Robust Estimation of Self-Exciting Point Process Models with Application to Neuronal Modeling

Abbas Kazemipour, *Student Member, IEEE,* Min Wu, *Fellow, IEEE,* and Behtash Babadi, *Member, IEEE*

***Abstract*—We consider the problem of estimating discrete self-exciting point process models from limited binary observations, where the history of the process serves as the covariate. We analyze the performance of two classes of estimators, namely the ℓ_1-regularized maximum likelihood and greedy estimators, for a canonical self-exciting point process and characterize the sampling tradeoffs required for stable recovery in the non-asymptotic regime. Our results extend those of compressed sensing for linear and generalized linear models with i.i.d. covariates to point processes with highly inter-dependent covariates. We further provide simulation studies as well as application to real spiking data from mouse's lateral geniculate nucleus and ferret's retinal ganglion cells which agree with our theoretical predictions.**

***Index Terms*—compressed sensing, point process models, sparsity, spontaneous activity, neural signal processing.**

I. Introduction

The theory of compressed sensing (CS) has provided a novel framework for measuring and estimating statistical models governed by sparse underlying parameters [1]–[6]. In particular, for linear models with random covariates and sparsity of the parameters, the CS theory provides sharp trade-offs between the number of measurement, sparsity, and estimation accuracy. Typical theoretical guarantees imply that when the number of random measurements are roughly proportional to sparsity, then stable recovery of these sparse models is possible.

Beyond those described by linear models, observations from binary phenomena form a large class of data in natural and social sciences. Their ubiquity in disciplines such as neuroscience, physiology, seismology, criminology, and finance has urged researchers to develop formal frameworks to model and analyze these data. In particular, the theory of point processes provides a statistical machinery for modeling and prediction of such phenomena. Traditionally, these models have been employed to predict the likelihood of self-exciting processes such as earthquake occurrences [7], [8], but have recently found applications in several other areas. For instance, these models have been used to characterize heart-beat dynamics [9], [10] and violence among gangs [11]. Self-exciting point process models have also found significant applications in analysis of neuronal data [12]–[18].

In particular, point process models provide a principled way to regress binary spiking data with respect to extrinsic stimuli and neural covariates, and thereby forming predictive statistical models for neural spiking activity. Examples include place cells in the hippocampus, spectro-temporally tuned cells in the primary auditory cortex, and spontaneous retinal or thalamic neurons spiking under tuned intrinsic frequencies. When fitted to neuronal data, these models exhibit three main features: first, the underlying parameters are nearly sparse or compressible; second, the covariates are often highly structured and correlated; and third, the input-output relation is highly nonlinear. Therefore, the theoretical guarantees of compressed sensing do not readily translate to prescriptions for point process estimation.

Self-exciting point processes have been utilized in neuroscience in order to assess the functional connectivity of neuronal ensembles. Estimation is typically carried out by regularized Maximum Likelihood (ML) estimation, where empirical methods, such as cross-validation, are employed to adjust regularization [19]. In the signal processing and information theory literature, sparse signal recovery under Poisson statistics has been considered in [20] with application to the analysis of ranking data. In [21], a similar setting has been studied, with motivation from imaging by photon-counting devices. Finally, in theoretical statistics, high-dimensional M-estimators with decomposable regularizers, such as the ℓ_1-norm, have been studied for Generalized Linear Models (GLM) [22].

A key underlying assumption in the existing theoretical analysis of estimating point process models is the independence and identical distribution (i.i.d.) of covariates. This assumption does not hold for self-exciting point processes, since the history of the process takes the role of the covariates. Nevertheless, regularized ML estimators show remarkable performance in fitting point process models to neuronal data with history dependence and highly non-i.i.d. covariates. In this paper, we close this gap by presenting new results on robust estimation of compressible point process models, relaxing the assumptions of i.i.d. covariates and exact sparsity common in CS.

In particular, we will consider a canonical discrete point process model and will analyze two classes of estimators for its underlying parameters: the ℓ_1-regularized maximum likelihood and greedy estimators. We will present theoretical guarantees that extend those of CS theory and characterize fundamental trade-offs between the number of measurements, model compressibility, and estimation error of point processes in the non-asymptotic regime. Our results reveal that when the number of measurements scale sub-linearly with the product of the ambient dimension and a generalized measure of sparsity (modulo logarithmic factors), then stable recovery of the underlying models is possible, even though the covariates

The authors are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: kaazemi@umd.edu; minwu@umd.edu; behtash@umd.edu).

Corresponding author: B. Babadi (e-mail: behtash@umd.edu).

solely depend on the history of the process. We will further discuss the extensions of these results to more general classes of point process models. Finally, we will present applications to simulated as well as real data from neuronal activity in mouse's lateral geniculate nucleus and ferret's retinal ganglion cells, which agree with our theoretical predictions. Aside from their theoretical significance, our results are particularly important in light of the technological advances in neural prostheses, which require robust neuronal system identification based on compressed data acquisition.

The rest of the paper is organized as follows: In Section II, we present our notational conventions, preliminaries and problem formulation. In Section III, we discuss the estimation procedures and state the main theoretical results of this paper. Section IV provides numerical simulations as well as application to real data. In Section V, we discuss the implications of our results and outline future research directions. Finally, we present the proofs of the main theoretical results, discuss their extensions, and give a brief background on relevant statistical tests in Appendices A, B, and C, respectively.

II. Premiliminaries and Problem Formulation

A. Self-Exciting Point Process Models

We first give a brief introduction to self-exciting point process models (see [23] for a detailed treatment). We will use the following notation throughout the paper. Parameter vectors are denote by bold-face Greek letters. For example, $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]'$ denotes a p-dimensional parameter vector, with $[\cdot]'$ denoting the transpose operator. We also use the notation x_i^j to represent the $(j-i+1)$-dimensional vector $[x_i, x_{i+1}, \cdots, x_j]'$ for any $i, j \in \mathbb{Z}$ with $i \leq j$.

We consider a sequence of observations in the form of binary spike trains obtained by discretizing continuous-time observations (e.g. electrophysiology recordings), using bins of length Δ. We assume that not more than one event fall into any given bin. In practice, this can always be achieved by choosing Δ small enough. The binary observation at bin i is denoted by x_i. The observation sequence can can be modeled as the outcome of conditionally independent Poisson or Bernoulli trials, with a spiking probability given by $\mathbb{P}(x_i = 1) =: \lambda_{i|H_i}$, where $\lambda_{i|H_i}$ is the spiking probability at bin i given the history of the process H_i up to bin i.

These models are widely-used in neural data analysis and are motivated by the continuous time point processes with history dependent conditional intensity functions [23]. For instance, given the history of a continuous-time point process H_t up to time t, a conditional intensity of $\lambda(t|H_t) = \lambda$ corresponds to the homogeneous Poisson process. As another example, a conditional intensity of $\lambda(t|H_t) = \mu + \int_{-\infty}^{t} \theta(t-\tau) dN(\tau)$ corresponds to a process known as the Hawkes process [24] with base-line rate μ and history dependence kernel $\theta(\cdot)$. Under the assumption of the orderliness of a continuous-time point process, a discretized approximation to these processes can be obtained by binning the process by bins of length Δ, and defining the spiking probability by $\lambda_i := \lambda(i\Delta|H_{i\Delta})\Delta + o(\Delta)$. In this paper, we consider discrete point processes characterized by the spiking probability $\lambda_{i|H_i}$, which are either inherently discrete or employed as an approximation to continuous-time point process models.

Throughout the rest of the paper, we drop the dependence of $\lambda_{i|H_i}$ on H_i to simplify notation, denote it by λ_i and refer to it by spiking probability. Given the sequence of observed data x_1^n, the negative log-likelihood function under the Poisson statistics can be expressed as:

$$\mathfrak{L}(\boldsymbol{\theta}) := -\frac{1}{n} \sum_{i=1}^{n} \left[x_i \log \lambda_i - \lambda_i \right], \tag{1}$$

where λ_i. Similarly, under the Bernoulli statistics, the negative log-likelihood takes the following form:

$$\mathfrak{L}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} x_i \log \lambda_i + (1 - x_i) \log(1 - \lambda_i). \tag{2}$$

When the discrete process viewed as an approximations to a continuous-time process, these log-likelihood functions are known as the Jacod log-likelihood approximations [23]. We will present our analysis for the negative log-likelihood given by (1), but our results can be extended to other statistics including (2) (See Appendix A for a discussion of extensions to other models).

Throughout this paper x_{-p+1}^n will be considered as the observed spiking sequence which will be used for estimation purposes. A popular class of models for λ_i is given by Generalized Linear Models (GLM). In its general form, a GLM consists of two main components: an observation model and an equation expressing some (possibly nonlinear) function of the observation mean as a *linear* combination of the covariates. In neural systems, the covariates consist of external stimuli as well as the history of the process. Inspired by spontaneous neuronal activity, we consider *fully* self-exciting processes, in which the covariates are only functions of the process history. As for a canonical discrete point process model inspired by the Hawkes process, we consider a GLM for which the spiking probability is a *linear* function of the process history:

$$\lambda_i := \mu + \boldsymbol{\theta}' x_{i-p}^{i-1}, \tag{3}$$

where μ is a positive constant representing the base-line rate, and $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]'$ is a parameter vector denoting the history dependence of the process. We refer to this process as the *canonical self-exciting process*. Other popular models in the computational neuroscience literature include the log-link model where $\lambda_i = \exp(\mu + \boldsymbol{\theta}' x_{i-p}^{i-1})$ and the logistic-link model where $\lambda_i = \frac{\exp(\mu + \boldsymbol{\theta}' x_{i-p}^{i-1})}{1 + \exp(\mu + \boldsymbol{\theta}' x_{i-p}^{i-1})}$. The parameter vector $\boldsymbol{\theta}$ can be thought of as the binary equivalent of autoregressive (AR) parameters in linear AR models.

When fitted to neuronal spiking data, the parameter vector $\boldsymbol{\theta}$ exhibits a degree of sparsity [19], [25]. That is, only certain lags in the history have a significant contribution in determining the statistics of the process. These lags can be thought of as the preferred or intrinsic delay in the spontaneous response of a neuron. To be more precise, for a sparsity level $s < p$, we denote by $S \subset \{1, 2, \cdots, p\}$ the support of the s highest elements of $\boldsymbol{\theta}$ in absolute value, and by $\boldsymbol{\theta}_S$ the best s-term approximation to $\boldsymbol{\theta}$. We also define

$$\sigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1 \tag{4}$$

and

$$\varsigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_2 \tag{5}$$

which capture the compressibility of the parameter vector $\boldsymbol{\theta}$ in the ℓ_1 and ℓ_2 sense, respectively. Note that by definition $\varsigma_s(\boldsymbol{\theta}) \leq \sigma_s(\boldsymbol{\theta})$. For a fixed $\xi \in (0,1)$, we say that $\boldsymbol{\theta}$ is (s,ξ)-*compressible* if $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$ [5]. Note that when $\xi = 0$, the parameter vector $\boldsymbol{\theta}$ is exactly s-sparse.

Finally, in this paper, we are concerned with the compressed sensing regime where $n \ll p$, i.e., the observed data has a much smaller length than the ambient dimension of the parameter vector. The main estimation problem of this paper is the following: *given observations x_{-p+1}^n from the canonical self-exciting process, the goal is to estimate the unknown baseline rate μ and the p-dimensional (s,ξ)-compressible history dependence parameter vector $\boldsymbol{\theta}$ in a stable fashion (where the estimation error is controlled) when $n \ll p$.*

III. Theoretical Results

In this section, we consider two estimators for $\boldsymbol{\theta}$, namely, the ℓ_1-regularized ML estimator and a greedy estimator, and present the main theoretical results of this paper on the estimation error of these estimators. Note that when μ is not known, the following results can be applied to the augmented parameter vector $[\mu, \boldsymbol{\theta}']'$. We analyze the case of known μ for simplicity of presentation.

