Dynamic and Succinct Statistical Analysis of Neuroscience Data

Sanggyun Kim, *Member, IEEE*, Christopher Quinn, *Student Member, IEEE*, Negar Kiyavash, *Senior Member, IEEE*, and Todd P. Coleman, *Senior Member, IEEE*

***Abstract*—**
Modern neuroscientific recording technologies are increasingly generating rich, multi-modal data that provide unique opportunities to investigate the intricacies of brain function. However, our ability to exploit the dynamic, interactive interplay amongst neural processes is limited by the lack of appropriate analysis methods. In this paper, some challenging issues in neuroscience data analysis are described, and some general-purpose approaches to address such challenges are proposed. Specifically, we discuss statistical methodologies with a theme of loss functions, and hierarchical Bayesian inference methodologies from the perspective of constructing optimal mappings. These approaches are demonstrated on both simulated and experimentally acquired neural data sets to assess causal influences and track time-varying interactions amongst neural processes on a fine time scale.

***Index Terms*—Loss function; minimax regret; directed information; prediction with expert advice; optimal transport theory; point processes; BRAIN Initiative; Human Brain Project**

I. Introduction

The brain is arguably the most complex dynamic system in nature, and understanding how it works is one of the greatest challenges in science. Recently, the developments of existing recording techniques and the advent of new measurement methods in neuroscience provide us rich amounts of data, which allow us to investigate fundamental neuroscience questions in an unprecedented manner [1], [2]. For example, the recent development of simultaneous recording of activity from multiple neurons provides us new opportunities to understand how complex function and computation arises from networks of interacting neurons [3]. These recording technologies will be accelerated by the BRAIN initiative [4], [5].

Most standard analysis methods are developed primarily for a specific modality such as continuous-valued data and designed for problems in which the structure in the data is static rather than dynamic [3]. As we collect rich datasets, across multiple modalities and time scales, these inadequacies in traditional methods will limit our ability to develop effective scientific conclusions or advance translational therapies. Here, we briefly describe some challenging analysis issues in neuroscience research.

- **Complexity**: The speed of evolution of neurotechnologies is growing the size of neural datasets that are acquired, and the BRAIN initiative will perhaps only accelerate this process. As such, management of complexity across multiple fronts will be of paramount importance. First, the complexity of the data acquired will render it useless to humans for interpretation unless appropriate simplification of the data is carried out. In some sense, there is a need to transform the 'big data' that is collected into 'small data' that can be easily visualized, and that balances ease of visualization with neurobiological relevance to the mechanism of interest. The details of this balance will undoubtedly be application-specific, but ideally a common set of core principles can be applied, by tying it to subsequent decision-making that will ensue. Secondly, the sheer amount of neural data of various types require the development of highly efficient algorithms which in some instances - e.g. neural prosthetics - will be needed to be implemented in real time [6], [7].
- **Dynamics**: The stochastic nature of ensemble activities of neural processes and the interaction among neural circuits, requires statistical analysis of ensemble recordings that succinctly reflect interaction and causal relationships. Thus, methodologies that can directly elucidate these interactions are urgently needed. In addition, in some instances, the time scales over which these interactions change are faster than what is required to effectively use statistical approaches assuming time-homogeneity. Such non-stationarities exacerbate the need to develop theoretical tools, and algorithms, that are able to characterize the dynamics of these neural patterns.
- **Uncertainty**: Because of the massive amount of data in ensemble recordings and their rich dynamics, the imperfections in our idealized models may lead to uncertainty in our predictions. As such, quantifying the uncertainty, and ensuring sufficient information aggregation that allows for reliable decision-making under uncertainty, will be a dominant theme moving forward.

Although there are individualized methods that are tailored

This work was supported in part by NSF CCF-1065022 and CCF-0939370, in part by AFOSR under Grants FA9550-11-1-0016 and FA9550-10-1-0573, and in part by the NSF under Grants CCF 10-54937 CAREER and CCF 10-65022. C. J. Quinn was supported by the Department of Energy Computational Science Graduate Fellowship, which is provided under Grant DE-FG02-97ER25308.

Sanggyun Kim is with the Department of Bioengineering, University of California San Diego, La Jolla, CA, 92093 USA e-mail: s2kim@ucsd.edu.

Christopher Quinn is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL, 61801 USA e-mail: quinn7@illinois.edu.

Negar Kiyavash is with the Department of Industrial and Enterprise Systems Engineering, University of Illinois, Urbana, IL, 61801 USA e-mail: kiyavash@illinois.edu.

Todd P. Coleman is with the Department of Bioengineering, University of California San Diego, La Jolla, CA, 92093 USA e-mail: tpcoleman@ucsd.edu.

Manuscript received October, 2013; revised.

TABLE I
THIS TABLE PROVIDES A CONCISE EXPLANATION OF THE CHALLENGES FOR NEURAL DATA ANALYSES THAT WE FORESEE AND SOME METHODS.

Challenges \ Methods	Robust Approximation	Causal Inference	Bayesian Inference
Complexity	Directed tree approximation		Convex optimization via optimal transport
Dynamics	Time-varying causal inference		
Uncertainty			

to one specific modality or one specific time scale of dynamics, there is an increasing need to develop a set of core theoretical principles that guide the philosophical underpinnings of algorithms that are then suited to specific physiological or mechanistic scenarios of neuroscientific interest. There is an increasing interest from the fields of classical statistics, control theory, and information theory at taking these core theoretical principles and tailoring them towards the analysis of complex neural data [8]. In this paper, we will develop a coherent philosophical framework that is rooted in the aforementioned disciplines, guides all of the procedures we develop, and is broadly applicable across modalities and time scales.

Specifically, in this paper, we will address a part of these challenges by developing efficient, quantitatively rigorous methods to track time-varying statistical dynamics of ensemble neural activity as well as parsimonious modeling and visualization tools to concisely describe multivariate neural responses. We will show all of these methods are fundamentally developed based on loss function and optimal transport theory [9], [10]. This framework enables us to understand the underlying dynamic mechanism of the data set in any modality and assess the important properties of complex neural systems with low complexity. Table I attempts to provide a concise explanation of the challenges we foresee, and how the methodologies and specific algorithms that are introduced in this paper provide attempts at ameliorating these challenges. When this framework is also physiologically guided, we can exploit the unique features of neuroscience data to develop new analysis tools to provide us new insight into them.

New recording technologies, combined with appropriate analysis methods, will have a significant impact on basic and clinical neuroscience research, and will have a great synergy with Human Brain Project to perform inference on existing neural data, simulate the human brain, and provide insights on future experiments or medical therapies [11], [12]. Succinct and dynamic representation of multiple neural data can be used to analyze the complex pattern of interconnected neuronal networks, and detect the origin or the direction of information propagation within the brain on fine time scales. Characterization of these complex networks will provide us a deeper understanding of the mechanism by which the brain works, leading to the improved diagnoses of neuropathologies, improved neural prostheses, and offering unique opportunities to explicitly link experimentation and computational modeling by using the information from the experiments to quantify better prediction from more complex models.

The rest of the paper is organized as follows. Section II suggests a general framework for addressing statistical challenges in neural data analysis from loss function perspective. Section III describes an efficient Bayesian inference with optimal maps to address computational issues in neural data analysis. Section IV shows the application of these frameworks to the analysis of multivariate neural spike trains. Section V concludes and discusses the paper.

II. ANALYSIS FROM LOSS FUNCTION PERSPECTIVE

Advances in recording technologies continue to provide richer, higher dimensional neuroscience data across a multitude of time scales and modalities. Although elementary statistical analysis methods such as cross-correlation [13] and joint peristimulus time histogram (JPSTH) [14] are still widely used in neural data analysis, there is a growing need to match the sophistication of experimentation with that of ways to characterize these dynamic neural processes. Specifically, as neuroscience experiments grow in their complexity, it is increasingly a challenge to disambiguate the variability across time scales and neural processes as an epiphenomenon, chance, or a reflection of a novel mechanism.

However, at the same time there is a need to balance a small set of philosophical approaches to modeling and inferring with data, with being neurobiologically plausible and relevant. We here demonstrate how measuring performances and designing algorithms from *loss function* perspective provide a foundation for capturing this variability while still embodying neurobiological plausibility.

A. Prediction and Loss Functions

In this section, we describe a conceptual approach towards developing a suite of robust, scalable, modality-agnostic methods for statistical analysis of neural data that are not only relevant now, but will continue to maintain relevance as the BRAIN and other initiatives create increasingly rich neural datasets.

Prediction is concerned with guessing an outcome that has not been observed yet. For example, one specifies a prediction p on the next outcome y of a random or unknown sequence given the past outcomes and possibly side information. As a general way to measure the performance of statistical modeling, we consider a *loss function*, $l(p, y)$, which defines the quality of the prediction and thus increases as the prediction p deviates more from the true outcome y. Although there are a variety of loss function for specific modalities, e.g. the squared error loss, absolute loss, $0-1$ loss, etc, we will primarily consider the *logarithmic* loss, which is applicable to any modality $y \in \mathsf{Y}$ where Y represents the set of possible outcomes. In this setting, the prediction p lies in the space of probability measures over y, i.e. $p = \{P(y) : y \in \mathsf{Y}\}$. Given a prediction p for an outcome $y \in \mathsf{Y}$, the log loss is represented by

$$l(p, y) = -\log p(y). \quad (1)$$

The log loss, also termed the 'self-information' loss, is the ideal 'Shannon' code length for compression of a symbol y drawn from distribution p and achieves the minimum expected total codelength for any uniquely decodable data compression scheme [15]. From the perspective of multi-modal neuroscience data analysis, the log loss has the desirable property that the prediction p lies in the space of beliefs over Y, so that we can develop neurophysiologically-specific classes of statistical models unique to each modality but still have a common way to perform statistical inference upon them.

