

# Network Estimation from Point Process Data

Benjamin Mark<sup>1,4</sup>, Garvesh Raskutti<sup>2,4</sup>, and Rebecca Willett<sup>3,4</sup>

<sup>1</sup>Department of Mathematics, <sup>2</sup>Department of Statistics,

<sup>3</sup>Department of Electrical and Computer Engineering

<sup>4</sup>Wisconsin Institute for Discovery

University of Wisconsin-Madison

February 15, 2018

## Abstract

Consider observing a collection of discrete events within a network that reflect how network nodes influence one another. Such data are common in spike trains recorded from biological neural networks, interactions within a social network, and a variety of other settings. Data of this form may be modeled as self-exciting point processes, in which the likelihood of future events depends on the past events. This paper addresses the problem of estimating self-excitation parameters and inferring the underlying functional network structure from self-exciting point process data. Past work in this area was limited by strong assumptions which are addressed by the novel approach here. Specifically, in this paper we (1) incorporate *saturation* in a point process model which both ensures stability and models non-linear thresholding effects; (2) impose general low-dimensional structural assumptions that include sparsity, group sparsity and low-rankness that allows bounds to be developed in the high-dimensional setting; and (3) incorporate long-range memory effects through moving average and higher-order auto-regressive components. Using our general framework, we provide a number of novel theoretical guarantees for high-dimensional self-exciting point processes that reflect the role played by the underlying network structure and long-term memory. We also provide simulations and real data examples to support our methodology and main results.

# 1 Introduction

In a variety of settings, our only glimpse of a network’s structure is through the lens of discrete time series observations. For instance, in a social network, we may observe a time series of members’ activities, such as posts on social media. In electrical systems, cascading chains of power failures reveal critical information about the underlying power distribution network. During epidemics, networks of computers or of a population are reflected by the time at which each node becomes infected. In biological neural networks, firing neurons can trigger or inhibit the firing of their neighbors, so that information about the network structure is embedded within spike train observations.

This paper focuses on estimating the *influence network* which models the extent to which one node’s activity stimulates or inhibits activity in another node. For instance, the network structure may indicate who is influencing whom within a social network [50, 55, 57, 26, 8, 65], the connectivity of neurons [9, 27, 12, 56, 32, 14, 49, 38], interactions among financial instruments [10, 4, 39], how power failures may propagate across the power grid [17], or patterns of criminal activity and military engagements [57, 8, 16, 37, 39]. The interactions between nodes are thus critical to a fundamental understanding of the underlying functional network structure and accurate predictions of likely future events.

Learning the influence network presents a number of challenges both in terms of formulating the model and developing suitable theory and methodology. First, in the applications described above the number of network nodes is typically large relative to the length of time they are observed, making the network parameter *high-dimensional*. Furthermore, the most natural model in these settings are multivariate *self-exciting point processes (SEPPs)*. While empirical work has demonstrated the efficacy of SEPP models in various applications (*cf.*, [16, 10, 37, 39, 17]), little is known about the statistical properties of these estimators. In this paper, we formulate a model and provide a general framework for estimating network parameters in discrete-time high-dimensional SEPP models.

Let  $M$  denote the number of nodes in the network and  $T$  the number of time intervals over which we collect data. We observe  $X_{t,m}$ , the number of events at node  $m$  during time period  $t$ , for  $m = 1, \dots, M$  and  $t = 1, \dots, T$ . We model these counts as

$$X_{t,m} \sim \text{Poisson}(\lambda_{t,m})$$

where the logarithm of  $\lambda_{t,m}$  is a function of the previous counts of events in the network and the interactions between nodes. For a simple example, we might have  $\log \lambda_{t,m} = \sum_{m'=1}^m A_{m,m'} X_{t-1,m'}$ . However, a fundamental challenge associated with SEPP models is that they can be highly unstable: due to the exponential link

function, the counts can diverge even when the interactions  $\{A_{m,m'}\}$  are small. In [23] the authors give extensive justification for the interest in these models from a neuroscience perspective, but also show how learned model parameters can result in generative models that are highly inconsistent with physiological measurements. Existing statistical learning bounds for SEPP models [28] guarantee stability by assuming all network interactions are inhibitory.

A major contribution of this work is learning guarantees for SEPPs without restrictive assumptions on the structure of the network or types of interactions among nodes. We will address stability issues by introducing saturation effects on the rate parameter  $\lambda_{t,m}$ . Saturated SEPP models were recently described in application-driven work without theoretical guarantees [17]. In contrast, this work aims to derive statistical learning guarantees for saturated point processes.

We study a fairly general class of saturated SEPPs whose parameters can be estimated via regularized maximum likelihood estimation. We assume that the number of possible interactions between nodes (*i.e.*, graph edges)  $M^2$  is large relative to the number of time points  $T$ , but that the network has an underlying low-dimensional structure that can be promoted via regularization. The question we address in our theory is how many time points  $T$  are needed to guarantee a desired level of statistical accuracy in terms of the number of nodes  $M$ , the underlying network structure, the regularizer used, and the type of saturation effects introduced?

## 1.1 Relationship to Prior Work

A number of works have studied *linear* SEPPs (where  $\lambda_{t,m}$  is a linear function of past events, in contrast to *log-linear* models, where  $\log \lambda_{t,m}$  is a linear function of past events) from a theoretical perspective. Examples include works on the Hawkes process [11, 15, 54, 29, 6]. In a multivariate Hawkes process setting, one frequently aims to learn the excitation matrix characterizing interactions within the network. In [18] the authors establish that learning the excitation matrix is sufficient for learning the directed information graph of the process. The linear Hawkes process is frequently studied under an assumption there are no inhibitory interactions, although recent work [11] was able to incorporate both inhibitory and stimulatory interactions. Prior work on learning parameters in discrete high-dimensional time series models requires linearity or Gaussianity assumptions (*cf.*, [7]) which do not hold in our model.

In contrast, we study log-linear SEPPs. Prior works have demonstrated the empirical value of log-linear SEPPs [41, 60] and these models are frequently used in the neuroscience community [23]. Moreover, log-linear point process models can be advantageous from the perspective of optimization [47] and naturally allow for inhibitory interactions. However, log-linear SEPPs can not easily model stimulatory

interactions while maintaining stability, and incorporating stimulatory interactions is a major contribution of this paper.

There is limited work on learning rates for log-linear SEPPs, and much of it is only applicable in the setting where  $M$  is small relative to  $T$  [20]. The most related work is our recent work [28] which considers a special case of our SEPP along with a sparsity assumption on the network and applies in the high-dimensional setting. This prior work is limited since the model only considers recent memory, sparsity regularization, and assumes only inhibitory influences to ensure stability and learnability.

## 1.2 Main Contributions

Our paper makes the following major contributions.

- We provide a general upper bound (Theorem 3.1) for developing theoretical guarantees for estimating SEPPs and build on the analysis in [28] in three significant ways. First, we incorporate saturation effects in our model by using a thresholding function in order to ensure stability, and account for these effects in our theory. Second, we provide learning rates for a class of processes which incorporate longer-range dependence effects in a variety of ways, improving upon [28] which only considers first-order auto-regressive models. Finally, we allow for several different regularization choices corresponding to various prior beliefs about the structure of the network.
- We apply our general upper bound to a number of different processes and regularization schemes. For processes with longer-range dependence, we prove that a restricted eigenvalue condition holds for the ARMA(1, 1) and AR(2) models in Lemmas 4.1 and 4.2 respectively.
- In terms of regularization schemes, we consider strict sparsity, group sparsity and low-rank regularization and provide three novel guarantees stated in Theorems 4.4, 4.6 and 4.8. All our mean-squared error bounds match the optimal bounds in the independent case up to log factors.
- A thorough simulation study in Section VI provides support for our theoretical mean-squared error bounds and also examines parameters associated with the magnitude of the entries of  $A^*$  and clipping thresholds.
- We further demonstrate the practical benefits of our regularized likelihood framework on three real data examples. The first involves modeling the interplay between crime events in different neighborhoods of Chicago, the

second modeling connections between different neurons in the brain within a rat during sleep and wake states, and the third involving meme-tracker data in social networks. The three examples illustrate the advantages of using different regularizers.

- Finally, we show that our SEPP framework can be viewed as a discretization of the widely studied Hawkes process, and discuss some advantages of considering point processes in discrete time.

### 1.3 Notation

For a matrix  $A$ , we let  $a_{m\cdot}$  denote the  $m^{\text{th}}$  row of  $A$  and  $a_{\cdot m}$  denote the  $m^{\text{th}}$  column of  $A$ . We then let  $\|a_{m\cdot}\|_{1+}$  denote the sum of the positive entries of  $a_{m\cdot}$  and  $\|a_{m\cdot}\|_{1-}$  denote the absolute value of the sum of the negative entries of  $a_{m\cdot}$ , so that

$$\|a_{m\cdot}\|_1 = \|a_{m\cdot}\|_{1+} + \|a_{m\cdot}\|_{1-}.$$

Given a norm  $\|\cdot\|_{\mathcal{R}}$  on a real vector space, we let  $\|\cdot\|_{\mathcal{R}^*}$  denote its dual norm defined by

$$\|v\|_{\mathcal{R}^*} = \sup_{\|u\|_{\mathcal{R}} \leq 1} \langle u, v \rangle$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard inner product. Throughout the paper, we work with mixed norms

$$\|A\|_{p,q} = \left( \sum_m \|a_{m\cdot}\|_q^p \right)^{\frac{1}{p}},$$

as well as the nuclear norm

$$\|A\|_* = \sum_{i=1}^M \sigma_i(A)$$

where  $\sigma_i(A)$  denotes the  $i$ th singular value of  $A$ , and the operator norm

$$\|A\|_{op} = \sup_{\|x\|_2 \leq 1} \|Ax\|_2.$$

The Frobenius norm, denoted by  $\|\cdot\|_F$ , is a special case of the  $\|\cdot\|_{p,q}$  norm with  $p = q = 2$ .

Finally, we let  $\|A\|_0$  denote the number of nonzero elements of a matrix  $A$ .

## 2 Model Formulation

In this section, we present a class of SEPPs and discuss how saturation effects can be included in order to ensure stability. Recall that  $X_{t,m}$  denotes the number of events

from node  $m$  during time period  $t$ . To start, consider the following model:

$$X_{t+1,m} \sim \text{Poisson}(\lambda_{t+1,m}) \quad (2.1)$$

$$\log(\lambda_{t+1,m}) = \nu_m + \sum_{s=1}^t \sum_{m'=1}^M h_{m,m'}[t-s] X_{s,m'}. \quad (2.2)$$

Here the logarithm of the rate for  $X_{t+1,m}$  is linear in all the previous observations. For each node  $m'$  in the network, that node's count  $X_{s,m'}$  at time  $s$  is scaled by an influence function  $h_{m,m'}$  evaluated at  $t-s$ . The influence function  $h_{m,m'}$  describes the relationship between nodes  $m$  and  $m'$ . As in [11], we assume each influence function can be written as the linear combination of  $K$  known basis functions  $\{\phi_k\}_{k=1}^K$ , *i.e.*,

$$h_{m,m'}[t] = \sum_{k=1}^K a_{m,m',k} \phi_k[t].$$

Hence estimating the network structure amounts to estimating the matrix  $A^* \in \mathbb{R}^{M \times MK}$  where the  $m^{\text{th}}$  row of  $A^*$  is  $\left( \left( a_{m,m',k} \right)_{m'=1}^M \right)_{k=1}^K$ . It will be convenient to rewrite (2.2) in matrix-vector form as

$$\log(\lambda_{t+1}) = \nu + A^* g(\mathcal{X}_t), \quad (2.3)$$

where  $\mathcal{X}_t = [X_1, \dots, X_t]$  denotes the history of the process up to time  $t$  and  $g(\mathcal{X}_T) \in \mathbb{R}^{MK \times 1}$  is the vector defined as follows. For  $k = 1, \dots, K$ , let

$$g_k(\mathcal{X}_T) := \begin{bmatrix} \sum_{s=1}^T X_{s,1} \phi_k[T-s] \\ \sum_{s=1}^T X_{s,2} \phi_k[T-s] \\ \vdots \\ \sum_{s=1}^T X_{s,M} \phi_k[T-s] \end{bmatrix} \quad (2.4)$$

and

$$g(\mathcal{X}_T) := \begin{bmatrix} g_1(\mathcal{X}_T) \\ g_2(\mathcal{X}_T) \\ \vdots \\ g_K(\mathcal{X}_T) \end{bmatrix}. \quad (2.5)$$

A number of commonly studied discrete time models can be realized in this manner. We briefly mention two which are discussed further in Section IV. As a first example,  $K = 1$  and  $\phi[t] = \alpha^t$  corresponds to an autoregressive moving average ARMA(1, 1) process. When  $K = p$  and  $\phi_k[t] = \mathbb{I}_{\{k=t\}}$  where  $\mathbb{I}_{\{B\}} :=$

$\begin{cases} 1, & B \text{ true} \\ 0, & \text{otherwise} \end{cases}$  is the indicator function, we recover the AR(p) process. This second example shows the value in assuming that  $h_{m,m'}$  is in the span of a collection of basis functions, rather than just one. Allowing for multiple basis functions allows us to study processes which incorporate higher order effects in more sophisticated ways than would be possible with only one basis function.

We let  $\nu_{\min}$  and  $\nu_{\max}$  be upper and lower bounds on the constant offset parameter  $\nu_m$  in (2.2) and we assume that  $A^*$  lies within a set  $\mathcal{A}$  which we define as follows. Let  $a_{\max}$  be an upper bound on  $\|a_m^*\|_{1+}$  and similarly let  $a_{\min}$  be an upper bound on  $\|a_m^*\|_{1-}$ . We let  $\mathcal{A}$  denote the set of  $M \times MK$  matrices with  $\|a_m\|_{1-} \leq a_{\min}$  and  $\|a_m\|_{1+} \leq a_{\max}$  for all  $m$ . With the assumption that  $A^* \in \mathcal{A}$  we can search for an estimate  $\hat{A}$  of  $A^*$  over the bounded set  $\mathcal{A}$ .

## 2.1 Saturation

As discussed in the introduction, point process models along the lines of (2.3) are widely used to describe count data in a variety of applications. However, due to instability issues inherent to SEPPs of this form, these models can be highly unstable and lead to unbounded counts. Hence, pure SEPPs make poor generative models (c.f., [23]) and are difficult to understand theoretically without making overly restrictive assumptions about  $A^*$  (c.f., [28]). We will address this problem by introducing saturation effects to the vector  $g(\mathcal{X}_t)$  defined in Equation (2.5). The application focused work [17] introduced saturated SEPPs, but to the best of our knowledge, this is the first work to study the theoretical properties of saturated models. To address stability issues we adjust the definition of  $g_k(\mathcal{X}_T)$  in (2.4) to the following:

$$g_k(\mathcal{X}_T) = \begin{bmatrix} \sum_{s=1}^T \min(X_{s,1}, \tilde{U}) \phi_k[T-s] \\ \sum_{s=1}^T \min(X_{s,2}, \tilde{U}) \phi_k[T-s] \\ \vdots \\ \sum_{s=1}^T \min(X_{s,M}, \tilde{U}) \phi_k[T-s] \end{bmatrix}. \quad (2.6)$$

That is, each past count which exceeds some threshold  $\tilde{U} \geq 1$  gets clipped to  $\tilde{U}$ . Further, we assume that

$$\sum_{s=1}^{\infty} \phi_k[s] \leq \tau < \infty$$

for each basis function, so that each entry of  $g(\mathcal{X}_t)$  in (2.5) is bounded by

$$\tilde{U} \sum_{s=1}^T \phi_k[s] \leq \tau \tilde{U} =: U.$$

In other words, with clipping we have  $\|g(\mathcal{X}_t)\|_\infty \leq U$ , guaranteeing the stability of our process.

In particular, this allows us to define the maximum and minimum Poisson rate from which each observation can be drawn. We denote the maximum and minimum rates by

$$R_{\max} = \exp(\nu_{\max} + a_{\max}U) \quad (2.7a)$$

$$R_{\min} = \exp(\nu_{\min} - a_{\min}U). \quad (2.7b)$$

Throughout this paper, we take  $\min(\cdot, \tilde{U})$  to be our saturation function for simplicity. However, our theory extends to other saturation functions provided that the function is bounded, which is crucial for our analysis. The details are provided in Proposition 2 in the appendix.

While this framework has advantages, a central question we need to address is how departing from the standard SEPP framework and incorporating non-linear saturation effects change our estimation errors.

## 2.2 Regularized optimization formulation

In the high-dimensional setting, the number of potential pairwise interactions,  $M^2$ , is large relative to the number of time periods,  $T$ , making standard maximum likelihood optimization techniques unsuitable. Instead, we assume some prior knowledge on the parameter  $A^*$ , which can be incorporated in estimation via a regularization term  $\|\cdot\|_{\mathcal{R}}$ . Specifically, we consider the estimator

$$\hat{A} = \arg \min_{A \in \mathcal{A}} \sum_{t=0}^T \sum_{m=1}^M \exp(\nu_m + a_m^\top g(\mathcal{X}_t)) - X_{t+1,m}(\nu_m + a_m^\top g(\mathcal{X}_t)) + \lambda \|A\|_{\mathcal{R}} \quad (2.8)$$

where the first two terms of (2.8) are the negative log-likelihood of the observed data given  $A$ . We discuss various choices of regularization  $\|\cdot\|_{\mathcal{R}}$  in the next section. Note that the optimization problem in (2.8) is convex. Further, it can easily be generalized to unknown  $\nu$ ; we omit this discussion here for simplicity of presentation.

