Inference of High-dimensional Autoregressive Generalized Linear Models

Eric C. Hall[*], Garvesh Raskutti[†], and Rebecca M. Willett[‡]

June 27, 2017

Abstract

Vector autoregressive models characterize a variety of time series in which linear combinations of current and past observations can be used to accurately predict future observations. For instance, each element of an observation vector could correspond to a different node in a network, and the parameters of an autoregressive model would correspond to the impact of the network structure on the time series evolution. Often these models are used successfully in practice to learn the structure of social, epidemiological, financial, or biological neural networks. However, little is known about statistical guarantees on estimates of such models in non-Gaussian settings. This paper addresses the inference of the autoregressive parameters and associated network structure within a generalized linear model framework that includes Poisson and Bernoulli autoregressive processes. At the heart of this analysis is a sparsity-regularized maximum likelihood estimator. While sparsity-regularization is well-studied in the statistics and machine learning communities, those analysis methods cannot be applied to autoregressive generalized linear models because of the correlations and potential heteroscedasticity inherent in the observations. Sample complexity bounds are derived using a combination of martingale concentration inequalities and modern empirical process techniques for dependent random variables. These bounds, which are supported by several simulation studies, characterize the impact of various network parameters on estimator performance.

1 Autoregressive Processes in High Dimensions

Imagine recording the times at which each neuron in a biological neural network fires or "spikes". Neuron spikes can trigger or inhibit spikes in neighboring neurons, and understanding excitation and inhibition among neurons provides key insight into the structure and operation of the underlying neural network [1, 2, 3, 4, 5, 6, 7]. A central question in the design of this experiment is *"for how long must I collect data before I can be confident that my inference of the network is accurate?"* Clearly the answer to this question will depend not only on the number of neurons being recorded, but also on what we may assume *a priori* about the network. Unfortunately, existing statistical and machine learning theory give little insight into this problem.

Neural spike recordings are just one example of a non-Gaussian, high-dimensional autoregressive processes, where the autoregressive parameters correspond to the structure of the underlying network. This paper examines a broad class of such processes, in which each observation vector is modeled using an exponential family distribution. In general, autoregressive models are a widely-used mechanism for studying time series in which each observation depends on the past sequence of observations. Inferring these dependencies is a key challenge in many settings, including finance, neuroscience, epidemiology, and sociology. A precise understanding of these dependencies facilitates more accurate predictions and interpretable models of the forces that determine the distribution of each new observation.

[*]E. C. Hall is with the Wisconsin Institute of Discovery, University of Wisconsin-Madison, Madison, WI, 53706, USA. e-mail: echall@wisc.edu

[†]G. Raskutti is with the Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706, USA. e-mail: raskutti@stats.wisc.edu

[‡]R. M. Willett is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA. e-mail: willett@discovery.wisc.edu. We gratefully acknowledge the support of the awards NSF CCF-1418976, NIH 1 U54 AI117924-01, 14-AFOSR-1103, and NSF DMS-1407028.

Much of the autoregressive modeling literature focuses on Gaussian noise and perturbation models, but in many settings Gaussian noise fails to capture the data at hand. This challenge arises, for instance, when observations correspond to count data – *e.g.,* when we collect data by counting individual events such as neurons spiking. Another example arises in epidemiology, where a common model involves infection traveling stochastically from one node in a network to another based on the underlying network structure in a process known as an "epidemic cascade" [8, 9, 10, 11]. These models are used to infer network structure based on the observations of infection time, which is closely related to the Bernoulli autoregressive model studied in this paper. Further examples arise in a variety of applications, including vehicular traffic analysis [12, 13], finance [14, 15, 16, 17], social network analysis [18, 19, 20, 21, 22], biological neural networks [1, 2, 3, 4, 5, 6, 7], power systems analysis [23], and seismology [24, 25].

Because of their prevalence across application domains, time series count data (*cf.* [26, 27, 28, 29, 30]) and other non-Gaussian autoregressive processes (*cf.* [31, 32, 33]) have been studied for decades. Although a substantial fraction of the this literature is focused on univariate time series, this paper focuses on multivariate settings, particularly where the vector observed at each time is high-dimensional relative to the duration of the time series. In the above examples, the dimension of the each observation vector would be the number of neurons in a neural network, the number of people in a social network, or the number of interacting financial instruments.

In this paper, we conduct a detailed investigation of a particular family of time series that we call the *vector generalized linear autoregressive* (GLAR) model. In addition, we examine our results for two members of this family: the Bernoulli autoregressive and the log-linear Poisson autoregressive (PAR) model. The PAR model has been explicitly studied in [34, 35, 36] and is closely related to the continuous-time Hawkes point process model [37, 38, 39, 40, 41] and the discrete-time INGARCH model [42, 43, 44, 45]. However, that literature does not contain the sample complexity results presented here. The INGARCH literature is focused on low-dimensional settings, typically univariate, whereas we are focused on the high-dimensional setting where the number of nodes or channels is high relative to the number of observations. Additionally, existing sample complexity bounds for Hawkes processes [40] focus on a linear (as opposed to log-linear) model with samples collected after reaching the stationary distribution. The log-linear model is largely used in practice both for numerical reasons and modeling efficacy for real world data. We note that linear models can predict inadmissible negative event rates, whereas the log-linear model enforces the feasibility of the predicted model. The log-linear and linear models exhibit very different behaviors in their properties and stationary distributions, making this work a significant step forward from the analysis of linear models. The extension of these prior investigations to the high-dimensional, non-stationary setting is non-trivial and requires the development of new theory and methods.

This paper focuses on estimating the parameters of a vector GLAR model from a time series of observations. We adopt a regularized likelihood estimation approach that extends and generalizes our previous work on Poisson inverse problems (*cf.* [46, 47, 48, 49]). While similar algorithms have been proposed in the above-mentioned literature, little is known about their *sample complexity* or *how inference accuracy scales with the key parameters such as the size of the network or number of entities observed, the time spent collecting observations, and the density of edges within the network or dependencies among entities.*

There has been a large body of work providing theoretical results for certain high-dimensional models under low-dimensional structural constraints (see *e.g.,* [50, 49, 51, 52, 53, 54, 55, 56, 41]). The majority of prior work has focused on the setting where samples are independent and/or follow a Gaussian distribution. In the GLAR setting, however, non-Gaussianity and temporal dependence among observations can make such analyses particularly challenging and beyond the scope of much current research in high-dimensional statistical inference (see [57] for an overview).

Perhaps the most closely related prior work to our setting in the high-dimensional setting is [58]. In [58], several performance guarantees are provided for different linear Gaussian problems with dependent samples including the Gaussian autoregressive model. Since [58] deals exclusively with linear Gaussian models, they exploit many properties of linear systems and Gaussian random variables that cannot be applied to non-Gaussian and non-linear autoregressive models. In particular, compared to standard autoregressive processes with Gaussian noise, in the GLAR setting the conditional variance of each observation is dependent on previous data instead of being a constant equal to the noise variance. Works such as [49, 50, 59] provide results for non-Gaussian models but still rely on independent observations. Weighted LASSO estimators for Hawkes processes address some of these challenges in a continuous-time setting [40].