B. Causal Inference: Reduction in Loss

Progress in neural recording technologies provides us multivariate time series neural data [1]. The number of simultaneously recorded neurons has been doubling every 7 years since 1950s, and with current multiple-electrode technology, hundreds of individual neurons can be recorded simultaneously [16]. What will be of paramount importance are succinct ways that humans can take this information, extract statistical meaning out of it, and characterize brain function or develop treatment options for intervention.

One class of methods to visualize statistical relationships between networks of random variables are traditional graphical models. Markov networks and Bayesian networks in particular represent two different perspectives on the structure of networks of random variables. Markov networks directly represent the dependence between each pair of variables, conditioned on all other variables. Bayesian networks represent factorizations of the joint distribution, so each variable potentially depends on preceding variables, and then the conditional terms are reduced. See [17] for an overview of graphical models.

Attempts to model the interactions of simultaneously recorded neural responses may be able to shed light on how the networks of neural processes represent and process information. If we have N random processes recorded across T time units, then visualization of statistical relationships in terms of a network of $N \times T$ random variables can be cumbersome and grows as our recording interval T increases. Moreover, such a representation will not aid with visualization of the structure of inter-dynamics of coupled time series, for instance, how the past of some processes affects the future of others.

A variety of quantitative techniques have been developed to elucidate the functional properties of complex neural network [3], [13], [14]; however, most methods that attempt to identify associations between neural responses offer little insight into the directional nature of the neural system, or sometimes give a misleading picture on the network. In this section, we discuss a general-purpose framework, resting upon the log loss, to develop an inference engine that uncovers the interactive nature of N time series as a directed graph on N nodes, where a directed edge encodes information about how the past of some processes affect the future of others (see Fig 1).

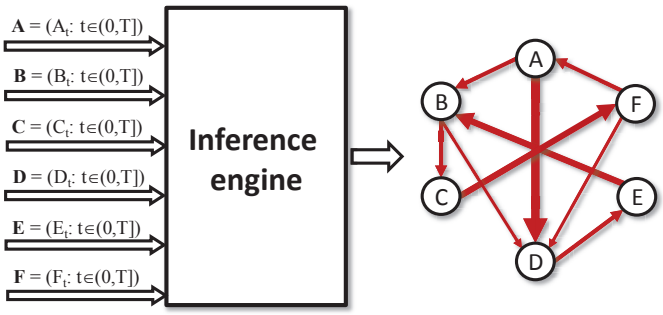

Fig. 1. An inference engine that uncovers the directed functional structure in the joint statistics underlying a collection of time series pertaining to neural recordings.

Within a context of a network, we aim to design an inference engine to produce a directed graph description to elucidate *causal* interactions between neural processes, over space and time, as shown in Fig. 1. Recently, Granger causality has proven to be an efficient method to infer directional relationships between sets of neural responses [18]–[20]. The basic idea of causality between time series was originally introduced by Wiener [21], and later Granger formalized it as follows [22]:

"we say that X is causing Y if we are better able to predict the future of Y using all available information than if the information apart from the past of X had been used."

Granger instantiated this idea for practical implementation using multivariate autoregressive (MVAR) models and linear regression. There is a class of graphical models developed to represent Granger's principle, known as Granger causality graphs [23]–[25]. These are mixed graphs (both directed and undirected graphs) for multivariate autoregressive time series. Nodes represent processes. The directed edges represent causal influences, as measured by Granger causality. The undirected edges represent instantaneous correlation. However, it is challenging to identify nonlinear relationships with this approach, and in other situations it is conceptually mismatched, for example, neural spike trains are a binary time series on a millisecond time scale: either one spike occurs or it does not. In [23], it is suggested that, conceptually, Granger causality graphs could be employed for nonlinear relationships. However, it is mentioned that some properties of the graphical model would not hold. They also suggested it would be impossible to infer structures where causal influences were nonlinear without assuming specific models.

Here, we discuss a sequential prediction framework that generalizes Granger's mathematical formulation of causality beyond autoregressive models to any modality. At its core, Granger's statement revolves around *prediction*. We test causal interaction from a neural process X to Y by comparing the performance of two predictors. One predictor specifies a prediction on the future of Y, Y_t, given the past of all processes up to time t-1, $\mathcal{H}^{t-1}$. The other predictor specifies a prediction on the future of Y, Y_t, given the past of all processes excluding the process X up to time t-1, $\mathcal{H}_X^{t-1}$. We then compare the performance of the two predictors in

terms of their loss, accumulated sequentially over time. On average, the predictor with less information incurs more loss. If their average accumulated losses are about the same, then we declare that X does not cause Y; otherwise, X causes Y. Here, we denote $y^t \triangleq \{y_1, ..., y_t\}$. Causal inference based on reduction in loss, accumulated over time, is assessed as follows:

Causal inference based on reduction in loss

1) Assign $\mathcal{H}^{t-1}$ as the past of all processes up to time $t-1$, and $\mathcal{H}_X^{t-1}$ as the past of all processes excluding process X, up to time $t-1$. For example, for three processes X, Y and Z, $\mathcal{H}^{t-1} = (x^{t-1}, y^{t-1}, z^{t-1})$ and $\mathcal{H}_X^{t-1} = (y^{t-1}, z^{t-1})$.
2) In parallel, update the two predictors: $p_t = P_{Y_t|H^{t-1}}(y_t|\mathcal{H}^{t-1})$, $\tilde{p}_t = \tilde{P}_{Y_t|H_X^{t-1}}(y_t|\mathcal{H}_X^{t-1})$;
3) y_t is revealed. The two predictors incur losses given by: $l(p_t, y_t)$, $l(\tilde{p}_t, y_t)$;
4) Quantify the reduction in loss: $r_t = l(\tilde{p}_t, y_t) - l(p_t, y_t)$.
5) Let $t = t+1$; go to 1.

From this, we can quantify the average reduction in loss as our measure of causality:

$$C_{X\to Y} = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]. \tag{2}$$

If $C_{X\to Y}$ is close to zero, it indicates that the past values of neural process X contain no significant information that would assist in predicting the activity of neural process Y. Thus, X has no causal influence on Y. On the contrary, if the reduction in loss is significantly greater than zero, it indicates that the past values of X contain information that improves the ability to predict the neural process Y. Thus X causes Y in the sense of Granger. We can perform this test for every possible directed edge for a collection of N recorded time series.

Using the log loss function, the causality measure $C_{X\to Y}$ in (2) becomes

$$\begin{aligned} C_{X\to Y} &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\log\frac{P_{Y_t|H^{t-1}}(y_t|\mathcal{H}^{t-1})}{\tilde{P}_{Y_t|H_X^{t-1}}(y_t|\mathcal{H}_X^{t-1})}\right] \qquad (3)\\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[D\Big(P_{Y_t|H^{t-1}}\|\tilde{P}_{Y_t|H_X^{t-1}}\Big)\right]\\ &= \frac{1}{T}\sum_{t=1}^{T} I(X^{t-1}; Y_t|\mathcal{H}_X^{t-1}),\\ &\triangleq \frac{1}{T} I(X\to Y|\mathcal{H}) \qquad (4)\end{aligned}$$

where $D(P\|Q)$ is the Kullback-Leibler (KL) divergence and $I(A;B|C)$ is the conditional mutual information between A and B given C [26]. Eqn. (4) turns out to be the causally conditioned directed information [19], [27], [28], with interpretations in control theory and feedback information theory. It has the key property that it is non-negative, and zero if and only if the future of Y is independent of the past of X given knowledge of all processes excluding X. Note that in general, unlike mutual information, directed information from X to Y is not necessarily equivalent to the directed information from Y to X, thus it provides the desirable property of direction of information flow across time. Moreover, this framework for assessing causal interaction is particularly desirable for neural data analysis because it works on arbitrary modalities and statistical models [19].

There were some recent works on the conceptual and theoretical link between Granger causality, conditional independence, and directed information theory [29]–[38]. They justified using directed information conceptually, motivated by equivalence of Granger causality and directed information in the case of jointly Gaussian processes [31], but did not identify properties of the graph. Independently, [35], [36] showed the relationship between Granger causality and transfer entropy. Directed information is the time average of transfer entropy. Transfer entropy was proposed by Schreiber [37], independently of directed information. Granger's original formulation of causality based on the linear regression modeling of stochastic processes is also a special case of this framework, when the distributions of neural responses are assumed to be Gaussian [38].

C. Robust Approximations for Massive Neural Data Analysis

Accessing the directed network of multiple neural processes can yield insight into the structures of a neural system and their functions. A large-scale network during motor maintenance behavior in awake monkeys, for instance, has been demonstrated using the causal network analysis [18], and strong local neighborhood structure between retinal ganglion cells has been found by modeling the interactions of these cells [39]. However, analyzing the high-dimensional neural data obtained using recent recording technologies becomes very challenging from both visualization and computational perspectives, even with directed network analysis, as the number of ensemble recordings grows.

Several methods have been developed that can reduce the complexity [40], [41]. They use simpler models but assume the simultaneously-recorded high-dimensional neural data is the product of a latent, low-dimensional state space.