## 3 Statistical learning bounds

### 3.1 Decomposable Regularizers

Our learning bounds apply to general decomposable regularizers introduced in [46]. Given a subspace  $\overline{\mathcal{M}} \subseteq \mathbb{R}^{M \times MK}$ , we define its orthogonal complement as

$$\overline{\mathcal{M}}^\perp = \{v \in \mathbb{R}^{M \times MK} \mid \langle u, v \rangle = 0 \text{ for all } u \in \overline{\mathcal{M}}\}.$$



Given a normed vector space  $(\mathbb{R}^{M \times MK}, \|\cdot\|_{\mathcal{R}})$  and subspaces  $\mathcal{M} \subseteq \overline{\mathcal{M}} \subseteq \mathbb{R}^{M \times MK}$ , we say  $\mathcal{R}$  is a decomposable regularizer with respect to  $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$  if for  $A \in \mathcal{M}$  and  $B \in \overline{\mathcal{M}}^\perp$  we have

$$\|A + B\|_{\mathcal{R}} = \|A\|_{\mathcal{R}} + \|B\|_{\mathcal{R}}.$$

This definition encompasses widely-studied regularizers including the  $l_1$  norm, nuclear norm, and the group sparsity inducing  $\|\cdot\|_{1,2}$  norm. We refer the reader to [46] for more details and intuition. While working in this general framework allows us to incorporate a wide variety of prior beliefs about the structure of our network, a fundamental question we need to address is how the specific choice of regularizer affects our learning rates. Due to the temporal dependence and non-linearities in SEPP models, deriving learning rates for various decomposable regularizers requires us to leverage martingale concentration inequalities.

## 3.2 Assumptions

In Section II we presented a class of SEPPs which depends on a choice of basis functions, and a general RMLE procedure which depends on a choice of regularization penalty. In this section, we introduce four assumptions which are needed for our theoretical guarantees. We then give examples where we show that for certain choices of basis functions  $\{\phi_1, \dots, \phi_K\}$  and regularizers  $\|\cdot\|_{\mathcal{R}}$  of interest the assumptions hold with high probability.

Our first assumption depends on the basis functions but is not related to the choice of regularizer.

**Assumption 1 (Restricted Eigenvalue).** There exists some  $\omega > 0$  and  $p \in \mathbb{N}$  such that smallest eigenvalue of  $\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-p}]$  is lower bounded by  $\omega$  for all  $t$ .

Assumption 1 is analogous to various restricted eigenvalue conditions in other works. However, in much of the literature, one needs to lower bound the eigenvalues of a sensing matrix whose columns are assumed to be independent. Dependence introduced in our autoregressive model makes this a more complex condition to verify. In past work on sparse autoregressive inference (*e.g.*, [7]), restricted eigenvalue conditions have been framed in terms of a stationary covariance matrix.

Informally, the value of  $\omega$  measures the strength of the intertemporal dependence of our process. If our network and basis functions are structured such that strong long-range dependencies exist, then the smallest eigenvalue can be near zero, leading to a poor bound on the error  $\|\hat{A} - A^*\|_F^2$ .

The RE condition must also account for the level of clipping in our process: if the network is so stimulatory that most observations are clipped, then the matrix

$\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-p}]$  will be nearly singular and  $\omega$  will be close to zero. Thus, to come up with an acceptable bound on  $\omega$ , we need to establish that our network is well-behaved enough that most observations will be unclipped. In other words, our theory suggests that introducing non-linear saturation effects will not ruin our ability to infer the structure of our network, provided that our network is not too stimulatory and is usually stable without clipping. A further discussion of the intuition behind the RE condition is provided in Example 1.

Next, we present assumptions which need to be verified in terms of the regularizer used. Recall that we assume the regularizer  $\|\cdot\|_{\mathcal{R}}$  used in Equation (2.8) is decomposable with respect to the pair of subspaces  $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ .

**Assumption 2** (Subspace Compatibility). There exists a constant  $\Psi(\overline{\mathcal{M}})$  satisfying

$$\sup_{A \in \overline{\mathcal{M}}} \frac{\|A\|_{\mathcal{R}}}{\|A\|_F} \leq \Psi(\overline{\mathcal{M}}).$$

Assumption 2 is a subspace compatibility condition as in [46], which controls how large the Frobenius norm can be relative to the  $\mathcal{R}$  norm on the subspace  $\overline{\mathcal{M}}$ .

**Assumption 3** (Cone Row Sparsity). Let  $A_{\overline{\mathcal{M}}}$  and  $A_{\overline{\mathcal{M}}^\perp}$  denote the projections of a matrix  $A$  onto the subspaces  $\overline{\mathcal{M}}$  and  $\overline{\mathcal{M}}^\perp$  respectively. Define

$$\mathcal{B}'_{\mathcal{R}} = \left\{ A \in \mathbb{R}^{M \times MK} : \|A_{\overline{\mathcal{M}}^\perp}\|_{\mathcal{R}} \leq 3\|A_{\overline{\mathcal{M}}}\|_{\mathcal{R}} \text{ and } \|A\|_F = 1 \right\}.$$

Then there exists a constant  $\mu_{\mathcal{R}}$  such that

$$\sup_{B \in \mathcal{B}'_{\mathcal{R}}} \|B\|_{2,1}^2 \leq \mu_{\mathcal{R}}.$$

Assumption 3 corresponds to assuming some notion of row sparsity on the error matrix  $\hat{A} - A^* =: \Delta$ . It is needed to apply the empirical process techniques from [28].

**Assumption 4** (Deviation Bound). Let

$$\epsilon_{t,m} = X_{t+1,m} - \exp(\nu_m + a_m^\top g(\mathcal{X}_t)),$$

then there exists a constant  $\lambda < \infty$  such that

$$\left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t g(\mathcal{X}_t)^\top \right\|_{\mathcal{R}^*} \leq \frac{\lambda}{2}.$$

Assumption 4 is similar to deviation bound conditions found in the literature. Due to the temporal dependence across observations, we must use martingale concentration inequalities under various norms in order to verify it.

### 3.3 General result

Provided our process and estimation procedure satisfy Assumptions 1-4 for reasonable constants, we can guarantee the learnability of our model.

**Theorem 3.1.** *Assume  $(X_t)_{t=1}^T$  is generated by (2.3) and satisfies Assumptions 1-4 and assume  $A^*$  is estimated according to (2.8) with a regularizer  $\|\cdot\|_{\mathcal{R}}$  that is decomposable with respect to the subspaces  $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ . Then*

$$\|\hat{A} - A^*\|_F^2 \leq \frac{36p\Psi(\overline{\mathcal{M}})\lambda^2}{R_{\min}^2\omega^2}$$

with probability at least  $1 - \frac{2}{M^2}$  for

$$T \geq \frac{128p^2U^4\mu_{\mathcal{R}}^2 \log M}{\omega^2}$$

for constants  $C, c > 0$  which are independent of  $M, T$  and  $\Psi(\overline{\mathcal{M}})$ .

Theorem 3.1 is a direct consequence of Theorem 1 in [46] combined with Theorem 1 in [28]. Specifically, [46] gives Theorem 3.1 in a general decomposable regularizer setup under a restricted strong convexity (RSC) assumption, which in our language states that the error  $\Delta =: \hat{A} - A^*$  satisfies

$$\frac{1}{T} \sum_t \sum_m (\Delta_m^\top g(\mathcal{X}_t))^2 \geq k\|\Delta\|_F^2$$

for some  $k > 0$ . Due to the fact that our process is neither linear nor Gaussian, many techniques, *e.g.*, [7] [44], used to establish an RSC condition directly are unworkable in our setting. Instead, we use similar techniques to Theorem 1 in [28] which uses empirical process results to turn the RSC assumption into the restricted eigenvalue (RE) condition in Assumption 1.

## 4 Examples

In order to use Theorem 3.1, we need to prove that the four assumptions hold for basis functions and regularizers of interest. First, we show that Assumption 1 is satisfied for different point process models. Second we show Assumptions 2-4 are satisfied for a class of regularizers. Finally, we combine the results from this section with Theorem 3.1 to give overall learning rates for ARMA(1, 1) and AR(2) processes under different regularization schemes.

Recall that the constants  $R_{\max}$  and  $R_{\min}$  from (2.7) denote the maximum and minimum Poisson rate for each observation.

## 4.1 Specific Point Process Models

**Example 1: ARMA(1, 1) process** First-order autoregressive moving average (ARMA(1, 1)) point process models have been studied in a variety of settings [37, 35]. Moreover, the corresponding continuous time model is one of the most frequently studied point process models [4, 13]. Consider the following saturated ARMA(1, 1) model with memory parameter  $\alpha \in [0, 1)$ :

$$\begin{aligned} X_{t+1} &\sim \text{Poisson}(\lambda_{t+1}) \\ \log(\lambda_{t+1}) &= \nu + A^* \min(X_t, \tilde{U}) + \alpha \log(\lambda_t). \end{aligned} \quad (4.1)$$

Algebraic manipulation shows that (4.1) is a special case of (2.3) with  $K = 1$  basis function corresponding to  $\phi[t] = \alpha^t$ . Here  $\alpha$  is a memory parameter which captures the strength of the long-range dependence in the process, and  $\|g(\mathcal{X}_t)\|_\infty \leq \frac{\tilde{U}}{1-\alpha} = U$  so that

$$\begin{aligned} R_{\max} &= \exp\left(\nu_{\max} + \frac{a_{\max}\tilde{U}}{1-\alpha}\right) \\ R_{\min} &= \exp\left(\nu_{\min} - \frac{a_{\min}\tilde{U}}{1-\alpha}\right). \end{aligned}$$

An AR(1) process, corresponding to (4.1) where  $\alpha = 0$ , was considered in [28]. However, due to the inherent instability of SEPPs without saturation, the authors were forced to assume  $a_{\max} = 0$ .

**Lemma 4.1.** *Suppose  $(X_t)_{t=1}^T$  is generated according to (4.1). Then Assumption 1 is satisfied with*

$$\omega = \min\left(\frac{1}{2}R_{\min}, \kappa\right)$$

where  $\kappa$  is a constant depending on  $\tilde{U}$ ,  $\alpha$  and  $a_{\max}$  but independent of  $M$ .

The proof of Lemma 4.1 requires us to account for the effects of clipping. We show that finding a lower bound on the eigenvalues of  $\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-p}]$  can be reduced to finding a lower bound on  $\text{Var}(\min(X_{t,m}, \tilde{U}) | \mathcal{X}_{t-1})$ , which simplifies the calculation since we rely on first-order dependence. Since we need to construct a lower bound on  $\text{Var}(\min(X_{t,m}, \tilde{U}) | \mathcal{X}_{t-1})$  we consider the two cases when the variance will be smallest.

In particular, if  $X_{t,m} \sim \text{Poisson}(R_{\min})$ , then its variance will be small because the variance and mean of a Poisson random variable are equal. Specifically, when  $X_{t,m} \sim \text{Poisson}(R_{\min})$  we lower bound the variance of  $\min(X_{t,m}, \tilde{U})$  by  $\frac{1}{2}R_{\min}$ .

On the other hand, when  $X_{t,m} \sim \text{Poisson}(R_{\max})$  and  $R_{\max}$  is large relative to  $\tilde{U}$ , then  $\min(X_{t,m}, \tilde{U})$  is likely to be  $\tilde{U}$  (clipped), so again the variance will be small.

We lower bound the variance by the constant  $\kappa$  that is the variance of a Bernoulli random variable, where one outcome corresponds to a Poisson random variable  $Z \sim \text{Poisson}(R_{\max})$  exceeding  $\tilde{U}$  (clipped) and the other outcome corresponds to  $Z < \tilde{U}$ .

One of these two worst case scenarios will give an absolute lower bound on the variance. In both cases we construct a lower bound on the variance independent of  $M$ . Note that  $\omega$  increases with  $R_{\min} = \exp(\nu_{\min} - \frac{a_{\min}\tilde{U}}{1-\alpha})$  so  $\omega$  grows inversely with  $\alpha$ . In other words, as the long range memory of the process increases, Lemma 4.1 suggests that network estimation becomes more difficult. This is consistent with prior work [7].

The value  $\kappa^{-2}$  may be viewed as a proxy for the rate of clipping, and the appearance of  $\kappa$  in Lemma 4.1 illustrates a tradeoff associated with clipped models. On one hand, clipping ensures a stable process. However, as clipping increases, it also increases the temporal dependencies among observations, leading to a smaller  $\kappa$  and larger error bound.

In Figure 1, we fix  $\alpha = 0$  and get a sense of the value of  $\kappa$  for varying  $a_{\max}$  and  $\tilde{U}$ . (Recall that larger  $\kappa$  corresponds to a better-posed estimation problem.) We see that for small  $\tilde{U}$ ,  $\kappa$  is not prohibitively small for a wide range of values of  $a_{\max}$ . However, as  $\tilde{U}$  increases the range of reasonable  $a_{\max}$  decreases, and as  $\tilde{U}$  approaches  $\infty$ , we approach the  $a_{\max} = 0$  setting from [28]. To understand this trend, we consider a special case of (4.1) with  $M = 1$ ,  $\nu = 0$ ,  $\alpha = 0$ ,  $A^* = \frac{3}{10}$  and  $\tilde{U} = 1000$ . In this case, the process follows:

$$X_{t+1} \sim \text{Poisson} \left( \exp \left( \frac{3 \min(X_t, 1000)}{10} \right) \right).$$

Given a small  $X_0$ , this process will not be clipped for the first few observations, but eventually the process will diverge and reach the clipping threshold of 1000. This will happen within the first 100 observations with probability  $\approx 1$ . Once an observation reaches  $X_t \geq 1000$  and is clipped, the next observation will follow

$$\begin{aligned} X_{t+1} &\sim \text{Poisson} \left( \exp \left( \frac{3 \min(X_t, 1000)}{10} \right) \right) \\ &= \text{Poisson}(\exp(300)). \end{aligned}$$

and so  $X_{t+1}$  is virtually guaranteed to be clipped as well. In other words, once we actually reach the clipping threshold, we enter a constant clipping regime which is reflected in the small value of  $\kappa$  for  $a_{\max} = .3$  and  $\tilde{U} = 1000$ .

On the other hand, if  $\tilde{U} = 6$  our process follows

$$X_{t+1} \sim \text{Poisson} \left( \exp \left( \frac{3 \min(X_t, 6)}{10} \right) \right).$$

Even when  $X_t \geq 6$ , the Poisson rate  $\exp(\frac{18}{10})$  is approximately 6, and so the next observation is reasonably likely to be unclipped and we do not enter the constant clipping loop as when  $\tilde{U} = 1000$ . In other words, over the long run we experience less clipping for smaller  $\tilde{U}$ , and thus  $\kappa$  is larger for smaller  $\tilde{U}$ .

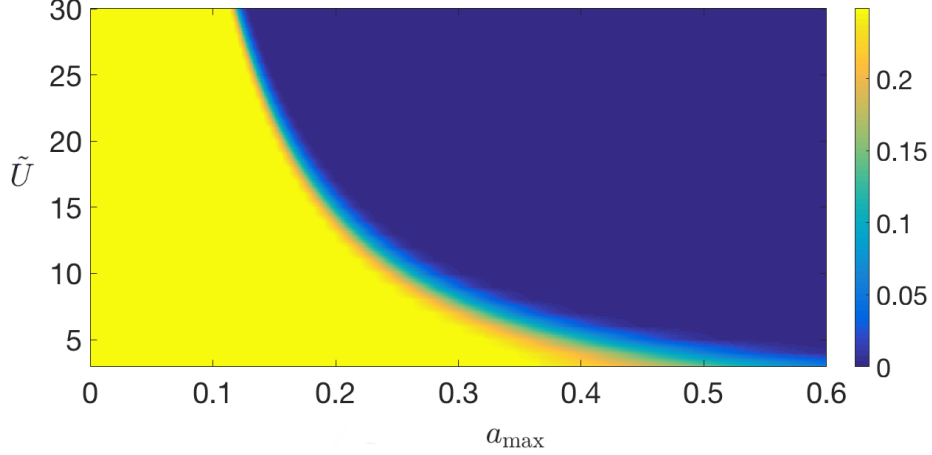


Figure 1: Colors indicate the value of  $\kappa$  for a given  $(a_{\max}, \tilde{U})$  pair. Note that due to our exponential link functions, elements of  $A^*$  above one would be unreasonably excitatory for many networks, and our process can be significantly stimulatory even with coefficients well below one.

**Example 2: AR(2) process** As a second example, we consider an AR process with two time lags:

$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1}) \quad (4.2)$$

$$\log(\lambda_{t+1}) = \nu + A_1^* \min(X_t, \tilde{U}) + A_2^* \min(X_{t-1}, \tilde{U}). \quad (4.3)$$

This process fits within the framework of (2.3) with two basis functions corresponding to:  $\phi_1[t] = \mathbb{I}_{\{t=1\}}$  and  $\phi_2[t] = \mathbb{I}_{\{t=2\}}$ ,  $\underbrace{A^*}_{M \times 2M} = [A_1^*, A_2^*]$ . This example illustrates

the benefit of considering a basis with more than one element to describe the influence functions  $h_{m,m'}$ . A richer class of higher order models can be expressed with multiple basis functions. Under this setup, the maximum and minimum possible Poisson rates are

$$R_{\max} = \exp(\nu_{\max} + a_{\max} \tilde{U})$$

and

$$R_{\min} = \exp(\nu_{\min} - a_{\min} \tilde{U}).$$

Learning rates for high-dimensional linear AR(p) processes with Gaussian noise were studied in [7]. However, the techniques used in that work to prove a restricted eigenvalue condition relied heavily on the Gaussianity of the process. We prove that the restricted eigenvalue condition in Assumption 1 holds for the AR(2) process in Lemma 4.2.

In order to state Lemma 4.2 we first need several definitions. A node  $m$  is said to be a *parent* of node  $m'$  if it influences  $m'$  through  $A_1^*$ , while  $m'$  is said to be a *child* of  $m$ . Furthermore, two nodes are said to be *siblings* if they share a parent node.

**Lemma 4.2.** *Suppose  $(X_t)_{t=1}^T$  is generated according to (4.3). Let  $\rho_m^{(p)}$  denote the number of parents of  $m$ , let  $\rho_m^{(c)}$  denote the children of  $m$  and let  $\rho_m^{(s)}$  denote the number of siblings of  $m$ . Then*

$$\lambda_{\min}(\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-2}]) \geq r_\rho > 0$$

for a constant  $r_\rho$  depending on  $R_{\max}$ ,  $R_{\min}$ ,  $\rho_m^{(p)}$ ,  $\rho_m^{(c)}$ ,  $\rho_m^{(s)}$  but independent of  $M$ .

The constant  $r_\rho$  scales inversely with  $\rho_m^{(p)}$ ,  $\rho_m^{(c)}$ ,  $\rho_m^{(s)}$  and  $R_{\max} - R_{\min}$ . In the high dimensional setting, this means a sparsity assumption on  $A^*$  is necessary for our bound to be useful.