To see why GLAR analysis can be challenging, consider momentarily a LASSO estimator of the autoregressive

parameters. In the classical LASSO setting, the accuracy of the estimate depends on characteristics of the Gram matrix associated with the design or sensing matrix. This matrix may be stochastic, but it is usually considered independent of the observations and performance guarantees for the estimator depend on the assumption that the matrix obeys certain properties (*e.g.,* the restricted eigenvalue condition [60]). In our setting, however, the "design" matrix is a function of the observed data, which in turn depends on the true underlying network or autoregressive model parameters. Thus a key challenge in the analysis of a LASSO-like estimator in the GLAR setting involves showing that the data- and network-dependent Gram matrix exhibits properties that ensure reliable estimates.

In this paper, we develop performance guarantees for the vector GLAR model that provide sample complexity guarantees in the high-dimensional setting under low-dimensional structural assumptions such as sparsity of the underlying autoregressive parameters. In particular, our main contributions are the following:

- Formulation of a maximum penalized likelihood estimator for vector GLAR models in high-dimensional settings with sparse structure.
- Mean-squared-error bounds on the proposed estimator as a function of the problem dimension, sparsity, and the number of observations in time for general GLAR models.
- Application of our general result to obtain sample complexity bounds for Bernoulli and Poisson GLAR models.
- Analysis techniques that simultaneously leverage martingale concentration inequalities, empirical risk minimization analysis, and covering arguments for high-dimensional linear regression.

This problem is substantially harder than the Gaussian case from a technical perspective because we can not exploit linearity and spectral properties of linear Gaussian time-series. In our case we have signal-dependent noise, and we can not exploit the same spectral properties. Additionally, with non-Gaussian noise, we are not guaranteed strong convexity of the objective function in the entire domain of possible solutions, and so extra care must be taken to define regions of strong convexity. Thus we have to develop new theoretical techniques, using new concentration bounds and a more refined analysis. The remainder of the paper is structured as follows: Section 2 introduces the generalized linear autoregressive model and Section 3 presents the novel risk bounds associated with the RMLE of the process. We then use our theory to examine two special cases (the Poisson and Bernoulli models) in Sections 3.1 and 3.2, respectively. The main proofs are provided in Section 4, while supplementary lemmas are deferred to the appendix. Finally, Section 5 contains a discussion of our results, their implications in different settings, and potential avenues for future work.

2 Problem Formulation

In this paper we consider the generalized linear autoregressive model:

$$X_{t+1,m}|X_t \sim p(\nu_m + a_m^{*\top} X_t), \tag{1}$$

where $X_{t+1,m}$ is the m^{th} observation of X_{t+1}, $(X_t)_{t=0}^{\infty}$ are M-variate vectors and $a^* \in [a_{\min}, a_{\max}]^M$ is an unknown parameter vector, $\nu \in [\nu_{\min}, \nu_{\max}]^M$ is a known, constant offset parameter, and p is an exponential family probability distribution. Specifically, $X \sim p(\theta)$ means that the distribution of the scalar X is associated with the density $p(x|\theta) = h(x)\exp[\phi(x)\theta - Z(\theta)]$, where $Z(\theta)$ is the so-called *log partition function*, $\phi(x)$ is the sufficient statistic of the data, and $h(x)$ is the base measure of the distribution. Distributions that fit such assumptions include the Poisson, Bernoulli, binomial, negative binomial and exponential. According to this model, conditioned on the previous data, the elements of X_t are independent of one another and each have a scalar natural parameter. The input of the function p in (1) is the natural parameter for the distribution, *i.e.,* $\nu + a_m^{*\top} X_t$ is the natural parameter of the conditional distribution at time $t+1$ for observation m. A similar, but low-dimensional, model appears in [36], but that work focuses on maximum likelihood and weighted least squares estimators in univariate settings that are known to perform poorly in high-dimensional settings (as is our focus). For these distributions it is straightforward to show when they have strongly convex log-partition functions, which will be crucial to our analysis. Note that this distribution has $\mathbb{E}[\phi(X_{t+1,m})|X_t] = Z'(\nu + a_m^{*\top} X_t)$ and $\mathrm{Var}(\phi(X_{t+1,m})|X_t) = Z''(\nu + a_m^{*\top} X_t)$, the first and second derivatives

of the log-partition function, respectively. Compared to standard autoregressive processes with Gaussian noise, the conditional variance is now dependent on previous data instead of being a constant equal to the noise variance.

We can state the conditional distribution explicitly as:

$$\mathbb{P}(X_{t+1}|X_t) = \prod_{m=1}^{M} h(X_{t+1,m}) \exp\left(\phi(X_{t+1,m})(\nu_m + a_m^{*\top} X_t) - Z(\nu_m + a_m^{*\top} X_t)\right),$$

where h is the base-measure of the distribution p. Using this equation and observations, we can find an estimate for the network A^* which is constructed row-wise by a_m^*. ($i.e.$ $a_m^{*\top}$ is the m^{th} row of A^*).

In general, we observe T samples $(X_t)_{t=0}^T$ and our goal is to infer the matrix A^*. In the setting where M is large, we need to impose structural assumptions on A^* in order to have strong performance guarantees. Let

$$\mathcal{S} := \{(\ell, m) \in \{1, \dots, M\}^2 : A^*_{\ell,m} \neq 0\}.$$

In this paper we assume that the matrix A^* is s-sparse, meaning that A^* belongs to the following class:

$$\mathcal{A}_s = \left\{A \in [a_{\min}, a_{\max}]^{M \times M} \mid \|A\|_0 \leq s\right\}.$$

where $\|A\|_0 := \sum_{\ell=1}^M \sum_{m=1}^M \mathbf{1}(|A_{\ell,m}| \neq 0)$ and $\mathbf{1}(\cdot)$ is the indicator function. That is, we assume $|\mathcal{S}| = s$. Furthermore, we define

$$\rho_m \triangleq \|a_m^*\|_0 \qquad \text{and} \qquad \rho \triangleq \max_m \rho_m,$$

so ρ is the maximum number of non-zero elements in a row of A^*.

We might like to estimate A^* via a constrained maximum likelihood estimator by solving the following optimization problem:

$$\arg\min_{A \in \mathcal{A}_s} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \left(Z(\nu_m + a_m^\top X_t) - a_m^\top X_t \phi(X_{t+1,m}) \right) \tag{2}$$

or its Lagrangian form

$$\arg\min_{A \in [a_{\min}, a_{\max}]^{M \times M}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \left(Z(\nu_m + a_m^\top X_t) - a_m^\top X_t \phi(X_{t+1,m}) \right) + \lambda \|A\|_0. \tag{3}$$

However, these are difficult optimization problems due to the non-convexity of the ℓ_0 norm. Therefore, we instead find an estimator using the element-wise ℓ_1 regularizer, the convex relaxation of the ℓ_0 function, along with the negative log-likelihood to create the following estimator:

$$\widehat{A} = \arg\min_{A \in [a_{\min}, a_{\max}]^{M \times M}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \left(Z(\nu_m + a_m^\top X_t) - a_m^\top X_t \phi(X_{t+1,m}) \right) + \lambda \|A\|_{1,1}, \tag{4}$$

where $\|\cdot\|_1$ is the ℓ_1 norm and $\|A\|_{1,1} = \sum_{m=1}^M \|a_m\|_1$. The above is the regularized maximum likelihood estimator (RMLE) for the problem, which attempts to find an estimate of A^* which both fits the empirical distribution of the data while also having many zero-valued elements. Notice that we assume the elements of A^* are bounded and we use these bounds in the estimator definition. One reason for this is that bounds on the elements of A^* can enforce stability. If the elements of A^* are allowed to be arbitrarily large, the system may become unstable and therefore impossible to make proper estimates. Knowing loose bounds facilitates our analysis but in practice does not appear to be necessary. In the experiment section we discuss choosing these bounds in the estimation process.