Within the context of directed network analysis, an approach was recently demonstrated to approximate the directed information flow among N processes (with N nodes and N^2 possible edges) with a directed tree, containing N nodes and $N-1$ possible edges [42]. Fig. 2 illustrates an example of a directed tree with six processes.

Each tree represents different dynamics. For example, suppose Y depends on X and Z, as described by $P(y_t|x_{t-1}, y_{t-1}, z_{t-1})$. In a directed tree, Y can only have one parent. If X is chosen, then the tree only represents the dynamics described by $P(y_t|x_{t-1}, y_{t-1})$. We measure how "close" this is to the original with the KL divergence, $D(P(y_t|x_{t-1}, y_{t-1}, z_{t-1})\|P(y_t|x_{t-1}, y_{t-1}))$.

Let $\mathcal{T}_{\mathrm{C}}$ denote the set of all directed trees. Consider any particular tree T. Let $\widehat{P}_T$ denote the distribution corresponding to T. The goal is to find the tree T that best represents the full dynamics, $\underset{T\in\mathcal{T}_{\mathrm{C}}}{\arg\min}\ D(P\|\widehat{P}_T)$, where P is the original

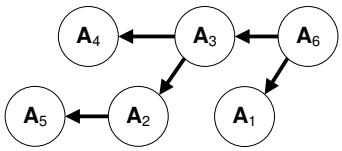

Fig. 2. Diagram of a tree generative model graph representing a sparse, approximate joint distribution.

TABLE II
VISUALIZATION AND INFERENCE COMPLEXITY OF AFOREMENTIONED APPROACHES IN TERMS OF NUMBER N OF PROCESSES AND NUMBER T OF TIME POINTS.

Description	Visualization complexity: nodes, edges	Inference complexity
Traditional graphical model	NT, $(NT)^2$	hard
Directed info graphs	N, N^2	medium
Tree directed info graph	N, N	easy

distribution. The best tree is the one that maximizes a sum of directed informations along the edges (X, Y) of the tree [42],

$$\arg\min_{T\in\mathcal{T}_C} D\Big(P\|\widehat{P}_T\Big) = \arg\max_{T\in\mathcal{T}_C} \sum_{(X,Y)\in T} I(X\to Y). \quad (5)$$

This can be solved by finding the directed information $I(X \to Y)$ for each pair of processes, and then using an efficient maximum-weight directed spanning tree algorithm, such as Edmunds' algorithm [43] with a complexity of $O(N^2)$, to find the best tree. Note that only N^2 calculations of directed informations are needed. Each calculation only uses statistics amongst two random processes, for which statistically consistent algorithms exist under appropriate assumptions [19], [44]. This result is analogous to that of Chow and Liu [45] for networks of random variables.

Table II provides a concise explanation of the reduction in complexity for both visualization and computation. It includes traditional Markov network graphical models, directed information graphs, and tree approximations. The inference complexity of traditional graphical models is deemed 'hard' because joint statistics on all processes, across all times, is required. Directed information graphs, instead, simply calculate causally conditioned directed information, which are moving averages of log likelihoods but nonetheless require joint statistics amongst all random processes. Lastly, the tree directed approximations only require pairwise statistics and have $O(N^2)$ total complexity.

Directed trees by definition do not have feedback, which is undoubtedly important in neuroscience. One purpose of a tree approximation of the network estimate is to identify the main path of the information flow in the network. For some neuroscience data sets such as those described in Section IV B, the scientific hypothesis to be tested involves understanding the main direction of information in the network. As such, this analysis method can efficiently elucidate some phenomena of interest, but generally speaking should be used with caution."

D. Dynamics and Non-stationarity: Minimax Regret

The anatomical connectivity between different regions or different neurons in a given brain region remain relatively stable over long periods of time, but the ensemble activity we can now record exhibits dynamic, changing, interactive relationships over faster time scales. So far, most existing methods for analyzing neural data are developed based on stationary joint statistics on the data generating mechanism. Thus, taking time averages as done before can blur out the dynamic, non-stationary aspects of brain function we may want to elucidate. This has typically been addressed in an ad-hoc manner, for example by using moving windows, whose length choice suffers from the bias-variance tradeoff [3]. This method did not track the evolution in a fine time scale, and constrained that the changing timings should be the same for between all pairs of neurons. Attempting to develop maximum-likelihood style approaches for time-varying parametric models are destined to fail, because the number of parameters grows with the number T of time points.

Here, we consider an approach from the theory of sequential prediction to build time-varying statistical models by combining reference forecasters, called *experts* [9]. The experts can be interpreted in different ways depending on various applications. It is possible to regard an expert as a black box of unknown computational power, possibly with access to private side information. In some cases the class of experts can be viewed as a parametric statistical model where each expert in the class is uniquely specified by a set of parameters and represents an optimal predictor for a given state of nature. For example, consider a family of models of binary activity. One expert corresponds to a Bernoulli probability model with probability of heads 0.5, while another expert corresponds to a probability model of heads probability being 0.1. In general, there can be a continuum of experts, in this case, being one-to-one correspondence with the [0,1] line. In our framework, each expert makes a prediction on a next outcome based upon all information it has had in the past. The predictions of each expert, and their performance in the past, are available to us. It is our job to combine these experts' advice and their past performances to provide one prediction that performs as best as possible on the new outcome that has not yet been revealed. A schematic diagram of sequential prediction with experts' advice is illustrated in Fig. 3. As shown in the figure, we will then design our own predictive strategy based on these experts' advice so that the cumulative loss will be close to that of the best expert in the class, in hindsight.

Let us denote the class of experts as E where each expert $e \in \mathsf{E}$ provides a prediction $e_t \in \mathsf{D}$ at each time t where D represents the set of possible decisions. We make the predictions of neural data in a sequential manner, and the performance of this sequential prediction is compared to that of a class of experts. The aforementioned framework of prediction with loss can be naturally viewed as the following repeated game between our own predictor p_t and environment to set the true outcomes y_t.

Sequential prediction with expert advice

For each round $t = 1, 2, \ldots, T$,

1) the environment chooses the next outcome $y_t \in \mathsf{Y}$ and expert advice $e_t \in \mathsf{D}$; the expert advice is revealed to the predictor;

Fig. 3. Sequential prediction with expert advice: (a) Our own predictive strategy that is designed based on experts' advice. Initially, experts' advice is combined with equal weights. (b) Optimal predictive strategy. For each round we strategize how to combine these advice for experts with uneven weights so that the cumulative loss is getting close to that of the best expert in the class.

2) the predictor p has y^{t-1} at its disposal and chooses a prediction $p_t \in \mathsf{D}$;
3) the environment reveals $y_t \in \mathsf{Y}$;
4) the predictor p incurs loss $l(p_t, y_t)$, and expert $e \in \mathsf{E}$ incurs loss $l(e_t, y_t)$.

The accumulated loss for the predictor p (or an expert e) on outcome sequence y^T is defined as

$$L_T(p, y^T) = \sum_{t=1}^{T} l(p_t, y_t). \tag{6}$$

The difference between the accumulated loss of our predictor p and that of an expert e is called as 'regret' for outcome sequence y^T, which is given by

$$R_T(p, e, y^T) = L_T(p, y^T) - L_T(e, y^T). \tag{7}$$

This measures how much our predictor experiences, in hindsight, of not having followed the advice of this particular expert [9]. It is our goal to design the predictor p such that its regret is close to that of the best expert in hindsight. By understanding the worst-case regret over all experts and all possible sequences, we desire to make a 'good' predictor to minimize the worst-case regret:

$$R_T(p, \mathsf{E}) \triangleq \sup_{y^T} \sup_{e \in \mathsf{E}} R_T(p, e, y^T). \tag{8}$$

This 'minimax' regret in the equation (8) measures the best possible performance guarantee one can have for a predictive algorithm that holds for all possible classes of experts in E and all possible sequences of outcomes of length T. It provides us a non-probabilistic guarantee on robust performance. This has been used in the classical statistics literature for development of probably good inference methods [9] and in the information theory literature for development of probably good model selection procedures [46], [47]. The minimizer of (8) is termed the normalized maximum likelihood estimator (NMLE), but its practical use is prohibitive because it requires solving an optimization problem whose complexity increases exponentially with time and does not have the prequential property [9], [48]. In the next section, we will develop efficient methods to asymptotically attain the minimax regret.

III. Optimal Transport and Bayesian Inference

Neuroscience data are increasingly being recorded from multiple modalities, which can operate on different spatial or temporal scales. For example, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) signals can be simultaneously recorded where the former has high temporal resolution but low spatial resolution, but the latter has the opposite. Although the objective of most current neurophysiological experiments is to relate relevant biological stimuli to multivariate neural data, the ability to simultaneously record multiple forms of activity such as neurophysiological, functional imaging and behavioral data is increasing. Developing appropriate statistical methods to analyze simultaneous multi-modality recordings will require innovative approaches to integrate information properly across the different temporal and spatial scales of various data sources. One natural way to perform this is from the perspective of Bayesian inference where likelihoods can link any specific modality to a latent mechanism of interest, and from which we can infer information across all modalities. The use of Bayesian inference methods within the context of learning, across heterogeneous modalities and times scales, for example, was accomplished within the context of learning in monkeys, in [49].

