity fields that would be caused by fly translation through space. We will investigate these issues using methods based on hierarchical sparse coding (**?**) and related computational methods for low rank decomposition.

Figure 2: Global motion and loom may generate similar local motion signals, but local motion signals must be integrated over space to distinguish the two stimulus types.

Proposed Research C: Detecting looming stimuli. Looming stimuli are created by objects as they approach the observer: the object becomes larger, and if it is on a collision course, opposite edges of the object will move in opposite directions on the retina. Thus, detecting loom requires integrating information over space, as any local motion detector cannot know if local motion is due to a looming object or is just wide-field motion. Two neuron types in *Drosophila* have recently been described as loom detectors, responding selectively to objects that grow larger in their receptive fields. The receptive fields of these cells to local motion have been characterized, but the relationship between these receptive fields, the nonlinear computations of these neurons, and the statistics of natural loom stimuli remain unclear. This is in large part due to the absence of good statistics on natural loom stimuli. It is also possible that these loom detectors use features beyond motion to detect approaching objects; for instance, a detector might integrate motion signals with light intensity information, since light intensity is correlated with distance. Here, one might also ask whether the motion detecting neurons upstream of the loom-sensitive neurons could convey information about stimulus features beyond just motion direction and speed. We will abstract loom detection as a statistical testing problem, building work on such as (**???**). Specifically, for a given object geometry such as a disk, we will study minimax rates for loom detection, in terms of the noise level and sparity of the number of boundary neuron measurements required. Fast hierarchical algorithms to achieve the minimax rates will be studied.

3. Data Representation (Objective 2)

The brain persistently stores the same information in several different ways across regions, each emphasizing different dimensions of the input. In doing so, different features become readily accessible to the computations implemented in these regions. More generally, by representing information in different ways, similar learning rules are able to automatically aggregate experience along different dimensions and thereby extract different and complementary knowledge. This can be contrasted with machine learning, both by the simple fact that data are stored on disk in one format and that finding relevant dimensions is often the output of the learning algorithm rather a natural consequence of how the input is structured. We will develop algorithms for machine representation of data that are informed by such understanding of how the brain represents information in different systems and leverage developments in embedding algorithms in machine learning for advanced processing of fMRI and cellular recordings and the information they contain.

Neuroscience Background. Multiple brain systems represent inputs from the world and store this information in different memory registers. Although these representations are sometimes redundant (e.g., across hemispheres in the same brain system), they often emphasize different aspects of the input that are extracted

via different neural pathways or as a result of transformations across brain regions. In visual cortex, for example, different inputs with similar appearance (e.g., the faces of siblings, the Grand Canyon from different vistas, bananas in a grocery store, etc.) will be stored together. In frontotemporal cortex, different inputs with the same conceptual, semantic, or functional meaning will be stored together despite differences in appearance (e.g., pieces of clothing, cooking utensils in a drawer, animals from an ecosystem, etc.). In the ventral striatum and orbitofrontal cortex, different inputs with the same reward value will be represented similarly despite appearances or meaning (although subective, e.g., favorite t-shirt and nostalgic meal, art show and music performance). Finally, in the hippocampus, overlap in appearance, meaning, and reward are discounted in favor of representing inputs that co-occur over space or time together (e.g., sequence of landmarks on a commute, the people in a social group, events on a memorable date, etc.). There are other dimensions that organize or dominate the representations in other brain systems, such as emotion, modality, tasks, and motor actions. However, here we will focus on visual cortex and the hippocampus, two brain systems with mature theories that address the nature of their representations.

Visual system

Hippocampal system

Computation and Inference Background. Exponential family embeddings are a new way to generalize classical methods of finding distributed representations in language (**?**). Consider a corpus of language $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, where each x_i is a word from a vocabulary of terms. An exponential family embedding has three components: (a) a notion of *context* for each data point, e.g., a window of observed words around each word (b) a form of the *conditional distribution*, e.g., for text a categorical distribution over V items is appropriate and (c) an *embedding structure* that describes how parameters are shared across data, e.g., for text we typically posit that each term (such as "walnut" or "bicycle") shares the same representations wherever it appears in the collection. An exponential family embedding posits two d-dimensional latent representations for each term v, one is the *embedding vector* ρ_v and the other is the *context vector* α_v, where d is a hyperparameter. The model asserts that each observation is drawn from a conditional distribution given its context. Exponential family embeddings generalize many existing methods for learning distributed representations, including continuous bag of words, negative sampling, and the many other variants of word2vec (**??**).