We note that while we assume that ν is a known constant vector, if we assume there is some unknown constant offset that we would like to estimate, we can fold it into the estimation of A. For instance, consider appending ν as an extra column of the matrix A^*, and appending a 1 to the end of each observation X_t. Then for indices $1, \dots, M$ the observation model becomes $X_{t+1,m}|X_t \sim p(a_m^{*\top} X_t)$ where a_m^* and X_t are the appended versions. We can then find the RMLE of this distribution to find both $\widehat{A}$ and $\widehat{\nu}$, but for clarity of exposition we assume a known ν.

Estimating the network parameters in the autoregressive setting with Gaussian observations can be formulated as a sparse inverse problem with connections to the well-known LASSO estimator. Consider the problem of estimating the a_m^*. Define

$$y_m = \begin{bmatrix} X_{2,m} \\ X_{3,m} \\ \vdots \\ X_{T,m} \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,M} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ X_{T-1,1} & X_{T-1,2} & \cdots & X_{T-1,M} \end{bmatrix},$$

where y_m is the time series of observed counts associated with the m^{th} node and $\mathbf{X}$ is a matrix of the observed counts associated with all nodes. Then $y_m = \mathbf{X}\mathbf{a}_\mathbf{m}^* + \epsilon_\mathbf{m}$, where $\epsilon_m := y_m - \mathbf{X}\mathbf{a}_\mathbf{m}^*$ is noise, and we could consider the LASSO estimator for each m:

$$\hat{a}_m = \arg\min_a \|y_m - \mathbf{X}\mathbf{a}\|_\mathbf{2}^\mathbf{2} + \lambda\|\mathbf{a}\|_\mathbf{1}.$$

However, there are two key challenges associated with the LASSO estimator in this context: (a) the squared residual term does not account for the non-Gaussian statistics of the observations and (b) the "design matrix" is data-dependent and hence a function of the unknown underlying network. In classical LASSO analyses, performance bounds depend on the design matrix satisfying the *restricted eigenvalue condition* or *restricted isometry property* or some related condition; it is relatively straightforward to ensure such a condition is satisfied when the design matrix is independent of the data, but much more challenging in the current context. As a result, despite the fact that we face a sparse inverse problem, the existing LASSO literature does not address the subject of this proposal.

3 Main Results

In this section, we turn our attention to deriving bounds for $\|\widehat{A} - A^*\|_F^2$, the difference in Frobenius norm between the regularized maximum likelihood estimator, $\widehat{A}$, and the true generating network, A^*, under the assumption that the true network is sparse. We assume that $A^* \in \mathcal{A}_s$. Recall $\rho \triangleq \max_m \|a_m^*\|_0$ is the maximum number of non-zero elements in a row of A^*. First we state assumptions on the GLAR process which are sufficient conditions to ensure the RMLE admits small errors.

Asumption 1. We assume that for any realization of the process defined by Equation 1 there exists a subset of observations $\{X_{\mathcal{T}_t}\}_{t=1}^{|\mathcal{T}|}$ for $\mathcal{T} \subseteq \{0, 1, \ldots, T-1\}$ that satisfies the conditions:

1. There exists a constant U such that $U \geq \max_{t\in\mathcal{T}} \|X_t\|_\infty$ where U is independent of T.
2. $Z(\cdot)$ is σ-strongly convex on a domain determined by U:

$$Z(x) \geq Z(y) + Z'(y)(x-y) + \frac{\sigma}{2}\|x-y\|_2^2$$

   for all $x, y \in [-\tilde{\nu} - 9\rho\tilde{a}U, \tilde{\nu} + 9\rho\tilde{a}U]$ where $\tilde{\nu} \triangleq \max(|\nu_{\min}|, |\nu_{\max}|)$, and $\tilde{a} \triangleq \max(|a_{\min}|, |a_{\max}|)$, where σ is independent of T.
3. The smallest eigenvalue of $\Gamma_t \triangleq \mathbb{E}[X_{\mathcal{T}_t} X_{\mathcal{T}_t}^\top | X_{\mathcal{T}_{t-1}}]$ is lower bounded by $\omega > 0$, which is independent of T.

We define the constant ξ as a constant such that $\xi \leq^{\triangle} |\mathcal{T}|/T$, which will be determined in part by the constant U, and can be set such that ξ is very close to 1.

For $\xi \approx 1$, Assumption 1 means most of the observed data is bounded independent of T. The assumption allows us to analyze time series in which the maximum of a series of iid random variables can grow with T, but any percentile is bounded by a constant. Our analysis will then be conducted on the bounded series $\{X_{\mathcal{T}_t}\}_{t=1}^{|\mathcal{T}|}$. The assumptions are proven to be true with high probability for the Bernoulli and Poisson cases in Sections 3.1 and 3.2, respectively, and the corresponding values of U, σ, ξ, and ω are computed explicitly.

Theorem 1. *Assume* $\lambda \geq \max_{1\leq m\leq M} \frac{2}{T}\left\|\sum_{t=0}^{T-1}\left(\phi(X_{t+1,m}) - \mathbb{E}[\phi(X_{t+1,m})|X_t]\right)X_t\right\|_\infty$*, and let* $\widehat{A}$ *be the RMLE for a process which obeys the conditions of Assumption 1. For any row of the estimator and for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\|\widehat{a}_m - a_m^*\|_2^2 \leq \frac{144}{\xi^2\sigma^2\omega^2}\rho_m\lambda^2$$

for $T \geq \frac{c\rho_m^2}{\omega^2}\left(\frac{\rho_m \log(2M)}{\omega^2} + \log(1/\delta)\right)$ *where* c *is independent of* M, T, ρ *and* s*. Furthermore,*

$$\|\widehat{A} - A^*\|_F^2 \leq \frac{144}{\xi^2\sigma^2\omega^2}s\lambda^2$$

with probability greater than $1-\delta$ *for* $T \geq \frac{c\rho^2}{\omega^2}\left(\left(\frac{\rho}{\omega^2}+1\right)\log(2M) + \log(1/\delta)\right)$.

To apply Theorem 1 to specific GLAR models, we need to provide bounds on λ, as well as σ, ω, U and ξ for Assumption 1. We do this in the next section for Bernoulli and Poisson GLAR models.

We can compare the results of Theorem 1 to the related results of [58]. In that work they arrive at rates for the Gaussian autoregressive process that are equivalent with respect to the sparsity parameter, number of observations and regularization parameter. However, we incur slightly different dependencies on ξ, σ and ω. These are due mainly to the fact that our bounds hold for a wide family of distributions and not just the Gaussian case, which has nice properties related to restricted strong convexity and specialized concentration inequalities. Additionally, the way λ is defined is very similar, but bounding λ for a non-Gaussian distribution will result in extra log factors. It is an open question whether this bound is rate optimal in the general setting.