A. ℓ_1-Regularized ML Estimation

Throughout the rest of the paper, we assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a closed convex feasible region for which $0 \leq \lambda_i \leq 1$ given by the conditions:

$$\begin{aligned} &1)\ 0 < \mathbf{1}'\boldsymbol{\theta} \leq c_1 < 1, \\ &2)\ 0 < \pi_{\min} \leq \mu - \|\boldsymbol{\theta}\|_1, \\ &3)\ \mu + \|\boldsymbol{\theta}\|_1 \leq \pi_{\max} < 1/2, \end{aligned} \tag{$\star$}$$

for some constants c_1, $\pi_{\min}$, and $\pi_{\max}$. These assumptions have been adopted mainly for technical reasons, and do not incur any loss of generality in practice (see Appendix A for details).

The natural estimator for the parameter vector is the ML estimator, which is widely used in neuronal modeling [25], and by virtue of (1) is given by:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathfrak{L}(\boldsymbol{\theta}), \tag{6}$$

In the regime of interest when $n \ll p$, the ML estimator is ill-posed and is typically regularized with a smooth norm. In order to capture the compressibility of the parameters, we consider the ℓ_1-regularized ML estimator:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \quad \mathfrak{L}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1. \tag{7}$$

where $\gamma_n > 0$ is a regularization parameter. It is easy to verify that the objective function and constraints in Eq. (7) are convex in $\boldsymbol{\theta}$ and hence $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$ can be obtained using standard numerical solvers. Note that the solution to (7) might not be unique. However, we will provide error bounds that hold for all possible solutions of (7), with high probability.

Input: $\mathfrak{L}(\boldsymbol{\theta}), s^\star$
Output: $\widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}}$
Initialization: { Start with the index set $S^{(0)} = \emptyset$; and the initial estimate $\widehat{\boldsymbol{\theta}}^{(0)}_{\mathsf{POMP}} = 0$
for $k = 1, 2, \cdots, s^\star$
  $j = \arg\max_i \left| \left(\nabla \mathfrak{L} \left(\widehat{\boldsymbol{\theta}}^{(k-1)}_{\mathsf{POMP}} \right) \right)_i \right|$
  $S^{(k)} = S^{(k-1)} \cup \{j\}$
  $\widehat{\boldsymbol{\theta}}^{(k)}_{\mathsf{POMP}} = \operatorname*{arg\,min}_{\mathrm{supp}(x) \subset S^{(k)}} \mathfrak{L}(\boldsymbol{\theta})$
end

TABLE I: Point Process Orthogonal Matching Pursuit (POMP)

It is known that ML estimates are asymptotically unbiased under mild conditions, and with p fixed, the solution converges to the true parameter vector as $n \to \infty$. However, it is not clear how fast the convergence rate is for finite n or when p is not fixed and is allowed to scale with n. This makes the analysis of ML estimators, and in general regularized M-estimators, very challenging [22]. Nevertheless, such an analysis has significant practical implications, as it will reveal sufficient conditions on n with respect to p as well as a criterion to choose γ_n, which result in a stable estimation of $\boldsymbol{\theta}$. Finally, note that we are fixing the ambient dimension p throughout the analysis. In practice, the history dependence is typically negligible beyond a certain lag and hence for a large enough p, point process models fit the data very well.

B. Greedy Estimation

Although there exist fast solvers to convex problems of the type given by Eq. (7), these algorithms are polynomial time in n and p, and may not scale well with high-dimensional data. This motivates us to consider greedy solutions for the estimation of $\boldsymbol{\theta}$. In particular, we will consider a generalization of the Orthogonal Matching Pursuit (OMP) [26], [27] for general convex cost functions. A flowchart of this algorithm is given in Table I, which we denote by the Point Process Orthogonal Matching Pursuit (POMP) algorithm. At each iteration, the component in which the objective function has the largest deviation is chosen and added to the current support. The algorithm proceeds for a total of $s^\star$ steps, resulting in an estimate with $s^\star$ components.

The main idea behind the generalized OMP is in the greedy selection stage, where the absolute value of the gradient of the cost function at the current solution is considered as the selection metric. Consider an estimate $\widehat{\boldsymbol{\theta}}^{(k-1)}$ at the $(k-1)$-st stage of the generalized OMP for a quadratic cost function of the form $\|\mathbf{b} - \mathbf{A}\boldsymbol{\theta}\|_2^2$, with $\mathbf{b}$ and $\mathbf{A}$ denoting the observation vector and covariates matrix, respectively. Then, the gradient takes the form $\mathbf{A}'(\mathbf{b} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{(k-1)})$ which is exactly the correlation vector between the residual error and the columns of $\mathbf{A}$ as in the original OMP algorithm.

C. Theoretical Guarantees

Recall that the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is assumed to be (s,ξ)-compressible, so that $\sigma_s(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1 = \mathcal{O}(s^{1-\frac{1}{\xi}})$, and the observed data are given by the vector

$x_{-p+1}^{n} \in \{0,1\}^{n+p-1}$, all in the regime of $s, n \ll p$. The main theoretical result regarding the performance of the ℓ_1-regularized ML estimator is given by the following theorem:

Theorem 1. *If $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, there exist constants d_1, d_2, d_3 and d_4 such that for $n > d_1 s^{2/3} p^{2/3} \log p$ and a choice of $\gamma_n = d_2\sqrt{\frac{\log p}{n}}$, any solution $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$ to (7) satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta}\right\|_2 \leq d_3\sqrt{\frac{s\log p}{n}} + \sqrt{d_3\sigma_s(\boldsymbol{\theta})}\sqrt[4]{\frac{\log p}{n}}, \quad (8)$$

with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d_4}}\right)$.

Similarly, the following theorem characterizes the performance bounds for the POMP estimate:

Theorem 2. *If $\boldsymbol{\theta}$ is (s,ξ)-compressible for some $\xi < 1/2$, there exist constants d_1', d_2', d_3' and d_4' such that for $n > d_1' s^{2/3} p^{2/3} (\log s)^{2/3} \log p$, the POMP estimate satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}} - \boldsymbol{\theta}\right\|_2 \leq d_2'\sqrt{\frac{s\log s\log p}{n}} + d_3'\frac{\log s}{s^{\frac{1}{\xi}-2}} \quad (9)$$

after $s^\star = \mathcal{O}(s\log s)$ iterations with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d_4'}}\right)$.

Remarks. An immediate comparison of the sufficient condition $n = \mathcal{O}(s^{2/3}p^{2/3}\log p)$ of Theorem 1 with those of [22] for GLM models with i.i.d. covariates given by $n = \mathcal{O}(s\log p)$ reveals that a loss of order $\mathcal{O}(p^{2/3}s^{-1/3})$ is incurred due to the inter-dependence of the covariates. However, the sample space of n i.i.d. covariates is np-dimensional, whereas in our problem the sample space is only $(n+p)$-dimensional. Hence, the aforementioned loss can be viewed as the price of self-averaging of the process accounting for the low-dimensional nature of the covariate sample space. To the best of our knowledge, the dominant loss of $\mathcal{O}(p^{2/3})$ in both theorems does not seem to be significantly improvable, as self-exciting processes are known to converge quite slowly to their ergodic state [28]. Surprisingly, the analysis of the sampling requirements of linear AR models reveals a loss of $\mathcal{O}(p^{2/3})$ in the number of measurements [29].

The sufficient condition of Theorem 2 given by $n = \mathcal{O}(s^{2/3}p^{2/3}(\log s)^{2/3}\log p)$ implies an extra loss of $(\log s)^{2/3}$ due to the greedy nature of the solution. Moreover, Theorem 2 requires a high compressibility level of the parameter vector $\boldsymbol{\theta}$ ($\xi < 1/2$), whereas Theorem 1 does not impose any extra restrictions on $\xi \in (0,1)$. Intuitively speaking, this comparison reveals the trade-off between computational complexity and compressibility requirements for convex optimization vs. greedy techniques, which are well-known for linear models [6].

As mentioned earlier, similar results hold for when μ is *unknown*, except with possibly slightly different constants (see Corollary 2 in Appendix A). The constants d_i, d_i', $i = 1,\cdots,4$, α and β are explicitly given in the proof of the theorems in Appendix A. As for a typical numerical example, for $\pi_{\min} = 0.05$, $\pi_{\max} = 0.5$, the constants of Theorem 1 can be chosen as $d_1 \approx 10^3, d_2 = 50, d_3 \approx 10^4$ and $d_4 = 4$. We will next give a sketch of the proof of these theorems. The full proofs are given in Appendix A.

Proof Sketch. The main ingredient in the proofs of Theorems 1 and 2 is inspired by the beautiful treatment of Negahban et al. in [22] in establishing the notion of Restricted Strong Convexity (RSC). By the convexity of the negative Jacod log-likelihood given by Eq. (1), it is clear that a small change in $\boldsymbol{\theta}$ results in a small change in the negative Jacod log-likelihood. However, the converse is not necessarily true. Intuitively speaking, the RSC condition guarantees that the converse holds: a small change in the log-likelihood implies a small change in the parameter vector, i.e., the log-likelihood is not too *flat* around the true parameter vector. A depiction of the RSC condition for $p = 2$, adopted from [22], is given in Figure 1. In Figure 1(a), the RSC does not hold since a change along θ_2 does not change the log-likelihood, whereas the log-likelihood in Figure 1(b) satisfies the RSC.

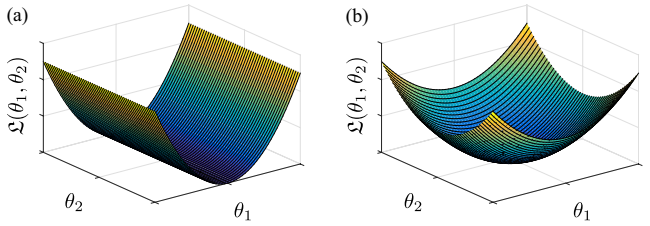


Fig. 1: Illustration of RSC (a) RSC does not hold (b) RSC does hold.

More formally, if the log-likelihood is twice differentiable at $\boldsymbol{\theta}$, the RSC is equivalent to existence of a lower quadratic bound on the negative log-likelihood:

$$\mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\Delta},\boldsymbol{\theta}) := \mathfrak{L}(\boldsymbol{\theta}+\boldsymbol{\Delta}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\Delta}'\nabla\mathfrak{L}(\boldsymbol{\theta}) \geq \kappa\|\boldsymbol{\Delta}\|_2^2, \quad (10)$$

for a positive constant $\kappa > 0$ and all $\boldsymbol{\Delta} \in \mathbb{R}^p$ in a carefully-chosen neighborhood of $\boldsymbol{\theta}$ depending on s and ξ. Based on the results of [22], when the RSC is satisfied, sufficient conditions akin to those in Theorems 1 and 2 can be obtained by estimating the Euclidean extent of the solution set around the true parameter vector (see Propositions 3 and 5 in Appendix A).