Fitting such models is difficult, and requires robust methods for computation and evaluation. We will lean on and adapt a line of research on black box variational inference methods (**?**), particularly for probabilistic programming (**??**), to be able to quickly develop and test our models. Variational inference approximates the posterior by fitting a family of distributions over the latent variables to be close in KL divergence (**??**). For simplicity, consider the problem of fitting embeddings. The variational distribution is $q(\alpha_{1:V}, \rho_{1:V}; \nu)$ and we fit the variational parameters ν to be close in KL divergence to $p(\alpha_{1:V}, \rho_{1:V} \mid \boldsymbol{x})$. Recent innovations in probabilistic programming also us to do this *generically* and *scalably*, easily fitting many types of models to large data sets. This enables the exploration of many variants of the models, e.g., different types of contexts, different values of d, different underlying conditional distributions. We support an empirical approach to making these choices, using cross-validation with the held-out predictive log likelihood (**?**). The intuition is that a model that provides a good density of the data is more likely to be one that is useful and interpretable.

Proposed Research A: Representations beyond co-occurrence statistics. Distributed embedding representations in machine learning are almost exclusively based on co-occurrence statistics. For instance, when constructing embeddings for words in text, names of colors ("red," "blue"...) will be embedded in nearby locations simply because they tend to be used together. How can a richer knowledge of representation in the brain be used to inspire algorithms for processing text, images, and audio? What are other brain-inspired features over which co-occurrence can be evaluated?

Proposed Research B: Embeddings for fMRI data. Exponential family embeddings open the door to developing more complex embeddings for problems in cognitive neuroscience: embeddings for fMRI and other types of data, embeddings with a complex notion of context, and embeddings that are themselves a part of a larger probabilistic model, such as where representations are shared or tied across tasks. For examples of some of these innovations see **?**. The approach will be based on the incorporation of a latent variable model, for which variational methods can be used for scalable inference, as described above.

Proposed Research C: Time-dependent representations. To capture dependence on time, we will build on one of the PIs earlier work on dynamic topic models (**?**). Dynamic topic models captured how the latent themes in a collection can grow and shrink and change over time. For example, the theme of "technology" in a scientific corpus might start with words about electricity and wires and end with words about computers and semiconductors. Dynamic topic models were developed specifically for language. We will generalize this idea to capture the evolution over time of distributed representations. Unlike dynamic topic models, the fitted model will capture multimodal data and the evolution of its latent characteristics. In the exponential family embedding framework, this amounts to placing a linear dynamic prior on the embedding vectors or the context vectors (or both).

4. Attentional Filtering (Objective 3)

Neuroscience Background. Models of attention are central understanding cognition in the brain. At higher levels of processing we actively select which dimensions to process based on their inherent salience and relevance to our goals. Consider the task of reconstructing a person's subjective experience of a movie from their brain data. Existing inverse modeling methods attempt to reproduce the pixels of the photos or movie frames being shown to the person, but if the human brain only represents a subset of this information these models have an inherent performance ceiling. Pixels are not ground truth for the brain, but the problem is that there is no principled way to figure out the internal "ground truth" (which can also vary from moment to moment). A useful public dataset for examining this is 17 subjects who watched a 90 minute episode of Sherlock while being scanned with fMRI. We have a frame-by-frame annotation of several dimensions of the movie, and so could try to reconstruct which of these dimensions are present in the brain data and how this attentional trace changes dynamically over time.

Computation and Inference Background. High-dimensional problems suffer from the curse of dimensionality. This has a precise mathematical characterization in standard models of statistical machine learning. The curse of dimensionality has two components, one statistical, the other computational. The statistical issue stems from the observation that in high dimensions, any local ball will contain very few data points. The computational curse is implied by the fact that searching over a sufficiently large space of models is often intractable. "Beating" the curse of dimensionality involves making assumptions about the structure of the learning problem. We will study mathematical formulations of attention as one approach to making assumptions for which learning is tractable in principle. In an attention-based model, the object of study is a lower-dimensional trace or curve through the high-dimensional input space.

In the deep learning literature, attention-based models were first developed in the setting of machine translation, using sequence-to-sequence algorithms based on recurrent neural networks (RNNs) (**?**). The attention mechanism is a type of alignment model, which is a key component of statistical translation methods (**?**). Attention has been applied to the problem of generating image descriptions by **?**.