3.1 Example 1: Bernoulli Distribution

For the Bernoulli distribution we have the following autoregressive model:

$$X_{t+1,m}|X_t \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\nu - a_m^{*\top}X_t)}\right). \tag{5}$$

The first observation about this model is that the sufficient statistic $\phi(x) = x$ and the log-partition function $Z(\theta) = \log(1+\exp(\theta))$, which is strongly convex when the absolute value of θ is bounded. One advantage of this model is that the observations are inherently bounded due to the nature of the Bernoulli distribution, so $\mathcal{T} = [0, 1, \ldots, T-1]$ and $\xi = 1$. Using this observation we derive the strong convexity parameter of Z on the bounded range, thus $\sigma = (3+\exp(\tilde{\nu}+9\rho\tilde{a}))^{-1}$.

To derive rates from Theorem 1, we must prove that Assumption 1 holds; this is shown with high-probability by Theorem 2.

Theorem 2. *For a sequence* X_t *generated from the Bernoulli autoregressive process with the matrix* A^* *with and the vector* ν*, we have the following properties:*

1. *The smallest eigenvalue of the matrix* $\Gamma_t = \mathbb{E}[X_t X_t^\top | X_{t-1}]$ *is lower bounded by* $\omega = (3+\exp(\tilde{\nu}+\rho\tilde{a}))^{-1}$.
2. *Assuming* $1 \leq t \leq T$ *and that* $T \geq 2$ *and* $\log(MT) \geq 1$*, then*

$$\max_{1\leq i,j\leq M}\frac{1}{T}\left|\sum_{t=0}^{T-1}X_{t-1,i}(X_{t,j} - \mathbb{E}[X_{t,j}|X_{t-1}])\right| \leq \frac{3\log(MT)}{\sqrt{T}}$$

with probability at least at least $1 - \frac{1}{MT}$.

Using these results we get the final sample error bounds for the Bernoulli autoregressive process.

Corollary 1. *The RMLE for the Bernoulli autoregressive process defined by Equation 5, and setting* $\lambda = \frac{6\log(MT)}{\sqrt{T}}$ *has error bounded by*

$$\|A^* - \widehat{A}\|_F^2 \leq C\left(3 + e^{\tilde{\nu}+9\rho\tilde{a}}\right)^4 \frac{s\log^2(MT)}{\xi^2 T}$$

with probability at least $1-\delta$ *for* $T \geq \max\left(\frac{2}{\delta M}, \frac{c\rho^2}{\omega^2}\left(\left(1+\frac{\rho}{\omega^2}\right)\log(2M) + \log(2/\delta)\right)\right)$ *for constants* $C, c > 0$ *which are independent of* M, T, s *and* ρ.

The lower bound on the number of observations T comes from needing to satisfy the conditions of both parts of Theorems 1 and 2. In order to get this statement we use a union bound over the high probability statements of Theorem 1 described in (9) and Theorem 2 which holds with probability greater than $1 - \frac{1}{MT}$.

3.2 Example 2: Poisson Distribution

In this section, we derive the relevant values to get error bounds for the vector autoregressive Poisson distribution. Under this model we have

$$X_{t+1,m}|X_t \sim \text{Poiss}(\exp(\nu + a_m^{*\top} X_t)).$$

We assume that $a_{\max} = 0$ for stability purposes, thus we are only modeling inhibitory relationships in the network. Deriving the sufficient statistic and log-partition function yields $\phi(x) = x$ and $Z(\theta) = \exp(\theta)$. The next important values are the bounds on the magnitude of the observations, which will both ensure the strong convexity of Z and the stability of the process.

Lemma 1. *For the Poisson autoregressive process generated with* $A^* \in [a_{\min}, 0]^{M\times M}$ *and constant vector* $\nu \in [\nu_{\min}, \nu_{\max}]$*:*

1. *If* $\log MT \geq 1$*, there exists constants* C *and* c *which depend on the value* $\nu_{\max}$*, but are independent of* T, M, s *and* ρ *such that* $0 \leq X_{t,m} \leq C\log(MT)$ *with probability at least* $1 - e^{-c\log(MT)}$ *for all* $1 \leq t \leq T$ *and* $1 \leq m \leq M$.

2. *For any* $\alpha \in (0,1)$ *such that* αMT *is an integer, there exist constants* U *and* c *which depend on the values of* $\nu_{\max}$ *and* α*, but independent of* T, M, s *and* ρ*, such that with probability at least* $1 - e^{-cMT}$*,* $0 \leq X_{t,m} \leq U$ *for at least* αMT *of the indices. We define* $\mathcal{T}$ *to be these* αMT *indices.*

As a consequence of Lemma 1, we have $\|X_t\|_\infty \leq U$ for at least ξT values of $t \in \{1, 2, \ldots, T\}$ where $\xi = 1 - (1-\alpha)M$. We additionally assume that U is large enough such that $\alpha > \frac{M-1}{M}$ and therefore $\xi \in (0,1)$.

Using this Lemma,we prove that Assumption 1 holds with high-probability, by deriving the strong convexity parameter of Z and a lower bound on the smallest eigenvalue of Γ_t. In the Poisson case, $Z(\cdot) = \exp(\cdot)$ and therefore the strong convexity parameter, $\sigma = \exp(-\tilde{\nu} + 9\rho a_{\min} U)$.

Theorem 3. *For a sequence* X_t *generated from the Poisson autoregressive process with the matrix* A^**, with all non-positive elements, and the vector* ν*, we have the following properties*

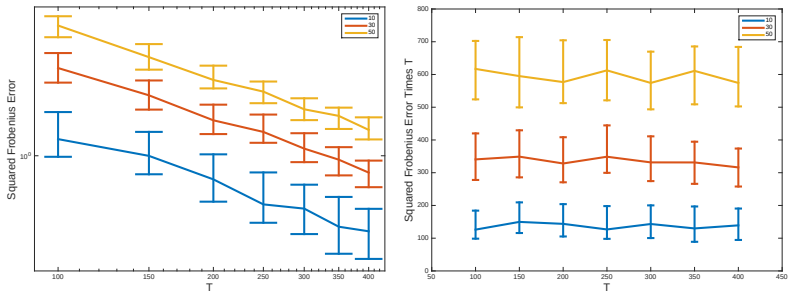
1. *The smallest eigenvalue of the matrix* $\Gamma_t = \mathbb{E}[X_{\mathcal{T}_t} X_{\mathcal{T}_t}^\top | X_{\mathcal{T}_{t-1}}]$*, for consecutive indices* $\mathcal{T}_t$ *and* $\mathcal{T}_{t-1}$ *in* $\mathcal{T}$ *as defined in Assumption 1, is lower bounded by* $\frac{4\xi}{5}\exp(\nu_{\min} + \rho a_{\min} U)$.