Generally speaking, Bayesian inference provides a foundation for learning from noisy and incomplete neural data; for instance, it offers a general approach to estimating the representation of biological information in neural observations [50]. However, when the latent variable θ is in a continuum, we have Bayes' rule as

$$f_{\Theta|Y}(\theta|y) \quad = \quad \frac{f_\Theta(\theta) f_{Y|\Theta}(y|\theta)}{\beta_y} \tag{9}$$

where $\beta_y \triangleq \int_\Theta f_\Theta(u) f_{Y|\Theta}(y|u) du$. Computing β_y or drawing samples from the posterior is one of the central challenges

in Bayesian inference. The typical approach for solving this problem is the use of Markov Chain Monte Carlo (MCMC) methods, where samples are drawn from a Markov chain whose invariant distribution is that of the posterior [51]. Despite popularity of MCMC method, when samples are generated from a Markov chain, they are statistically dependent, leading to smaller effective sample sizes. More importantly, the number of iterations of running the chain that is necessary for the system to converge is not well understood.

An alternative approach for Bayesian inference to avoid Markov chain simulation was recently proposed in [52], which is inspired by optimal transport theory [10]. The main idea of this approach is to find a map that transforms a random variable distributed according to the prior to a random variable distributed according to the posterior. However, in general, solving an optimal transport problem is hard. Recently, we showed that for a large class of Bayesian inference problems (e.g. with log-concave likelihoods and priors, which can model the various kinds of neural data), one class of variational problems over maps leads to an efficient (e.g. convex) optimization problem that only requires drawing independent, identically distributed (iid) samples from the prior [53]. Because there is a rich theory of convergence for iid time series and the prior is typically easy to sample from, and because there are many methods to solve convex optimization problems [54], this provides an alternative approach, with decades-old convergence criteria, for solving these classes of log-concave Bayesian inference problems.

A. Jacobian Equation and Optimal Transport

We now provide some notation relevant to development of our efficient Bayesian inference algorithms where the latent variable is in a continuum. Consider a set $\mathsf{W} \subset \mathbb{R}^d$ for some d, and define the space of all probability measures on W as $\mathsf{P}(\mathsf{W})$. Given a $P \in \mathsf{P}(\mathsf{W})$ and a $Q \in \mathsf{P}(\mathsf{W})$, we seek a map $S : \mathsf{W} \to \mathsf{W}$ to *push forward* P to Q (denoted as $S_{\#}P = Q$) if a random variable W distributed with P results in $Z \triangleq S(W)$ distributed with Q. We say that $S : \mathsf{W} \to \mathsf{W}$ is a 'diffeomorphism' on W if S is invertible and both S and S^{-1} are differentiable. For any such diffeomorphism S assuming that p (q) is the density of P (Q), then we have from the Jacobian equation that

$$p(u) = q(S(u))|\det J_S(u)| \tag{10}$$

where J_S is the Jacobian of the map S. From the theory of optimal transport [10], for any p and q, there always exists a *monotonic* map S (for which $\det J_S(u) > 0$) such that $S_{\#}P = Q$. Thus, without loss of generality, by defining the set of monotonic diffeomorphisms on W as $\mathcal{S}(\mathsf{W})$, we have that for any such $S \in \mathcal{S}(\mathsf{W})$:

$$p(u) = q(S(u)) \det J_S(u). \tag{11}$$

Within the context of Bayesian inference, P represents the given prior P_Θ and Q represents the posterior $P_{\Theta|Y=y}$ we are trying to develop. We seek a monotone diffeomorphism map S_y^* for which ${S_y^*}_{\#}P_\Theta = P_{\Theta|Y=y}$ to push forward the prior distribution P_Θ to the posterior distribution $P_{\Theta|Y=y}$, giving rise to the following equation:

$$\begin{aligned} f_\Theta(\theta) &= f_{\Theta|Y=y}(S_y^*(\theta)) \det\Big(J_{S_y^*}(\theta)\Big) &(12)\\ &= \frac{f_{Y|\Theta}(y|S_y^*(\theta)) f_\Theta\left(S_y^*(\theta)\right)}{\beta_y} \det\Big(J_{S_y^*}(\theta)\Big) &(13)\end{aligned}$$

where (13) follows from (9). Then, for an arbitrary monotone diffeomorphism $S_y \in \mathcal{S}(\mathsf{W})$ instead of S_y^*, a new operator T can be defined as

$$\begin{aligned} T(S_y, \theta) &\triangleq \log f_{Y|\Theta}(y|S_y(\theta)) + \log f_\Theta\left(S_y(\theta)\right)\\ &+ \log\det\Big(J_{S_y}(\theta)\Big) - \log f_\Theta(\theta). \end{aligned} \tag{14}$$

When we consider any other diffeomorphism S_y instead of S_y^* in (13), we note that either ${S_y}_{\#}P_\Theta = \tilde{P}_{\Theta|Y=y}$ where $\tilde{P}_{\Theta|Y=y}$ need not equal the true posterior $P_{\Theta|Y=y}$, or equivalently the inverse S_y^{-1} satisfies ${S_y^{-1}}_{\#}P_{\Theta|Y=y} = \tilde{P}_\Theta$ where $\tilde{P}_\Theta$ need not equal the true posterior P_Θ. This is shown in Fig. 4. For any diffeomorphism S_y, the Kullback-Leibler (KL) divergence is given by

$$\begin{aligned} D\Big(P_\Theta \| \tilde{P}_\Theta\Big) &\triangleq \int_{\theta\in\mathsf{\Theta}} f_\Theta(\theta) \log \frac{f_\Theta(\theta)}{\tilde{f}_\Theta(\theta)} d\theta\\ &= \log\beta_y - \int_{\theta\in\mathsf{\Theta}} f_\Theta(\theta) T(S_y, \theta) d\theta. \end{aligned}$$

The KL divergence is non-negative and clearly there exists a monotone diffeomorphism S_y^* satisfying ${S_y^*}_{\#}P_\Theta = P_{\Theta|Y=y}$, for which the KL divergence is exactly zero. Thus an equivalent problem to solve is to minimize a KL divergence, or equivalently, maximize the expectation of the T operator:

$$\begin{aligned} (P1)\quad S_y^* &= \underset{S_y\in\mathcal{S}(\mathsf{W})}{\arg\min}\, D\Big(P_\Theta \| \tilde{P}_\Theta\Big) &(15)\\ &= \underset{S_y\in\mathcal{S}(\mathsf{W})}{\arg\max} \int_{\theta\in\mathsf{\Theta}} f_\Theta(\theta) T(S_y, \theta) d\theta. &(16)\end{aligned}$$

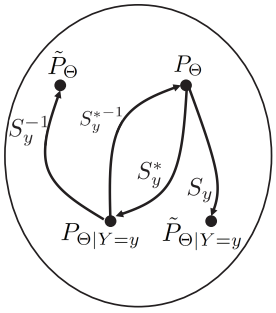

Fig. 4. Bayesian inference with optimal maps. We design a map that *pushes forward* the prior distribution to the posterior distribution. We begin with a prior distribution P_Θ. Upon an observation $Y = y$, it is our objective to find $P_{\Theta|Y=y}$. Up front, because of the difficulty in computing β_y, $P_{\Theta|Y=y}$ is unknown; but we know it exists. A 'desirable' diffeomorphism S_y^* pushes the prior P_Θ to the posterior $P_{\Theta|Y=y}$; equivalently, ${S_y^*}^{-1}$ pushes the posterior $P_{\Theta|Y=y}$ to the prior P_Θ. An arbitrary diffeomorphism S_y will push P_Θ to some distribution $\tilde{P}_{\Theta|Y=y}$ that is not necessarily $P_{\Theta|Y=y}$; equivalently, ${S_y}^{-1}$ pushes the posterior $P_{\Theta|Y=y}$ to some distribution $\tilde{P}_\Theta$ that is not necessarily P_Θ.

The optimization problem to find a map S_y is a search over an infinite-dimensional space of monotone diffeomorphisms.

From here, we can transform the problem into searching for coefficients of an orthogonal basis of functions using the Wiener-Askey polynomial expansion [55]–[57]. For example, if $\Theta = [-1, 1]$ and P_Θ is uniformly distributed, then $\phi^{(j)}(\theta)$ are the Legendre polynomials. If $\Theta = \mathbb{R}$ and P_Θ is Gaussian, then $\phi^{(j)}(\theta)$ are the Hermite polynomials. Rather than optimizing over functions, we can perform functional analysis and approximate any $S \in \mathcal{S}(\Theta)$ as a linear combination of basis functions:

$$S(\theta) = \sum_{j \in \mathcal{J}} \mathbf{g}_j \phi^{(j)}(\theta), \tag{17}$$

where $\phi^{(j)}(\theta) \in \mathbb{R}$ are d-variate polynomials and $\mathbf{g}_j \in \mathbb{R}^d$ are the expansion coefficients, with d being the dimension of W. By assembling the set $\{\mathbf{g}_j\}_{j \in \mathcal{J}}$ into a matrix $F = [\mathbf{g}_1, ..., \mathbf{g}_K]$ of size $d \times K$ where $K = |\mathcal{J}|$, and every polynomial $\{\phi^{(j)}(\theta)\}_{j \in \mathcal{J}}$ into a column vector $A(\theta) = [\phi^{(1)}(\theta), ..., \phi^{(K)}(\theta)]^T$ of size $K \times 1$, the map is then represented as

$$S(\theta) = FA(\theta). \tag{18}$$

Under the basis expansion in (18), we have $J_S(\theta) = FD_\Theta A(\theta)$ of size $d \times d$, and we define the analogous T operator for the coefficients of the basis as:

$$\begin{aligned} \tilde{T}(F, \theta) &\triangleq \log f_{Y|\Theta}(y|FA(\theta)) + \log f_\Theta\left(FA(\theta)\right) \\ &+ \log\det\left(FJ_A(\theta)\right) - \log f_\Theta(\theta). \end{aligned} \tag{19}$$