The major technical challenge for the canonical self-exciting process, as opposed to the GLM models with i.i.d. covariates in [22], lies in the fact that the covariates are highly inter-dependent as they are formed by the history of the process. Hence, it is not straightforward to establish RSC with high probability, as the large deviation techniques used for i.i.d. random vectors does not hold. We establish the RSC for the canonical self-exciting process in two steps (see Lemma 1 in Appendix A-A). First, we show that RSC holds for the expected value of the negative log-likelihood $\mathbb{E}[\mathfrak{L}(\boldsymbol{\theta})]$, and then by invoking results on concentration of dependent random variables show that the negative log-likelihood $\mathfrak{L}(\boldsymbol{\theta})$ resides in a sufficiently small neighborhood of $\mathbb{E}[\mathfrak{L}(\boldsymbol{\theta})]$ with high probability, and hence satisfies the RSC.

The rest of the proof of Theorem 1 (given in Appendix A-B) establishes that upon satisfying the RSC, the estimation error can be suitably bounded. Similarly, Theorem 2 is proven using

the RSC of the canonical self-exciting process together with the results adopted from [26] on the performance of OMP for convex cost functions (see Appendix A-C).

IV. Application to Simulated and Real Data

In this section, we study the performance of the conventional ML estimator, the ℓ_1-regularized ML estimator, and the POMP estimator on simulated data as well as real spiking data recorded from mouse's lateral geniculate nucleus (LGN) neurons.

A. Simulation Studies

In order to simulate spiking data governed by the canonical self-exciting process, we use the commonly-used thinning technique [30]. Thinning is a standard technique for generating inhomogeneous point process data from arbitrary spiking probabilities. Suppose that the spiking probability is upper bounded as $\lambda_i \leq \lambda_{\max}$ for some $\lambda_{\max}$ almost surely for $i = 1, 2, \cdots$. The thinning method first generates a *homogeneous* point process (Poisson) with rate $\lambda_{\max}$, which we will denote by x_i^h, for $i = 1, 2, \cdots$. Then, starting at $i = 1$, the spiking probability λ_i is computed using Eq. (3) in order to generate a Bernoulli random variable $b_i \sim \text{Bernoulli}\left(\frac{\lambda_i}{\lambda_{\max}}\right)$. Then, the point process given by $x_i := b_i x_i^h$ has the spiking probability given by Eq. (3).

Figure 2 shows the first 500 samples of the canonical self-exciting process of length $n = 1000$ generated using a history dependence parameter vector of length $p = 50$ shown in Figure 3(a) with $\mu = 0.1$. The parameter vector $\boldsymbol{\theta}$ is compressible with a sparsity level of $s = 3$ and $\sigma_3(\boldsymbol{\theta}) = 0.35$. A value of $\gamma_n = 0.03$ is used to obtain the ℓ_1-regularized ML estimate, which is slightly tuned around the theoretical estimate given by Theorem 1. Figures 3(b), 3(c), and 3(d) show the estimated history dependence parameter vectors using ML, ℓ_1-regularized ML, and POMP, respectively. It can be readily visually observed that regularized ML and POMP significantly outperform the ML estimate.

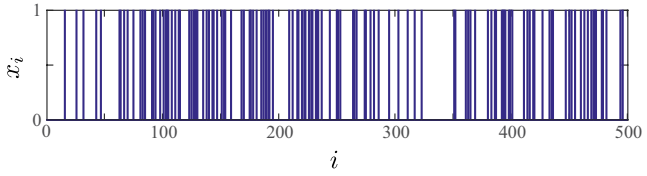
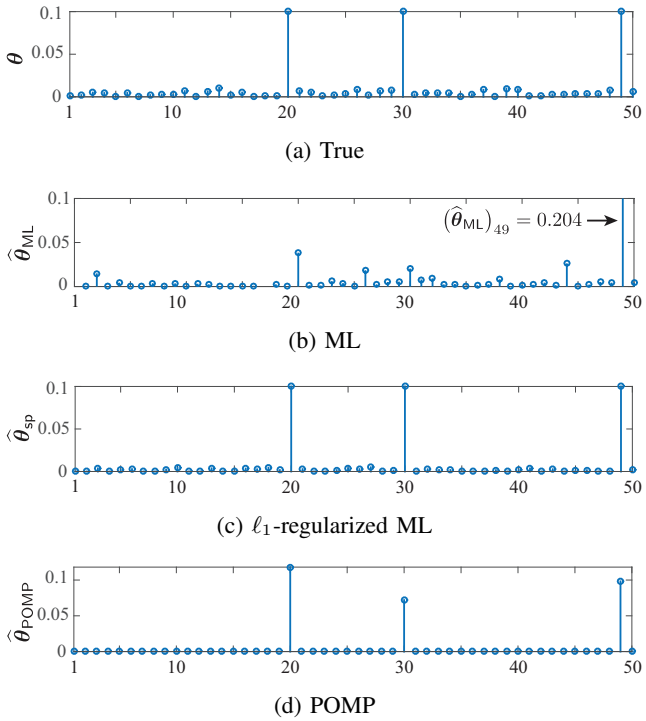


Fig. 2: A sample of the simulated canonical self-exciting process.

In order to quantify this performance gain, we repeated the same experiment by generating realizations corresponding to randomly chosen supports of size s for $\boldsymbol{\theta}$ and spiking events of length n, using a range of sparsity levels $s = 2, 3, 5$ and 10 as well as $10^2 \leq n \leq 10^6$. In each case, the magnitudes of the components of $\boldsymbol{\theta}$ were chosen to satisfy the assumptions ($\star$). For a given $\boldsymbol{\theta}$, the mean-square-error (MSE) of the estimate $\widehat{\boldsymbol{\theta}}$ is defined as $\widehat{\mathbb{E}}\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\}$, where $\widehat{\mathbb{E}}\{\cdot\}$ is the sample average over the realizations of the process. A comparison of the MSE of the estimators is shown in Figure 4. As it can be inferred from Figure 4, the ℓ_1-regularized ML and POMP have a systematic performance gain over the ML estimate, with the former outperforming the rest.



(a) True

(b) ML

(c) ℓ_1-regularized ML

(d) POMP

Fig. 3: (a) True parameters vs. (b) ML, (c) ℓ_1-regularized ML, and (d) POMP estimates.

The MSE comparison in Figure 4 requires one to know the true parameters. In practice, the true parameters are unknown, and statistical tests are typically used to assess the goodness-of-fit of the estimates to the observed data. We use the Kolmogorov-Smirnov (KS) test and the autocorrelation function (ACF) test to assess the goodness-of-fit. These tests are based on the time-rescaling theorem for point processes [31], which states that if the time axis is rescaled using the conditional intensity of an inhomogeneous Poisson process, the resulting point process is a homogeneous Poisson process with unit rate. Thereby, using the estimated conditional intensities, one can test for the validity of the time-rescaling theorem via two statistical tests: the KS test reveals how close the empirical quantiles of the time-rescaled point process to the true quantiles of a unit rate Poisson process, and the ACF test reveals how close the ISI values of the time-rescaled process are to the true ISI distribution of a unit rate Poisson process. Details of these tests are given in Appendix C.

Figure 5 shows the KS and ACF tests (at 95% and 99% confidence levels, respectively) for the ML ℓ_1-regularized ML, and the POMP estimates from Figure 3. The yellow shades mark the regions below the specified confidence levels. The ML estimate fails to pass either test, while the regularized and POMP estimates satisfy both tests.

B. Application to the analysis of LGN spiking activity

In this section, we compare the performance of the ML, ℓ_1-regularized ML, and POMP estimators in modeling the spontaneous spiking activity recorded from the lateral geniculate nucleus (LGN) neurons. The LGN is part of the thalamus in the brain, which acts as a relay from the retina to the primary visual cortex [32]. The data were recorded at $1ms$ resolution

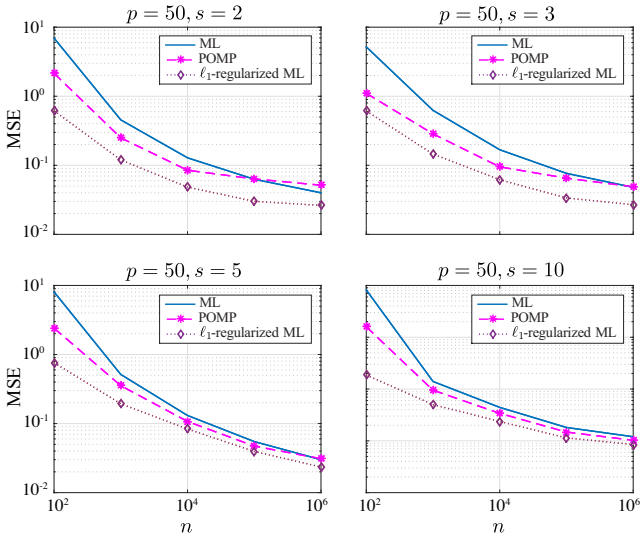


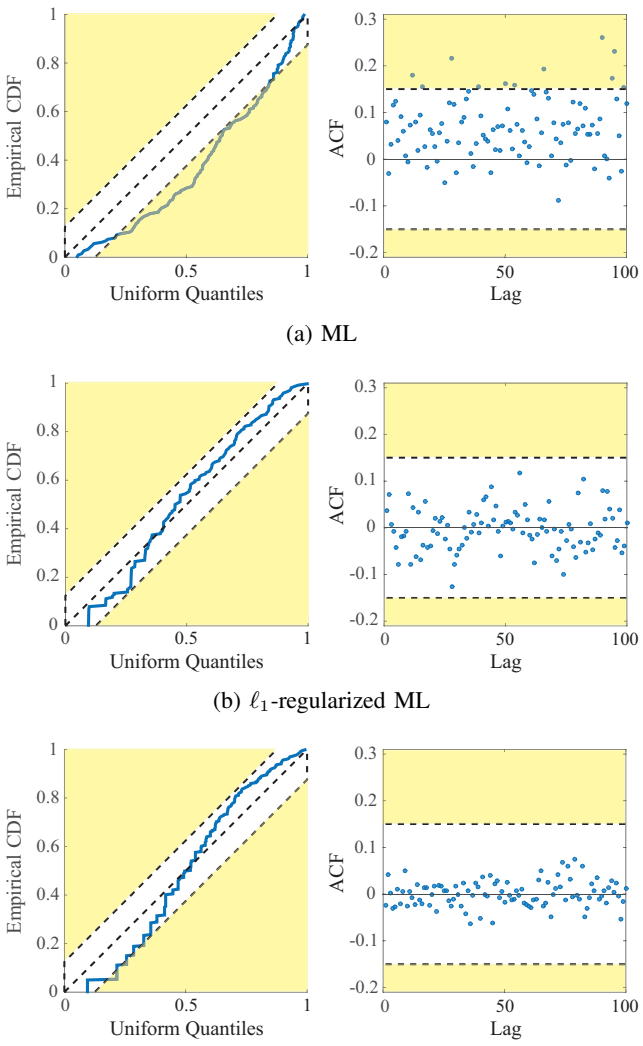
Fig. 4: MSE performance of the ML, ℓ_1-regularized ML and POMP estimators.

from the mouse LGN neurons using single-unit recording [33]. We used about 5 seconds of data from one neuron for the analysis. In order to capture the history dependence governing the spontaneous spiking activity of the LGN neuron, we model the spiking probability using the canonical self-exciting process model with $p = 100$ ($\Delta = 1ms$). Figure 6 shows the spiking data used in the analysis.