Recall from Section **??** that an exponential family embedding model uses a latent representation of the

input and output variables. Consider the case of high dimensional density estimation, where the goal is to estimate a density $p(x)$ for $x \in \mathbb{R}^d$. A latent variable embedding model takes the form

$$p(x) = \int p(z) \exp(\rho(x)^T \alpha(z) - \Psi_{\rho,\alpha})\, dz$$

where $z \in \mathbb{R}^m$ is a latent Gaussian vector, $\rho : \mathbb{R}^d \to \mathbb{R}^K$ is an embedding of the input space, and $\alpha : \mathbb{R}^m \to \mathbb{R}^K$ is an embedding of the latent Gaussian, and $\Psi_{\rho,\alpha}$ is a normalizing constant. We will investigate attention-based versions of embedding models. Many possibilities exist for doing so. For instance, consider a parameterized curve $t \mapsto (Z_t, S_t)$ where Z_t is a Gaussian vector and $S_t \subset \{1, \ldots, d\}$ is a subset of the input variables. Define the inner product $\langle x, z \rangle_{\rho,\alpha} = \int_0^1 \rho\left(x_{S(t)}\right)^T \alpha(Z_t) dt$, and the corresponding embedding model $p(x) = \int p(z) \exp\left(\langle x, z \rangle_{\rho,\alpha} - \Psi_{\rho,\alpha}\right) dz$. In another version, consider a graph over $\{1, \ldots, d\}$ with edge set E, and define the inner product $\langle x, z \rangle_{\rho,\alpha} = \sum_{(j,k)\in E} \rho(x_j, x_k)^T \alpha(Z_j, Z_k)$ where Z is now a vector-valued Gaussian random field, and the model can be viewed as a nonparametric graphical model (**?**).

These are thought of as attention-based models as the "energy" $\langle x, z \rangle_{\rho,\alpha}$ is localized to subsets of the sample space. Discriminative or regression-based versions of these models are defined similarly. The project will develop variational inference algorithms for this family of models. When the mapping $\alpha(z)$ is defined in terms of, for instance, a feed-forward neural network, this becomes an instance of a variational auto-encoder (**?**).

Proposed Research A: Algorithms using variational inference.

Proposed Research B: Learning complexity under attention.

Proposed Research C: Applications to fMRI movie data.

Proposed Research D: Experiments with fMRI to test attention models.

5. Results from Prior NSF Support

John Lafferty was previously supported as co-PI under NSF grant IIS-1116730, "III: Small: Nonparametric Structure Learning for Complex Scientific Datasets," from August 1, 2011 to July 31, 2014. The PI of the grant was Han Liu (Princeton University) and the co-PIs were Lafferty and Larry Wasserman (Carnegie Mellon University). The total award amount was $499,344; the amount awarded to the University of Chicago was $118,750.

Intellectual Merit. The project focused on developing scalable methods for finding structure in complex scientific datasets, without making strong distributional assumptions. The project explored several aspects of nonparametric structure learning, including methods, theory, large-scale computing, and applications, with five concrete aims: (1) nonparametric structure learning in high dimensions, (2) nonparametric conditional structure learning, (3) regularization parameter selection, (4) parallel and online nonparametric learning, and (5) minimax theory for nonparametric structure learning problems. The outcomes included practical models and algorithms; application areas included genomics, cognitive neuroscience, climate science, astrophysics, and language processing. Publications resulting from this grant include (**??**).

Broader Impact. The broader impact of the project included interdisciplinary training for graduate students from biostatistics, computer science, statistics, and medical schools, strengthening the collaboration

and interdisciplinary infrastructure between Carnegie Mellon, Johns Hopkins, the University of Chicago and Princeton, and broadly disseminating the results from this research in journals from all relevant fields. The research had impact outside of machine learning and statistics. In a genomic study, PI Liu applied structured nonparametric methods to analyze high dimensional genomic data, identifying several gene mutations associated with autism. These results were published in Nature (**?**), and reported in the New York Times. In another neuroscience study, the PI developed an effective algorithm for predicting Attention Deficit Hyperactive Disorder (ADHD) disease. The research led to several statistical software packages in R, including (**???**), all of which are freely available on CRAN.

Lafferty is currently supported as PI under NSF grant DMS-1513594, "Constrained Statistical Estimation and Inference: Theory, Algorithms and Applications," from June 29, 2015 to July 1, 2018. The total award amount was $320,000. After two years of the project, the remainder of the funds were transferred from the University of Chicago to Yale University, where the PI moved in July 2017.