We now define a problem where we approximate an expectation of $\tilde{T}(F, \theta)$ by a weighted sum of iid samples, and we approximate the set of all functions using a truncated polynomial chaos expansion [55]–[57]:

$$\begin{aligned} (P2) \qquad F^* &= \underset{F \in \mathbb{R}^{d \times K} : FJ_A(\Theta_1) \succ 0, ..., FJ_A(\Theta_N) \succ 0}{\arg\max} V(F), \\ V(F) &\triangleq \frac{1}{N} \sum_{i=1}^{N} \tilde{T}(F, \Theta_i) \end{aligned} \tag{20}$$

where $\Theta_1, \Theta_2, \ldots, \Theta_N$ are drawn iid from P_Θ. This leads to the following theorem [53]:

Theorem III.1. *Problem (P2) solves the Bayesian inference problem, and if $f_\Theta(\theta)$ is log-concave and $f_{Y|\Theta}(y|\theta)$ is log-concave in θ, then this problem is convex and thus efficiently solvable.*

This result is dependent upon log-concavity of the prior and likelihood, but for most neural data sets, this assumption holds: Many common statistical models in neural data sets satisfy the prior and likelihood assumptions in Theorem III.1. For example, Gaussian/Laplace/Uniform priors on θ [58] and generalized linear model (GLM) likelihood functions for $f(y|\theta)$ [59], all fall within the class of log-concave distributions.

Fig. 5 illustrates an example of Bayesian inference using an optimal map in a 2-dimensional compact space. Each color of color maps represents a specific set of parameters in the 2-dimensional space. That is, each color uniquely specifies a particular expert. Initially we assume a uniform distribution over these parameters as shown in the top plot of Fig. 5 (a). All the parameters (colors) are uniformly distributed in the 2-dimensional compact space as shown in Fig. 5 (b). The boxed-in region represents an area, which shows the relatively high likelihood values given a next outcome. The likelihood function is illustrated in the top plot of Fig. 5 (c). This likelihood function places more of its mass over this area. Fig. 5 (d) shows the 'visual effect' of the nonlinear mapping of the 2-dimensional parameter space using a designed optimal map. We can see a large increase in resolution (i.e., an increase in probability) over the green/yellow/orange space of interest at the expense of the remaining parameter space. Note, however, none of the colors have been removed.

B. Relation to Minimax Sequential Prediction

Bayesian inference plays an important role in designing mixture forecaster with experts' advice in Section II-D. One implementable approach pertaining to a sublinear minimax regret is considered. A natural predicting strategy is based on computing a weighted average of experts' predictions as illustrated in Fig. 3 (b). Since our goal is to minimize the regret, it is reasonable to decide the weights according to the regret up to time $t-1$. For example, if the regret is large, then we give more weight to the corresponding expert, and vice versa. That is, we weight more those experts whose cumulative loss is small, and thus we regard the weight as an arbitrary function of the experts' loss. This leads to the class of *mixture* forecasters that are more easily implementable while still satisfying sublinear regret.

Suppose that each expert e is uniquely specified by a parameter θ such as $e(y_t|y^{t-1}, \theta)$. We also define for notational convenience: $l(e, y^{t-1}) \triangleq l(\theta, y^{t-1})$. We define a weight $w_{\theta,t}$ for each expert e at time t as

$$w_{\theta,t} = \frac{e^{-\eta L_{t-1}(\theta, y^{t-1})}}{\int_\theta e^{-\eta L_{t-1}(\theta, y^{t-1})} d\theta} \tag{21}$$

where η is a positive number. So the weight of an expert e depends on its past performance $L_{t-1}(\theta, y^{t-1})$, and implies we listen more to the advice of the experts, whose recent loss functions are relatively small.

To define the mixture forecaster, a non-negative number $\pi_0(\theta) \geq 0$ is assigned to each expert such that $\int_\theta \pi_0(\theta) d\theta = 1$ as a prior information. Then the mixture forecaster becomes a weighted average of experts' predictions, which is represented by

$$p^*_{\mathsf{E}}(y_t|y^{t-1}) = \int_\theta \pi_0(\theta) e(y_t|y^{t-1}, \theta) w_{\theta,t-1} d\theta. \tag{22}$$

When we select $\eta = 1$, it is expressed by

$$\begin{aligned} p^*_{\mathsf{E}}(y_t|y^{t-1}) &= \frac{\int_\theta \pi_0(\theta) e(y_t|y^{t-1}, \theta) e(y^{t-1}|\theta) d\theta}{\int_\theta \pi_0(\theta) e(y^{t-1}|\theta) d\theta} \quad (23) \\ &= \int_\theta \pi_{t-1}(\theta) e(y_t|y^{t-1}, \theta) d\theta \quad (24) \end{aligned}$$

where

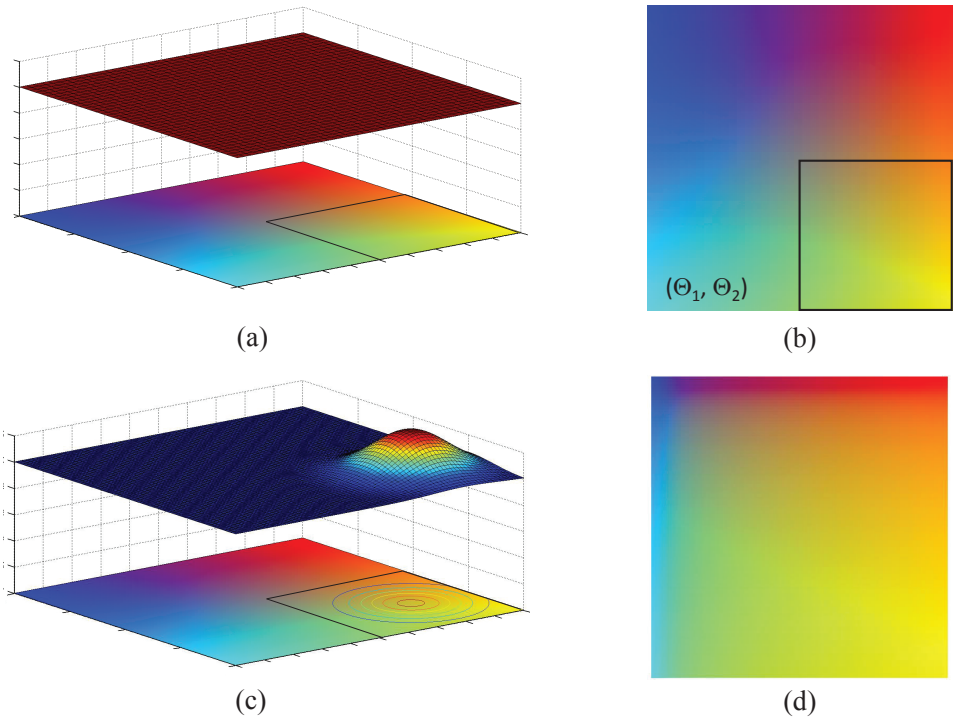

Fig. 5. An example of Bayesian inference with optimal maps in 2-dimensional compact parameter space. (a) Uniform prior on the parameters in the space. The red rectangular on top represents a uniform prior on the parameters. Different colors on bottom represent different values in the parameter space, i.e., different experts. (b) Distribution of the parameters under the uniform prior. The boxed-in region represents an area of interest given a next outcome. (c) More weights on the area with high likelihood values given the outcome. The 2-dimensional distribution on top represents the likelihood values given the next outcome. The parameters on the area with high likelihood values will be more weighted. (d) The effect of the nonlinear optimal mapping on the parameter space.

$$\pi_{t-1}(\theta) = \frac{\pi_0(\theta)e(y^{t-1}|\theta)}{\int_\theta \pi_0(\theta)e(y^{t-1}|\theta)d\theta}. \tag{25}$$

The mixture forecaster in (24) builds a predictive strategy as a weighted average of experts' advice and each weight at time t is determined by a posterior probability of expert given the observation up to time $t-1$. It is a natural way to combine the experts' advice, since the posterior distribution represents one's state of knowledge about each expert. It can also be calculated by efficient Bayesian inference method with optimal transport and convex optimization as described in Section III A. Using the law of conditional probability it can be rewritten as

$$p^*_{\mathsf{E}}(y_t|y^{t-1}) = \frac{e(y_t|y^{t-1},\theta)\pi_{t-1}(\theta)}{\pi_t(\theta)}. \tag{26}$$

The mixture forecaster in (26) is a 'good' predictor satisfying $R_T(p^*_{\mathsf{E}}, \mathsf{E}) = o(T)$. When we think of π_0 as a *prior* on E, then this is a Bayesian approach, and there is a natural way to select the prior based on Jeffrey's prior [60]. Under general conditions, Jeffrey's prior, denoted as $f^*_\theta(u)$, is the unique prior, for which minimax optimality holds [48]. It is given by

$$f^*_\theta(u) \propto \sqrt{\det(J(u))} \tag{27}$$

where $J(u)$ is Fisher information with respect to the likelihood function [15]. Jeffrey's prior is log-concave for the point process GLM of neural spiking activity. Thus, performing inference on this class of models with minimax optimal regret is efficient.