Figure 7 shows the estimated history dependence parameter vectors using the three methods. The regularized parameter γ_n was chosen using a two-fold cross-validation refinement around the value obtained from our theoretical results. Both the regularized ML (Figure 7(b)) and POMP (Figure 7(c)) estimates capture significant history dependence components around a lag of $90 - 95$ ms (marked by the upward arrows). In [34], an intrinsic neuronal oscillation frequency of around 10 Hz has been reported in around 30% of all classes of mouse retinal cells under experiment, using combined two-photon imaging and patch-clamp recording. Our results are indeed consistent with the above mentioned findings about the intrinsic spiking frequency of retinal neurons. To see this, we consider the power spectral density of the canonical self-exciting process given by:

$$S(\omega) = \frac{1}{2\pi}\left(\pi_\star^2\delta(\omega) + \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2\,|1-\Theta(\omega)|^2}\right), \quad (11)$$

where $\Theta(\omega)$ is the discrete-time Fourier transform of $\boldsymbol{\theta}$ and $\pi_\star = \frac{\mu}{1-\mathbf{1}'\boldsymbol{\theta}}$ denotes the stationary distribution probability of spiking. The derivation of the power spectral density is given in Appendix A. The power spectral density of the canonical self-exciting process resembles the Bartlett spectrum of the Hawkes process [24], [35], [36], whose peaks correspond to the significant oscillatory components of the underlying point process. Our estimated parameter vectors $\boldsymbol{\theta}$ using the regularized ML and POMP have significant nonzero components around lags of $90 \leq k \leq 95$. As a result, $S(\omega)$ peaks at $\omega = \frac{2\pi}{k\Delta}$. Hence, $f = \frac{1}{k\Delta}$ is an estimate of the significant intrinsic frequency of the underlying self-exciting process. Using the estimated numerical values, the intrinsic frequency



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Fig. 5: KS and ACF tests at 95% and 99% confidence levels, respectively, for the ML, ℓ_1-regularized ML and POMP estimates.

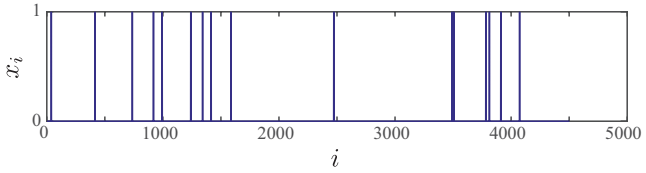
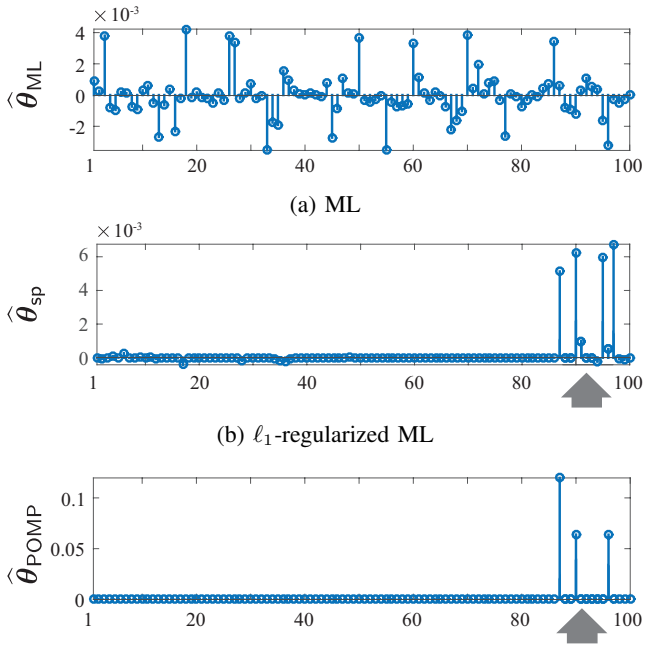


Fig. 6: The LGN spiking data used in the analysis.

is around $10.5 - 11$Hz, which is consistent with experimental findings of [34]. Compared to the method in [34], our estimates are obtained using much shorter recordings of spiking activity and provide a principled framework to study the oscillatory behavior of LGN neurons using the theory of point processes.

Note that there is a difference in the orders of magnitudes of the POMP estimate compared to the ML and regularized ML estimates. This is due to the fact that the POMP estimate is exactly s-sparse, whereas the ML and regularized ML estimates consist of $p = 100$ non-zero values. In order to assess the goodness-of-fit of these estimates, we invoke the KS and ACF tests. Figure 8 shows the corresponding KS and ACF test plots. As it is implied from Figure 8(a), the ML estimate



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Fig. 7: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the LGN spiking parameters.

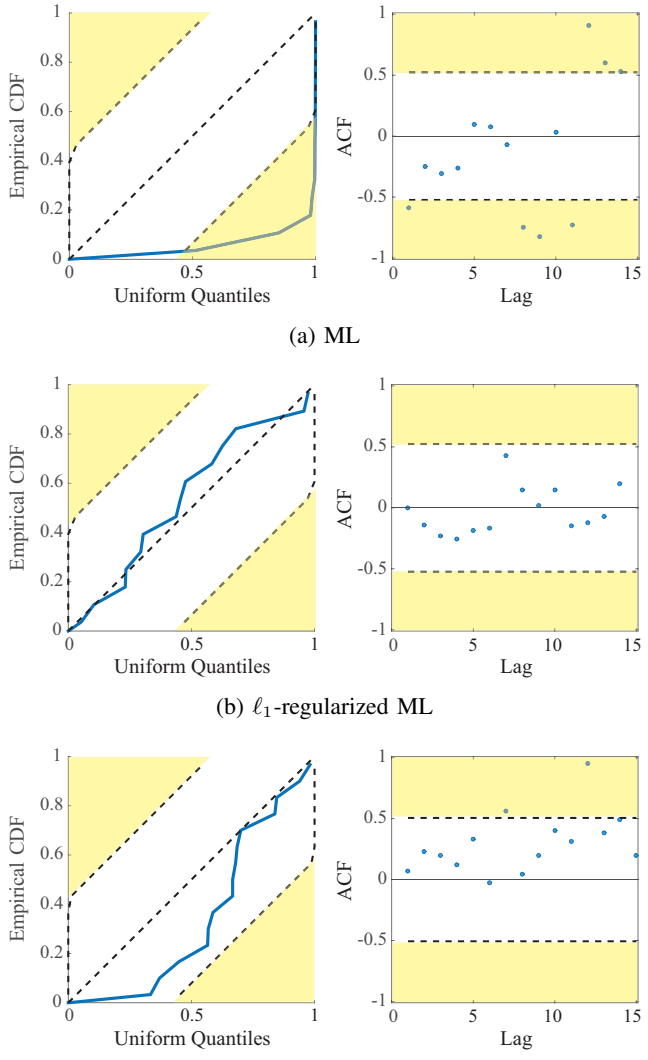
fails both tests due to overfitting, whereas the regularized ML (Figure 8(b)) passes both tests at the specified confidence levels. The POMP estimate (Figure 8(c)), however, passes the KS test while marginally failing the ACF test. The latter observation implies that the seemingly negligible components of the parameter vector captured by the regularized ML estimate seem to be important in explaining the statistics of the observed data.

C. Application to the analysis of retinal gangllion cell spiking activity

In this section, we apply the ML, ℓ_1-regularized ML, and POMP estimators to spontaneous spiking data recorded from the retinal ganglion cells (RGC) of neonatal and adult ferrets [37]. The retinal ganglion cells are located in the innermost layer of the retina. They integrate information from photoreceptors and project them into the brain [38]. The data were recorded using a multi-electrode array from the ferret retina at 50 μs [37]. We used 2.5 seconds of data from one neuron for the analysis (neuron 2, session 1, adult data set, CARMEN data base [39]). In order to capture the history dependence governing the spontaneous spiking activity of the RGC neuron, we model the spiking probability using a logistic link model of the form $\lambda_i = \frac{\exp(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})}{C+\exp(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})}$, with $C = 100$ and $p = 50$ ($\Delta = 25\ ms$).

Figure 9 shows the spiking data used in our analysis. The RGC activity in the adult ferret is characterized by bursts of activity with a mean firing rate of 9 ± 7 Hz, which are separated by $0.5 - 1\ s$ intervals [37].

Figure 10 shows the estimated history dependence parameter vectors using the three methods. The regularized parameter γ_n was chosen using a two-fold cross-validation



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Fig. 8: KS and ACF tests at 99% confidence level, for the ML, ℓ_1-regularized ML and POMP estimates.

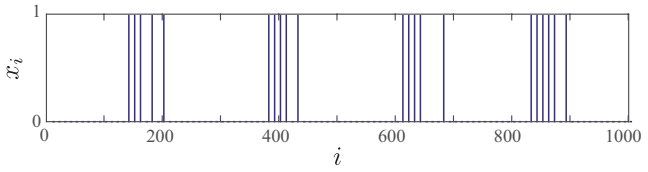
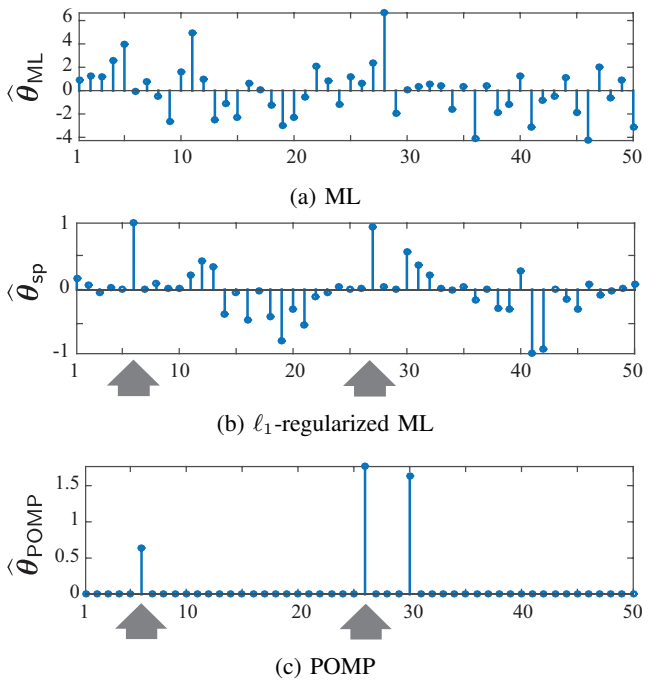


Fig. 9: The RGC spiking data used in the analysis.

refinement around the value obtained from our theoretical results. Both the regularized ML (Figure 10(b)) and POMP (Figure 10(c)) estimates capture significant self-exiting history dependence components around the lags of 150 ms and $0.65 - 0.75\ s$ (marked by the upward arrows). These self-exciting components are consistent with the aforementioned empirical estimates of [37], as they indicate that the data can be characterized by a combination of $\frac{1}{150\ ms} = 6.66$ Hz bursts separated by gaps of length $0.65 - 0.75\ s$.

Figure 11 shows the KS and ACF tests for the three methods. As shown in Figure 11(a), the ML estimate fails the KS test due to overfitting, whereas the regularized ML (Figure 11(b)) and POMP (Figure 11(c)) pass both tests at the specified confidence levels. In order to further inspect the



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Fig. 10: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the RGC spiking parameters.