2. *Assuming* $X_{t,m} \leq C\log(MT)$ *for all* $1 \leq m \leq M$ *and* $1 \leq t \leq T$ *and that* $T \geq 2$ *and* $\log(MT) \geq 1$*, then*

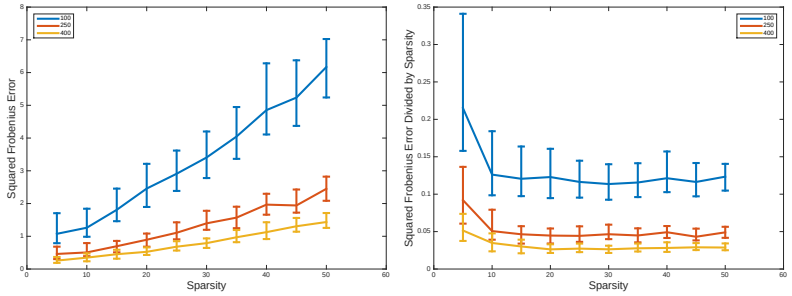
$$\max_{1\leq i,j\leq M} \frac{1}{T}\left|\sum_{t=0}^{T-1} X_{t-1,i}(X_{t,j} - \mathbb{E}[X_{t,j}|X_{t-1}])\right| \leq 4C^2 e^{\nu_{\max}} \frac{\log^3(MT)}{\sqrt{T}}$$

with probability at least at least $1 - \exp(-c\log(MT))$ *for some* $c > 0$ *independent of* ρ, s, M *and* T.

Using Theorem 3, we can find the error bounds for the PAR process by using the result of Theorem 1.



(a) MSE vs. T (b) MSE $\cdot T$ vs T



(c) MSE vs s (d) MSE/s vs s

Figure 1: The top row of plots shows the MSE behavior over a range of T values, from 100 to 400 all less than or equal to $M^2 = 400$, where (a) is the MSE and (b) is the MSE multiplied by T to show that the MSE is behaving as $1/T$. The bottom row shows the MSE behavior over a range of s values, where (c) shows MSE and (d) shows MSE divided by s to show that the MSE is linear is s. In all plots the median value of 100 trials is shown, with error bars denoting the middle 50 percentile.

Corollary 2. *Using the results of Theorem 1 and using the Poisson autoregressive model with A^* with all non-positive values, the RMLE admits the overall error rate of*

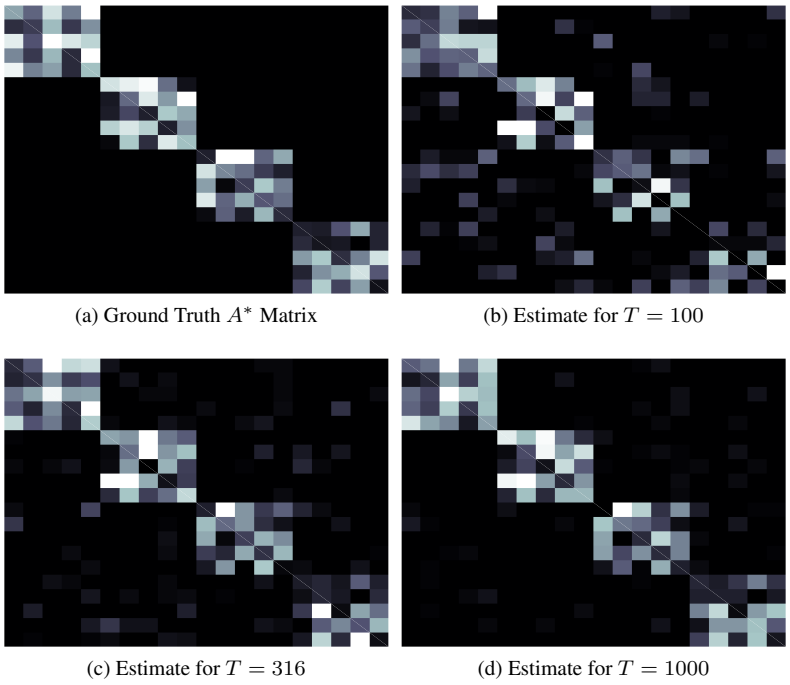
$$\|\hat{A} - A^*\|_F^2 \leq C \exp(20|a_{\min}|U\rho) \frac{s \log^6(MT)}{\xi^3 T}$$

with probability at least $1 - \delta$ for $T \geq \max\left(\left(\frac{4}{\delta M}\right)^c, \frac{c\rho^2}{\omega^2}\left(\left(\frac{\rho}{\omega^2} + 1\right)\log(2M) + \log(4/\delta)\right)\right)$ for constants $C, c > 0$ which are independent of M, T, s and ρ

Again, the lower bound on the number of observations comes from combining the high probability statements of each of the constituent parts of the corollary in the same way as was done in the Bernoulli case. In this case all of Theorem 1, both parts of Lemma 1 and Theorem 3 need to hold.

3.3 Experimental Results

We validate our theoretical results with experimental results performed on synthetically generated data using the Poisson autoregressive process. We generate many trials of synthetic data with known underlying parameters and then compare the estimated values. For all trials the constant offset vector ν is set identically at 0, and the 20×20 matrices A^* are set such that s randomly assigned values are in the range $[-1, 0]$ and with constant $\rho = 5$. Data is then generated according the process described in Equation 1 with the Poisson distribution. X_0 is chosen as a 20 dimensional vector drawn randomly from $\text{Poisson}(1)$, then T observations are used to perform the estimation. The parameters s and T are then varied over a wide range of values. For each (s, T) pair 100 trials are performed, the



(a) Ground Truth A^* Matrix (b) Estimate for $T = 100$

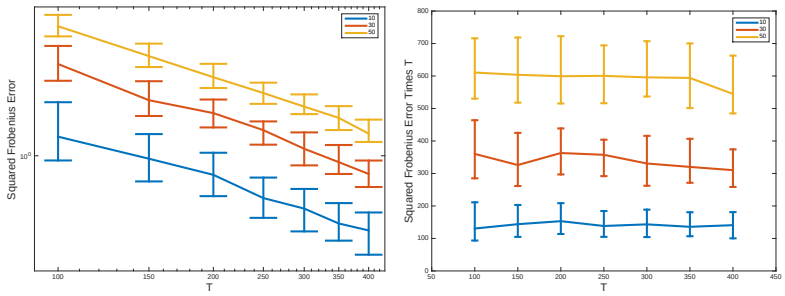
(c) Estimate for $T = 316$ (d) Estimate for $T = 1000$

Figure 2: These images show the ground truth A^* matrix (a) and 3 different estimates of the matrix created using increasing amounts of data. We observe that even for a relatively low amount of data we have picked out most of the support but with several spurious artifacts. As the amount of data increases, fewer of the erroneous elements are estimated. All images are scaled from 0 (dark) to -1 (bright).

regularized maximum likelihood estimate $\hat{A}$ is calculated with $\lambda = 0.1/\sqrt{T}$ and the MSE is recorded. The MSE curves are shown in Figure 1. Notice that the true values of A^* are bounded by -1 and 0, but in our implementation we do not enforce these bounds (we set $a_{\min} = -\infty$ and $a_{\max} = \infty$ in Equation 4). While $a_{\min} = -\infty$ would cause the theoretical bounds to be poor, the theory can be applied with the smallest and largest elements of the matrix estimated from the unconstrained optimization. In other words, the theory depends on having an upper and lower bound on the rates, but mostly as a theoretical convenience, while the estimator can be computed in an unconstrained way.