We prove Lemma 4.2 by showing that the matrix  $\text{Cov}(g(\mathcal{X}_t)|\mathcal{X}_{t-2})$  is strictly diagonally dominant. A matrix  $B$  is said to be strictly diagonally dominant if there exists a constant  $\omega > 0$  such that  $b_{i,i} - \sum_{j \neq i} |b_{i,j}| \geq \omega$  for all  $i$ , and the eigenvalues of a symmetric strictly diagonally dominant matrix are lower bounded by  $\omega$ . With a sparsity assumption on  $A^*$ , almost all of the off diagonal terms in  $\text{Cov}(\mathcal{X}_t)$  will be zero, and the remaining terms can be controlled with the techniques from Lemma 4.1 and appropriate assumptions on the size of  $R_{\min}$  and  $R_{\max}$  relative to the sparsity constants  $\rho_m^{(p)}$ ,  $\rho_m^{(c)}$ ,  $\rho_m^{(s)}$ .

## 4.2 Regularization Examples

In this subsection, we verify Assumptions 2-4 under various regularization schemes.

**Example 1: Element-wise Sparsity Regularization** We first explore sparsity regularization for these processes that accounts for the sparsity of  $A^*$  natural to many application domains. For the remainder of the section, we assume

$$\|A^*\|_0 = s \ll M^2.$$

Sparse models of network structure encapsulate essential aspects of many common statistical network models [24], and have connections to stochastic block models,

exponential random graph models, and various graphical models. We consider the regularizer

$$\|A\|_{1,1} = \sum_i \sum_j |a_{i,j}|$$

along with its dual

$$\|A\|_{\infty,\infty} = \max_i \max_j |a_{i,j}|.$$

To see that  $\|\cdot\|_{1,1}$  is decomposable we first define the set

$$S = \{(i, j) \in \{1, \dots, M\} \times \{1, \dots, MK\} : A_{i,j}^* \neq 0\},$$

and next define

$$\mathcal{S} = \{s \in \mathcal{R}^{M \times MK} : s_{i,j} = 0 \text{ for all } (i, j) \notin S\}.$$

Then  $\|\cdot\|_{1,1}$  is decomposable with respect to the pair  $(\mathcal{S}, \mathcal{S}^\perp)$ .

Note that the optimization problem corresponding to  $\|\cdot\|_{1,1}$  regularization is convex and can be solved with a variety of sparse regularization solvers. Furthermore, it can trivially be parallelized across the rows of  $A$ .

**Lemma 4.3.** *Suppose  $(X_t)_{t=1}^T$  is generated according to (2.3) with  $\|A^*\|_0 = s$ . Further, assume  $A^*$  is estimated according to (2.8) using  $\|\cdot\|_{1,1}$  regularization. Then*

(a) *Assumption 2 is satisfied with*

$$\Psi(\bar{S}) = 4\sqrt{s}.$$

(b) *Assumption 3 is satisfied with*

$$\mu_{(1,1)} = 4\sqrt{s}.$$

(c) *Assumption 4 is satisfied with*

$$\left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t g(\mathcal{X}_t)^\top \right\|_{\infty,\infty} \leq CR_{\max} \frac{\log^3(MT)}{\sqrt{T}}$$

*with probability at least  $1 - \frac{1}{(MT)^c}$  for constants  $C, c > 0$  which are independent of  $M, T$  and  $s$ .*



## Proof Overview

- To verify Assumption 4, we rely on the fact that the Poisson rate can never exceed  $R_{\max}$ . This allows us to bound the largest recorded observation by  $C \log(MT)$  with high probability. From here, we are in a position to use martingale concentration inequalities developed in [33] to establish the deviation bound.
- Assumptions 2 and 3 are straightforward consequences of the relation between  $l_1$  and  $l_2$  norms.

Combining Lemma 4.3, Theorem 3.1 and the restricted eigenvalues results from the previous subsection gives overall bounds for sparse SEPPs which are applicable in the high-dimensional setting.

**Theorem 4.4. (*Learning rates for  $l_1$  regularization*)** Suppose  $(X_t)_{t=1}^T$  follows the SEPP framework of (2.3) and  $A^*$  is estimated using sparsity regularization.

(a) If  $(X_t)_{t=1}^T$  is generated according to the ARMA(1, 1) model in (4.1) then

$$\|\hat{A} - A^*\|_F^2 \leq C \frac{R_{\max}^2}{R_{\min}^2 \min(\frac{1}{2}R_{\min}, \kappa)^2} \frac{s \log^6(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 s \frac{\log M}{\min(\frac{1}{2}R_{\min}, \kappa)^2}$$

for constants  $C, c > 0$  which are independent of  $M, T$  and  $s$ .

(b) If  $(X_t)_{t=1}^T$  is generated according to the AR(2) model in (4.3) then

$$\|\hat{A} - A^*\|_F^2 \leq C \frac{R_{\max}^2}{R_{\min}^2 r_\rho^2} \frac{s \log^6(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 s \frac{\log M}{r_\rho^2}.$$

The mean-squared error bound  $\frac{s \log^6(MT)}{T}$  matches the minimax optimal rate in the independent case [52] up to log factors. Theorem 4.4 extends results in Hall et al. [28] to ARMA(1,1) and AR(2) processes.

**Example 2: Group Sparsity** Group lasso regularization is a popular tool used to estimate a sparse parameter where one has prior knowledge on the structure of the sparsity (see *e.g.*, [62] for more details). We consider a special case of group lasso regularization where the groups are the columns of the matrix. Let  $a_{\cdot m}$  denote the  $m^{\text{th}}$  column vector of a matrix  $A$ . Our structured sparsity assumption is that only  $s_G \ll M$  columns of  $A^*$  contain nonzero entries.

In terms of network structure, this means that only a small number of hub nodes have influence on other nodes in the network. To estimate networks of this form, we consider  $l_2$  penalization on the columns vectors, followed by  $l_1$  penalization on the resulting  $l_2$  norms. In other words, we have

$$\|A\|_G = \|A^\top\|_{1,2} = \sum_m \|a_{\cdot m}\|_2.$$

The dual of this norm is

$$\|A\|_{G^*} = \|A^\top\|_{\infty,2} = \max_m \|a_{\cdot m}\|_2.$$

Let

$$S_G = \{i : a_{\cdot i}^* \neq 0\}.$$

Then  $\|\cdot\|_G$  is decomposable with respect to the subspaces

$$\mathcal{S}_G = \{A : a_{\cdot j} = 0 \text{ for all } j \notin S_G\}$$

and

$$\mathcal{S}_G^\perp = \{A : a_{\cdot j} = 0 \text{ for all } j \in S_G\}.$$

We show Assumptions 2-4 hold in Lemma 4.5 below.

**Lemma 4.5.** *Suppose  $(X_t)_{t=1}^T$  is generated according to (2.3) where only  $s_G$  columns of  $A^*$  contain nonzero entries. Further, assume  $A^*$  is estimated according to (2.8) using  $\|\cdot\|_G$  regularization. Then*

(a) *Assumption 2 is satisfied with*

$$\Psi(\mathcal{S}_G) = 4\sqrt{s_G}.$$

(b) *Assumption 3 is satisfied with*

$$\mu_G = 16s_G.$$

(c) *Assumption 4 is satisfied with*

$$\left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t g(\mathcal{X}_t)^\top \right\|_{G^*} \leq C R_{\max} \log^2(MT) \sqrt{\frac{M}{T}}$$

*with probability at least  $1 - \frac{1}{(MT)^c}$ .*

## Proof Overview

- For Assumption 4 we construct a high probability bound on the  $l_2$  norm of each column of our noise matrix and take a union bound over all the columns to get a final bound on the  $\|\cdot\|_{\infty,2}$  norm. To bound the norm of each individual column, we rely on [51] which provides martingale concentration inequalities for 2-smooth norms.
- For Assumption 3 we derive an error row sparsity constant which depends only on  $s_G$  rather than  $M$ . The  $\|\cdot\|_{2,1}$  norm can be large relative to the Frobenius norm in cases where the matrix is row-dense. In this case, the  $l_1$  norm of each row can be on the order of  $\sqrt{M}$  larger than the  $l_2$  norm. However, we only need to derive a compatibility constant on the cone

$$\mathcal{B}_G = \{B \in \mathbb{R}^{M \times MK} : \|B_{\overline{\mathcal{S}}_G^\perp}\|_G \leq 3\|B_{\overline{\mathcal{S}}_G}\|_G\}.$$

Since elements of  $\overline{\mathcal{S}}_G$  has at most  $s_G$  nonzero entries in each row, we can think of all matrices in the cone  $\mathcal{B}_G$  as being “almost row sparse” and so the  $\|\cdot\|_{2,1}$  norm should not be  $O(\sqrt{M})$  larger than the Frobenius norm on the cone.

- Assumption 2 follows from the relationship between the  $l_1$  and  $l_2$  norms.

Combining Lemma 4.5, Theorem 3.1 and the restricted eigenvalue conditions from the previous subsection gives the following result.

### **Theorem 4.6. (Learning rates for group lasso regularization)**

Suppose  $(X_t)_{t=1}^T$  follows the SEPP framework of (2.3) and  $A^*$  is estimated using column group lasso regularization.

(a) If  $(X_t)_{t=1}^T$  is generated according to the ARMA(1, 1) model in (4.1) then

$$\|\hat{A} - A^*\|_F^2 \leq \frac{C}{R_{\min}^2 \min(\frac{1}{2}R_{\min}, \kappa)^2} \frac{s_G M \log^4(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 s_G^2 \frac{\log M}{\min(\frac{1}{2}R_{\min}, \kappa)^2}.$$

for constants  $C, c > 0$  which are independent of  $M, T$  and  $s_G$ .

(b) If  $(X_t)_{t=1}^T$  is generated according to the AR(2) model in (4.3) then

$$\|\hat{A} - A^*\|_F^2 \leq \frac{C}{R_{\min}^2 r_\rho^2} \frac{s_G M \log^4(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 s_G^2 \frac{\log M}{r_\rho^2}.$$

**Example 3: Low-rank Regularization** Estimation of high-dimensional but low-rank matrices is a widely studied problem with numerous applications [44, 53, 64, 36, 19]. Low-rank models can be seen as a generalization of sparse models, where the matrix is sparse in an unknown basis. A standard technique to estimate a low-rank matrix is to take a convex relaxation of an  $l_0$  penalty on the singular values [19]: the nuclear norm penalty

$$\|A\|_* = \sum_{i=1}^M \sigma_i(A),$$

where  $\sigma_i(A)$  denotes the  $i$ th singular value of  $A$ . The dual to the nuclear norm is the operator norm

$$\|A\|_{op} = \sup_{\|x\|_2=1} \|Ax\|_2.$$

As discussed in [46], the nuclear norm is decomposable with respect to the subspaces

$$\mathcal{W} = \{A \in \mathbb{R}^{M \times MK} : \text{row}(A) \subseteq \text{row}(A^*) \text{ and } \text{col}(A) \subseteq \text{col}(A^*)\}$$

and

$$\overline{\mathcal{W}}^\perp = \{A \in \mathbb{R}^{M \times MK} : \text{row}(A) \subseteq \text{row}(A^*)^\perp \text{ and } \text{col}(A) \subseteq \text{col}(A^*)^\perp\},$$

where  $\text{row}(A)$  and  $\text{col}(A)$  denote the row and column spaces of  $A$  respectively. Unlike the previous two examples, here  $\mathcal{W} \neq \overline{\mathcal{W}}$ .

In this low-rank setup, there is no limitation on the number of nodes which can influence a given node. This introduces challenges in establishing Assumption 3, which guarantees near row sparsity of the error. In order to get around this, we impose a technical assumption on  $\|A^*\|_{2,1}$  in Lemma 4.7. An area of interest for future work is to examine whether our estimation procedure is flawed when one node can have many nodes influence it, or whether the need for this assumption is an artifact of our analysis.

**Lemma 4.7.** *Suppose  $(X_t)_{t=1}^T$  is generated according to (2.3) with  $\text{rank}(A^*) = r$  and  $\|A^*\|_{2,1}^2 = D\sqrt{M}$  for a universal constant  $D$ . Further, assume  $A^*$  is estimated according to (2.8) over the ball  $\{A : \|A\|_{2,1}^2 \leq D\sqrt{M}\}$  using nuclear norm regularization. Then*

(a) *Assumption 2 is satisfied with*

$$\Psi(\overline{\mathcal{W}}) = \sqrt{2r}.$$

(b) *Assumption 3 is satisfied with*

$$\mu_* = 2D\sqrt{M}.$$

(c) *Assumption 4 is satisfied with*

$$\left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t g(\mathcal{X}_t)^\top \right\|_{op} \leq \log^4(MT) \sqrt{\frac{M}{T}}$$

*with probability at least  $1 - \frac{1}{(MT)^c}$ .*

## Proof Overview

- The main challenge in Lemma 4.7 comes in the proof of the deviation bound condition, which depends on the concentration properties of vector-valued martingales. The concentration properties of 2-smooth norms was studied in a number of works, including [21, 22]. We leverage recent work in [34] which extends the concentration results for 2-smooth norms to operator norms.
- Assumption 3 follows from assuming  $\|A^*\|_{2,1}$  and  $\|\widehat{A}\|_{2,1}$  are on the order of  $\sqrt{M}$ . Without this assumption, we could potentially have  $\mu_* = O(M)$ . This would give us a final bound in Corollary 4.8 which is only applicable when  $T \geq M^2$ , so this technical assumption is crucial in constructing a meaningful bound.
- The subspace compatibility constant in Assumption 2 was shown in [53]. In the sparsity case, this condition is trivial because  $\mathcal{S} = \overline{\mathcal{S}}$  and thus  $\Delta_{\overline{\mathcal{S}}}$  is known to lie in a subspace where every element is  $s$ -sparse. The condition is more subtle in the nuclear norm regularization case because  $\mathcal{W} = \overline{\mathcal{W}}$  if and only if  $A^*$  is symmetric. We do not assume symmetry of  $A^*$  so  $\Delta_{\overline{\mathcal{W}}}$  need not lie in the subspace  $\mathcal{W}$  where each element has rank at most  $r$ . However, [53] shows that  $\overline{\mathcal{W}}$  only contains matrices of rank at most  $2r$ .

Combining Lemma 4.7, Theorem 3.1 and the restricted eigenvalues results gives the following Theorem.

**Theorem 4.8. (Learning rates for nuclear norm regularization)**

Suppose  $(X_t)_{t=1}^T$  follows the SEPP framework of (2.3) and  $A^*$  is estimated using nuclear norm regularization over the ball  $\{A : \|A\|_{2,1}^2 \leq D\sqrt{M}\}$ .

(a) If  $(X_t)_{t=1}^T$  is generated according to the ARMA(1, 1) model in (4.1) then

$$\|\hat{A} - A^*\|_F^2 \leq \frac{C}{R_{\min}^2 \min(\frac{1}{2}R_{\min}, \kappa)^2} \frac{rM \log^8(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 M \frac{\log M}{\min(\frac{1}{2}R_{\min}, \kappa)^2}$$

for constants  $C, c > 0$  which are independent of  $M, T$  and  $r$ .

(b) If  $(X_t)_{t=1}^T$  is generated according to the AR(2) model in (4.3) then

$$\|\hat{A} - A^*\|_F^2 \leq \frac{C}{R_{\min}^2 r_\rho^2} \frac{rM \log^8(MT)}{T}$$

with probability at least  $1 - \frac{1}{(MT)^c}$  for  $T, M$  satisfying

$$T \geq 128U^4 M \frac{\log M}{r_\rho^2}.$$

Once again the mean-squared error bound  $\frac{rM \log^8(MT)}{T}$  matches the minimax optimal rate for independent design [45] up to log factors.

## 5 Numerical experiments

We validate our methodology and theory using a simulation study and real data examples. The focus of the simulation study is to confirm that the rates in mean-squared error in terms of  $s, r, T$  and  $\kappa$  scale as the theory predicts. We generate data according to the ARMA(1, 1) model from 4.1.

Our focus with real data experiments is to demonstrate that the models we analyze are sufficiently complex to capture real-world phenomena and enhance prediction performance relative to naive models. Others have successfully used more

complex, difficult to analyze models (*cf.*, [39, 17]) which are similar in spirit to those analyzed here. Our claim is not that our approach leads to uniformly better empirical performance than previous methods, but rather that our models capture essential elements of all these approaches and hence our theoretical work provides insights into a variety of approaches.

Our first real data example shows that our model and estimation procedure determines interactions among shooting events across different communities of Chicago that obeys sensible spatial structure (even though the algorithm does not use any spatial information). Our second real data example looks at neuron firing patterns in the brain of a rat and shows that our model can differentiate between the firing patterns during a sleep period and the patterns during a wake period. Finally, we examine a data set consisting of articles posted by different news websites and we try to determine influences between the sites using a variety of different regularization techniques. We implement the estimation procedure in (2.8) using the SpaRSA algorithm from [61].

## 5.1 Simulation study

We generate data according to (4.1) with  $\nu = 0$ ,  $M = 50$ ,  $\tilde{U} = 6$ ,  $\alpha = .25$  and varying values for  $T$  and  $s$ . Recall that  $\nu$  controls the background rate,  $M$  is the number of nodes,  $\tilde{U}$  is the clipping threshold,  $\alpha$  is the memory of the process in (4.1),  $T$  is the number of time steps observed, and  $s$  is the number of edges in the network.

Each time we generate a matrix, we randomly select  $s$  entries to be nonzero, and assign each value uniformly in  $[-.7, .3]$ . The sparsity ranges between 10 and 50. With these parameters, our process is usually stable on its own, and only occasionally relies on the clipping function to dampen the observations. Even at  $s = 50$ , only around 1% of the observations exceed 6, and the clipping percentage is even lower for smaller  $s$ . For each choice of  $s, T$ , we run 100 trials with  $\lambda = .1/\sqrt{T}$ . In the  $i$  trial we form a ground truth matrix  $A_i^*$ , compute  $\hat{A}_i$ , and measure the mean squared error (MSE) as  $\|A_i^* - \hat{A}_i\|_F^2$ .

In Figure 2(a), we plot MSE vs  $T$  for several different values of  $s$ , and in Figure 2(b) we plot MSE vs  $s$  for several values of  $T$ . The plots agree with our theory, which suggests that the error scales linearly in  $s$  and  $\frac{1}{T}$ .