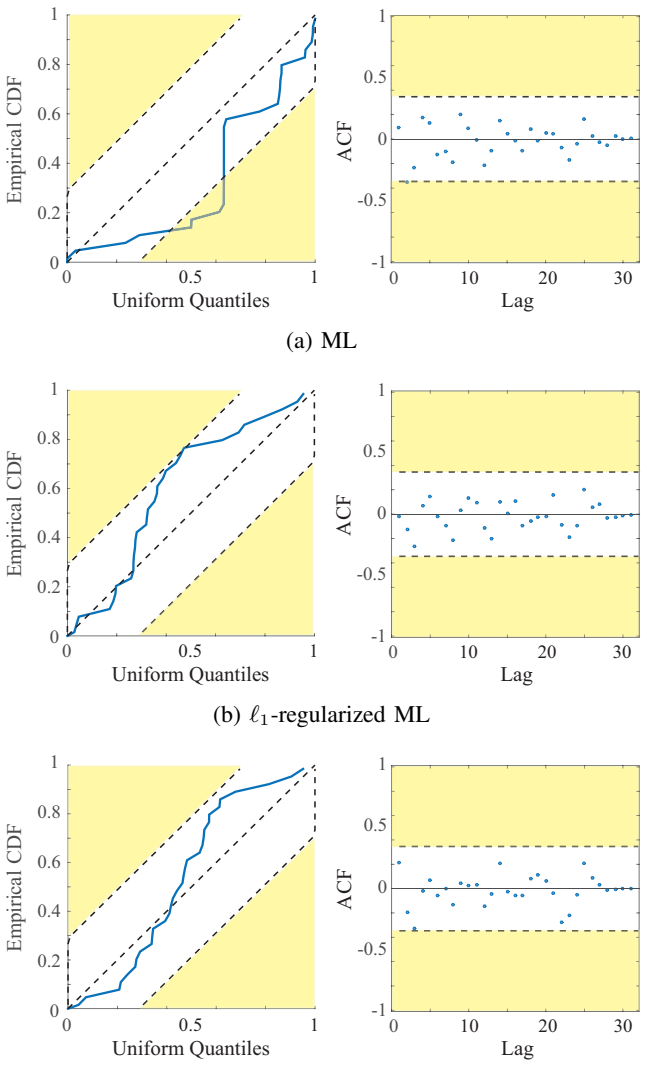
goodness-of-fit of these methods, we plot the estimated spiking rates in Figure 12. The ML estimate shown in Figure 12(a) overfits the spiking events by rapidly saturating the rate to either 0 and 1, which results in undesired high rate estimates where there are no spikes. On the contrary, the regularized ML (Figure 12(b)) and POMP (Figure 12(c)) provide a more reliable estimate of the rates consistent with the spiking events.

V. Discussion and Future Work

In this paper, we studied the sampling properties of ℓ_1-regularized ML and greedy estimators for a canonical self-exciting process. The main theorems provide non-asymptotic sampling bounds on the number of measurements, which lead to stable recovery of the parameters of the process. To the best of our knowledge, our results are the first of this kind, and can be readily generalized to various other classes of self-exciting point processes, such as processes with logarithmic or logistic link.

Compared to the existing literature, our results bring about two major contributions. First, we provide a theoretical underpinning for the advantage of ℓ_1-regularization in ML estimation as well as greedy estimation in problems involving point process observations. These methods have been used in neuroscience in an ad-hoc fashion. Our results establish the utility of these techniques by characterizing the underlying sampling trade-offs. Second, our analysis relaxes the widely-assumed hypotheses of i.i.d. covariates. This assumption is often violated when working with history-dependent data such as neural spiking data.

We also verified the validity of our theoretical results through simulations studies as well as application to real neuronal spiking data from mouse's LGN and ferret's RGC neurons. These results show that both the regularized ML



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Fig. 11: KS and ACF tests at 95% confidence level, for the ML, ℓ_1-regularized ML and POMP estimates.

and the greedy estimates significantly outperform the widely-used ML estimate. In particular, through making a connection with the spectrum of discrete point processes, we were able to quantify the estimation of the intrinsic firing frequency of LGN neurons. Our future work includes generalization of our analysis to multivariate point process models, which will allow to infer network properties from multi-unit recordings of neuronal ensembles.

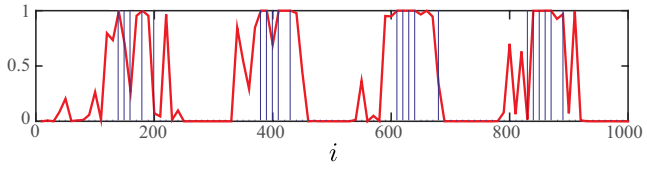
VI. Acknowledgement

We would like to thank L. A. Kontorovich for helpful discussions regarding reference [40].

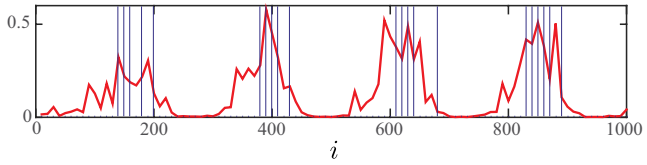
Appendix A
Proofs of Main Theorems

Recall that if the log-likelihood is twice differentiable with respect to $\boldsymbol{\theta}$, the Restricted Strong Convexity (RSC) property implies the existence of a lower quadratic bound on the negative log-likelihood:

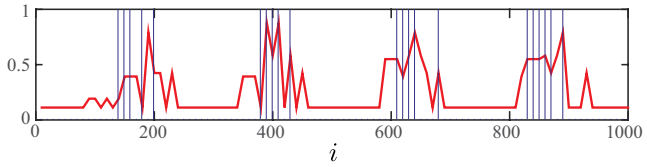
$$\mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\Delta},\boldsymbol{\theta}) := \mathfrak{L}(\boldsymbol{\theta}+\boldsymbol{\Delta}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\Delta}'\nabla\mathfrak{L}(\boldsymbol{\theta}) \geq \kappa\|\boldsymbol{\Delta}\|_2^2, \quad (12)$$



(a) ML



(b) ℓ_1-regularized ML



(c) POMP

Fig. 12: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the RGC spiking rate. Blue vertical lines show the locations of the spikes, and red traces show the estimated rate.

for a positive constant $\kappa > 0$ and all $\boldsymbol{\Delta} \in \mathbb{R}^p$ satisfying:

$$\sigma_S(\boldsymbol{\Delta}) \leq 3\|\boldsymbol{\Delta}_S\|_1 + 4\sigma_S(\boldsymbol{\theta}). \tag{13}$$

for any index set $S \subset \{1, 2, \cdots, p\}$ of cardinality s. The latter condition is known as the cone constraint.

The canonical self-exciting process can be viewed as a Markov chain with states $X_i = x_{i-p}^{i-1}$. Since each x_i has two possible values, there are 2^p possible states. This Markov chain is irreducible since transition from any state to any other state is possible in at most p steps. Also, transition from an all-zero state to itself is possible. Hence the chain is aperiodic as well. This implies that there exists a stationary distribution for the Markov chain. We also know that if $\{X_i\}_{i=1}^{\infty}$ is a stationary Markov Chain, then for any functional $f(.)$, $\{f(X_i)\}_{i=1}^{\infty}$ is a strictly stationary stochastic process (SSS). Therefore the canonical self-exciting process and the spiking probability sequence λ_1^n are both SSS. In particular, we have

$$\pi_\star := \mathbb{E}[x_i] = \mathbb{E}\left[\mathbb{E}\left[x_i|\lambda_i\right]\right] = \mathbb{E}[\lambda_i] = \mu + \pi_\star \mathbf{1}'\boldsymbol{\theta}.$$

Hence, the stationary probability $\pi_\star$ satisfies:

$$\pi_\star = \frac{\mu}{1 - \mathbf{1}'\boldsymbol{\theta}}. \tag{14}$$

Note that by (14), in order for the process to be stationary it is necessary that $\mathbf{1}'\boldsymbol{\theta} < 1$. The gap $1 - \mathbf{1}'\boldsymbol{\theta}$ will play an important role in controlling the convergence rate of the aforementioned Markov chain to its stationary distribution (see the proof of Proposition 2 below).

Recall the technical assumptions $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given by ($\star$):

$$\begin{aligned} &1)\ 0 < \mathbf{1}'\boldsymbol{\theta} \leq c_1 < 1, \\ &2)\ 0 < \pi_{\min} \leq \mu - \|\boldsymbol{\theta}\|_1, \\ &3)\ \mu + \|\boldsymbol{\theta}\|_1 \leq \pi_{\max} < 1/2. \end{aligned} \tag{$\star$}$$

The first assumption results in the stationarity of the canonical self-exciting process. The second and third assumptions make sure that the process does not become all-zero, and that it has a sufficiently fast mixing (See the proof of Proposition 2 below). For simplicity we will also use the notation

$$\mathbb{S}_p(t) := \{\boldsymbol{\nu} \mid \|\boldsymbol{\nu}\|_p = t\}.$$

to denote the p-norm ball of radius t.

A. A Key Lemma

The proofs of Theorems 1 and 2 are mainly based on the following key lemma establishing the Restricted Strong Convexity condition for the canonical self-exciting process:

Lemma 1 (Restricted Strong Convexity of the canonical self-exciting process)**.** *Let x_{-p+1}^n denote a sequence of samples from the canonical self-exciting process with parameters $\{\mu, \boldsymbol{\theta}\}$ satisfying the conditions given by ($\star$). Then, for $n \geq d_1 s^{2/3} p^{2/3} \log p$, the negative log-likelihood function $\mathfrak{L}(\boldsymbol{\theta})$ satisfies the RSC property with a positive constant $\kappa > 0$ with probability at least $1 - 2\exp\left(-\frac{c\kappa^2 n^3}{s^2 p^2}\right)$, for some constant c, and both κ and c are only functions of d_1, c_1, $\pi_{\min}$, and $\pi_{\max}$.*

Proof of Lemma 1. The proof is inspired by the elegant treatment of Negahban et al. [22]. The major difficulty in the proof lies in the high inter-dependence of the covariates and observations.

A second order Taylor expansion of the negative log-likelihood (1) around $\boldsymbol{\theta}$ yields:

$$\begin{aligned} \mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\Delta}, \boldsymbol{\theta}) &= \mathfrak{L}(\boldsymbol{\theta} + \boldsymbol{\Delta}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\Delta}'\nabla\mathfrak{L}(\boldsymbol{\theta}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \frac{\left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2}{\left(\mu + \boldsymbol{\theta}' x_{i-p}^{i-1} + \nu(\boldsymbol{\Delta}' x_{i-p}^{i-1})\right)^2} \\ &\geq \frac{1}{\pi_{\max}^2} \frac{1}{n}\sum_{i=1}^{n} x_i \left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2, \end{aligned}$$

for some $\nu \in [0, 1]$. The inequality follows from the fact that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \boldsymbol{\Delta}$ satisfy ($\star$), and hence:

$$\pi_{\min} < \mu + \boldsymbol{\theta}' x_{i-p}^{i-1} + \nu \boldsymbol{\Delta}' x_{i-p}^{i-1} < \pi_{\max}.$$

For simplicity of notation, we define the n-sample empirical expectation as follows:

$$\hat{\mathbb{E}}_n\{f(x.)\} := \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

for any measurable function $f(x.)$. Note that the subscript $x.$ refers to an index in the set $\{1, 2, \cdots, n\}$. The result of the lemma is equivalent to proving that

$$\hat{\mathbb{E}}_n\left[x_. \left(\boldsymbol{\Delta}' x_{.-p}^{.-1}\right)^2\right] \geq \kappa \|\boldsymbol{\Delta}\|_2^2, \tag{15}$$

holds with probability greater than $1 - 2\exp\left(-\frac{c\kappa^2 n^3}{s^2 p^2}\right)$. Since both sides of (15) are quadratic in $\boldsymbol{\Delta}$, the statement is equivalent to proving

$$\hat{\mathbb{E}}_n\left[x_.(\boldsymbol{\Delta}' x_{.-p}^{.-1})^2\right] \geq \kappa,$$

for all $\|\boldsymbol{\Delta}\|_2 \in \mathbb{S}_2(1)$. We establish this in two steps:

Step 1. First, we show that the statement holds for the true expectation:

$$\mathbb{E}\Big[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\Big] \geq \kappa_l > 0 \tag{16}$$

for some κ_l which will be specified below, for all $\|\boldsymbol{\Delta}\|_2 \in \mathbb{S}_2(1)$. To establish the inequality (16), we need the following lemma:

Lemma 2. *Let $R \in \mathbb{R}^{p\times p}$ be the $p \times p$ covariance matrix of a stationary process with power spectral density $S(\omega)$, and denote its maximum and minimum eigenvalues by $\lambda_{\max}(p)$ and $\lambda_{\min}(p)$ respectively then $\lambda_{\max}(p)$ is increasing in p, $\lambda_{\min}(p)$ is decreasing in p and we have*

$$\lambda_{\mathsf{min}}(p) \downarrow \inf_{\omega} S(\omega), \tag{17}$$

and

$$\lambda_{\mathsf{max}}(p) \uparrow \sup_{\omega} S(\omega). \tag{18}$$

Proof. This is a standard result in spectral analysis. See for example [41]. □

We can therefore lower-bound $\mathbb{E}\left[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right]$ by:

$$\begin{aligned}
&\mathbb{E}\left[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right] = \mathbb{E}\left[\mathbb{E}[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2 | x_{\cdot-p}^{\cdot-1}]\right] \\
&= \mathbb{E}\left[\left(\mu + \boldsymbol{\theta}' x_{\cdot-p}^{\cdot-1}\right)\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right] > \pi_{\min}\mathbb{E}\left[\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right] \\
&= \pi_{\min}\boldsymbol{\Delta}'\mathbb{E}\left[x_{\cdot-p}^{\cdot-1} x_{\cdot-p}^{\cdot-1\prime}\right]\boldsymbol{\Delta} = \pi_{\min}\boldsymbol{\Delta}' R \boldsymbol{\Delta} \\
&\geq \pi_{\min}\lambda_{\min}(p) \geq \pi_{\min}\inf_{\omega} S(\omega),
\end{aligned}$$

where the first inequality follows from the first assumption in $(\star)$ and the second inequality follows from Lemma 2. Finally we will show that the power spectral density is bounded away from zero. Let $r_{-\infty}^{\infty}$ and $c_{-\infty}^{\infty}$ denote the autocorrelation and autocovariance values of the process, respectively. By the stationarity of the process we have:

$$\begin{aligned}
r_k &= \mathbb{E}\left[x_{\cdot+k} x_{\cdot}\right] = \mathbb{E}\left[x_k x_0\right] = \mathbb{E}\left[\mathbb{E}\left[x_k x_0 | x_{-\infty}^{k-1}\right]\right] \\
&= \mathbb{E}\left[\mu x_0 + \boldsymbol{\theta}' x_{k-p}^{k-1} x_0\right] = \mu\pi_\star + \boldsymbol{\theta}' \boldsymbol{r}_{k-p}^{k-1}.
\end{aligned}$$

for $k > 0$. Similarly, by subtracting the means we have the following identity for the autocovariance:

$$c_k = \boldsymbol{\theta}' \boldsymbol{c}_{k-p}^{k-1}. \tag{19}$$

A straightforward calculation gives $c_0 = \pi_\star - \pi_\star^2$. Eq. (19) resembles the Yule-Walker equations for an AR process of order p with parameter $\boldsymbol{\theta}$ and the innovations variance given by $\sigma^2 = \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2}$. Thus, the power spectral density of the canonical self-exciting process can be expressed as:

$$S(\omega) = \frac{1}{2\pi}\left(\pi_\star^2 \delta(\omega) + \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2 |1 - \Theta(\omega)|^2}\right). \tag{20}$$

Thus the bound of Eq. (16) is established with $\kappa_l := \pi_{\min}\pi_\star(1-\pi_\star)/8\pi$.

Step 2. We now show that the empirical and the true expectations of $x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2$ are close enough to each other. Let

$$\mathfrak{D}_{\Delta,n} := \hat{\mathbb{E}}_n\left[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right] - \mathbb{E}\left[x_{\cdot}\left(\boldsymbol{\Delta}' x_{\cdot-p}^{\cdot-1}\right)^2\right].$$

and

$$\mathfrak{D}_n := \sup_{\boldsymbol{\Delta}\in\mathbb{S}_2(1)} |\mathfrak{D}_{\Delta,n}|$$

We will prove that

$$\mathbb{P}\left[\mathfrak{D}_n \geq \frac{\kappa_l}{4}\right] \leq 2\exp\left(-\frac{c\kappa_l^2 n^3}{s^2 p^2}\right), \tag{21}$$

for some constant c, from which the statement of the lemma follows by taking $\kappa = \kappa_l/4$. In order to establish the concentration inequality of Eq. (21), we need to invoke a result from concentration of dependent random variables. First, we define the normalized Hamming metric between two sequences x_1^n and y_1^n is defined as $d(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i \neq y_i)$.

We will now show that (23) implies that $\mathfrak{D}_n$ has Lipschitz continuity with respect to the normalized Hamming distance. This is key to the rest of our proof for using a concentration inequality for dependent random variables. In order to complete our proof we need the following lemma which gives us the a concentration inequality for dependent random variables. The normalized Hamming metric between two sequences x_1^n and y_1^n is defined as $d(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i \neq y_i)$. The following proposition establishes that the empirical expectation appearing in Eq. (15) is is $\mathcal{O}(\frac{sp}{n})$-Lipschitz with respect to the normalized Hamming metric:

Proposition 1. *$\mathfrak{D}_n$ is $\mathcal{O}(\frac{sp}{n})$-Lipschitz with respect to the normalized Hamming metric.*

Proof. First, by evaluating the first order optimality conditions of the solution $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$, it can be shown that the error vector $\boldsymbol{\Delta} = \widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta}$ satisfies the inequality given by (13)

$$\sigma_S(\boldsymbol{\Delta}) \leq 3\|\boldsymbol{\Delta}_S\|_1 + 4\sigma_S(\boldsymbol{\theta}).$$

with S denoting the support of the best s-term approximation to $\boldsymbol{\theta}$ (see for example [22]). By the assumption of $\sigma_S(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, we can choose a constant c_0 such that $\sigma_S(\boldsymbol{\theta}) \leq c_0\sqrt{s}$. Hence,

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_1 &\leq 4\|\boldsymbol{\Delta}_S\|_1 + \sigma_S(\boldsymbol{\theta}) \\
&\leq (4+c_0)\sqrt{s}\|\boldsymbol{\Delta}_S\|_2 \leq (4+c_0)\sqrt{s}
\end{aligned} \tag{22}$$

where we have used the fact that $\|\boldsymbol{\Delta}_S\|_1 \leq \sqrt{s}\|\boldsymbol{\Delta}_S\|_2 \leq \sqrt{s}$ for all $\boldsymbol{\Delta} \in \mathbb{S}_2(1)$. Therefore for all $i \in \{1, 2, \cdots, n\}$, we have:

$$0 \leq x_i\left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2 \leq \left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2 \leq \|\boldsymbol{\Delta}\|_1^2 \leq (4+c_0)^2 s. \tag{23}$$

We first prove the claim for $\mathfrak{D}_{\Delta,n}$. To establish the latter, we need to prove

$$\frac{1}{n}\left|\sum_{i=1}^n x_i\left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2 - y_i\left(\boldsymbol{\Delta}' y_{i-p}^{i-1}\right)^2\right| \leq C d(x_{-p+1}^n, y_{-p+1}^n),$$

for some $C = \mathcal{O}(\frac{sp}{n})$, or equivalently

$$\left|\sum_{i=1}^n x_i\left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2 - y_i\left(\boldsymbol{\Delta}' y_{i-p}^{i-1}\right)^2\right| \leq C' \sum_{i=-p+1}^n \mathbf{1}(x_i \neq y_i), \tag{24}$$

for some $C' = \mathcal{O}(s)$. Let us start by setting the values of x_{-p+1}^n equal to those of y_{-p+1}^n and iteratively change x_j to $1-$

x_j for all indices j where $x_j \neq y_j$ to obtain the configuration given by x^n_{-p+1}. For each such change (say x_j to $1-x_j$), the left hand side changes by at most

$$\begin{aligned}&\left|\sum_{i=1}^{n} x_i \left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2_{|x_j=1} - x_i \left(\boldsymbol{\Delta}' x_{i-p}^{i-1}\right)^2_{|x_j=0}\right| \\ &\leq x_j \left(\boldsymbol{\Delta}' x_{j-p}^{j-1}\right)^2 + 2\sum_{i\neq j} x_i |\Delta_{i-j}| \|\boldsymbol{\Delta}\|_1 \\ &\leq 3\|\boldsymbol{\Delta}\|_1^2 \leq 3(4+c_0)^2 s,\end{aligned}$$

where we have used the inequality given by Eq. (23). Hence, the C can be taken as $3(4+c_0)^2 sp/n$ and the claim of the proposition for $\mathfrak{D}_{\Delta,n}$ follows. A very similar argument can be used to extend the claim to $\mathfrak{D}_n$. Let $\boldsymbol{\Delta}^\star := \boldsymbol{\Delta}^\star(x^n_{-p+1})$ be the $\boldsymbol{\Delta}$ for which the supremum in the definition of $\mathfrak{D}_n$ is achieved (such a choice of $\boldsymbol{\Delta}$ exists by the Weierstrass extreme value theorem). Since $\boldsymbol{\Delta}^\star$ also satisfies (22), a similar argument shows that $\mathfrak{D}_n$ is $\mathcal{O}(\frac{sp}{n})$-Lipschitz (with possibly different constants). □

Now, let $H = [x_{i-p}^{i-2}, 1]$ and $\widehat{H} = [x_{i-p}^{i-2}, 0]$ be two vectors (history components) of length p which only differ in their last component, and let the mixing coefficient $\bar{\eta}_{ij}$ for $j \geq i$ be defined as:

$$\bar{\eta}_{ij} = \|p(x_j^n|H) - p(x_j^n|\widehat{H})\|_{TV}, \tag{25}$$

with $\|\cdot\|_{TV}$ denoting the total variation difference of the probability measures induced on $\{0,1\}^{n-j+1}$. Also, let

$$\eta_{ij} = \sup_{H,\widehat{H}} \bar{\eta}_{ij},$$

And

$$Q_{n,i} := 1 + \eta_{i,i+1} + \cdots + \eta_{i,n}.$$

We can know invoke Theorem 1.1 of [40] and state the following proposition:

Proposition 2. *If $\mathfrak{D}_n$ is C-Lipschitz and $q := \max_{1\leq i\leq n} Q_{n,i}$, then*

$$\mathbb{P}\left[|\mathfrak{D}_n - \mathbb{E}[\mathfrak{D}_n]| \geq t\right] \leq 2\exp\left(\frac{-2nt^2}{qC^2}\right).$$

Proof. The proof is identical to the beautiful treatment of [40] when specializing the underlying function of the variables x^i_{-p+1} to be $\mathfrak{D}_n$. □

Proposition 1 establishes that $C = C'sp/n$, for some constant C'. Now, we have

$$\eta_{ij} \leq 2^{n-j+1}|\pi_{\max}^{n-j+1} - \pi_{\min}^{n-j+1}| \leq (2\pi_{\max})^{n-j+1},$$

where we have used the fact that each element of the measures $p(x_j^n|H)$ and $p(x_j^n|\widehat{H})$ satisfies the second assumption in ($\star$) and that the size of the state space $\{0,1\}^{n-j+1}$ is given by 2^{n-j+1}. Since additionally we have $\pi_{\max} < 1/2$, $\eta_{ij} \leq \rho^{n-j+1}$ for $\rho := 2\pi_{\max} < 1$. Hence, $Q_{n,i} \leq \frac{1}{1-\rho}$ for all i, and $q \leq \frac{1}{1-\rho}$ by definition.