We show a series of plots which compare the MSE versus increasing behavior of T and s, as well as comparing the behavior of MSE$\cdot T$ and of MSE/s. Plotted in each figure is the median of 100 trials for each (s, T) pair, with error bars denoting the middle 50 percentile. These plots show that setting λ proportional to T^{-1} gives us the desired T^{-1} error decay rate. Additionally, we see that the error increases approximately linearly in the sparsity level s, as predicted by the theory. Finally, Figure 2 shows one specific example process and the estimates produced. The first image is the ground truth matrix, generated to be block diagonal, in order to more easily visualize support structure whereas in the first experiment the support is chosen at random. One set of data is generated using this matrix, and then estimates are constructed using the first $T = 100, 316$ and 1000 data points. The figure shows how with more data, the estimates become closer to the original, where much of the error comes from including elements off the support of the true matrix.

One important characteristic of the our results is that it does not depend on any assumptions about the stationarity or the mixing time of the process. To show that this is truly a property of the system and not just our proof technique, we repeat the experimental process described above, but for each set of observations of length T, we first generate 10,000 observations to allow the process to mix. In other words, for every matrix A we generate $T + 10,000$ observations, but only use the last T to find the RMLE. The plots in figure 3 show the results of this experiment. The important observation is that the results both scale the same way, and have approximately the same magnitude as the experiment when no mixing was done.



(a) MSE vs. T (b) MSE $\cdot T$ vs T

(c) MSE vs s (d) MSE/s vs s

Figure 3: Repeat of experimental set up from Figure 1, but now allowing for mixing. The top row of plots shows the MSE behavior over a widely varying range of T values, from 100 to 400, where (a) is the MSE and (b) is the MSE multiplied by T to show that the MSE is behaving as $1/T$. The bottom row shows the MSE behavior over a range of s values, where (c) shows MSE and (d) shows MSE divided by s to show that the MSE is linear is s. In all plots the median value of 100 trials is shown, with error bars denoting the middle 50 percentile. Most importantly, the behavior and magnitude of errors in this plot matches the results with no mixing.

4 Proofs

4.1 Proof of Theorem 1

Proof. We start the proof by making an important observation about the estimator defined in Equation 4: this loss function can be completely decoupled by a sum of functions on rows. Therefore we can bound the error of a single row of the RMLE and add the errors to get the final bound. For each row we use a standard method in empirical risk minimization and the definition of the minimizer of the regularized likelihood for each row:

$$\frac{1}{T}\sum_{t=0}^{T-1} Z(\nu_m + \widehat{a}_m^\top X_t) - \widehat{a}_m^\top X_t \phi(X_{t+1,m}) + \lambda\|\widehat{a}_m\|_1$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1} Z(\nu_m + a_m^{*\top} X_t) + a_m^{*\top} X_t \phi(X_{t+1,m}) + \lambda\|a_m^*\|_1.$$

We define $\epsilon_{t,m} \triangleq \phi(X_{t+1,m}) - \mathbb{E}[\phi(X_{t+1,m})|X_t]$, which is conditionally zero mean random variable. By using a moment generating function argument, we know that $\mathbb{E}[\phi(X_{t+1,m})|X_t] = Z'(\nu_m + a_m^{*\top} X_t)$, and therefore

$\phi(X_{t+1,m}) = Z'(\nu_m + a_m^{*\top} X_t) + \epsilon_{t,m}$. Hence

$$\frac{1}{T}\sum_{t=0}^{T-1} Z(\nu_m + \widehat{a}_m^\top X_t) - \widehat{a}_m^\top X_t (Z'(\nu_m + a_m^{*\top} X_t) + \epsilon_{t,m}) + \lambda \|\widehat{a}_m\|_1$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1} Z(\nu_m + a_m^{*\top} X_t) - a_m^{*\top} X_t (Z'(\nu_m + a_m^{*\top} X_t) + \epsilon_{t,m}) + \lambda \|a_m^*\|_1.$$

Now we use the definition of a Bregman divergence to lower bound the left hand side. An important property of Bregman divergences is that if they are induced by a strongly convex function, then the Bregman can be lower bounded by a scaled ℓ_2 difference of its arguments. This is where our squared error term will come.

$$\frac{1}{T}\sum_{t=0}^{T-1} \big(Z(\nu_m + \widehat{a}_m^\top X_t) - Z(\nu_m + a_m^{*\top} X_t) - Z'(\nu_m + a_m^{*\top} X_t)(\widehat{a}_m^\top X_t - a_m^{*\top} X_t)\big)$$
$$\leq \left| \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m} \Delta_m^\top X_t \right| + \lambda(\|a_m^*\|_1 - \|\widehat{a}_m\|_1),$$

where $\Delta_m = \widehat{a}_m - a_m^*$. Let $B_Z(\cdot\|\cdot)$ denote the Bregman divergence induced by Z. Hence

$$\frac{1}{T}\sum_{t=0}^{T-1} B_Z(\nu_m + \widehat{a}_m^\top X_t \| \nu_m + a_m^{*\top} X_t) \quad \leq \quad \left| \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m} \Delta_m^\top X_t \right| + \lambda(\|a_m^*\|_1 - \|\widehat{a}_m\|_1).$$

First we upper bound the right-hand side of the inequality as follows:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} B_Z(\nu_m + \widehat{a}_m^\top X_t \| \nu_m + a_m^{*\top} X_t) &\leq \left| \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m} \Delta_m^\top X_t \right| + \lambda(\|a_m^*\|_1 - \|\widehat{a}_m\|_1) \\
&= \left| \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m} \Delta_m^\top X_t \right| + \lambda(\|a_{m,\mathcal{S}}^*\|_1 - \|\widehat{a}_{m,\mathcal{S}}\|_1 - \|\widehat{a}_{m,\mathcal{S}^c}\|_1) \\
&\leq \left| \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m} \Delta_m^\top X_t \right| + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1 \\
&\leq \|\Delta_m\|_1 \left\| \frac{1}{T}\sum_{t=0}^{T-1} X_t \epsilon_{t,m} \right\|_\infty + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1.
\end{aligned}$$

In the above, we use the defintion of $\mathcal{S}$ as the true support of A^* and have used the decomposability of $\|\cdot\|_1$. The decomposability of the norm means that we have the property

$$\|x\|_1 = \|x_{\mathcal{S}}\|_1 + \|x_{\mathcal{S}^C}\|_1.$$

Note that $\left\| \frac{1}{T}\sum_{t=0}^{T-1} X_t \epsilon_{t,m} \right\|_\infty \leq \max_{1\leq m\leq M} \left\| \frac{1}{T}\sum_{t=0}^{T-1} X_t \epsilon_{t,m} \right\|_\infty$. Under the assumption that $\max_{1\leq m\leq M} \left\| \frac{1}{T}\sum_{t=0}^{T-1} X_t \epsilon_{t,m} \right\|_\infty \leq \lambda/2$ and by the non-negativity of the Bregman divergence on the left hand side of the inequality, we have that

$$0 \leq \frac{\lambda}{2}\|\Delta_m\|_1 + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1.$$

Using the decomposability of the ℓ_1 norm, this inequality implies that for all rows $1 \leq m \leq M$, we have that $\|\Delta_{m,\mathcal{S}^c}\|_1 \leq 3\|\Delta_{m,\mathcal{S}}\|_1$. Since $\|\Delta_{m,\mathcal{S}^c}\|_1 \leq 3\|\Delta_{m,\mathcal{S}}\|_1$, $\|\Delta_m\|_1 \leq 4\|\Delta_{m,\mathcal{S}}\|_1$ and consequently

$$\|\Delta_m\|_1 \leq 4\sum_{j\in\mathcal{S}} |\Delta_{m,j}| \leq 8\rho_m \tilde{a}$$

where the final inequality follows since $|\Delta_{m,j}| \leq 2\tilde{a}$ for all j. Using this inequality and the fact that $\|a_m^*\|_1 \leq \rho_m \tilde{a}$ implies that $\|\hat{a}_m\|_1 \leq 9\rho_m\tilde{a}$, and therefore for all $t \in \mathcal{T}$ the range of both $\nu_m + a_m^{*\top} X_t$ and $\nu_m + \hat{a}_m^\top X_t$ are in $[-\tilde{\nu} - 9\rho\tilde{a}, \tilde{\nu} + 9\rho\tilde{a}]$.