Next, instead of assuming that  $A^*$  only has  $s$  non-zero entries, we assume that  $A^*$  has rank  $r$ , and measure MSE as a function of  $r$ . We hold the remaining parameters the same. To generate a rank  $r$  matrix  $A^*$ , we randomly generate a  $M \times r$  matrix and multiply it by a randomly generated  $r \times M$  matrix where the entries of both matrices are uniformly drawn from  $[-.7, .3]$ . We then normalize the resulting matrix so that  $a_{\max}$  is approximately .3. For all choices of rank considered, less than 5% of

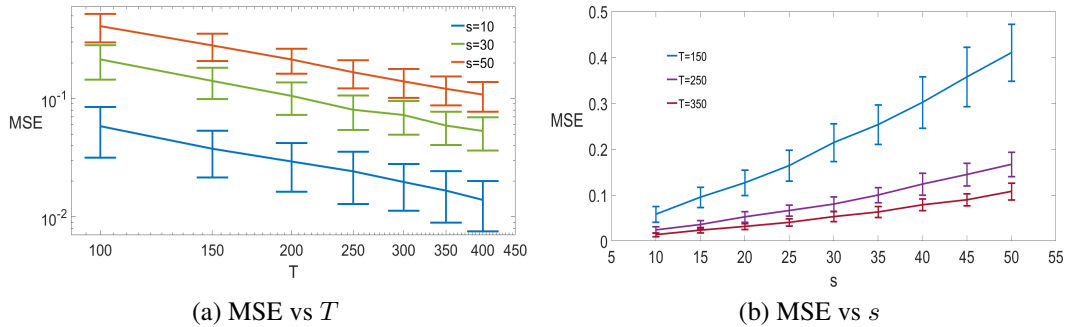


Figure 2: (a) shows MSE vs  $T$  for varying values of  $s$ , while (b) shows MSE vs  $s$  for varying values of  $T$ . Plots agree with theory which suggests that MSE scales linearly in  $s$  and  $\frac{1}{T}$ . Median of 100 trials are shown, and error bars denote the standard deviation.

observations are clipped. We set  $\lambda = .1\sqrt{\frac{M}{T}}$  as guided by our theory, run 100 trials for each  $(r, T)$  pair, and plot the median in Figure 3. The simulations agree with Lemma 4.7 which suggests that the MSE should scale linearly in  $r$  and  $\frac{1}{T}$ .

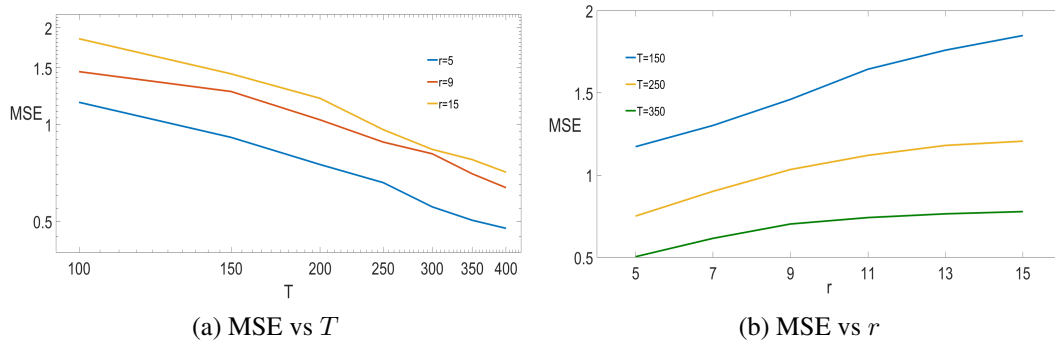


Figure 3: (a) shows MSE vs  $T$  for varying values of  $r$ , while (b) shows MSE vs  $r$  for varying values of  $T$ . Plots agree with theory which suggests that MSE scales linearly in  $r$  and  $\frac{1}{T}$ . Median of 100 trials are shown.

We now examine the relationship between  $a_{\max}$ , our theoretical MSE, and simulated MSE. For the remainder of the section fix  $\tilde{U} = 6$  and  $\alpha = 0$  and  $T = 400$ . Recall from Figure 1 that there is a stark phase transition where  $\kappa$  goes from reasonably large to minuscule. For  $\tilde{U} = 6$ , this transition occurs between  $a_{\max} = .3$  and  $a_{\max} = .4$ . Our MSE bound scales with  $\kappa^{-2}$ , so this phase transition controls where our bound is reasonably small.



To get a sense of how tight the bound is, we consider two different methods to generate a  $50 \times 50$  matrix  $A^*$ . The first is a block design, where  $A^*$  is zero outside of five  $10 \times 10$  blocks on the diagonal. Within the blocks, each row has five nonzero entries picked at random with values equal to  $\frac{a_{\max}}{5}$ . Matrices with this structure have strong feedback loops, where large observations from one node stimulate other nodes which are likely to feedback to the original node. In other words, with this block design method it is likely that many observations will actually be drawn close to the maximum possible rate  $R_{\max}$ , so we expect the MSE to align closely with our theoretical bound. We estimate  $A^*$  using  $l_1$  regularization, for varying values of  $a_{\max}$ .

As a second method, we consider a low-rank design and estimate  $A^*$  using nuclear norm regularization. We choose the first two rows of  $A^*$  to be orthogonal, both with row sums equal to  $a_{\max}$ . We then let each remaining row be a random convex combination of the first two rows. In this case, feedback loops are less of an issue; if one node has a large observation, the nodes it stimulates are less likely to have strong connections feeding back to the original node. Since this design method is not particularly stimulatory, we expect that most observations will not be drawn close to the maximum rate  $R_{\max}$ . Our theoretical bound is potentially loose in this case for the following reason. Our bound on  $\kappa$  which captures the amount of clipping is worst case based on the size of the coefficients of  $A^*$ , but does not take into account the structure of  $A^*$ . If the coefficients of  $A^*$  are large but are structured such that there aren't a lot of feedback loops then our bound on  $\kappa$  will be loose.

We randomly generate 50 different  $A^*$  over various  $a_{\max}$  for both design choices and then evaluate their efficiency by plotting the fraction of trials for which the MSE is above one. The results are shown in Figure 4. The simulated results also exhibit strong phase transitions, with the fraction of accurate trials shifting from one to zero with small changes in  $a_{\max}$ . This suggests that our theoretical results capture a real phenomenon of our model. In the block case, the phase transition occurs almost exactly where predicted by the MSE, whereas in the low-rank case there is a small lag in the phase transition. In other words, while our theoretical results are fairly tight for very stimulatory network structures, there appears to be some flexibility for networks with weaker feedback loops.

Finally, we consider the block matrix design under a wide range of  $\tilde{U}$ . We consider  $\tilde{U}$  between 3 and 30, and  $a_{\max}$  between 0 and .6 in increments of .02. For each  $(a_{\max}, \tilde{U})$  pair, we generate 20 matrices according to the block matrix design strategy outlined above and estimate  $\hat{A}$  via sparsity regularization. In Figure 5 we plot a heat map displaying the fraction of trials for which the MSE is below one. The red line shows the  $\kappa = .01$  contour, so Lemma 4.3 suggests our model will be hard to learn above the boundary line. Figure 5 generally resembles the heat map displaying values of  $\kappa$  in Figure 1, suggesting that the role of  $\kappa$  in our theory reflects a true

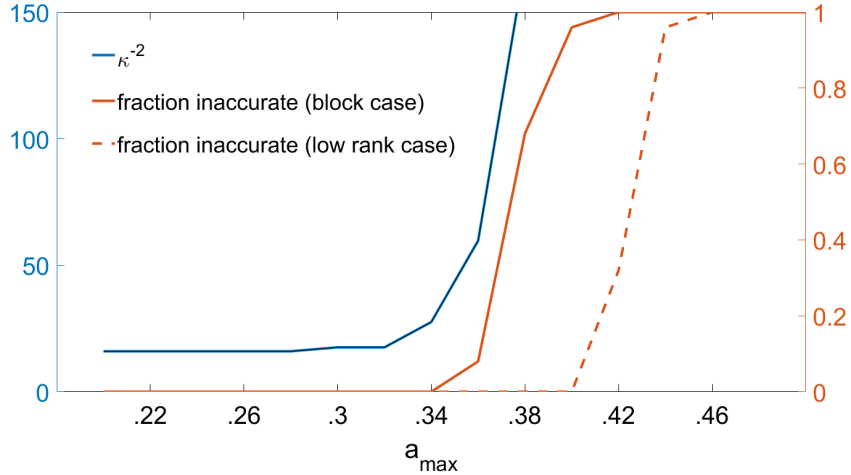


Figure 4: In blue,  $\kappa^{-2}$  as a function of  $a_{\max}$ . In orange, the fraction of 50 trials which have MSE above 1 for a block matrix design strategy and a low-rank matrix design strategy. Our theoretical bound on the MSE scales with  $\kappa^{-2}$ , suggesting stark phase transitions in the MSE which are confirmed in the simulated results. The block matrices correspond to more stimulatory networks than the low-rank ones, and therefore align more closely with our worst case bounds.

phenomenon that when  $a_{\max}$  is sufficiently large and clipping is frequent, then the model becomes difficult to learn.

## 5.2 Real Data Example – Chicago Crime

A number of studies have used various self-exciting point processes to predict crime, including [57, 43, 42]. We test our model on a data set [2] consisting of burglaries in Chicago since 2004, broken down by the  $M = 77$  community areas in the city. In [3], the authors fit self-exciting processes to the Chicago homicide data broken down by community area and performed clustering on the areas as we do below. We estimate the network based on the data from January 2004 to August 2010 and test it on the data from September 2010 to March 2017. To test results, we compare the log-likelihood of events using our learned matrix on the test set data, with that for a constant Poisson process. This gives approximately 600 time periods for both our training and test sets. We set  $\lambda = .1/\sqrt{T}$  using our theory as a guide. We show results for a half-day time discretization period, with  $\tilde{U} = 7$  and  $\alpha = .2$ .

The test set log-likelihood of our learned matrix shows an improvement over the test set log-likelihood for a learned constant process, where  $\lambda_{t,m} = \bar{\lambda}_m$  for all  $t$  ( $-6.62 \times 10^5$  compared to  $-1.09 \times 10^6$ ).

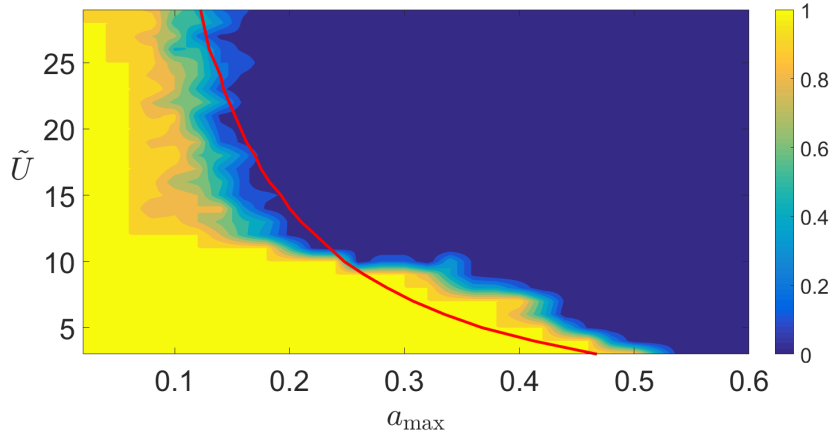


Figure 5: Heat map displaying the fraction of trials for which the MSE is below one for different  $(a_{\max}, \tilde{U})$  pairs. Red line shows  $\kappa = .01$  contour. Above this line our theory predicts difficulty in learning, agreeing with the heat map which shows inaccurate recovery of  $A^*$  for  $(a_{\max}, \tilde{U})$  pairs above the line. For each  $(a_{\max}, \tilde{U})$  pair, we run 20 trials with a block matrix design.

To examine the structure of our learned matrix, we treat the positive coefficients of the matrix as edges of the adjacency matrix of a graph. We then perform spectral clustering with four clusters. The results are shown in Figure 6, with colors indicating cluster membership. We note that our data contains no information about the geospatial location of the areas aside from index (not location) of the community area. However, there are clear geographic patterns in the clusters, providing some validation to the estimated influences between communities.

Finally, we test whether modeling these crime patterns as a multi-dimensional point processes leads to stronger results than modeling the patterns as a collection of independent univariate point processes, where

$$\log \lambda_{t+1,m} = \nu_m + a_m^* \min(X_{t,m}, \tilde{U}) + \alpha \log(\lambda_{t,m}).$$

Specifically, we compare finding  $\hat{A}$  as in (2.8) with choosing  $\hat{A}$  to be the solution to the optimization problem in (2.8) over the set of all diagonal matrices. We then perform the log-likelihood analysis described above for  $\alpha$  varying from 0 to .6 in increments of .2 and for the time-discretization period varying from half a day to three days in increments of half a day. The results are shown in Table 1. Note that the multivariate model outperforms the univariate one whenever the discretization period is at least a full day but does worse for the half day discretization period.

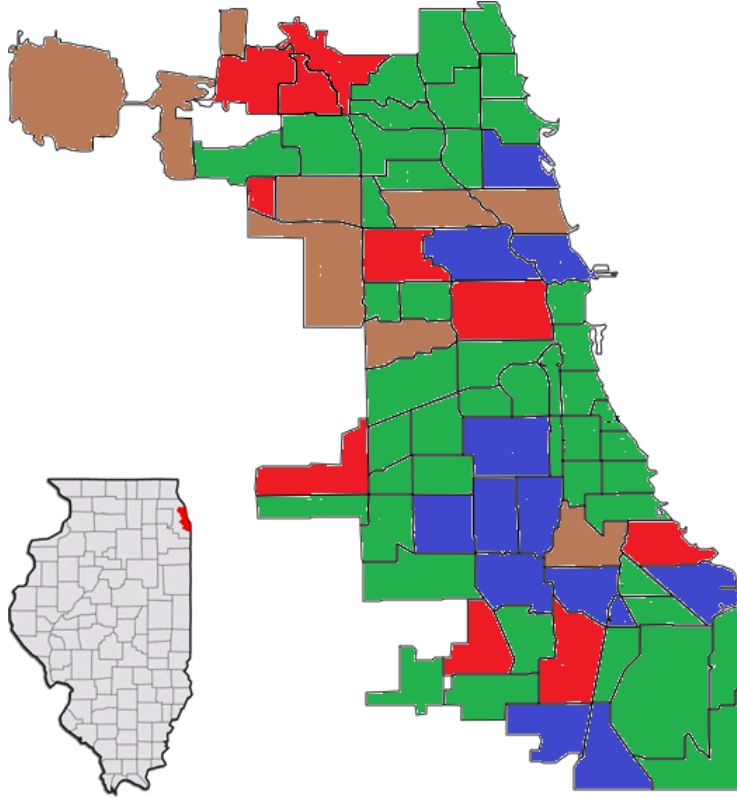


Figure 6: Clusters learned from crime data with a half-day time discretization. The clusters are overlaid on a map of community areas in Chicago. The data contained no geospatial information, but clusters show geographical patterns.

### 5.3 Real Data Example – Spike Train Data

SEPPs have been widely used in neuroscience to describe neuron spike train data [38, 60, 41, 40]. In this section, we analyze a multi-neuron spike train dataset from [58, 59]. The dataset consists of spike trains recorded from 51 neurons in the brain of a rat. The recordings were divided into a wake period and a sleep period. Using the data from the first half of the wake period, we learn a matrix  $A_{\text{wake}}$  using equation (4.1), a 100ms discretization period,  $\alpha = .7$  and  $\tilde{U} = 5$ . We then follow the same process to learn  $A_{\text{sleep}}$ . We get a sense of the structure of the matrices  $A_{\text{wake}}$  and  $A_{\text{sleep}}$  in Figure 7. We note that connections are much stronger during the wake period, during which there is more frequent neural firing.

In previous work [35] the authors use a similar SEPP to analyze neural spikes and discuss the significance of the time discretization period in more depth. In particular, they conclude that while models at this discretization length may have strong

Table 1: Difference between Log-Likelihood of multivariate process and univariate process

Discretization	$\alpha = 0$	$\alpha = .2$	$\alpha = .4$
.5 days	$-5.4 \times 10^4$	$-6.5 \times 10^4$	$-6.6 \times 10^4$
1 day	$1.2 \times 10^5$	$7.6 \times 10^4$	$6.2 \times 10^4$
1.5 days	$1.8 \times 10^5$	$1.1 \times 10^5$	$1.4 \times 10^5$
2 days	$1.8 \times 10^5$	$1.2 \times 10^5$	$2.3 \times 10^5$
2.5 days	$2.1 \times 10^5$	$1.1 \times 10^5$	$2.7 \times 10^5$
3 days	$1.9 \times 10^5$	$1.1 \times 10^5$	$2.8 \times 10^5$

predictive power, the discretization period is sufficiently large that the connections learned are not direct physical effects. In other words, if the connection between neuron A and neuron B is negative, this suggests that neuron B is less likely to fire in a 100ms interval after neuron A fires. However, there could be a complex chain of interactions causing this effect, and it does not mean there is a direct physical connection between neuron A and neuron B.

To validate  $A_{\text{wake}}$ , we compute the log-likelihood of events for the second half of the wake period using both  $A_{\text{wake}}$  and  $A_{\text{sleep}}$  as the ground truth matrix. We find  $\log p(X_{\text{wake}}|\hat{A}_{\text{wake}}) = -6.6 \times 10^4$  while  $\log p(X_{\text{sleep}}|\hat{A}_{\text{wake}}) = -7.4 \times 10^4$ .

Following the same process for  $A_{\text{sleep}}$ , we find that  $\log p(X_{\text{sleep}}|\hat{A}_{\text{sleep}}) = -2.34 \times 10^5$  while  $\log p(X_{\text{wake}}|\hat{A}_{\text{sleep}}) = -3.04 \times 10^5$ . This suggests that our model is capable of differentiating firing patterns in different sleep states. The log-likelihood of events for a constant process was orders of magnitude smaller in both cases.

## 5.4 Real Data Example – Memetracker Data

As a final example, we consider a data set [1] which consists of metadata for a collection news articles and blog posts. We only consider the time and website from which each post occurs but omit all other data such as the content of the post and other websites to which the post links. Further, we consider only articles posted by 198 popular news sites from <http://www.memetracker.org/lag.html>. Low-rank models have been applied in social network settings in a number of different works [65, 66, 63]; in particular, the work [65] proposes low-rank regularization of a point process model on this same data set.