With these dynamic, time-varying predictions from expert advice, we separately compute the sequential predictor $p^*_{\mathsf{E}}(y_t|\mathcal{H}^{t-1})$ and $p^*_{\mathsf{E}}(y_t|\mathcal{H}^{t-1}_X)$ and then compute

$$C_{X\to Y}(t) \triangleq D\left(p^*_{\mathsf{E}}(y_t|\mathcal{H}^{t-1})||p^*_{\mathsf{E}}(y_t|\mathcal{H}^{t-1}_X)\right) \tag{28}$$

at time t. This provides a time-varying measure of causality where each individual predictor at time t is computed efficiently with our Bayesian inference methodology and minimax integration of each expert's advice.

IV. Applications

In this section, we demonstrate the application of the aforementioned methodologies to the analysis of simultaneously recorded spiking activity from multiple neurons. Methods based on reduction in loss function were used to infer the causal network of ensemble neural spiking activity using point process model. This approach was tested first on simulated data, and subsequently applied to neural activity recorded from the primary motor cortex (M1) of a monkey. Some examples of the approximate estimated network topologies with reduced complexity are shown. Moreover, a time-varying causal inference extension of our methodology was performed on the same data using the sequential prediction framework.

A. Network Analysis of Ensemble Neural Spiking Activity

A general framework for analyzing the causality network between multiple neural processes was introduced based on reduction in loss function in the subsection II-B. In this subsection, we will show how this framework is applied to estimate the causality network between multiple neural spike trains using the point process models [19], [20]. The discrete, all-or-nothing nature of a sequence of action potentials together with their stochastic structure suggests that neural spike trains may be regarded as point processes [61]–[63]. Let $N_{i,t}$ denote the sample path that counts the number of spikes of neuron i in the time interval $(0,t]$ for $t \in (0,T]$ for $i = 1,...,M$ recorded neurons. A point process model of a spike train for neuron i can be completely characterized by its conditional intensity function (CIF), $\lambda_i(t|\mathcal{H}^t)$, defined as

$$\lambda_i(t|\mathcal{H}^t) = \lim_{\Delta \to 0} \frac{\Pr[N_{i,t+\Delta} - N_{i,t} = 1|\mathcal{H}^t]}{\Delta} \tag{29}$$

where $\mathcal{H}^t$ denotes the spiking history of all the neurons in the ensemble up to time t [64]. The CIF represents the instantaneous firing probability and serves as a fundamental block for constructing the likelihoods and probability distributions for point process data. It is a history dependent function, and reduces to a Poisson process if it is independent of the history. To simplify the notation we denote $\lambda_i(t|\mathcal{H}^t)$ as $\lambda_i(t)$. In the GLM framework to model the relationship between the spiking activity and its covariates (the spiking history) [65], the logarithm of the CIF is modeled as a linear combination of the functions of the covariates that describe the neural activity dependencies, and thus is expressed as

$$\log \lambda_i(t) = \theta_{i,0} + \sum_{m=1}^{M} \theta_{i,m} \cdot \mathbf{h}_m^t. \tag{30}$$

Here, $\theta_{i,0}$ relates to a background level of the activity of neuron i, $\theta_{i,m}$ is a d-dimensional vector of parameters to relate the past spiking of neuron m to the future spiking of neuron i, and $\mathbf{h}_m^t$ is a d-dimensional vector whose each element represents the spikes in the spiking history of neuron m up to time t. The '$\cdot$' represents the dot product between vectors.

To test the causal interaction from neuron j to i, we developed two 'predictors', one class of point process GLM models that $P(N_i^T)$ that has the past of all neurons as the covariates for the CIF, and another class given by $\tilde{Q}(N_i^T)$ that has the past of all except for neuron j. The point process likelihood is given, up to a normalization constant, by [64]:

$$P(N_i^T) = \exp\left\{ \int_0^T \log \lambda_i(t) dN_{i,t} - \int_0^T \lambda_i(t) dt. \right\}. \tag{31}$$

Note that for $P(N_i^T)$, $\lambda_i(t)$ includes the past spiking of all neurons. The other point process likelihood, $\tilde{P}(N_i^T)$, is given by the same equation (31), but with $\lambda_i(t)$ replaced as $\tilde{\lambda}_i(t)$:

$$\log \tilde{\lambda}_i(t) = \tilde{\theta}_{i,0} + \sum_{\substack{m=1 \\ m \neq j}}^{M} \tilde{\theta}_{i,m} \cdot \mathbf{h}_m^t, \tag{32}$$

which excludes the effect of the past spiking of neuron j. Model parameters for $P(N_i^T;\theta_i)$ and for $\tilde{P}(N_i^T;\tilde{\theta}_i)$ were fitted by maximium likellihood and then the causality measure from j to i is defined using the expected value of reduction in log loss functions, which is given by

$$C_{j\to i} = \frac{1}{T}\mathbb{E}\left[\log \frac{P(N_i^T;\theta_{\mathbf{i}}^*)}{\tilde{P}(N_i^T;\tilde{\theta}_{\mathbf{i}}^*)}\right] \tag{33}$$

$$= \mathrm{D}(P(N_i^T;\theta_{\mathbf{i}}^*)||\tilde{P}(N_i^T;\tilde{\theta}_{\mathbf{i}}^*)). \tag{34}$$

If the history spiking of neuron j helps predict the spiking activity of neuron i, the directed information should be greater than zero, and then we say that neuron j 'Granger-causes' i [66]. The equality of the causally conditioned directed information holds when neuron j has no causal influence on i. Excitatory and inhibitory influences from neuron j to i are distinguished by the sign of $\sum_\tau \theta_{i,j}(\tau)$ in the equation (30), which represents an averaged influence of the past spiking of j on i.

The directed information measure, $C_{j\to i}$, of the equation (33) given by the log-likelihood ratio provides an indication of the relative strength of causal interaction, but little insight into whether or not it is statistically significant. We use the goodness-of-fit (GOF) statistics based on the log-likelihood ratio test to address this issue. We denote the deviance obtained using the model parameter θ_i^* as $D0$, and the deviance obtained using $\tilde{\theta}_i^*$ as D1. The deviance difference between two models is equivalent to 2 times log-likelihood ratio given by $\Delta D = D0 - D1 = 2C_{j\to i}$ [67]. If both models describe the data well, then the deviance difference may be asymptotically described by $\Delta D \approx \chi_d^2$ where d is the difference in dimensionality of two models [67], [65]. Thus, if the value of ΔD is consistent with the χ_d^2 distribution, the causal influence is not statistically significant. On the contrary, if the value of ΔD is in the critical region, i.e., greater than the upper tail $100(1-\alpha)\%$ of the χ_d^2 where α determines false positive rates, then the causal influence is determined as statistically significant. When we use the common statistical thresholds to detect statistically significant causal interactions between possibly many pairs of neurons, we will suffer from unacceptably large false positives [68]. Here we used a multiple-hypothesis testing error measure called false discovery rate (FDR) to address the multiple comparison problem [69].

In Fig. 6, the network estimates of cross-correlation based and causal inference methods were compared using synthetically generated neural spike trains. Simulated spike train data were generated based on the three-neuron network of Fig. 6 (a). The blue and red arrows represented the inhibitory (causing a decrease in firing rate) and excitatory (causing an increase in firing rate) interactions, respectively. In the network, there were directed excitatory interactions from neuron 1 to 2 and 2 to 3 but there was no excitatory interaction from 1 to 3. The neurons had inhibitory effects on each other in a counterclockwise direction. Specific point process models that were used to generate spike trains based on this network were described in the Simulation section of [20]. Examples of generated neural spike trains during the first 5 sec are

illustrated in Fig. 6 (b). It is hard to visually estimate the underlying network between neurons from this raster plot. Fig. 6 (c) shows the estimated network using the cross-correlogram method. Note that this method determined a direct excitatory connection from neuron 1 to 3 as well as direct excitatory connections from 1 to 2 and 2 to 3. However, truly, there was no direct excitatory connection from 1 to 3. It failed to detect all inhibitory interactions either. Fig. 6 (d) presents the estimated network using the causal inference method. The estimated pattern matched the original network exactly. This estimated network did not show the excitatory connection from neuron 3 to 1 and succeeded in detecting all inhibitory interactions.

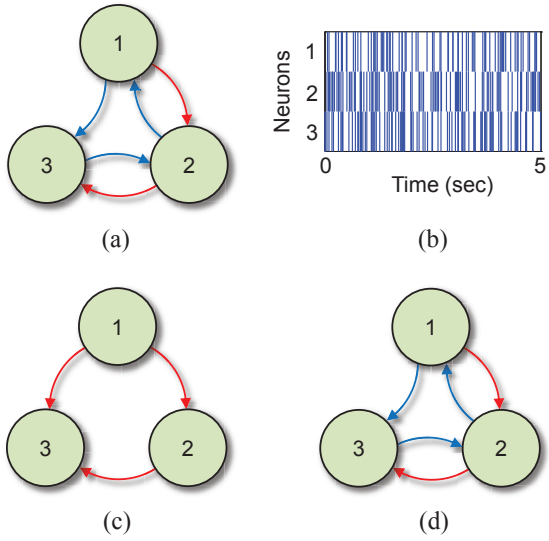

Fig. 6. Comparison of network estimates using synthetic neural spike trains: (a) Three-neuron network used to generate synthetic neural spike trains. The blue and red arrows represent the inhibitory (causing a decrease in firing rate) and excitatory (causing an increase in firing rate) interactions, respectively. (b) Examples of generated neural spike trains during the first 5 sec. (c) Network estimate based on cross-correlation. (d) Network estimate based on causal inference.