Using the statement of Proposition 2, we get:

$$\mathbb{P}\left[\mathfrak{D}_n \geq \mathbb{E}[\mathfrak{D}_n] + \frac{\kappa_l}{2}\right] \leq 2\exp\left(\frac{-n^3\kappa_l^2(1-\rho)}{2C's^2p^2}\right). \tag{26}$$

It only remains to show that the expectation in (26) can be suitably bounded. Note that we have

$$\begin{aligned}\mathbb{E}[\mathfrak{D}_n] = \mathbb{E}[|\mathfrak{D}_{\Delta^\star,n}|] &= \int_0^\infty (1 - F_{\mathfrak{D}_{\Delta^\star,n}}(t))dt \\ &\leq \int_0^\infty 2\exp\left(-\frac{2(1-\rho)n^3t^2}{C's^2p^2}\right)dt = 2\sqrt{\frac{C'\pi}{(1-\rho)}}\frac{ps}{n^{3/2}}.\end{aligned}$$

Thus choosing $n \geq d_1 s^{2/3}p^{2/3}\log p$, for some positive constant d_1, $\mathbb{E}[\mathfrak{D}_n]$ drops as $1/\log^{3/2} p$, and will be smaller than $\kappa_l/4$ for large enough p. Hence, combined with (26) and by defining $c := \frac{1-\rho}{2C'}$ we have:

$$\mathbb{P}\left[\mathfrak{D}_n \geq \frac{\kappa_l}{4}\right] \leq 2\exp\left(\frac{-cn^3\kappa_l^2}{s^2p^2}\right)$$

which establishes the claim of the lemma. □

Lemma 1 can be viewed as the key result in the proofs of Theorems 1 and 2 which follow next.

B. Proof of Theorem 1

We first restate the main result of [22] concerning RSC and its implications in controlling the estimation error for GLMs:

Proposition 3. *For a negative log-likelihood $\mathfrak{L}(\boldsymbol{\theta})$ which satisfies the RSC with parameter κ, every solution to the convex optimization problem (7) satisfies*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta}\right\|_2 \leq \frac{2\gamma_n\sqrt{s}}{\kappa} + \sqrt{\frac{2\gamma_n\sigma_s(\boldsymbol{\theta})}{\kappa}} \tag{27}$$

with a choice of the regularization parameter

$$\gamma_n \geq 2\left\|\nabla\mathfrak{L}(\boldsymbol{\theta})\right\|_\infty. \tag{28}$$

Proof. The proof is a special case of Theorem 1 of [22]. □

The first term in the bound (28) is increasing in s and corresponds to the estimation error of the s largest components of $\boldsymbol{\theta}$ in magnitude, whereas the second term is decreasing in s and represents the cost of replacing $\boldsymbol{\theta}$ with its s-sparse approximation. Note that by (27) and (28), we need to have

$$\mathbb{E}\left[\nabla\mathfrak{L}(\boldsymbol{\theta})\right] = \mathbf{0},$$

in order for the upper bound (27) to tend to 0 as $n \to \infty$.

Next, we will establish a suitable upper bound on $\|\nabla\mathfrak{L}(\boldsymbol{\theta})\|_\infty$ which holds with high probability and provides the appropriate scaling of γ_n. From Eq. (1), we have

$$\nabla\mathfrak{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[x_i - (\mu + \boldsymbol{\theta}' x_{i-p}^{i-1})\right]\frac{x_{i-p}^{i-1}}{\mu + \boldsymbol{\theta}' x_{i-p}^{i-1}}. \tag{29}$$

We proceed in two steps:

Step 1. We first show that

$$\mathbb{E}\left[\nabla\mathfrak{L}(\boldsymbol{\theta})\right] = \mathbf{0}. \tag{30}$$

To see this, we use the law of iterated expectations on the ith term as follows:

$$\begin{aligned}
&\mathbb{E}\left[[x_i-(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})]\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}' x_{i-p}^{i-1}}\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[x_i-(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}' x_{i-p}^{i-1}}\middle| x_{i-p}^{i-1}\right]\right]\\
&=\mathbb{E}\left[\underbrace{\mathbb{E}\left[x_i-(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})\middle| x_{i-p}^{i-1}\right]}_{0}\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}' x_{i-p}^{i-1}}\right]=0 \quad (31)
\end{aligned}$$

Summing over i, establishes (30).

Step 2. We next show that the summation given by (29) is concentrated around its mean. The iterated expectation argument used in establishing (31) implies that

$$(\nabla\mathfrak{L}(\boldsymbol{\theta}))_i=\left[x_i-(\mu+\boldsymbol{\theta}' x_{i-p}^{i-1})\right]\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}' x_{i-p}^{i-1}}$$

is a martingale with respect to the filtration given by

$$\mathcal{F}_i=\sigma\left(x_{-p+1}^{i}\right),$$

where $\sigma(\cdot)$ denote the sigma-field generated by the random variables in its argument. We will now state the following concentration result for sums of bounded and dependent random variables [42]:

Proposition 4. *Fix $n\geq 1$. Let Z_i's be bounded $\mathcal{F}_i$-measurable random variables, satisfying for each $i=1,2,\cdots,n$,*

$$\mathbb{E}\left[Z_i|\mathcal{F}_{i-1}\right]=0, \quad \textit{almost surely}.$$

Then there exists a constant c such that for all $t>0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}Z_i-\mathbb{E}[Z_i]\geq t\right)\leq\exp\left(-cnt^2\right).$$

Proof. This result is a special case of Theorem 2.5 of [42] for *bounded* random variables. □

Proposition 4 implies that

$$\mathbb{P}\left(|(\nabla\mathfrak{L}(\boldsymbol{\theta}))_i|\geq t\right)\leq\exp(-cnt^2). \quad (32)$$

By union bound, we get:

$$\mathbb{P}\Big(\left\|\nabla\mathfrak{L}(\boldsymbol{\theta})\right\|_\infty\geq t\Big)\leq\exp(-ct^2n+\log p). \quad (33)$$

Choosing $t=\sqrt{\frac{1+\alpha_1}{c}}\sqrt{\frac{\log p}{n}}$ for some $\alpha_1>0$ yields

$$\begin{aligned}
\mathbb{P}\left(\left\|\nabla\mathfrak{L}(\boldsymbol{\theta})\right\|_\infty\geq\sqrt{\frac{1+\alpha_1}{c}}\sqrt{\frac{\log p}{n}}\right)&\leq 2\exp(-\alpha_1\log p)\\
&\leq\frac{2}{n^{\alpha_1}}. \quad (34)
\end{aligned}$$

Hence, a choice of $\gamma_n=d_2\sqrt{\frac{\log p}{n}}$ with $d_2:=\sqrt{\frac{1+\alpha_1}{c}}$ satisfies (28) with probability at least $1-\frac{2}{n^{\alpha_1}}$. Combined with the result of Lemma 1 for $n>d_1s^{2/3}p^{2/3}\log p$, we have that the RSC is satisfied with a constant κ with a probability at least $1-\frac{1}{p^{\alpha_2}}\geq 1-\frac{1}{n^{\alpha_2}}$ for some constant α_2. Finally, Proposition 3 establishes the claim of Theorem 1.

C. Proof of Theorem 2

The proof is mainly based on the following proposition, adopted from Theorem 2.1 of [26], stating that the greedy procedure is successful in obtaining a reasonable $s^\star$-sparse approximation, if the cost function satisfies the RSC:

Proposition 5. *Suppose that $\mathfrak{L}(\boldsymbol{\theta})$ satisfies RSC with a constant $\kappa>0$. Let $s^\star$ be a constant such that*

$$s^\star\geq\frac{4s}{\pi_{\min}^2\kappa}\log\frac{20s}{\pi_{\min}^2\kappa}=\mathcal{O}(s\log s), \quad (35)$$

Then, we have

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)}-\boldsymbol{\theta}_S\right\|_2\leq\frac{\sqrt{6}\epsilon_{s^\star}}{\kappa},$$

where $\epsilon_{s^\star}$ satisfies

$$\epsilon_{s^\star}\leq\sqrt{s^\star+s}\|\nabla\mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty. \quad (36)$$

Proof. The proof is a specialization of the proof of Theorem 2.1 in [26] to our setting. □

Lemma 1 establishes the RSC for the negative log-likelihood function. In order to complete the proof of Theorem 2, we need to bound $\|\nabla\mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty$. We have

$$\begin{aligned}
\mathbb{E}\left[\nabla\mathfrak{L}(\boldsymbol{\theta}_S)\right]&=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[x_i-(\mu+\boldsymbol{\theta}_S' x_{i-p}^{i-1})\right]\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}_S' x_{i-p}^{i-1}}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[(\boldsymbol{\theta}-\boldsymbol{\theta}_S)' x_{i-p}^{i-1}\middle| x_{i-p}^{i-1}\right]\frac{x_{i-p}^{i-1}}{\mu+\boldsymbol{\theta}_S' x_{i-p}^{i-1}}\right]\\
&\leq c_2\sigma_s(\boldsymbol{\theta})\mathbf{1}.
\end{aligned}$$

where we have used the fact that $0\leq x_i\leq 1$ for all i, and $c_2>0$ is a positive constant. Invoking the result of Proposition 4 together with the union bound yields:

$$\mathbb{P}\left(\|\nabla\mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty\geq c_1\sqrt{\frac{\log p}{n}}+c_2\sigma_s(\boldsymbol{\theta})\right)\leq\frac{2}{n^{\beta_1}}.$$

for some constants c_1 and β_1. Hence, we get the following concentration result for $\epsilon_{s^\star}$:

$$\mathbb{P}\left(\epsilon_{s^\star}\geq\sqrt{s^\star+s}\left(c_1\sqrt{\frac{\log p}{n}}+c_2\sigma_s(\boldsymbol{\theta})\right)\right)\leq\frac{2}{n^{\beta_1}}. \quad (37)$$

Noting that by (35) we have $s^\star+s=\mathcal{O}(s\log s)\leq c_0 s\log s$, for some constant c_0, and invoking the result of Lemma 1, we get:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)}-\boldsymbol{\theta}_S\right\|_2&\leq d_2'\sqrt{\frac{s\log s\log p}{n}}+d_3' s\log s\sigma_s(\boldsymbol{\theta})\\
&\leq d_2'\sqrt{\frac{s\log s\log p}{n}}+d_3'\frac{\log s}{s^{\frac{1}{\xi}-2}},
\end{aligned}$$

where $d_2'=\sqrt{c_0}c_1$ and $d_3'=\sqrt{c_0}c_2$. with probability $\left(1-\exp\left(-\frac{c\kappa^2n^3}{s^2(\log s)^2p^2}\right)\right)\left(1-\frac{2}{n^{\beta_1}}\right)$. Choosing $n>d_1's^{2/3}(\log s)^{2/3}p^{2/3}\log p$ establishes the claimed success probability of Theorem 2. Finally, we have:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)}-\boldsymbol{\theta}\right\|_2&=\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)}-\boldsymbol{\theta}_S+\boldsymbol{\theta}_S-\boldsymbol{\theta}\right\|_2\\
&\leq\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)}-\boldsymbol{\theta}_S\right\|_2+\|\boldsymbol{\theta}_S-\boldsymbol{\theta}\|_2.
\end{aligned}$$

Using $\|\boldsymbol{\theta}_S - \boldsymbol{\theta}\|_2 \leq \sigma_s(\boldsymbol{\theta}) = \mathcal{O}\left(\frac{1}{s^{\frac{1}{\xi}-1}}\right)$ completes the proof.