To test the model, we collect all articles posted by 198 popular new websites during October 2008. Using a one hour discretization period, we divide the month into a training set and a test set, giving  $T = 500$  training periods and 500 test periods.

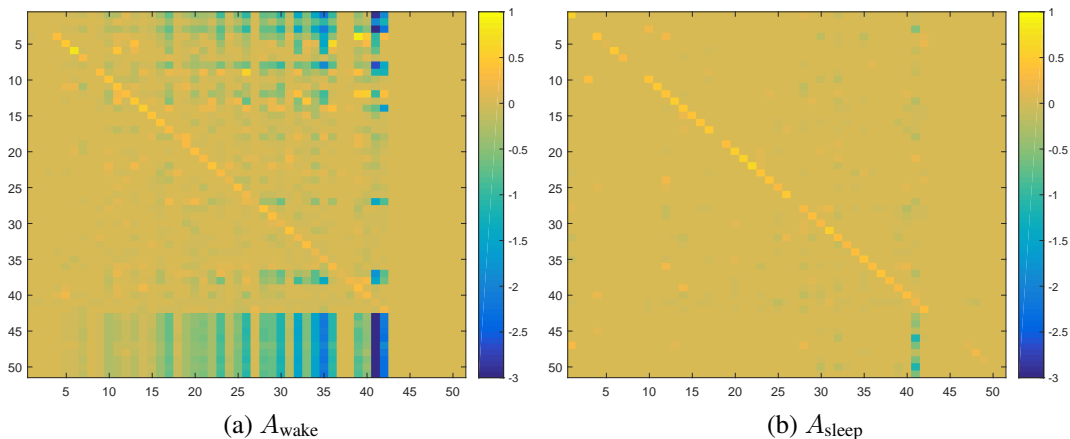


Figure 7: (a) shows  $A_{\text{wake}}$  matrix charting estimated relationships between neurons in a rat’s brain during a wake period, while (b) shows  $A_{\text{sleep}}$  matrix charting estimated relationships between neurons in a rat’s brain during a sleep period.

We train our model using the following regularization techniques. We perform the  $l_1$  regularization and nuclear norm regularization schemes described in Section 4.2, as well as a low-rank plus sparse model where we use the regularizer

$$\|\cdot\|_{\mathcal{R}} = \|\cdot\|_{1,1} + \|\cdot\|_{*}.$$

This last model is optimized using alternating descent. Finally, we learn a multi-dimensional model with no regularization, where we simply use the negative log-likelihood as our loss function, and a one-dimensional model where all interactions between different nodes are set to zero. The results are shown in Figure 8. The low-rank model performs best, followed by the low-rank plus sparse model, suggesting that the interactions between websites exhibit some low-rank behavior.

## 6 Connections to Hawkes Process

In this section, we observe that the model in (2.3) can be seen as a discretized version of the multivariate Hawkes process, in which there is much long-standing and recent interest (*e.g.*, [30, 31, 13, 11, 65, 48, 25]). By formulating our discrete time model in this manner we aim to highlight the connections between the two classes of models. There are advantages to analyzing point process models from both the continuous and discrete perspective. Some advantages of the discrete perspective include:

1. Real world data comes inherently discretized. In some cases, *e.g.*, social media posts, one might record data accurately up to very fine time windows. However,

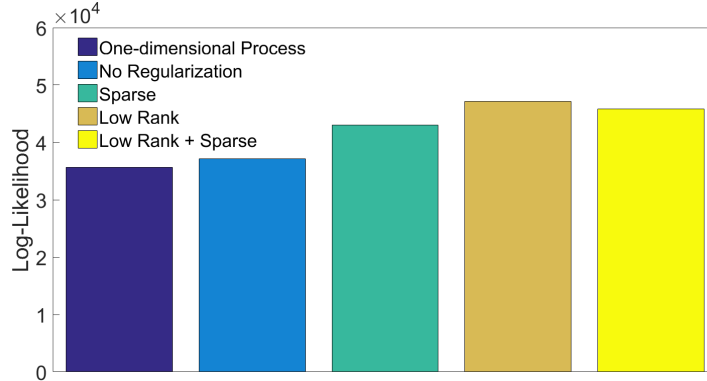


Figure 8: Log-likelihood of events on test set for matrices learned using memetracker data set under a variety of structure assumptions. Data set consists of timestamps for articles posted by 198 popular news websites and blogs during October 2008.

in other problems, the data collection process forces a coarse discretization. For example in [37], which used Hawkes processes to model civilian deaths in Iraq, reliable data was only obtained for the day on which attacks occurred.

2. In many works the authors discuss continuous Hawkes models, but their algorithms work with discretized data for computational efficiency. Examples include [4, 37, 41]. This provides additional motivation for our decision to study the ARMA(1, 1) model because its continuous version is one of the most widely studied types of Hawkes process.

## 6.1 Continuous Hawkes

In a multivariate Hawkes process, point process observations are drawn using an intensity function  $\lambda(\mathcal{X}_\tau)$ , where  $\mathcal{X}_\tau$  is the collection of all events up to (continuous) time  $\tau$ . Each event  $i$  is associated with two components:  $(\tau_i, m_i)$ , where  $\tau_i$  is the time of the event and  $m_i$  is the node or channel associated with the event.  $N_\tau$  denotes the number of events before time  $\tau$ .

We model the log-linear Hawkes process intensity at node  $m$  as<sup>1</sup>

$$\log \lambda_m^{(c)}(\mathcal{X}_\tau) = \nu_m + \sum_{i=1}^{N_\tau} h_{m, m_i}(\tau - \tau_i), \quad (6.1)$$

where the  $(c)$  superscript denotes continuous time.

<sup>1</sup> $\lambda_m^{(c)}(\mathcal{X}_\tau)$  would be more precisely written as  $\lambda_m^{(c)}(\tau; \mathcal{X}_\tau)$ ; we let the dependence on  $\tau$  be understood for simplicity of presentation.

Here each function  $h_{m,m'}(\tau)$  measures the influence of node  $m'$  on node  $m$  after  $\tau$  seconds since the event on  $m'$ . This model is standard in the point process literature. We write each of these functions as a linear combination of the basis functions  $\phi_1(\tau), \dots, \phi_K(\tau)$ :

$$h_{m,m'}(\tau) = \sum_{k=1}^K a_{m,m',k}^* \phi_k(\tau), \quad (6.2)$$

yielding

$$\begin{aligned} \log \lambda_m^{(c)}(\tau) &= \nu_m + \sum_{i=1}^{N_\tau} \sum_{k=1}^K a_{m,m_i,k}^* \phi_k(\tau - \tau_i) \\ &= \nu_m + \sum_{m'=1}^M a_{m,m'}^* \left[ \sum_{\substack{i < N_\tau: \\ m_i = m'}} \sum_{k=1}^K \phi_k(\tau - \tau_i) \right] \\ &= \nu_m + \sum_{m'=1}^M \sum_{k=1}^K a_{m,m',k}^* g_{m,k}^{(c)}(\mathcal{X}_\tau), \end{aligned} \quad (6.3)$$

where

$$g_{m,k}^{(c)}(\mathcal{X}_\tau) := \sum_{\substack{i < N_\tau: \\ m_i = m}} \phi_k(\tau - \tau_i).$$

Vectorizing across nodes and letting

$$\begin{aligned} \lambda^{(c)}(\mathcal{X}_\tau) &:= [\lambda_1^{(c)}(\mathcal{X}_\tau), \dots, \lambda_M^{(c)}(\mathcal{X}_\tau)]^\top \in \mathbb{R}_+^M \\ \nu &:= [\nu_1, \dots, \nu_M]^\top \in \mathbb{R}^M \\ g^{(c)}(\mathcal{X}_\tau) &:= (g_{m,k}^{(c)}(\mathcal{X}_\tau))_{m \in \{1, \dots, M\}, k \in \{1, \dots, K\}} \in \mathbb{R}^{MK} \\ A^* &:= (a_{m,m',k})_{m, m' \in \{1, \dots, M\}, k \in \{1, \dots, K\}} \in \mathbb{R}^{M \times MK}, \end{aligned}$$

we have

$$\log \lambda^{(c)}(\mathcal{X}_\tau) = \nu + A^* g^{(c)}(\mathcal{X}_\tau)$$

which exhibits the same general form as (2.3).

In order to formalize the connection between the multivariate Hawkes process in (6.3) and the SEPP in (2.3) we first describe our sampling process and the Hawkes and Poisson log likelihoods needed to prove Proposition 1: The Hawkes process can be discretized by sampling  $\lambda^{(c)}(\mathcal{X}_\tau)$  at  $\tau = t\Delta$  for some sampling period  $\Delta > 0$  and letting

$$X_{t,m} = \sum_{i=N_{(t-1)\Delta}+1}^{N_{t\Delta}} \mathbb{I}_{m=m_i} \quad (6.4)$$



for  $t = 1, \dots, T$ .

Here  $\mathbb{I}_E$  is the indicator function which returns 1 if  $E$  is true and 0 if  $E$  is false and  $X_{t,m}$  is the number of events on node  $m$  during the sampling interval  $[(t-1)\Delta, t\Delta)$ . Overloading notation somewhat, let  $\mathcal{X}_t = \{X_{s,m}\}_{s=1, \dots, t, m=1, \dots, M}$  be the history of event counts up to time  $t$ . The log-likelihood of the original Hawkes process observations up to time  $T\Delta$  is

$$\begin{aligned} \ell_H(\mathcal{X}_{T\Delta} | \{\lambda_m^{(c)}\}_m) &= \sum_{i=1}^{N_{T\Delta}} \log \lambda_{m_i}^{(c)}(\tau_i) - \sum_{m=1}^M \int_0^{T\Delta} \lambda_m^{(c)}(\tau) d\tau \\ &= \sum_{m=1}^M \sum_{i=1}^{N_{T\Delta, m}} \log \lambda_{m_i}^{(c)}(\tau_i) - \int_0^{T\Delta} \lambda_m^{(c)}(\tau) d\tau. \end{aligned}$$

Further note that if  $X_{t,m} \sim \text{Poisson}(\lambda_{t,m})$ , then the Poisson log likelihood is proportional to

$$\ell_p(\mathcal{X}_T | \{\lambda_m(\mathcal{X}_t)\}_{t,m}) = \sum_{m=1}^M \sum_{t=1}^T [X_{t,m} \log(\lambda_m(\mathcal{X}_t)) - \lambda_m(\mathcal{X}_t)].$$

We consider in Proposition 1 a SEPP with the intensity

$$\lambda_{t,m} = \Delta \lambda_m^{(c)}(\mathcal{X}_{\Delta t}) \equiv \Delta \lambda_m^{(c)}(\Delta t; \mathcal{X}_{\Delta t}), \quad (6.5)$$

where the last equality makes the sampling time explicit. We now present a proposition which formalizes the connections between the SEPP and the log-linear Hawkes process.

**Proposition 1.** The likelihood of the discretized multivariate Hawkes data in (6.4) can be approximated by the likelihood of the Poisson autoregressive model (2.3) with intensity (6.5), modulo terms independent of the unknown  $\lambda^{(c)}$ , where the approximation error depends on the sampling period  $\Delta$ .

This proposition suggests that the models and analysis we develop for SEPPs also provides insight into related Hawkes process models provided that the sampling period  $\Delta$  is sufficiently small.

## 7 Conclusion

The proposed saturated SEPP allows us to analyze statistical learning rates for a large class of point processes, including discretized Hawkes processes, with long-range memory and saturation or clipping effects common in real-world systems.

The analysis presented in this paper addresses instability issues present in prior works and incorporates a wide variety of structural assumptions on the ground truth processes by allowing for arbitrary decomposable regularizers. The proposed bounds provide novel insight not only into sample complexity bounds, but also into phase transition boundaries dictated by stability and saturation effects that are supported by simulation results. In addition, experiments on data from neuroscience, criminology, and social media suggest that the models considered in this paper exhibit sufficient complexity to model real-world phenomena.

## References

- [1] Memetracker. 2008. <http://www.memetracker.org/data.html>.
- [2] City of Chicago. 2017. <https://data.cityofchicago.org/Public-Safety/Crimes-2001-topresent/ijzp-q8t2>.
- [3] Ryan Adams and Scott. Linderman. Discovering latent network structure in point process data. In *Proc. International Conference on Machine Learning (ICML)*, 2014.
- [4] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. A. Laeven. Modeling financial contagion using mutually exciting jump processes. Technical report, National Bureau of Economic Research, 2010.
- [5] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.
- [6] E. Bacry and J.-F. Muzy. First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62:2184–2202, 2016.
- [7] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567, 2015.
- [8] C. Blundell, K. A. Heller, and J. M. Beck. Modelling reciprocating relationships with Hawkes processes. In *Proc. NIPS*, 2012.
- [9] 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*, 7(5):456–461, 2004.

- [10] V. Chavez-Demoulin and J. A. McGill. High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance*, 36(12):3415–3426, 2012.
- [11] S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten. The multivariate Hawkes process in high dimensions: Beyond mutual excitation. 2017. <https://arxiv.org/abs/1707.04928>.
- [12] T. P. Coleman and S. Sarma. Using convex optimization for nonparametric statistical analysis of point processes. In *Proc. ISIT*, 2007.
- [13] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes, Vol. I: Probability and its Applications*. Springer-Verlag, New York, second edition, 2003.
- [14] M. Ding, CE Schroeder, and X. Wen. Analyzing coherent brain networks with Granger causality. In *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pages 5916–8, 2011.
- [15] J. Muzy E. Bacry, S. Gaïffas. A generalization error bound for sparse and low-rank multivariate hawkes processes. *arXiv:1501.00725*, 2015.
- [16] 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*, 3:72–94, 2010.
- [17] Ş. Ertekin, C. Rudin, T. H. McCormick, et al. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.
- [18] J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal. Learning network of multivariate Hawkes processes: A time series approach. *arXiv preprint arXiv:1603.04319*, 2016.
- [19] M Fazel. Matrix rank minimization with applications. 2002. <http://faculty.washington.edu/mfazel/thesis-final.pdf>.
- [20] K. Fokianos and D. Tjøstheim. Log-linear poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578, 2011.
- [21] D.H. Garling. Functional central limit theorems in banach spaces. *Annals of Probability*, 4(4):600–611, 1976.

- [22] D.H. Garling. Convexity, smoothness and martingale inequalities. *Israel Journal of Mathematics*, 29:189–198, 1978.
- [23] F. Gerhard, M. Deger, and W. Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process glms. *PLoS Computation Biology*, 13(2):1069–1097, 2017.
- [24] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.
- [25] H. Zha H. Xu, M. Farajtabar. Learning granger causality for hawkes processes. *Proceedings of Machine Learning Research*, 48:1717–1726, 2016.
- [26] E. Hall and R. Willett. Dynamical models and tracking regret in online convex programming. In *Proc. International Conference on Machine Learning (ICML)*, 2013. <http://arxiv.org/abs/1301.1254>.
- [27] E. Hall and R. Willett. Online learning of neural network structure from spike trains. In *Proceedings of the 7th International IEEE EMBS Neural Engineering Conference (NER'15)*, 2015.
- [28] E. C. Hall, G. Raskutti, and R. Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.
- [29] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. LASSO and probabilistic inequalities for multivariate point processes. *arXiv preprint arXiv:1208.0570*, 2012.
- [30] A. G. Hawkes. Point spectra of some self-exciting and mutually-exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:83–90, 1971.
- [31] A. G. Hawkes. Point spectra of some mutually-exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971.
- [32] M. Hinne, T. Heskes, and M. A. J. van Gerven. Bayesian inference of whole-brain networks. *arXiv:1202.1696 [q-bio.NC]*, 2012.
- [33] Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer, 2003.

- [34] A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. 2008. <https://arxiv.org/abs/0809.0813>.
- [35] R. Kelly, R. Kass, M. Smith, and T. Lee. Accounting for network effects in neuronal responses using l1 regularized point process models. *NIPS*, 23(2):1099–1107, 2010.
- [36] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. 2009. <https://arxiv.org/abs/1602.07389>.
- [37] E. Lewis, E. Mohler, P. J. Brantingham, and A. L. Bertozzi. Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264, 2012.
- [38] S. Linderman, R. Adams, and J. Pillow. Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*, 2016.
- [39] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. In *ICML*, pages 1413–1421, 2014. arXiv:1402.0914.
- [40] S. Mensi. A new mathematical framework to understand single neuron computations. 2014. PhD Thesis.
- [41] S. Mensi, R. Naud, and W. Gerstner. From stochastic nonlinear integrate-and-fire to generalized linear models. *NIPS*, 2011(1):1377–1385, 2011.
- [42] G. Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497, 2014.
- [43] G. Mohler, M. Short, P. Brantingham, F. Schoenber, and G. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2012.
- [44] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.
- [45] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion. 13:1665–1697, 2012.

- [46] Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Statistical Science*, volume 27, pages 538–557, 2010.
- [47] A. K. Oh, Z. T. Harmany, and R. M. Willett. To  $e$  or not to  $e$  in Poisson image reconstruction. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2014.
- [48] S. A. Pasha and V. Solo. Hawkes-Laguerre reduced rank model for point processes. In *ICASSP*, 2013.
- [49] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454:995–999, 2008.
- [50] M. Raginsky, R. Willett, C. Horn, J. Silva, and R. Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8):5544–5562, 2012.
- [51] A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. 2015. <https://arxiv.org/abs/1510.03925>.
- [52] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls. *IEEE Trans. on Information Theory*, 57(10):6976–6994, 2011.
- [53] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [54] P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *Annals of Statistics*, 38(5):2781–2822, 2010.
- [55] J. Silva and R. Willett. Hypergraph-based anomaly detection in very large networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):563–569, 2009.
- [56] A. C. Smith and E. N. Brown. Estimating a state-space model from point process observations. *Neural Computation*, 15:965–991, 2003.
- [57] A. Stomakhin, M. B. Short, and A. Bertozzi. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27(11), 2011.