B. Directed Graph Representations of Ensemble Neural Data

Multiple neural spike trains were simultaneously recorded from the M1 of a monkey during visuomotor task; The monkey was trained to move a cursor on a horizontal screen that was aligned to the monkey's hand to the position of a target. When the monkey successfully reached the current target, a new target was displayed at a random location within a workspace while the current target disappeared. The monkey received a juice reward after successfully acquiring five or seven consecutive targets. Multiple single unit spiking activities from the M1 in a monkey were then recorded using an Utah microelectrode array. More details can be found in [70].

Fig. 7 (a) depicts a directed network graph estimated by applying the causal inference method in (34) to 37 high firing neurons recorded in the M1 of the monkey [42]. The blue arrow represents the dominant direction of the edges, which is along the rostro-caudal axis (or anterior-posterior axis), which is a straight line as an axis that has at the upper end the nose, followed by the tail. This direction is consistent with the beta wave propagation direction of local field potentials in the motor cortex, which researchers surmise mediates the information transfer between different brain regions [71]. Fig. 7 (b) illustrates that the directed tree approximation methodology from Section II-C enabled more succinct visualization with a directed tree approximation, and but still preserved relevant information for analysis of mechanistic neurobiological phenomena.

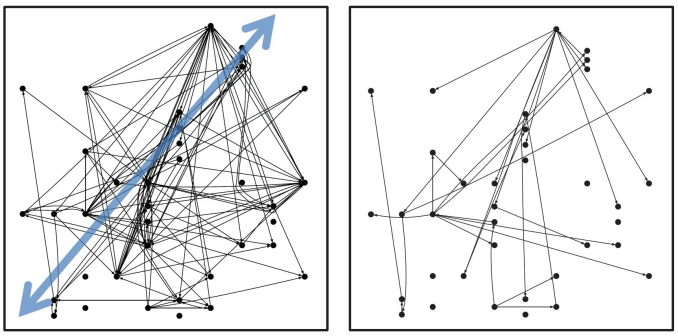

Fig. 7. Graphical structures of statistically significant directed information and its causal dependence tree approximation. The blue arrow depicts a dominant orientation of the edges. The relative positions of neurons correspond to those of the recording electrodes. (a) Graphical structure of directed information values. (b) Causal dependence tree approximation. The figure is from Quinn et al., 2013 [42].

C. Dynamic Analysis of Ensemble Neural Data

In this section, we demonstrate the use of minimax time-varying causality measures from Section III-B elucidate how simultaneously recorded, motor cortical neurons in non-human primates spatially coordinates their spike activity during a visuomotor task using a two-link exoskeleton manipulandum [72].

Fig. 8 shows the spatiotemporal patterns of network connectivity during the visuomotor task obtained using static and dynamic methods, respectively. The top plot presents the network connectivity estimated using the static causal inference method at different timings in relation to the visual cue onset at 0 ms: time window 1 for $[-100, 50]$ ms, 2 for $[50, 200]$ ms, and 3 for $[200, 350]$ ms, respectively [73]. As shown, most functional connectivity was detected for time window 2 than other two intervals.

Fig. 8 (b) shows the time-varying causal interactions between some pairs of neurons at every 1 ms. It could track time-varying causality networks and observed more interactions after visual stimulus, which is consistent with the findings in Fig. 8 (a). It also provided the timing information on when the causal influences occurred and disappeared in relation to visual cue onset. Fig. 8 (b) elucidates more details about the dynamic, non-stationary causal influences, and is consistent with the static analysis (for example, the directed edge $8 \rightarrow 1$ is absent in the first panel, and present in the next two).

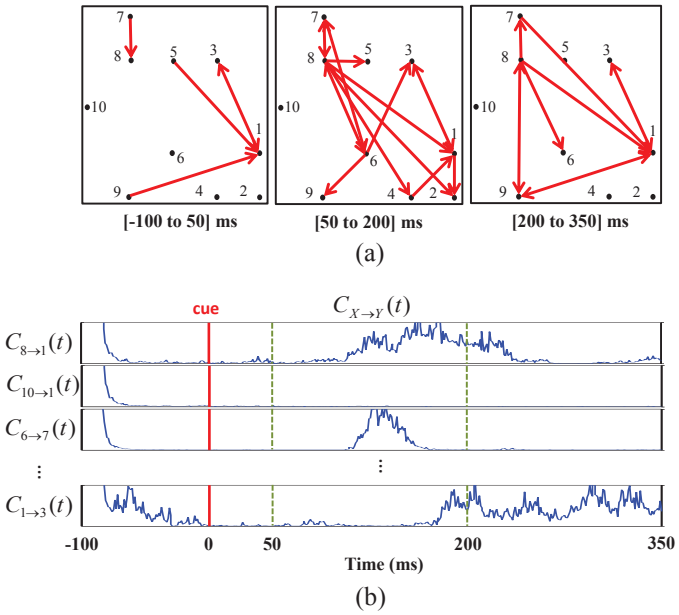

Fig. 8. Spatiotemporal patterns of network connectivity. The dynamics of effective network in primary motor cortex of a monkey using real neural spike train data is tracked: (a) Three snapshots of time-varying networks are obtained every 150 ms using conventional approach. Red arrow represents functional connections. The black dots represent the relative positions of the electrodes on the array where the neurons were detected. (b) The evolution of network dynamics is tracked every 1 ms using the proposed approach. Red vertical bar represents the onset time of visual cue.

V. Conclusion

We have developed a framework to develop scalable, multimodal methods that address the key challenges that are of increasing importance in neuroscience data analysis. It was our attempt to balance having classical statistics, information theory, and control theory underpinnings with the agility to be applicable to specific neuroscientific scenarios where physiological constraints can be embedded within the framework.

Our approach based on general loss function perspective enables us to extend the analysis of neuroscience data to high-dimensionality, dynamics, and harness robustness to uncertainty. The use of optimal transport theory provided us a tool for efficiently solving Bayesian inference problems with convex optimization, which by itself is of importance in many statistical analysis settings, but more specifically, enables our ability to develop robust methods for dynamic analysis of ensemble neural processes. Although our examples were specific to causal inference for ensemble neural spiking activity, we note that our general purpose exposition elucidates how it can be applicable more generally.

In the future, we believe that uncertainty due to large dynamic data sets will lead to new theoretical statistics and algorithms that are specifically tailored to these massive, dynamic data sets. What will be important is a balance developing theoretical frameworks that have a 'forest' perspective and have common underpinnings, with having the features and extensibility to be applicable to neurophysiology of interest. In addition, we believe that the uncertainty due to these massive datasets will lead to the need for novel methods of performing sequential experimental design: providing canonical frameworks to extract information from experiments, characterize uncertainty, and if necessary provide suggestions on subsequent interventional experiments to refine uncertainty as efficiently as possible. There is reason to suggest that newly developed principles and algorithms that lie at the intersection of (i) sequential transmission of a message point in a continuum over a noisy channel with feedback [74]–[76] and (ii) 'observability' and 'filter stability' in stochastic systems [77]–[80] can play an important role in this setting.

Acknowledgment

The authors would like to thank N. G. Hatsopoulos for providing neural spike train data, and D. Mesa and R. Ma for providing useful comments and figures.

References

[1] M. A. Nicolelis, *Methods for neural ensemble recordings*. CRC press, 2007.
[2] A. P. Alivisatos, M. Chung, G. M. Church, R. J. Greenspan, M. R. Roukes, and R. Yuste, "The brain activity map project and the challenge of functional connectomics," *Neuron*, vol. 74, no. 6, pp. 970–974, Jun 2012.
[3] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature Neuroscience*, vol. 7, pp. 456–461, 2005.
[4] "Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Interim Report," National Institute of Health, Interim Report, Sep. 2013.
[5] B. He, T. Coleman, G. M. Genin, G. Golver, X. Hu, N. Johnson, T. Liu, S. Makeig, P. Sajda, and K. Ye, "Grand challenges in mapping the human brain: NSF workshop report," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 11, pp. 2983–2992, 2013.
[6] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, 2006.
[7] J. C. Kao, S. D. Stavisky, D. Sussillo, P. Nuyujukian, and K. V. Shenoy, "Information systems opportunities in brain-machine interfaces," *Proceedings of the IEEE*, 2014.
[8] O. Milenkovic, G. Alterovitz, G. Battail, T. Coleman, J. Hagenauer, S. Meyn, N. Price, M. Ramoni, I. Shmulevich, and W. Szpankowski, "Introduction to the special issue on information theory in molecular biology and neuroscience," *IEEE Transactions on Information Theory*, vol. 56, pp. 649–652, 2010.
[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.
[10] C. Villani, *Topics in optimal transportation*. AMS, 2003.
[11] A. Abbott, "Brain-simulation and graphene projects win billion-euro competition," *Nature News*, 2013.
[12] A. H. Ropper *et al.*, "Brain in a box," *The New England journal of medicine*, vol. 367, no. 26, pp. 2539–2541, 2012.
[13] C. D. Brody, "Correlations without synchrony," *Neural computation*, vol. 11, no. 7, pp. 1537–1551, 1999.
[14] G. L. Gerstein and D. H. Perkel, "Simultaneously recorded trains of action potentials: analysis and functional interpretation," *Science*, vol. 164, no. 3881, pp. 828–830, 1969.
[15] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[16] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," *Nature Neuroscience*, vol. 14, no. 2, pp. 139–142, 2011.
[17] D. Kollar and N. Friedman, *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.
[18] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality," *Proc Natl Acad Sci*, vol. 101, pp. 9849 – 9854, 2004.

[19] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[20] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, March 2011.