APPENDIX B
EXTENSIONS OF THE MAIN RESULTS

For simplicity and clarity of presentation, we have opted to present the proofs for the case of known μ and for the canonical self-exciting process as a canonical discrete point process model. The following corollary extends our results to the case of unknown μ.

Corollary 1. *The claims of Theorems 1 and 2 hold when μ is not known, except for possibly slightly different constants.*

Proof. The claim is a direct consequence of the boundedness of covariates and can be treated by replacing $\boldsymbol{\theta}$ with the augmented parameter vector $[\mu, \boldsymbol{\theta}']'$ and augmenting the covariate vectors with an initial component of 1. The reader can easily verify that all the proof steps can be repeated in the same fashion. □

The canonical self-exciting process can be generalized to a larger class of point processes by generalization of its spiking probability function. In a more general form we can consider a spiking probability function given by

$$\lambda_i = \phi\left(\boldsymbol{\theta}' x_{i-p}^{i-1}\right), \tag{38}$$

where $\phi(\cdot)$ is a possibly nonlinear function. In their continuous form, such processes are referred to as the *nonlinear Hawkes process* [43]. It has been shown that [44] for an α-Lipschitz function $\phi(.)$ such that

$$\alpha \mathbf{1}'\boldsymbol{\theta} < 1,$$

a unique stationary process satisfying (38) exists (even when $p \to \infty$). Two of the commonly-used models in neural data analysis are the log-link and logistic-link models corresponding to Poisson and Bernoulli statistics, respectively, which we discussed in Section II (See also [25] and [20]). Our prior numerical studies in [45] revealed a similar performance improvement of the ℓ_1-regularized ML and the greedy solution over the ML estimate for the log-link model. Indeed, due to the boundedness of the covariates, the function $\phi(\cdot)$ for these models will be Lipschitz, and hence the resulting point processes will be stationary. The latter fact is key to extending our proofs to other models and is summarized by the following corollary:

Corollary 2. *The claims of Theorems 1 and 2 hold when the spiking probability is given by $\lambda_i = \phi\left(\boldsymbol{\theta}' x_{i-p}^{i-1}\right)$ for some continuous and twice-differentiable function $\phi(\cdot)$ (e.g., $\phi(x) = \exp(x)$ or $\phi(x) = \mathsf{logit}^{-1}(x)$), except for possibly slightly different constants.*

Proof. The claim is a direct consequence of the boundedness of covariates which results in $\phi(\cdot)$ being Lipschitz and hence the stationarity of the underlying process. Moreover, for twice-differentiable $\phi(\cdot)$, the proof of Lemma 1 can be generalized in a straightforward fashion. The reader can easily verify that all the remaining portions of the proofs of the main theorems can be repeated for such $\phi(\cdot)$ in a similar fashion to that of the canonical self-exciting process. □

APPENDIX C
GOODNESS-OF-FIT TESTS FOR POINT PROCESS MODELS

In this appendix, we will give an overview of the statistical tools used to assess the goodness-of-fit of point process models. A detailed treatment can be found in [25].

A. The Time-Rescaling Theorem

Let $0 < t_1 < t_2 < \cdots$ be a realization of a continuous point process with conditional intensity $\lambda(t) > 0$, i.e. t_k is the first instance at which $N(t_k) = k$. Define the transformation

$$z_k := Z(t_k) = \int_{t_{k-1}}^{t_k} \lambda(t)dt. \tag{39}$$

Then, the transformed point process with events occurring at $t'_k = \sum_{i=1}^{k} z_k$ corresponds to a homogeneous Poisson process with rate 1. Equivalently, $z_1, z_2, \cdots$ are *i.i.d exponential* random variables. The latter can be used to construct statistical tests for the goodness-of-fit.

B. The Komlogorov-Smirnov Test for Homogeneity

Suppose that we have obtained the rescaled process through (39) with the *estimated* conditional intensity. When applying the time-rescaling theorem to the discretized process, if the estimated conditional intensity is close to its true value, the rescaled process is expected to behave as a homogeneous Poisson process with rate 1. The Kolmogorov-Smirnov (KS) test can be used to check for the homogeneity of the process. Let z_k's be the rescaled times and define the transformed rescaled times by the inverse exponential CDF:

$$u_k := 1 - e^{-z_k}.$$

If the true conditional intensity was used to rescale the process, the random variables u_k must be i.i.d. $\mathsf{Uniform}(0, 1]$ distributed. The KS test plots the empirical qualities of u_k's versus the true quantiles of the uniform density given by $b_k = \frac{k-1/2}{J}$, where J is the total number of observed spikes. If the conditional intensity is well estimated, the resulting curve must lie near the 45° line. The asymptotic statistics of the KS distribution can be used to construct confidence intervals for the test. For instance, the 95% and 99% confidence intervals are approximately given by $\pm\frac{1.36}{\sqrt{J}}$ and $\pm\frac{1.63}{\sqrt{J}}$ hulls around the 45° line, respectively.

C. The Autocorrelation Function Test for Independence

In order to check for the independence of the resulting rescaled intervals z_k, the following transformation is used:

$$v_k = \Phi^{-1}(u_k)$$

where Φ is the standard Normal CDF. If the true conditional intensity was used to rescale the process, then v_k's would be i.i.d. Gaussian and their uncorrelatedness would imply independence. The Autocorrelation Function (ACF) of the variables v_k must then be close to the discrete delta function. The 95% and 99% confidence intervals can be considered using the asymptotic statistics of the sample ACF, approximately given by $\pm\frac{1.96}{\sqrt{J}}$ and $\pm\frac{2.575}{\sqrt{J}}$, respectively.

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians Madrid, August 22–30*, 2006, pp. 1433–1452.

[3] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[4] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[5] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[6] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[7] Y. Ogata, "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 9–27, 1988.

[8] D. Vere-Jones, "Stochastic models for earthquake occurrence," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–62, 1970.

[9] R. Barbieri, E. C. Matten, A. A. Alabi, and E. N. Brown, "A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 1, pp. H424–H435, 2005.

[10] G. Valenza, L. Citi, E. P. Scilingo, and R. Barbieri, "Point-process nonlinear models with Laguerre and Volterra expansions: Instantaneous assessment of heartbeat dynamics," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2914–2926, 2013.

[11] M. Egesdal, C. Fathauer, K. Louie, J. Neuman, G. Mohler, and E. Lewis, "Statistical and stochastic modeling of gang rivalries in Los Angeles," *SIAM Undergraduate Research Online*, vol. 3, pp. 72–94, 2010.

[12] E. N. Brown, D. P. Nguyen, L. M. Frank, M. A. Wilson, and V. Solo, "An analysis of neural receptive field plasticity by point process adaptive filtering," *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12 261–12 266, 2001.

[13] A. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Comp.*, vol. 15, no. 5, pp. 965–991, 2003.

[14] 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, no. 5, pp. 456–461, 2004.

[15] L. Paninski, "Maximum likelihood estimation of cascade point-process neural encoding models," *Network: Comp. in Neural Systems*, vol. 15, no. 4, pp. 243–262, 2004.

[16] 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.

[17] L. Paninski, J. Pillow, and J. Lewi, "Statistical models for neural encoding, decoding, and optimal stimulus design," *Progress in brain research*, vol. 165, pp. 493–507, 2007.

[18] J. W. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multi-neuron spike trains," *Neural Comp.*, vol. 23, no. 1, pp. 1–45, 2011.

[19] Z. Chen, D. F. Putrino, S. Ghosh, R. Barbieri, and E. N. Brown, "Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 2, pp. 121–135, 2011.

[20] D. Motamedvaziri, M. H. Rohban, and V. Saligrama, "Sparse signal recovery under Poisson statistics," *arXiv preprint arXiv:1307.4666*, 2013.

[21] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 3990–4002, 2010.

[22] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[23] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2007.

[24] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[25] 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.

[26] T. Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6215–6221, 2011.

[27] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[28] P. Reynaud-Bouret and E. Roy, "Some non asymptotic tail estimates for Hawkes processes," *Bulletin of the Belgian Mathematical Society-Simon Stevin*, vol. 13, no. 5, pp. 883–896, 2007.

[29] A. Kazemipour, B. Babadi, and M. Wu, "Sufficient conditions for stable recovery of sparse autoregressive models," in *50th Conference on Information Sciences and Systems (CISS), Princeton, NJ*, 2016.

[30] P. A. Lewis and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics Quarterly*, vol. 26, no. 3, pp. 403–413, 1979.

[31] E. Brown, R. Barbieri, V. Ventura, R. Kass, and L. Frank, "The time-rescaling theorem and its application to neural spike train data analysis," *Neural computation*, vol. 14, no. 2, pp. 325–346, 2002.

[32] E. G. Jones, M. Steriade, and D. McCormick, *The thalamus*. Plenum Press New York, 1985.

[33] B. Scholl, A. Y. Tan, J. Corey, and N. J. Priebe, "Emergence of orientation selectivity in the mammalian visual pathway," *The Journal of Neuroscience*, vol. 33, no. 26, pp. 10 616–10 624, 2013.

[34] J. Borowska, S. Trenholm, and G. B. Awatramani, "An intrinsic neural oscillator in the degenerating mouse retina," *The Journal of Neuroscience*, vol. 31, no. 13, pp. 5000–5012, 2011.

[35] M. S. Bartlett, "Statistical estimation of density functions," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 245–254, 1963.

[36] M. Bartlett, "The spectral analysis of point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 264–296, 1963.

[37] R. O. Wong, M. Meister, and C. J. Shatz, "Transient period of correlated bursting activity during development of the mammalian retina," *Neuron*, vol. 11, no. 5, pp. 923–938, 1993.

[38] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience*. Lippincott Williams & Wilkins, 2007, vol. 2.

[39] S. J. Eglen, M. Weeks, M. Jessop, J. Simonotto, T. Jackson, and E. Sernagor, "A data repository and analysis framework for spontaneous neural activity recordings in developing retina," *GigaScience*, vol. 3, no. 1, p. 3, 2014.

[40] L. A. Kontorovich, K. Ramanan *et al.*, "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, vol. 36, no. 6, pp. 2126–2158, 2008.

[41] U. Grenander and G. Szegö, *Toeplitz forms and their applications*. Univ of California Press, 1958, vol. 321.

[42] S. A. van de Geer, "On Hoeffding's inequality for dependent random variables," in *Empirical Process Techniques for Dependent Data*, H. Dehling and W. Philipp, Eds. Springer, 2001.

[43] L. Zhu, "Nonlinear Hawkes processes," *arXiv preprint arXiv:1304.7531*, 2013.

[44] P. Brémaud and L. Massoulié, "Stability of nonlinear Hawkes processes," *The Annals of Probability*, pp. 1563–1588, 1996.

[45] A. Kazemipour, B. Babadi, and M. Wu, "Sparse estimation of self-exciting point processes with application to LGN neural modeling," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 478–482.