- [58] B. Watson, D. Levenstein, J. Greene, J. Gelinas, and G. Buzaki. Multi-unit spiking activity recorded from rat frontal cortex (brain regions mpfc, ofc, acc, and m2) during wake-sleep episode wherein at least 7 minutes of wake are followed by 20 minutes of sleep. 2016. Dataset:CRCNS.org. <http://dx.doi.org/10.6080/K02N506Q>.
- [59] B. Watson, D. Levenstein, J. Greene, J. Gelinas, and G. Buzaki. Network homeostasis and state dynamics of neocortical sleep. *Neuron*, 90(4):839–852, 2016.
- [60] A. Weber and J. Pillow. Capturing the dynamical repertoire of single neurons with generalized linear models. 2016. <https://arxiv.org/abs/0903.4742>.
- [61] S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57, 2009.
- [62] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. 68:49–67, 02 2006.
- [63] J. Zhang, J. Chen, S. Zhi, Y. Chang, P. Yu, and J. Han. Link prediction across aligned networks with sparse and low rank matrix estimation. *ICDE*, 33, 2017.
- [64] Liu. Zhang and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.*, 31(3):1235–1256, 2009.
- [65] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.
- [66] J. Zhuang, T. Mei, S. Hoi, X. Hua, and Y. Zhang. Community discovery from social media by low-rank matrix recovery. *TIST*, 5(4), 2015.

## A Proofs

### A.1 Proof of Theorem 3.1

Theorem 3.1 is the combination of results from [46] and [28]. We give a proof for the sake of completeness but claim no originality of techniques. For the first part of

the proof, we follow Theorem 1 from [46]. By the definition of  $\hat{A}$  and properties of Bregman divergence for strongly convex functions, we have

$$\frac{R_{\min}}{2T} \sum_m \sum_t (\Delta_m^\top g(\mathcal{X}_t))^2 \leq \frac{1}{T} \left| \sum_m \sum_t \epsilon_{t,m} \Delta_m^\top g(\mathcal{X}_t) \right| + \lambda(\|A^*\|_{\mathcal{R}} - \|\hat{A}\|_{\mathcal{R}}).$$

where  $R_{\min}$  is a strong convexity parameter for  $e^x$  on the domain  $x \in [R_{\min}, R_{\max}]$ . Next note that

$$\begin{aligned} \sum_m \sum_t \epsilon_{t,m} \Delta_m^\top g(\mathcal{X}_t) &= \sum_m \sum_{m'} \Delta_{m,m'} \sum_t X_{t,m'} \epsilon_{t,m} \\ &= \langle \Delta, \sum_t \epsilon_t g(\mathcal{X}_t)^\top \rangle \\ &\leq \|\Delta\|_{\mathcal{R}} \left\| \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right\|_{\mathcal{R}^*}. \end{aligned}$$

Thus, assuming  $\lambda/2 > \frac{1}{T} \left\| \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right\|_{\mathcal{R}^*}$  we have

$$\frac{1}{T} \left| \sum_m \sum_t \epsilon_{t,m} \Delta_m^\top g(\mathcal{X}_t) \right| + \lambda(\|A^*\|_{\mathcal{R}} - \|\hat{A}\|_{\mathcal{R}}) \leq \frac{\lambda}{2} \|\Delta\|_{\mathcal{R}} + \lambda\|A^*\|_{\mathcal{R}} - \lambda\|\hat{A}\|_{\mathcal{R}}. \quad (\text{A.1})$$

Then

$$\|\hat{A}\|_{\mathcal{R}} = \|A^* + \Delta\|_{\mathcal{R}} = \|A^* + \Delta_{\overline{\mathcal{M}}^\perp} + \Delta_{\overline{\mathcal{M}}}\|_{\mathcal{R}}.$$

Since  $\mathcal{R}$  is decomposable with respect to the subspaces  $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ , we have

$$\|\hat{A}\|_{\mathcal{R}} \geq \|A^*\|_{\mathcal{R}} + \|\Delta_{\overline{\mathcal{M}}^\perp}\|_{\mathcal{R}} - \|\Delta_{\overline{\mathcal{M}}}\|_{\mathcal{R}}.$$

Thus

$$\frac{\lambda}{2} \|\Delta\|_{\mathcal{R}} + \lambda\|A^*\|_{\mathcal{R}} - \lambda\|\hat{A}\|_{\mathcal{R}} \leq \frac{3\lambda}{2} \|\Delta_{\overline{\mathcal{M}}}\|_{\mathcal{R}} - \frac{\lambda}{2} \|\Delta_{\overline{\mathcal{M}}^\perp}\|_{\mathcal{R}} \leq \frac{3\lambda}{2} \|\Delta_{\overline{\mathcal{M}}}\|_{\mathcal{R}}. \quad (\text{A.2})$$

Recalling that  $\Psi(\overline{\mathcal{M}})$  is the subspace compatibility constant, we have

$$\|\Delta_{\overline{\mathcal{M}}}\|_{\mathcal{R}} \leq \Psi(\overline{\mathcal{M}}) \|\Delta_{\overline{\mathcal{M}}}\|_F \leq \Psi(\overline{\mathcal{M}}) \|\Delta\|_F.$$

It follows that

$$\frac{\lambda}{2} \|\Delta\|_{\mathcal{R}} + \lambda\|A^*\|_{\mathcal{R}} - \lambda\|\hat{A}\|_{\mathcal{R}} \leq \frac{3\lambda}{2} \Psi(\overline{\mathcal{M}}) \|\Delta\|_F.$$

Let  $\|\Delta\|_T^2 = \frac{1}{T} \sum_m \sum_t (\Delta_m^\top g(\mathcal{X}_t))^2$  and therefore

$$\|\Delta\|_T^2 \leq \frac{3\lambda}{R_{\min}} \Psi(\overline{\mathcal{M}}) \|\Delta\|_F.$$



From here, we reduce the lower bound into the restricted eigenvalue condition. Denote the subsets

$$\mathcal{B}_{\mathcal{R}} = \{B \in \mathbb{R}^{M \times MK} : \|B_{\overline{\mathcal{M}}^\perp}\|_{\mathcal{R}} \leq 3\|B_{\overline{\mathcal{M}}}\|_{\mathcal{R}}\}$$

and

$$\mathcal{B}'_{\mathcal{R}} = \{B \in \mathcal{B}_{\mathcal{R}} : \|B\|_F = 1\}.$$

Note that Equation (A.2) implies that  $\Delta \in \mathcal{B}_{\mathcal{R}}$ . Let  $\mathcal{T} = \{p, 2p, \dots, T\}$  so  $|\mathcal{T}|/|T| = \frac{1}{p}$  (here we assume  $\frac{T}{p}$  is an integer for simplicity). By Assumption 1 we have

$$\begin{aligned} \|\Delta\|_T^2 &\geq \frac{1}{T} \sum_{t \in \mathcal{T}} \sum_m \Delta_m^\top \mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-p}] \Delta_m \\ &\quad - \sup_{B \in \mathcal{B}_{\mathcal{R}}} \frac{1}{T} \sum_{t \in \mathcal{T}} \sum_m (b_m^\top g(\mathcal{X}_t))^2 - \mathbb{E}[(b_m^\top g(\mathcal{X}_t))^2 | \mathcal{X}_{t-p}] \\ &\geq \frac{\omega}{p} \|\Delta\|_F^2 - \sup_{B \in \mathcal{B}_{\mathcal{R}}} \frac{1}{T} \sum_{t \in \mathcal{T}} \sum_m (b_m^\top g(\mathcal{X}_t))^2 - \mathbb{E}[(b_m^\top g(\mathcal{X}_t))^2 | \mathcal{X}_{t-p}]. \end{aligned}$$

We want to show

$$\sup_{B \in \mathcal{B}_{\mathcal{R}}} \frac{1}{T} \sum_{t \in \mathcal{T}} \sum_m (b_m^\top g(\mathcal{X}_t))^2 - \mathbb{E}[(b_m^\top g(\mathcal{X}_t))^2 | \mathcal{X}_{t-p}] \leq \frac{\omega \|\Delta\|_F^2}{2p} \quad (\text{A.3})$$

with high probability, and we note that it suffices to show this for all  $B \in \mathcal{B}'_{\mathcal{R}}$ .

Now we define the matrix  $G \in \mathbb{R}^{M \times M}$  as follows:

$$G := \frac{1}{T} \sum_{t \in \mathcal{T}} (g(\mathcal{X}_t)g(\mathcal{X}_t)^\top - \mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-p}]).$$

Following the definition of  $G$ , and denoting each entry  $G_{m,m'}$ ,

$$\begin{aligned} \sup_{B \in \mathcal{B}'_{\mathcal{R}}} \frac{1}{T} \sum_t \sum_m (b_m^\top g(\mathcal{X}_t))^2 - \mathbb{E}[(b_m^\top g(\mathcal{X}_t))^2 | \mathcal{X}_{t-p}] &= \sup_{B \in \mathcal{B}'_{\mathcal{R}}} \sum_{m=1}^M b_m^\top G b_m \\ &\leq \sup_{B \in \mathcal{B}'_{\mathcal{R}}} \|B\|_{2,1}^2 \max_{m,m'} |G_{m,m'}|. \end{aligned}$$

Recall that  $\sup_{B \in \mathcal{B}'_{\mathcal{R}}} \|B\|_{2,1}^2 \leq \mu_{\mathcal{R}}$  by Assumption 3. Hence

$$\sup_{B \in \mathcal{B}'_{\mathcal{R}}} \frac{1}{T} \sum_t \sum_m (b_m^\top g(\mathcal{X}_t))^2 - \mathbb{E}[(b_m^\top g(\mathcal{X}_t))^2 | \mathcal{X}_{t-p}] \leq \mu_{\mathcal{R}} \max_{m,m'} |G_{m,m'}|.$$

Note that each entry  $G_{m,m'}$  is a martingale and  $|G_{m,m'}| \leq 2U^2$ . Therefore we can apply the Azuma-Hoeffding inequality [5]. For completeness, we state the Azuma-Hoeffding inequality as Theorem A.5 in Section A.9. If we let

$$Y_n := \frac{1}{T} \sum_{t=0}^n (g(\mathcal{X}_t)g(\mathcal{X}_t)^T - \mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^T | \mathcal{X}_{t-p}]),$$

where  $n = 0, 1, 2, \dots, |\mathcal{T}|$ , and we set  $t = \frac{\omega}{2\mu_{\mathcal{R}}p}$  and  $c_n = \frac{2U^2}{T}$  as in Theorem A.5 in Appendix A.9, we have

$$\mathbb{P}\left(|G_{m,m'}| \geq \frac{\omega}{2\mu_{\mathcal{R}}p}\right) \leq 2 \exp\left(-\frac{T\omega^2}{32U^4p^2\mu_{\mathcal{R}}^2}\right).$$

Applying a union bound,

$$\mathbb{P}\left(\max_{m,m'} |G_{m,m'}| \geq \frac{\omega}{2\mu_{\mathcal{R}}p}\right) \leq 2M^2 \exp\left(-\frac{T\omega^2}{32U^4p^2\mu_{\mathcal{R}}^2}\right).$$

Hence if we set

$$T > \frac{128U^4p^2\mu_{\mathcal{R}}^2 \log M}{\omega^2},$$

(A.3) holds with probability at least

$$1 - \frac{2}{M^2},$$

guaranteeing that  $\|\Delta\|_T^2 \geq \frac{\omega}{2p}\|\Delta\|_F^2$  with this same probability. Putting everything together, we have

$$\|\Delta\|_F^2 \leq \frac{36p\Psi(\overline{\mathcal{M}})^2\lambda^2}{R_{\min}^2\omega^2}$$

with probability at least  $1 - \frac{2}{M^2}$  for

$$T \geq \frac{128p^2U^4\mu_{\mathcal{R}}^2 \log M}{\omega^2}.$$

□

## A.2 Proof of Lemma 4.1

Before we prove Lemma 4.1, we first need the following supporting lemma.

**Lemma A.1.** *Let  $Z = \min(\lfloor \lambda \rfloor, \tilde{U})$ . Define the random variables  $\overline{X} \sim \text{Poisson}(\lambda)$ ,  $X = \min(\overline{X}, \tilde{U})$  and*

$$Y = \begin{cases} 0 & \text{if } X \leq Z \\ 1 & \text{if } X > Z \end{cases}.$$

Then

$$\text{Var}(Y) \leq \text{Var}(X).$$

**Proof** We write

$$X = (Y + Z) + (X - Y - Z).$$

Since  $Z$  is a constant we have

$$\text{Var}(X) = \text{Var}(Y) + \text{Var}(X - Y) + 2\text{Cov}(Y, X - Y)$$

and it suffices to show  $\text{Cov}(Y, X - Y) \geq 0$ . Conditioning on  $Y$  gives

$$\begin{aligned} \mathbb{E}[(Y - \mathbb{E}[Y])(X - Y - \mathbb{E}[X - Y])] &= \\ & p(Y = 1)(1 - \mathbb{E}[Y])\mathbb{E}_{X,Y|Y=1}[X - Y - \mathbb{E}[X - Y]] \quad (\text{A.4}) \\ & + p(Y = 0)(-\mathbb{E}[Y])\mathbb{E}_{X,Y|Y=0}[X - Y - \mathbb{E}[X - Y]]. \end{aligned}$$

Now observe that  $Y = 1$  implies  $X - Y \geq Z$  (where we rely on the fact that  $X$  and  $Z$  are both integers) while  $Y = 0$  implies  $X - Y \leq Z$ ; then

$$\mathbb{E}[X - Y|Y = 1] \geq \mathbb{E}[X - Y] \quad (\text{A.5})$$

and

$$\mathbb{E}[X - Y|Y = 0] \leq \mathbb{E}[X - Y]. \quad (\text{A.6})$$

We argue that both terms in the sum in (A.4) are non-negative. For the first term, we have

$$\mathbb{E}_{X,Y|Y=1}[X - Y - \mathbb{E}[X - Y]] = \mathbb{E}_{X,Y|Y=1}[X - Y] - \mathbb{E}[X - Y] \geq 0$$

by (A.5).

$$\mathbb{E}_{X,Y|Y=0}[X - Y - \mathbb{E}[X - Y]] = \mathbb{E}_{X,Y|Y=0}[X - Y] - \mathbb{E}[X - Y] \leq 0$$

by (A.6). Finally note that  $\mathbb{E}[Y] \in (0, 1)$  so that both terms in (A.4) are indeed non-negative, and therefore  $\text{Cov}(Y, X - Y) \geq 0$  as claimed.  $\square$

We now prove Lemma 4.1. Note that

$$\text{Cov}(\min(X_{t,m}, \tilde{U}), \min(X_{t,m'}, \tilde{U})|\mathcal{X}_{t-1}) = 0$$

for  $m \neq m'$ . We have

$$\begin{aligned} \mathbb{E}[\min(X_t, \tilde{U}) \min(X_t, \tilde{U})^\top |\mathcal{X}_{t-1}] &= \\ & \mathbb{E}[\min(X_t, \tilde{U})|\mathcal{X}_{t-1}]\mathbb{E}[\min(X_t, \tilde{U})|\mathcal{X}_{t-1}]^\top + \text{Diag}(\text{Var}(\min(X_t, \tilde{U})|\mathcal{X}_{t-1})) \end{aligned}$$

where the first matrix is positive semi-definite because it is the outer product of a vector with itself. Thus, to come up with a lower bound for our original matrix, we just need to lower bound the smallest element of  $\text{Var}(\min(X_t, \tilde{U})|\mathcal{X}_{t-1})$ . This amounts to lower bounding the variance of  $\min(X_\lambda, \tilde{U})$  where  $X_\lambda$  is a Poisson random variable with mean  $\lambda \in [R_{\min}, R_{\max}]$ . Define

$$Y_\lambda = \begin{cases} 0 & \text{if } \min(X_\lambda, \tilde{U}) \leq \min(\lfloor \lambda \rfloor, \tilde{U}) \\ 1 & \text{if } \min(X_\lambda, \tilde{U}) > \min(\lfloor \lambda \rfloor, \tilde{U}). \end{cases}$$

By Lemma A.1,  $\text{Var}(\min(X_\lambda, \tilde{U})) \geq \text{Var}(Y_\lambda)$  so our problem reduces to lower bounding the variance of  $Y_\lambda$ . We argue that

$$\text{Var}(Y_\lambda) \geq \min(\text{Var}(Y_{R_{\min}}), \text{Var}(Y_{R_{\max}}))$$

by considering two cases. When analyzing these cases, we use the fact that  $\text{Var}(Y_\lambda)$  will be minimized when the probability of outcome (0) is either maximized or minimized. We take  $R_{\min} \leq \frac{1}{5}$  to make the exposition clearer. At the end of the proof we discuss the  $R_{\min} > \frac{1}{5}$  scenario which is virtually identical.

**Case 1:**  $\lambda \in [R_{\min}, \tilde{U})$  where  $\tilde{U}$  may be either greater than or less than  $R_{\max}$

In this scenario

$$Y_\lambda = \begin{cases} 0 & \text{if } X_\lambda \leq \lfloor \lambda \rfloor \\ 1 & \text{if } X_\lambda > \lfloor \lambda \rfloor. \end{cases}$$

We claim  $\text{Var}(Y_{R_{\min}}) \leq \text{Var}(Y_\lambda)$  for  $\lambda \in [R_{\min}, \tilde{U})$ . To do this, we look at two subcases.

First, if  $1 \leq \lambda \leq \tilde{U}$ , then basic properties of the median of the Poisson distribution imply that the probability of outcome (0) will be between  $\frac{1}{5}$  and  $\frac{4}{5}$  and so  $\text{Var}(Y_\lambda) \geq \frac{4}{25}$ . For the second case where  $R_{\min} \leq \lambda < 1$ , outcome (0) corresponds to

$$\mathbb{P}(X = 0 | X \sim \text{Poisson}(\lambda)) = \exp(-\lambda) \leq \exp(-R_{\min}).$$

Since  $R_{\min} \leq \frac{1}{5}$  we get that  $\exp(-R_{\min}) > \frac{4}{5}$ . Combining the two cases, we have concluded that  $\text{Var}(Y_\lambda)$  is minimized on  $\lambda \in [R_{\min}, \tilde{U}]$  at  $\lambda = R_{\min}$ . Now, when  $\lambda = R_{\min}$ ,

$$\text{Var}(Y_\lambda) = \exp(-R_{\min})(1 - \exp(-R_{\min})).$$

Since

$$f(x) = \frac{e^{-x}(1 - e^{-x})}{x}$$

decreases monotonically on the interval  $(0, \frac{1}{5}]$ , we have

$$\min_{x \in (0, \frac{1}{5}]} f(x) \geq f\left(\frac{1}{5}\right) \geq \frac{1}{2}.$$

Using the fact that  $R_{\min} \in (0, \frac{1}{5}]$  we conclude

$$\text{Var}(Y_{R_{\min}}) = \exp(-R_{\min})(1 - \exp(-R_{\min})) \geq \frac{1}{2}R_{\min}. \quad (\text{A.7})$$

**Case 2:**  $\lambda \in [\tilde{U}, R_{\max}]$

Next we consider the variance when  $\lambda \geq \tilde{U}$ . By the same reasoning as in Case 1, outcome (0) can have probability no larger than  $\frac{4}{5}$ . It remains to consider when outcome (1) can have probability larger than  $\frac{4}{5}$ . It is clear that for  $\lambda \in [\tilde{U}, R_{\max}]$ , outcome (1) will be maximized for  $\lambda = R_{\max}$ . When  $\lambda = R_{\max}$ , we directly compute the variance as

$$\text{Var}(Y_{R_{\max}}) = \sum_{i=0}^{\tilde{U}-1} \frac{R_{\max}^i \exp^{-R_{\max}}}{i!} \left(1 - \sum_{i=0}^{\tilde{U}-1} \frac{R_{\max}^i \exp^{-R_{\max}}}{i!}\right) = \kappa. \quad (\text{A.8})$$

Combining Case 1 and Case 2, we get that

$$\text{Var}(Y_\lambda) \geq \min(\text{Var}(R_{\min}), \text{Var}(R_{\max}))$$

and combining Equations (A.7) and (A.8) gives the final result.