[21] N. Wiener, *The theory of prediction*. In: Beckenbach EF, editors. Modern mathematics for engineers. New York: McGraw-Hill., 1956.

[22] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[23] R. Dahlhaus and M. Eichler, "Causality and graphical models in time series analysis," *Oxford Statistical Science Series*, pp. 115–137, 2003.

[24] R. Dahlhaus, "Graphical interaction models for multivariate time series 1," *Metrika*, vol. 51, no. 2, pp. 157–172, 2000.

[25] M. Eichler, "Granger causality and path diagrams for multivariate time series," *Journal of Econometrics*, vol. 137, no. 2, pp. 334–353, 2007.

[26] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[27] J. Massey, "Causality, feedback and directed information," in *Intl. Symp. on Info. Th. and its Applications*. Citeseer, 1990, pp. 27–30.

[28] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, University of Manitoba, Canada, 1998.

[29] P.-O. Amblard and O. J. Michel, "The relation between granger causality and directed information theory: a review," *Entropy*, vol. 15, no. 1, pp. 113–143, 2012.

[30] P. Amblard and O. Michel, "On directed information theory and Granger causality graphs," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

[31] ——, "Relating Granger causality to directed information theory for networks of stochastic processes," *Arxiv preprint arXiv:0911.2873*, 2009.

[32] C. Gourieroux, A. Monfort, and E. Renault, "Kullback causality measures," *Annals of Economics and Statistics*, pp. 369–410, 1987.

[33] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series (corresp.)," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 598–601, 1987.

[34] C. W. Granger, "Some recent development in a concept of causality," *Journal of econometrics*, vol. 39, no. 1, pp. 199–211, 1988.

[35] L. Barnett, A. Barrett, and A. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

[36] L. Barnett and T. Bossomaier, "Transfer entropy as a log-likelihood ratio," *Physical Review Letters*, vol. 109, no. 13, p. 138105, 2012.

[37] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, pp. 461–464, 2000.

[38] S. Kim and E. N. Brown, "A general statistical framework for assessing granger causality," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 2222–2225.

[39] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[40] M. M. Churchland, B. M. Yu, M. Sahani, and K. V. Shenoy, "Techniques for extracting single-trial activity patterns from large-scale neural recordings," *Current opinion in neurobiology*, vol. 17, no. 5, pp. 609–618, 2007.

[41] M. Y. Byron, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity," *Journal of neurophysiology*, vol. 102, no. 1, pp. 614–635, 2009.

[42] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Efficient methods to compute optimal tree approximations of directed information graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 12, pp. 3173–3182, 2013.

[43] P. Humblet, "A distributed algorithm for minimum weight directed spanning trees," *IEEE Transactions on Communications*, vol. 31, no. 6, pp. 756–762, 1983.

[44] L. Zhao, H. Permuter, Y. Kim, and T. Weissman, "Universal estimation of directed information," in *2010 IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2010, pp. 1433–1437.

[45] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[46] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[47] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[48] P. Grünwald, *The minimum description length principle*. The MIT Press, 2007.

[49] T. P. Coleman, M. Yanike, W. A. Suzuki, and E. N. Brown, "A mixed-filter algorithm for dynamically tracking learning from multiple behavioral and neurophysiological measures," *The dynamic brain: an exploration of neuronal variability and its functional significance*, pp. 1–16, 2011.

[50] Y. Ahmadian, J. Pillow, and L. Paninski, "Efficient markov chain monte carlo methods for decoding neural spike trains," *Neural Computation*, vol. 23, no. 1, pp. 46–96, 2011.

[51] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, 2008.

[52] T. El Moselhy and Y. Marzouk, "Bayesian inference with optimal maps," *Journal of Computational Physics*, vol. 231, no. 23, pp. 7815–7850, 2012.

[53] S. Kim, R. Ma, D. Mesa, and T. P. Coleman, "Efficient Bayesian Inference Methods via Convex Optimization and Optimal Transport," in *IEEE International Symposium on Information Theory (ISIT)*, 2013.

[54] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[55] R. Ghanem and P. D. Spanos, *Stochastic finite elements: a spectral approach*. DoverPublications, 2003.

[56] D. Xiu and G. Karniadakis, "The Wiener-Askey polynomial chaos for stochastic differential equations," *SIAM journal on scientific computing*, vol. 24, no. 2, pp. 619–644, 2003.

[57] N. Wiener, "The homogeneous chaos," *American Journal of Mathematics*, vol. 60, no. 4, pp. 897–936, 1938.

[58] I. H. Stevenson, J. M. Rebesco, N. G. Hatsopoulos, Z. Haga, L. E. Miller, and K. P. Kording, "Bayesian inference of functional connectivity and network structure from spikes," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 203–213, 2009.

[59] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.

[60] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 2002.

[61] E. N. Brown, "Theory of point processes for neural systems," *Methods and models in neurophysics*, pp. 691–726, 2005.

[62] E. N. Brown, R. Barbieri, U. T. Eden, and L. M. Frank, "Likelihood methods for neural spike train data analysis," *Computational neuroscience: A comprehensive approach*, pp. 253–286, 2003.

[63] R. E. Kass, V. Ventura, and E. N. Brown, "Statistical issues in the analysis of neuronal data," *Journal of Neurophysiology*, vol. 94, no. 1, pp. 8–25, 2005.

[64] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure*. Springer, 2007, vol. 2.

[65] A. J. Dobson, *An introduction to generalized linear models*. CRC press, 2010.

[66] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 1969.

[67] P. MacCullagh and J. A. Nelder, *Generalized linear models*. CRC press, 1989, vol. 37.

[68] R. G. Miller, *Simultaneous statistical inference*. McGraw-Hill New York, 1966.

[69] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[70] J. Reimer and N. G. Hatsopoulos, "Periodicity and evoked responses in motor cortex," *The Journal of Neuroscience*, vol. 30, no. 34, pp. 11 506–11 515, 2010.

[71] D. Rubino, K. A. Robbins, and N. G. Hatsopoulos, "Propagating waves mediate information transfer in the motor cortex," *Nature neuroscience*, vol. 9, no. 12, pp. 1549–1557, 2006.

[72] S. H. Scott, "Apparatus for measuring and perturbing shoulder and elbow joint positions and torques during reaching," *Journal of Neuroscience Methods*, vol. 89, no. 2, pp. 119 – 127, 1999.

[73] S. Kim, K. Takahashi, N. G. Hatsopoulos, and T. P. Coleman, "Information transfer between neurons in the motor cortex triggered by visual cues," in *IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2011, pp. 7278–7281.

[74] C. Omar, A. Akce, M. Johnson, T. Bretl, R. Ma, E. Maclin, M. McCormick, and T. Coleman, "A feedback information-theoretic approach to the design of brain–computer interfaces," *Intl. Journal of Human-Computer Interaction*, vol. 27, no. 1, pp. 5–23, 2010.
[75] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, 2011.
[76] R. Ma and T. Coleman, "Generalizing the posterior matching scheme to higher dimensions via optimal transportation," in *Allerton Conference*, 2011.
[77] R. Van Handel, "Observability and nonlinear filtering," *Probability theory and related fields*, vol. 145, no. 1-2, pp. 35–74, 2009.
[78] R. van Handel, "When do nonlinear filters achieve maximal accuracy?" *SIAM Journal on Control and Optimization*, vol. 48, no. 5, pp. 3151–3168, 2009.
[79] ——, "Nonlinear filtering and systems theory," in *19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010)*, 2010.
[80] S. Gorantla and T. P. Coleman, "Equivalence between reliable feedback communication and nonlinear filter stability," in *2011 IEEE International Symposium on Information Theory (ISIT)*, 2011, pp. 164–168.

PLACE PHOTO HERE

Sanggyun Kim (S'02-M'09) received the Ph.D. degree in electrical engineering and computer science (EECS) from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2008. He joined the Statistical Learning for Signal Processing Laboratory in EECS at KAIST in 2001. From January 2009 to June 2010 he was a postdoctoral researcher at the Department of Brain and Cognitive Sciences (BCS), Massachusetts Institute of Technology (MIT), Cambridge, MA.

He is currently a postdoctoral scholar in Bioengineering at the University of California San Diego (UCSD), USA. His research interests include statistical signal processing, machine learning and information theory with applications to neuroscience and multimedia data.

PLACE PHOTO HERE

Christopher J. Quinn (S'10) received a B.S. in Engineering Physics from Cornell University and M.S. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2008 and 2010 respectively.

He is currently a Ph.D. student in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. His research interests include information theory, network science, and statistical signal processing.

PLACE PHOTO HERE

Negar Kiyavash (S'99-M'06) is an assistant professor in the Department of Industrial and Enterprise Systems Engineering (ISE) at the University of Illinois at Urbana-Champaign, USA. She received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2006. Her research interests include information theory and statistical signal processing with applications to security and network inference. Dr. Kiyavash is a recipient of the NSF CAREER and AFOSR YIP awards.

PLACE PHOTO HERE

Todd P. Coleman (S'01-M'05-SM'11) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 2005.

He was a postdoctoral scholar in neuroscience at MIT and MGH during the 2005-2006 academic year. He was an Assistant Professor in ECE and Neuroscience at the University of Illinois from 2006-2011. He is currently an Associate Professor in Bioengineering and director of the Neural Interaction Laboratory at UCSD. His research is highly interdisciplinary and lies at the intersection of bio-electronics, neuroscience, medicine, and applied mathematics. Dr. Coleman is a science advisor for the Science & Entertainment Exchange (National Academy of Sciences).