Now to lower bound the Bregman divergence in terms of the Frobenius norm, we use the first condition of Assumption 1. Inherently, the RMLE will admit estimates which should converge to the true matrix A^* under a Bregman divergence induced by the log-partition function, but we are interested in convergence of the Frobenius norm. Therefore, to convert from one to the other, we require the log-partition function to be strongly convex. This issue is side-stepped in the Gaussian noise case, due to the fact that the Bregman in question would identically be the Frobenius norm. By Assumption 1, Z is σ-strongly convex, and therefore on $\mathcal{T}$ it is true that $B_Z(\nu + \widehat{a}_m^\top X_t \| \nu_m + a_m^{*\top} X_t) \geq \frac{\sigma}{2}(\Delta_m^\top X_t)^2$ and $B_Z(\nu_m + \widehat{a}_m^\top X_t \| \nu_m + a_m^{*\top} X_t) \geq 0$ on the rest of the time indices.

Therefore

$$\frac{1}{T}\sum_{t=0}^{T-1} B_Z(\nu_m + \widehat{a}_m^\top X_t \| \nu_m + a_m^{*\top} X_t) \quad \leq \quad \frac{\lambda}{2}\|\Delta_m\|_1 + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1,$$

implies

$$\frac{\sigma}{2T}\sum_{t\in\mathcal{T}} (\Delta_m^\top X_t)^2 \quad \leq \quad \frac{\lambda}{2}\|\Delta_m\|_1 + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1.$$

Define $\|\Delta_m\|_T^2 = \frac{1}{T}\sum_{t\in\mathcal{T}} (\Delta_m^\top X_t)^2$ for any $\Delta \in \mathbb{R}^{M\times M}$, then we have the bound:

$$\frac{\sigma}{2}\|\Delta_m\|_T^2 \quad \leq \quad \frac{\lambda}{2}\|\Delta_m\|_1 + \lambda\|\Delta_{m,\mathcal{S}}\|_1 - \lambda\|\Delta_{m,\mathcal{S}^c}\|_1 \leq \frac{3\lambda}{2}\|\Delta_{m,\mathcal{S}}\|_1.$$

Therefore we can define the cone on which the vector Δ_m must be defined:

$$\mathcal{B}_{m,\mathcal{S}} := \{\Delta \in [a_{\min} - a_{\max}, a_{\max} - a_{\min}]^M \mid \|\Delta_{m,\mathcal{S}^c}\|_1 \leq 3\|\Delta_{m,\mathcal{S}}\|_1\},$$

and restrict ourselves to studying properties of vectors in that set. Since $\|\Delta_{m,\mathcal{S}}\|_1 \leq \sqrt{\rho_m}\|\Delta_m\|_2$ where ρ_m is the number of non-zeros of a_m^*, we have that

$$\|\Delta_m\|_T^2 \quad \leq \quad \frac{3}{\sigma}\lambda\sqrt{\rho_m}\|\Delta_m\|_2 = \delta_m\|\Delta_m\|_2, \tag{6}$$

where $\delta_m \triangleq \frac{3}{\sigma}\lambda\sqrt{\rho_m}$. Now we consider three cases: if $\|\Delta_m\|_T \geq \|\Delta_m\|_2$, then $\max(\|\Delta_m\|_T, \|\Delta_m\|_2) \leq \delta_m$. On the other hand if $\|\Delta_m\|_T \leq \|\Delta_m\|_2$ and $\|\Delta_m\|_2 \leq \delta_m$, then $\max(\|\Delta_m\|_T, \|\Delta_m\|_2) \leq \delta_m$.

Hence the final case we need to consider is $\|\Delta_m\|_T \leq \|\Delta_m\|_2$ and $\|\Delta_m\|_2 \geq \delta_m$. Now we follow a similar proof technique to that used in Raskutti et al. [55] adapted to dependent sequences, to understand this final scenario. Let us define the following set:

$$\mathcal{B}_m(\delta_m) := \{\Delta_m \in \mathcal{B}_{m,\mathcal{S}} \mid \|\Delta_m\|_2 \geq \delta_m\}. \tag{7}$$

Further, let us define the alternative set:

$$\mathcal{B}'_m(\delta_m) := \{\Delta_m \in \mathcal{B}_{m,\mathcal{S}} \mid \|\Delta_m\|_2 = \delta_m\}. \tag{8}$$

We wish to show that for $\Delta_m \in \mathcal{B}_m(\delta_m)$, we have $\|\Delta_m\|_T^2 \geq \kappa\|\Delta_m\|_2^2$ for some $\kappa \in (0,1)$ with high probability, and therefore Equation 6 would imply that $\max(\|\Delta_m\|_T, \|\Delta_m\|_2) \leq \delta_m/\kappa$. We claim that it suffices to show that $\|\Delta_m\|_T^2 \geq \kappa\|\Delta_m\|_2^2$ is true on $\mathcal{B}'_m(\delta_m)$ with high probability. In particular, given an arbitrary non-zero $\Delta_m \in \mathcal{B}_m(\delta_m)$, consider the re-scaled vector $\tilde{\Delta}_m = \frac{\delta_m}{\|\Delta_m\|_2}\Delta_m$. Since $\Delta_m \in \mathcal{B}_m(\delta_m)$, we have $\tilde{\Delta}_m \in \mathcal{B}_m(\delta_m)$ and $\|\tilde{\Delta}_m\|_2 = \delta_m$ by construction. Together, these facts imply $\tilde{\Delta}_m \in \mathcal{B}'_m(\delta_m)$. Furthermore, if $\|\tilde{\Delta}_m\|_T^2 \geq \kappa\|\tilde{\Delta}_m\|_2^2$ is true, then $\|\Delta_m\|_T^2 \geq \kappa\|\Delta_m\|_2^2$ is also true. Alternatively if we define the random variable $\mathcal{Z}_T(\mathcal{B}'_m) = \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{\delta_m^2 - \|\Delta_m\|_T^2\}$, then it suffices to show that $\mathcal{Z}_T(\mathcal{B}'_m) \leq (1-\kappa)\delta_m^2$.