If  $R_{\min} \geq \frac{1}{5}$  an identical argument shows that for Case 1 where  $\lambda \in [R_{\min}, \tilde{U})$ ,  $\text{Var}(Y_\lambda) \geq \frac{4}{25}$  and for Case 2 where  $\lambda \in [\tilde{U}, R_{\max}]$ ,  $\text{Var}(Y_\lambda) \geq \kappa$ . Hence, a lower bound on  $\text{Var}(Y_\lambda)$  covering all possible values of  $R_{\min}$  would be  $\min(\frac{1}{2}R_{\min}, \kappa, \frac{4}{25})$ . In main body of the paper we present the bound for the  $R_{\min} \leq \frac{1}{5}$  scenario in order to make the statement more interpretable. □

### A.3 Proof of Theorem 4.2

Recall that the AR(2) model is a special case of (2.3) with  $\phi_1[t] = \mathbb{I}_{\{t=1\}}$  and  $\phi_2[t] = \mathbb{I}_{\{t=2\}}$ . With these choices of basis functions,

$$g(\mathcal{X}_t) = [\min(X_1, \tilde{U}), \min(X_2, \tilde{U})]^\top.$$

A computation shows that if we choose to condition on  $\mathcal{X}_{t-1}$  as in the proof of Lemma 4.1 we get a singular matrix. However, Assumption 1 allows us to condition

on  $\mathcal{X}_{t-p}$  for any  $p > 0$  and so for this example it will be easiest to condition on  $\mathcal{X}_{t-2}$ . We have

$$\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-2}] = \mathbb{E}[g(\mathcal{X}_t) | \mathcal{X}_{t-2}]\mathbb{E}[g(\mathcal{X}_t) | \mathcal{X}_{t-2}]^\top + \text{Cov}(g(\mathcal{X}_t) | \mathcal{X}_{t-2}).$$

The first matrix is an outer product of a vector with itself so it is positive semi-definite and it suffices to lower bound the eigenvalues of the covariance matrix  $\text{Cov}(g(\mathcal{X}_t) | \mathcal{X}_{t-2})$ . Recall that a matrix  $B$  is said to be strictly diagonally dominant if

$$b_{i,i} - \sum_{j \neq i} |b_{i,j}| \geq \omega > 0$$

for all  $i$ , and the eigenvalues of a symmetric strictly diagonally dominant matrix are lower bounded by  $\omega$ . To lower bound the eigenvalues of the covariance matrix, we will show it is strictly diagonally dominant. We break the rows up into two cases. The first case corresponds to rows where the diagonal depends on a lagged count  $X_{t-1,m}$  while the second case corresponds to rows where diagonal depends on a count  $X_{t,m}$  without a lag.

**Case 1: Rows 1 through  $M$**  The first  $M$  rows of  $\text{Cov}(g(\mathcal{X}_t) | \mathcal{X}_{t-2})$  have their diagonal of the form  $\text{Var}(X_{t-1,m} | \mathcal{X}_{t-2}) \geq R_{\min}$ . We have  $\text{Cov}(X_{t-1,m}, X_{t-1,m'} | \mathcal{X}_{t-2}) = 0$  for all  $m' \neq m$ . If node  $m$  is not a parent of  $m'$ , then  $X_{t-1,m}$  and  $X_{t,m'}$  are independent conditioned on  $\mathcal{X}_{t-2}$ , so  $\text{Cov}(X_{t-1,m}, X_{t,m'} | \mathcal{X}_{t-2}) = 0$ . All that remains is to control  $\text{Cov}(X_{t-1,m}, X_{t,m'} | \mathcal{X}_{t-2})$  for the  $\rho_m^{(c)}$  children of  $m$ .

To do this, recall the decomposition  $X_{t,m'} = \exp(\nu_{m'} + a_{m'}^\top g(\mathcal{X}_{t-1})) + \epsilon_{t,m'}$ . For the remainder of the proof we let  $f_{m'}(\mathcal{X}_t) = \exp(\nu_{m'} + a_{m'}^\top g(\mathcal{X}_t))$  for notational simplicity and note that the Poisson noise term  $\epsilon_{t,m}$  is zero mean conditioned on  $\mathcal{X}_{t-1,m}$ . Hence

$$\text{Cov}(X_{t-1,m}, X_{t,m'} | \mathcal{X}_{t-2}) = \text{Cov}(X_{t-1,m}, f_{m'}(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2}).$$

Since  $f_{m'}(\mathcal{X}_{t-1})$  takes values in the interval  $[R_{\min}, R_{\max}]$ , the variance of  $f_{m'}(\mathcal{X}_{t-1})$  is bounded by a scaled Bernoulli random variable which takes values 0 with probability  $\frac{1}{2}$  and  $R_{\max} - R_{\min}$  with probability  $\frac{1}{2}$ . This variance is equal to  $\frac{(R_{\max} - R_{\min})^2}{4}$  and therefore

$$\begin{aligned} \text{Cov}(X_{t-1,m}, f_{m'}(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2}) &\leq \\ &\sqrt{\text{Var}(X_{t-1,m} | \mathcal{X}_{t-2}) \text{Var}(f_{m'}(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2})} \leq \frac{\sqrt{R_{\max}}(R_{\max} - R_{\min})}{2}. \end{aligned}$$

Hence the off diagonal entries sum to at most

$$\frac{\rho_m^{(c)} \sqrt{R_{\max}}(R_{\max} - R_{\min})}{2}$$

so for these rows of the covariance matrix we have

$$\text{Cov}(g(\mathcal{X}_t))_{i,i} - \sum_{j \neq i} \text{Cov}(g(\mathcal{X}_t))_{i,j} \geq R_{\min} - \frac{\rho_m^{(e)} \sqrt{R_{\max}} (R_{\max} - R_{\min})}{2}. \quad (\text{A.9})$$

**Case 2: Rows  $M + 1$  through  $2M$**  We next consider the final  $M$  rows of the covariance matrix whose diagonal is of the form  $\text{Var}(X_{t,m} | \mathcal{X}_{t-2}) \geq R_{\min}$ . We know  $\text{Cov}(X_{t,m}, X_{t-1,m'} | \mathcal{X}_{t-2})$  will be zero whenever node  $m'$  is not a parent of  $m$ , and for the  $\rho_m^{(p)}$  parents of  $m$ , the covariance is bounded below by

$$\frac{\sqrt{R_{\max}} (R_{\max} - R_{\min})}{2}$$

just as in the previous paragraph.

Finally, we need to consider  $\text{Cov}(X_{t,m}, X_{t,m'} | \mathcal{X}_{t-2})$ . When  $m$  and  $m'$  are not siblings this covariance will be zero. When they do share a parent, we again recall the decomposition  $X_{t,m} = f_m(\mathcal{X}_{t-1}) + \epsilon_{t,m}$  and  $X_{t,m'} = f_{m'}(\mathcal{X}_{t-1}) + \epsilon_{t,m'}$  and note that the  $\epsilon_{t,m}$  and  $\epsilon_{t,m'}$  are zero mean conditioned on  $X_{t,m'}$  and  $X_{t,m}$  respectively. Therefore

$$\text{Cov}(X_{t,m}, X_{t,m'} | \mathcal{X}_{t-2}) = \text{Cov}(f_m(\mathcal{X}_{t-1}), f_{m'}(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2})$$

and using the fact that each  $f_i(\mathcal{X}_t)$  takes values in the interval  $[R_{\min}, R_{\max}]$  it follows that this covariance is bounded by

$$\sqrt{\text{Var}(f_m(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2}) \text{Var}(f_{m'}(\mathcal{X}_{t-1}) | \mathcal{X}_{t-2})} \leq \frac{(R_{\max} - R_{\min})^2}{4}.$$

Recall that  $\rho_m^{(s)}$  denotes the number of siblings of  $m$ . Overall we have concluded that the sum of the off diagonal entries for the first  $m$  rows is at most

$$\frac{\rho_m^{(p)} \sqrt{R_{\max}} (R_{\max} - R_{\min})}{2} + \frac{\rho_m^{(s)} (R_{\max} - R_{\min})^2}{4}$$

so that for these rows we have

$$\begin{aligned} \text{Cov}(g(\mathcal{X}_t))_{i,i} - \sum_{j \neq i} \text{Cov}(g(\mathcal{X}_t))_{i,j} &> \\ R_{\min} - \frac{\rho_m^{(p)} \sqrt{R_{\max}} (R_{\max} - R_{\min})}{2} - \frac{\rho_m^{(s)} (R_{\max} - R_{\min})^2}{4}. \end{aligned} \quad (\text{A.10})$$

We conclude that the smallest eigenvalue of the covariance matrix is lower bounded by the minimum of the two lower bounds in Equations (A.9) and (A.10), and we define this minimum to be  $r_\rho$ .  $\square$

## A.4 Proof of Lemma 4.3

We know from Equation (A.2) that  $\|\Delta_{\mathcal{S}^\perp}\|_1 \leq 3\|\Delta_{\mathcal{S}}\|_1$  and since

$$\|\Delta\|_1 = \|\Delta_{\mathcal{S}^\perp}\|_1 + \|\Delta_{\mathcal{S}}\|_1$$

it follows that  $\|\Delta\|_1 \leq 4\|\Delta_{\mathcal{S}}\|_1$ . Recall that  $\|v\|_1 \leq \sqrt{s}\|v\|_2$  for any  $s$ -sparse vector  $v$ . Thus we have

$$\|\Delta\|_1 \leq 4\|\Delta_{\mathcal{S}}\|_1 \leq 4\sqrt{s}\|\Delta_{\mathcal{S}}\|_F \leq 4\sqrt{s}\|\Delta\|_F.$$

For Assumption 4 we use a concentration result due to [33] in a similar manner as in [28]. The result is restated as Theorem A.4 below. Define  $Y_n = \frac{1}{T} \sum_{t=1}^n g(\mathcal{X}_t)_i \epsilon_{t,j}$  and note the following values

$$Y_n - Y_{n-1} = \frac{g(\mathcal{X}_n)_i}{T} \epsilon_{n,j}$$

and

$$M_n^k = \sum_{t=1}^n \mathbb{E} \left[ \left( \frac{g(\mathcal{X}_t)_i}{T} \epsilon_{t,j} \right)^k \middle| \mathcal{X}_{t-1} \right].$$

Thus  $Y_n$  is a martingale. We have  $g(\mathcal{X}_t)_i \leq U$ , and by Lemma 1 in [28],

$$\epsilon_{t,j} \leq X_{t,j} \leq C \log(MT)$$

for all  $t, j$  with probability at least  $1 - \exp(-cMT)$ . Thus,

$$|Y_n - Y_{n-1}| \leq \frac{CU \log(MT)}{T} =: B$$

with this same probability. Next, note that

$$\begin{aligned} M_n^2 &= \sum_{t=1}^n \mathbb{E} \left[ \frac{g(\mathcal{X}_t)_i^2}{T^2} \epsilon_{t,j}^2 \middle| \mathcal{X}_{t-1} \right] \\ &\leq \frac{1}{T^2} \sum_{t=1}^n U^2 R_{\max} \\ &= \frac{n}{T^2} U^2 R_{\max} =: \widehat{M}_n^2. \end{aligned}$$

Here we use that  $\mathbb{E}[\epsilon_{t,j}^2 | \mathcal{X}_{t-1}]$  is the variance of a Poisson random variable with mean bounded by  $R_{\max}$ , so it must also be bounded by  $R_{\max}$ . Next, we bound  $M_n^k$ :

$$M_n^k := \sum_{i=1}^n \mathbb{E} \left[ \left( \frac{g(\mathcal{X}_i)_m}{T^2} (X_{i,l} - \mathbb{E}[X_{i,l} | \mathcal{X}_{i-1}]) \right)^k \middle| \mathcal{X}_{i-1} \right] \leq B^{k-2} M_n^2.$$



In the language of Theorem A.4,

$$\begin{aligned} D_n &:= \sum_k \frac{\gamma^k}{k!} M_n^k \\ &\leq \frac{\widehat{M}_n^2}{B^2} \sum_k \frac{\gamma^k B^k}{k!} =: \widehat{D}_n. \end{aligned}$$

Let  $\tilde{D}_n$  corresponds to the negative sequence of  $D_n$ , and so it is still bounded by  $\widehat{D}_n$ . Using Markov's inequality, we get

$$\begin{aligned} \mathbb{P}(|Y_n| \geq y) &= \mathbb{P}(Y_n \geq y) + \mathbb{P}(-Y_n \leq y) \\ &\leq \mathbb{E}[\exp(\gamma Y_n)] \exp(-\gamma y) + \mathbb{E}[\exp(-\gamma Y_n)] \exp(-\gamma y) \\ &\leq \mathbb{E}[\exp(\gamma Y_n - D_n)] \exp(\widehat{D}_n - \gamma y) + \mathbb{E}[\exp(-\gamma Y_n - \tilde{D}_n)] \exp(\widehat{D}_n - \gamma y). \end{aligned}$$

Using Theorem A.4 we conclude that

$$\mathbb{E}[\exp(\gamma Y_n - D_n)] \leq 1$$

and  $\mathbb{E}[\exp(\gamma Y_n - \tilde{D}_n)] \leq 1$  so

$$\mathbb{P}(|Y_n| \geq y) \leq 2 \exp(\widehat{D}_n - \gamma y).$$

We set

$$\gamma = \frac{1}{B} \log \left( 1 + \frac{yB}{\widehat{M}_n^2} \right)$$

and to simplify things note that  $(1+x) \log(1+x) - x \geq \frac{3x^2}{2(x+3)}$ . Putting everything together gives

$$\begin{aligned} \mathbb{P}(|Y_T| \geq y) &\leq 2 \exp \left( \frac{-3y^2}{2yB + 6\widehat{M}_n^2} \right) \\ &= 2 \exp \left( \frac{3y^2 T}{2UC \log(MT)y + 6R_{\max}} \right). \end{aligned}$$

Now, recall that

$$\frac{\lambda}{2} = \frac{4CR_{\max}U^2 \log^2(MT)}{\sqrt{T}}$$

and setting  $y = \frac{\lambda}{2}$  gives

$$\begin{aligned} \mathbb{P}(Y_T \geq \frac{\lambda}{2}) &\leq 2 \exp\left(\frac{48C^2U^4 \log^4(MT)}{2C^2U^3 \log^3(MT)/\sqrt{T} + 6U^2R_{\max}}\right) \\ &= 2 \exp\left(\frac{48U \log(MT)}{2/\sqrt{T} + \frac{6R_{\max}}{C^2U \log^3(MT)}}\right) \\ &\leq 2 \exp\left(\frac{48U \log(MT)}{8}\right). \end{aligned}$$

Taking a union bound over all  $i, j$  gives us

$$\begin{aligned} \mathbb{P}\left(\max_{i,j} \frac{1}{T} \left| \sum_{t=1}^T g(\mathcal{X}_t)_{i\epsilon_{t,j}} \right| \geq 4U^2 \log^2(MT)/\sqrt{T}\right) \\ \leq \exp(\log(2M^2) - 6U \log(MT)) \\ \leq \exp(3 \log(MT) - 6U \log(MT)) \\ = \exp(-c \log(MT)) \end{aligned}$$

for  $c = 6U - 3$  which is positive since  $U \geq 1$ . In the final statement of the proof we assume  $C \log(MT) \geq U$  and replace  $U$  with  $C \log(MT)$  in order to limit the number of constants and make the crucial dependencies clear. This assumption should hold for reasonable choices of  $U$  in the settings we imagine in practice, but if not, a factor of  $\log^2(MT)$  can be replaced by  $U^2$  in the final bound.  $\square$

## A.5 Proof of Lemma 4.5

For Assumption 2, note that any  $A \in \overline{\mathcal{S}}_G$  satisfies  $A_{.c} = 0$  for  $i \notin S_G$ . Therefore

$$\begin{aligned} \|\Delta\|_G &\leq 4\|\Delta_{\overline{\mathcal{S}}_G}\|_G = 4 \sum_{i \in S_G} \|\Delta_{.i}\|_2 \\ &\leq 4\sqrt{s_G} \|\Delta_{\overline{\mathcal{S}}_G}\|_F \\ &\leq 4\sqrt{s_G} \|\Delta\|_F. \end{aligned}$$

For Assumption 3, we have  $\|\Delta\|_G^2 \leq 16s_G \|\Delta\|_F^2$  from the previous paragraph, and we claim  $\|A\|_{2,1} \leq \|A\|_G$  for any matrix  $A$ . To see this, we compute

$$\|A\|_G^2 = \left( \sum_c \sqrt{\sum_r a_{r,c}^2} \right)^2 = \sum_c \sum_{c'} \sqrt{(\sum_r a_{r,c}^2)(\sum_r a_{r,c'}^2)}$$

while

$$\|A\|_{2,1}^2 = \sum_r (\sum_c |a_{r,c}|)^2 = \sum_c \sum_{c'} \sum_r |a_{r,c}| |a_{r,c'}|.$$

To complete the proof, we fix  $c, c'$  and need to show

$$\sum_r |a_{r,c}| |a_{r,c'}| \leq \sqrt{\left(\sum_r a_{r,c}^2\right) \left(\sum_r a_{r,c'}^2\right)},$$

or equivalently that

$$\left(\sum_r |a_{r,c}| |a_{r,c'}|\right)^2 \leq \left(\sum_r a_{r,c}^2\right) \left(\sum_r a_{r,c'}^2\right).$$

We have

$$\left(\sum_r |a_{r,c}| |a_{r,c'}|\right)^2 = \sum_r \sum_{r'} |a_{r,c} a_{r,c'} a_{r',c} a_{r',c'}|.$$

Let  $\mathcal{J}$  denote all two element combinations of  $M$  and we can write

$$\sum_r \sum_{r'} |a_{r,c} a_{r,c'} a_{r',c} a_{r',c'}| = \sum_r (a_{r,c} a_{r,c'})^2 + \sum_{(i,j) \in \mathcal{J}} 2 |a_{i,c} a_{j,c} a_{i,c'} a_{j,c'}|.$$

On the other hand,

$$\begin{aligned} \left(\sum_r a_{r,c}^2\right) \left(\sum_r a_{r,c'}^2\right) &= \sum_r \sum_{r'} a_{r,c}^2 a_{r',c'}^2 \\ &= \sum_r (a_{r,c} a_{r,c'})^2 + \sum_{(i,j) \in \mathcal{J}} (a_{i,c} a_{j,c'})^2 + (a_{j,c} a_{i,c'})^2. \end{aligned}$$

The proof follows from noting that

$$(a_{i,c} a_{j,c'})^2 + (a_{j,c} a_{i,c'})^2 \geq 2 |a_{i,c} a_{j,c} a_{i,c'} a_{j,c'}|$$

for any real numbers  $a_{i,c}, a_{i,c'}, a_{j,c}, a_{j,c'}$ .