Intellectual Merit. The project is studying constraints that are present in complex scientific data analysis problems, but that have not been thoroughly studied in traditional approaches. Different aspects of theory, algorithms, and applications of statistical procedures, with constraints imposed on the storage, runtime, shape, energy or physics of the estimators and applications. The overall goal of the research is to develop theory and tools that can help scientists to conduct more effective data analysis. Publications under this grant have included (**???????????**)

Broader Impact. The broader impact of the project is aimed in three directions. First is the development of flexible and principled large scale data analysis tools that can benefit many scientific domains. Second, is the development of software that is widely distributed, allowing others to build on the work. The third is to education, to allow the research to impact the training of students at both the graduate and undergraduate levels.

Lafferty was previously PI of NSF grant DMS-1547396, "RTG: Computational and Applied Mathematics in Statistical Science" from July 1, 2016 to July 1, 2017. This grant did not transfer to Yale University; the current PI is Jonathan Weare at the University of Chicago. The total award amount is $1,697,320.

Intellectual Merit. This Research Training Group (RTG) project supports creation of a dynamic, interactive, and vertically integrated community of students and researchers working together in computational and applied mathematics and statistics. The work is motivated by the growing need to train the next generation of statisticians and computational and applied mathematicians in new ways, to confront data-centric problems in the natural and social sciences.

Broader Impact. The broader impact includes vertical integration of education and training from undergraduate to postdoctoral researchers, including activities at Toyota Technological Institute at Chicago and Argonne National Laboratory. Participants in the RTG will receive an educational experience that provides them with strong preparation for positions in industry, government, and academics, with an ability to adopt approaches to problem solving that are drawn from across the computational, mathematical, and statistical sciences.

6. Results from TRIPODS Project

Jeff Brock is currently supported as PI under NSF TRIPODS grant CCF-1740741, "Foundations of Model Driven Discovery from Massive Data," from September 1, 2017 to August 31, 2020. The co-PIs of the grant are Stuart Geman, Eli Upfal, Bjorn Sandstede, and Joseph Hogan (Brown University). The total

award amount was $1,482,177.

The TRIPODS grant, in its first year, sought to bring researchers from diverse communities together for workshop activities, and likewise engaged remote researchers to come to campus for seminar talks and research engagement with affiliated faculty and students. We have recently hired 3 postdoctoral fellows to start on July 1, who will assist with achieving the goals of the project.

A fall 2017 workshop entitled Geometry and Topology of Data took place at ICERM in December 2017. This workshop was intended to bridge communities working in Topological Data Analysis, and methods of Diffusion Geometry to study high dimensional data sets. Impacts from workshop were extensive, leading to future collaborations that are the topic of three of our TRIPODS + X proposals, in neuroscience, in data science education, and in gene regulatory networks. We are tracking collaborations going forward, but list one paper of Oudot and Solomon, on which Oudot spoke at the conference (arXiv:1712.03630).

Brown's TRIPODS Institute has been focused on developing models for center related activities that build on intensive visits from outside scientists working in the focus areas of the Institute. In particular, the TRIPODS Institute sponsored short visits from researchers in the theme of topological data analysis to the TRIPODS sponsored Applied Topology Seminar. Highlights include visits from, Justin Curry, Attila Gyulassy, Lori Ziegelmeier, Carina Curto, Steve Oudot, Anthea Monod and Andrew Blumberg.

Training Activities. The visit from Steve Oudot, of INRIA, served further the collaboration with Brock's graduate student Isaac Solomon. Oudot spoke on their work at the workshop, and Solomon is speaking at the Ohio State TRIPODS workshop on their joint work in late May 2018. According to Solomon, "The general goal of our project is to use techniques from applied topology to analyze intrinsic metric structures, in particular metric graphs... Steve presented our ongoing work at the workshop, and I have presented on it at various conferences, including the upcoming TRIPODS workshop at Ohio next week."

Bjorn Sandstede's student Melissa McGuirl met with many of the TRIPODS visitors, who provided valuable research insights as well as career mentorship. In particular, she met with her co-Advisor, Andrew Blumberg, from U. Texas to discuss her thesis work. Melissa also had substantial interactions with many of the speakers including Atilla Gyulassy of whom she says, "We met to discuss ways to extract features from images. In particular we talked about Morse theory applications. We also discussed the current software tools available for this method."

The PI has engaged in planning a week long summer bootcamp with Isaac Solomon, Melissa McGuirl and Henry Adams of Colorado State, who will spend a week in early August presenting the framework of topological data analysis in the context of machine learning. This weeklong workshop (to run at ICERM) will focus on bringing graduate students up to speed with the tools and techniques of TDA, and bring in outside speakers to present their applications in to machine learning.

Finally, the TRIPODS Institute will sponsor a two day workshop in mid-August on building community in theoretical foundations of data science, focused on regional collaboration and interaction between researchers in the theoretical aspects of data science.