For Assumption 4, we rely on Theorem 1 in [51] which is restated as Theorem A.3. In our setup, we need to bound the  $l_2$  norm of the  $m$ th column of  $\frac{1}{T} \sum_t \epsilon_t g(\mathcal{X}_t)^\top$ . Note that the  $l_2$  norm is 2-smooth, because for any  $x, y \in \mathbb{R}^m$  we have

$$\begin{aligned} \|x + y\|^2 + \|x - y\|^2 &= \langle x + y, x + y \rangle + \langle x - y, x - y \rangle \\ &= 2\langle x, x \rangle + 2\langle y, y \rangle. \end{aligned}$$

In the language of Theorem A.3, for a fixed  $m$  we form a martingale difference sequence  $\{Z_t\}$  with

$$Z_t = \frac{1}{T} \left( \epsilon_t g(\mathcal{X}_t)^\top \right)_{.m}$$

so that

$$\|Z_t\|_2 = \frac{1}{T} \sqrt{\sum_{m'} (g(\mathcal{X}_t)_{m} \epsilon_{t,m'})^2} = \frac{g(\mathcal{X}_t)_m}{T} \sqrt{\sum_{m'} \epsilon_{t,m'}^2}.$$

We know  $g(\mathcal{X}_t)_m \leq U$  and by Lemma 1 from [28]  $\epsilon_{t,m'} \leq C \log(MT)$  with probability at least  $1 - \exp(-cMT)$ . We conclude

$$\|Z_t\|_2 \leq \frac{1}{T} U \sqrt{MC \log^2(MT)}$$

and thus

$$\sum_{t=1}^T \|Z_t\|_2^2 \leq CU^2 \log^2(MT) \frac{M}{T}.$$

To compute the constant  $Q_{\max}$  appearing in Theorem A.3 we let  $R(x) = x^\top x$  so that  $\nabla R(x) = x$ . Then for any  $x, y$  in the unit ball with respect to the  $\|\cdot\|_2$  norm, we have

$$D_R(x, y) = \|x\|_2^2 - \|y\|_2^2 - \langle y, x - y \rangle \leq \|x\|_2^2 + \|y\|_2 \|x - y\|_2 \leq 3.$$

by Cauchy-Schwarz. Thus we can take  $Q_{\max} = \sqrt{3}$ . To simplify, we note  $W_n \leq V_n$  and

$$(\mathbb{E}[\sqrt{V_n + W_n}])^2 \leq \mathbb{E}[V_n + W_n] \leq 2V_n.$$

Further,  $2.5Q_{\max}(\sqrt{V_n} + 1) \leq 5\sqrt{V_n}$ . With these simplifications, Theorem 1 from [51] says that

$$\mathbb{P}\left(\frac{1}{T} \left\| \left( \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right)_{.m} \right\|_2 > (5 + 2u)V_n\right) \leq \sqrt{2} \exp\left(-\frac{u^2}{16}\right).$$

Setting  $u = \log(T)$  and plugging in our values for  $V_n$  we conclude that

$$\mathbb{P}\left(\frac{1}{T} \left\| \left( \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right)_{.m} \right\|_2 > CU \log^2(MT) \sqrt{\frac{M}{T}}\right) \leq \sqrt{2} \exp(-\log^2(MT)).$$

Taking a union bound over all  $m$ , we get that

$$\frac{1}{T} \left\| \left( \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right)_{.m} \right\|_2 \leq CU \log^2(MT) \sqrt{\frac{M}{T}}$$

for all  $m$  with probability at least  $1 - \sqrt{2} \exp(-\log(MT)) = 1 - \frac{\sqrt{2}}{MT}$ .  $\square$

## A.6 Proof of Lemma 4.7

For Assumption 3, from the statement of Lemma 4.7 we have  $\|A^*\|_{2,1}^2 \leq D\sqrt{M}$  and we search for  $\hat{A}$  over the ball  $\{A : \|A\|_{2,1}^2 \leq D\sqrt{M}\}$ . Thus

$$\sup_{B \in \mathcal{B}'_{\mathcal{R}}} \|B\|_{2,1}^2 \leq 2D\sqrt{M} = \mu_{\mathcal{R}}.$$

In contrast to the sparsity case, Assumption 2 is nontrivial to verify in the low-rank case because  $\mathcal{W} \neq \overline{\mathcal{W}}$ . However, this condition was shown in Lemma 3.4 of [53].

For Assumption 4 we rely on the notion of a  $k$ -regular normed vector space defined in Section A.9 as well as Theorem 2.1 from [34] which is stated in Theorem A.2. Further, Example 3.1 in [34] establishes that  $(\mathbb{R}^{M \times MK}, \|\cdot\|_*)$  is  $k$ -regular for  $k = 3 \log(\min(M, N))$ . In the language of Theorem A.2 we form a martingale difference sequence  $\{\zeta_t\}$  with  $\zeta_t = \frac{1}{T} \epsilon_t g(\mathcal{X}_t)^\top$  and then

$$\|\zeta_t\|_{op} = \frac{\epsilon_t^\top g(\mathcal{X}_t)}{T} = \frac{1}{T} \sum_{m=1}^M \epsilon_{t,m} g(\mathcal{X}_t)_m.$$

Consider the random variable

$$\sum_{m=1}^M \epsilon_{t,m} g(\mathcal{X}_t)_m. \quad (\text{A.11})$$

We have  $g(\mathcal{X}_t)_m \leq U$  and  $\epsilon_{t,m} \leq C \log(MT)$  for all  $t, m$  with probability at least  $e^{-cMT}$  by Lemma 1 from [28]. Further, conditioned on  $\mathcal{X}_t$ , the  $\epsilon_{t,m} g(\mathcal{X}_t)_m$  are all independent, so (A.11) is a sum of zero mean independent random variables bounded by  $CU \log(MT)$ . Hence, we have

$$\sum_{m=1}^M \mathbb{E}[\epsilon_{t,m} g(\mathcal{X}_t)_m | \mathcal{X}_{m-1}] \leq CM \log^2(MT) U^2,$$

and applying Bernstein's inequality gives

$$\mathbb{P}\left(|\epsilon_t^\top g(\mathcal{X}_t)| > \sqrt{M} \log^2(MT)\right) \leq 2 \exp\left(-\frac{\log^4(MT)/2}{C \log^2(MT) U^2 + \frac{CU \log(MT)}{3\sqrt{M}}}\right).$$

Therefore

$$\begin{aligned} & \mathbb{P}\left(|\epsilon_t^\top g(\mathcal{X}_t)| > \sqrt{M} \log^2(MT) \text{ for at least one } t\right) \\ & \leq 2 \exp\left(\log(T) - \log^4(MT)\right) / (2C \log^2(MT) U^2 + 1) \\ & \leq \exp\left(-\frac{\log^4(MT)}{4C \log^2(MT) U^2 + 2}\right). \end{aligned}$$

We apply Theorem A.2 with  $k = 3 \log(M)$ ,

$$\sum_{i=1}^T \sigma_i^2 = \frac{M \log^4(MT)}{T}$$

and  $\gamma = \log(T)$ . This gives

$$\begin{aligned} \mathbb{P}\left(\frac{1}{T} \left\| \sum_t \epsilon_t g(\mathcal{X}_t)^\top \right\|_{op} > (3\sqrt{2} \log(M) + \sqrt{2} \log(T)) \log^4(MT) \sqrt{\frac{M}{T}}\right) \\ \leq \exp\left(\frac{-\log^2(T)}{2}\right). \end{aligned}$$

□

## A.7 Proof of Proposition 1

In this proof, we take  $\lambda_m^{(c)}(\tau)$  to mean  $\lambda_m^{(c)}(\tau; \mathcal{X}_\tau)$ . Using the approximation

$$\int_{(t-1)\Delta}^{\Delta t} \lambda_m^{(c)}(\tau) d\tau \approx \Delta \lambda_m^{(c)}(\Delta t)$$

we derive an approximate sampled Hawkes (SH) log-likelihood proportional to

$$\ell_H(\mathcal{X}_{T\Delta} | \{\lambda_m^{(c)}\}_m) \approx \sum_{m=1}^M \sum_{t=1}^T [X_{t,m} \log \lambda_m^{(c)}(\Delta t) - \Delta \lambda_m^{(c)}(\Delta t)] =: \ell_{SH}(\mathcal{X}_T | \{\lambda_m^{(c)}\}_m).$$

If  $X_{t,m}$  were generated according to (2.3) with intensity (6.5) for  $T = 1, \dots, T$ , then, ignoring terms independent of  $A^*$ ,

$$\ell_P(\mathcal{X}_t | \{\Delta \lambda_m^{(c)}(\mathcal{X}_{\Delta t})\}_{t,m}) := \sum_{m=1}^M \sum_{t=1}^T [X_{t,m} \log \Delta \lambda_m^{(c)}(\Delta t) - \Delta \lambda_m^{(c)}(\Delta t)].$$

Note that

$$\ell_P(\mathcal{X}_T | \{\Delta \lambda_m^{(c)}(\mathcal{X}_{\Delta t})\}_{t,m}) = \ell_{SH}(\mathcal{X}_T | \{\lambda_m^{(c)}\}_m) + C$$

where the constant  $C$  depends on  $\Delta$  but is independent of  $\lambda_m^{(c)}$ . □

## A.8 Extension to more general saturation functions

In the main body of the paper the only saturation function we consider is  $f(x) = \min(x, \tilde{U})$  for purposes of simplicity, but our theory extends to a larger class of saturation functions  $f$ . However, for our analysis it is crucial to assume that the function  $f$  is bounded so that we can define the maximum and minimum rates,  $R_{\min}$  and  $R_{\max}$ , from which each observation is drawn. The only place where we rely on the structure of  $f$  beyond its boundedness is in proving the restricted eigenvalue condition in Assumption 1. In the case of the ARMA(1, 1) model, we show our results extend to monotonic differentiable functions in Proposition 2 below.

**Proposition 2.** Suppose  $(X_t)_{t=1}^T$  is generated according to the ARMA(1, 1) model in (4.1) with a general saturation function  $f$  applied entrywise to the vector  $X_t$ . Suppose  $f$  is bounded on  $\mathbb{R}$ , monotonically increasing, and differentiable with  $f'(x) \geq c$  on  $[0, R_{\max}]$ . Then

$$\lambda_{\min}[\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-1}]] \geq c^2 \min\left(\frac{1}{2}R_{\min}, \kappa\right).$$

**Proof** We have

$$\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^\top | \mathcal{X}_{t-1}] = \mathbb{E}[g(\mathcal{X}_t) | \mathcal{X}_{t-1}]\mathbb{E}[g(\mathcal{X}_t) | \mathcal{X}_{t-1}]^\top + \text{Diag}(\text{Var}(g(\mathcal{X}_t) | \mathcal{X}_{t-1}))$$

where the first matrix is positive semi-definite because it is the outer product of a vector with itself. Thus, to come up with a lower bound for our original matrix, we just need to lower bound the smallest element of  $\text{Var}(g(\mathcal{X}_t) | \mathcal{X}_{t-1})$ . This amounts to lower bounding the variance of  $f(X)$  where  $X$  is a Poisson random variable with mean  $\lambda \in [R_{\min}, R_{\max}]$ .

Let  $p = \mathbb{P}(X \leq \lfloor \lambda \rfloor)$  so  $1 - p = \mathbb{P}(X \geq \lceil \lambda \rceil)$ . Consider the random variable  $X'$  which takes the value  $\lfloor \lambda \rfloor$  with probability  $p$  and  $\lceil \lambda \rceil$  with probability  $1 - p$ . Since  $f$  is monotonic, the argument from Lemma A.1 shows that  $\text{Var}(f(X')) \leq \text{Var}(f(X))$  so we reduce our problem to lower bounding the variance of  $f(X')$ .

Note that this variance is equal to the shifted random variable  $X''$  defined by  $f(X') = 0$  with probability  $p$  and  $f(\lceil \lambda \rceil) - f(\lfloor \lambda \rfloor)$  with probability  $1 - p$  which is a scaled Bernoulli random variable with variance

$$(f(\lceil \lambda \rceil) - f(\lfloor \lambda \rfloor))^2 p(1 - p).$$

Since  $f'(x) \geq c$  on  $[0, R_{\max}]$ ,

$$f(\lceil \lambda \rceil) - f(\lfloor \lambda \rfloor) \geq c$$

and so the lower bound on our variance becomes  $c^2 p(1 - p)$ . Finally, by Lemma 4.1 we have  $p(1 - p) \geq \min(\frac{1}{2}R_{\min}, \kappa)$  which completes the proof.  $\square$

## A.9 Supplemental Theorems

**Definitions** Before introducing martingale concentration results, we give the following definitions.

**Definition 1.** A Banach space  $(E, \|\cdot\|)$  is *s-smooth* if there exists  $C > 0$  satisfying

$$\|x + y\|^s + \|x - y\|^s \leq 2\|x\|^s + 2C^s\|y\|^s$$

for all  $x, y \in E$ .

Note that  $(\mathbb{R}^M, \|\cdot\|_2)$  is 2-smooth with  $C = 1$  because

$$\|x + y\|^2 + \|x - y\|^2 = \langle x + y, x + y \rangle + \langle x - y, x - y \rangle = 2\langle x, x \rangle + 2\langle y, y \rangle.$$

**Definition 2.** A Banach space  $(E, \|\cdot\|)$  is  $k$ -regular if there exists  $k_+ \in [1, k]$  along with a norm  $\|\cdot\|_+$  such that  $(E, \|\cdot\|_+)$  is  $k_+$ -smooth and

$$\|x\|^2 \leq \|x\|_+^2 \leq \frac{k}{k_+} \|x\|^2$$

for all  $x \in E$ .

By Example 3.3 from [34], the space  $(\mathbb{R}^{M \times N}, \|\cdot\|_*)$  is  $k$ -regular for  $k = 3 \log(\min(M, N))$ .

**Theorem A.2.** (Theorem 2.1.iii in [34])

Let  $(E, \|\cdot\|)$  be  $k$ -regular and let  $\zeta_i$  be an  $E$ -valued martingale difference sequence with  $\|\zeta_i\| \leq \sigma_i$ . Let  $S_N = \sum_{i=1}^N \zeta_i$ . Then

$$\mathbb{P} \left( \|S_N\| \geq (\sqrt{2k} + \sqrt{2}\gamma) \sqrt{\sum_{i=1}^N \sigma_i^2} \right) \leq \exp\left(-\frac{\gamma^2}{2}\right).$$

**Theorem A.3.** (Theorem 1 in [51]) Let  $(E, \|\cdot\|)$  be a 2-smooth Banach space. Let  $R$  be a function which is 1-strongly convex on the unit ball in the dual norm of  $\|\cdot\|$ .

$$D_R : B_* \times B_* \rightarrow \mathbb{R}$$

be the Bregman divergence with respect to  $R$ , and finally let  $Q_{\max}^2 = \sup_{x, y \in B_*} D_R(f, g)$ . Let  $Z_1, \dots, Z_n$  be a martingale difference sequence with  $V_n = \sum_{t=1}^n \|Z_t\|^2$  and  $W_n = \sum_{t=1}^n \mathbb{E}_{t-1} \|Z_t\|^2$ . Then

$$\begin{aligned} \mathbb{P} \left( \left\| \sum_{t=1}^n Z_t \right\| > 2.5Q_{\max}(\sqrt{V_n} + 1) + u \sqrt{V_n + W_n + (\mathbb{E}[\sqrt{V_n + W_n}])^2} \right) \\ \leq \sqrt{2} \exp\left(-\frac{u^2}{16}\right). \end{aligned}$$

**Theorem A.4.** (Lemma 3.3 in [33]) Let  $(Y_n)$  be a martingale and let

$$M_n^k = \sum_{i=1}^n \mathbb{E}[(Y_i - Y_{i-1})^k | \mathcal{Y}_{i-1}].$$

Let  $\gamma$  be such that for all  $i \leq n$ , we have

$$\mathbb{E}[\exp(|\gamma(Y_i - Y_{i-1})|)] \leq \infty.$$



Then

$$\epsilon_n = \exp\left(\gamma Y_n - \sum_{k \geq 2} \frac{\gamma^k}{k!} M_n^k\right)$$

is a super-martingale. Moreover, if  $Y_0 = 0$  then  $\mathbb{E}[\epsilon_n] \leq 1$ .

**Theorem A.5.** (Azuma-Hoeffding inequality) Let  $(Y_n)$  be a martingale and  $|Y_n - Y_{n-1}| < c_n$ . Then

$$\mathbb{P}(|Y_N - Y_0| \geq t) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{n=1}^N c_n^2}\right).$$