For this step we use some recent concentration bounds [61] and empirical process techniques [62] for martingale random variables. Recall that the empirical norm is $\|\Delta_m\|_T^2 = \frac{1}{T}\sum_{t\in\mathcal{T}}(\Delta_m^T X_t)^2$. Further let $(t_i)_{i=1}^{|\mathcal{T}|}$ denote the indices in $\mathcal{T}$. Next we define the conditional expectation

$$Y_T := \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\mathbb{E}\big[(\Delta_m^T X_{t_i})^2 | X_{t_1}, X_{t_2}, \ldots, X_{t_i-1}\big].$$

Then we have

$$\mathcal{Z}_T(\mathcal{B}'_m) = \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{\delta_m^2 - \|\Delta_m\|_T^2\} \leq \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{\delta_m^2 - Y_T\} + \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}.$$

To bound the first quantity, $\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{\delta_m^2 - Y_T\}$, we first note that

$$\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{\delta_m^2 - Y_T\} \leq \delta_m^2 - \delta_m^2\omega = (1-\omega)\delta_m^2$$

by Assumption 1 and the fact that $\|\Delta_m\|_2^2 = \delta_m^2$ since $\Delta_m \in \mathcal{B}'_m(\delta_m)$. Thus

$$\mathcal{Z}_T(\mathcal{B}'_m) \leq (1-\omega)\delta_m^2 + \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}.$$

Now we focus on bounding $\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}$. First, we use a martingale version of the bounded difference inequality using Theorem 2.6 in [61] (see Appendix 7.4):

$$\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\} \leq \mathbb{E}[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}] + \frac{\omega\delta_m^2}{4},$$

with high probability. Recall that on $\mathcal{T}$, we have $0 \leq (\Delta_m^\top X_t)^2 \leq \|\Delta_m\|_1^2\|X_t\|_\infty^2 \leq U^2\|\Delta_m\|_1^2$. Because $\Delta_m \in \mathcal{B}'_m(\delta_T)$, it is true that $\|\Delta_m\|_1 \leq 4\|\Delta_{m,\mathcal{S}}\|_1$. We then use the the relationship between the ℓ_1 and ℓ_2 norms to say $\|\Delta_{m,\mathcal{S}}\|_1 \leq \sqrt{\rho_m}\|\Delta_{m,\mathcal{S}}\|_2 \leq \sqrt{\rho_m}\|\Delta_m\|_2$ where ρ_m is the number of non-zeros in the m^{th} row of the true matrix A^*. Putting these together means $(\Delta_m^\top X_t)^2 \leq 16U^2\rho_m\delta_m^2$. In particular, we apply Theorem 4 in Appendix 7.4 with $Z_T = \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}$, $a = \frac{\omega\delta_m^2}{4}$, $L_t = -\frac{16U^2\rho_m\delta_m^2}{T}$ and $U_t = \frac{16U^2\rho_m\delta_m^2}{T}$, and therefore $C_T^2 = \frac{32^4U^4\rho_m^2\delta_m^4}{T}$. Therefore, applying Theorem 4

$$\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\} \leq \mathbb{E}[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}] + \frac{\omega\delta_m^2}{4},$$

with probability at least $1-\exp(-\frac{2T}{32^4U^4\rho_m^2})$. Since $T \geq 32^4U^4\rho_m^2\log(M)$, the above statement holds with probability at least $1-\frac{1}{M^2}$. Hence

$$\mathcal{Z}_T(\mathcal{B}'_m) \leq (1-\omega)\delta_m^2 + \frac{\omega\delta_m^2}{4} + \mathbb{E}[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}].$$

Now we bound $\mathbb{E}[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}]$. Here we use a recent symmetrization technique adapted for martingales in [62]. To do this, we introduce the so-called *sequential Rademacher complexity* defined in [62]. Let $(\epsilon_t)_{t=1}^T$ be independent Rademacher random variables, that is $\mathbb{P}(\epsilon_t = +1) = \mathbb{P}(\epsilon_t = -1) = \frac{1}{2}$. For a function class $\mathcal{F}$, the sequential Rademacher complexity $\mathcal{R}_T(\mathcal{F})$ is:

$$\mathcal{R}_T(\mathcal{F}) := \sup_{X_1,X_2,\ldots,X_T}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T\epsilon_t f(X_t(\epsilon_1,\epsilon_2,\ldots,\epsilon_{t-1}))\right].$$

Note here that X_t is a function of the previous independent random variables $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})$. Using Theorem 2 in [62] (also stated Appendix 7.4) with $f(X_t) = (\Delta_m^T X_t)^2$ and noting that even though we use the index set $\mathcal{T}$, $(X_t)_{t\in\mathcal{T}}$ is still a martingale, it follows that:

$$\mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}\right] \leq 2 \sup_{X_{t_1},X_{t_2},\ldots,X_{|\mathcal{T}|}} \mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)} \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}(\Delta_m^T X_{t_i})^2\right].$$

Additionally since $|\Delta_m^\top X_t| \leq 4U\sqrt{\rho_m}\delta_m$ by the argument above and using the symmetry of Rademacher random variables

$$\begin{aligned}\mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}\right] &\leq 2 \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)} \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}\Delta_m^T X_{t_i}|\Delta_m^T X_{t_i}|\right] \\ &\leq 8U\sqrt{\rho_m}\delta_m \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)} \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}\Delta_m^T X_{t_i}\right]\end{aligned}$$

The final step is to upper bound the sequential Rademacher complexity $\mathcal{R}_T = \mathbb{E}[\sup_{\Delta\in\mathcal{B}'_m(\delta_m)} \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_t\Delta_m^\top X_{t_i}]$ where X_{t_i} is a function of $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t_i-1})$. Clearly:

$$\frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_t\Delta_m^\top X_{t_i} \leq \left\|\frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_t X_{t_i}\right\|_\infty \|\Delta_m\|_1.$$

Because $\Delta_m \in \mathcal{B}'_m(\delta_m)$ we have $\|\Delta_m\|_1 = \|\Delta_{m,\mathcal{S}}\|_1 + \|\Delta_{m,\mathcal{S}^c}\|_1 \leq 4\|\Delta_{m,\mathcal{S}}\|_1$ and $\|\Delta_{m,\mathcal{S}}\|_1 \leq \sqrt{\rho_m}\|\Delta_{m,\mathcal{S}}\|_2 \leq \sqrt{\rho_m}\|\Delta_m\|_2 = \sqrt{\rho_m}\delta_m$.

$$\begin{aligned}\mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}\right] &\leq 8U\sqrt{\rho_m}\delta_m \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)} \frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}\Delta_m^T X_{t_i}\right] \\ &\leq 8U\sqrt{\rho_m}\delta_m \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \left\|\frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}X_{t_i}(\epsilon_1,\ldots,\epsilon_{t_i-1})\right\|_\infty \sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\|\Delta_m\|_1 \\ &\leq 32U^2\rho_m\delta_m^2 \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \left\|\frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}X_{t_i}(\epsilon_1,\ldots,\epsilon_{t_i-1})\right\|_\infty.\end{aligned}$$

Finally, we use Lemma 6 applied to the index set $\mathcal{T}$:

$$\begin{aligned}\mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}\right] &\leq 32U^2\rho_m\delta_m^2 \sup_{X_1,X_2,\ldots,X_{|\mathcal{T}|}} \left\|\frac{1}{T}\sum_{i=1}^{|\mathcal{T}|}\epsilon_{t_i}X_{t_i}(\epsilon_1,\ldots,\epsilon_{t_i-1})\right\|_\infty \\ &\leq 128U^4\rho_m\delta_m^2\frac{\log(MT)}{\sqrt{T}},\end{aligned}$$

with probability at least $1 - \frac{1}{(MT)^2}$. Now if we set $T \geq \frac{256^2U^8\rho_m^2\log^2(MT)}{\omega^2}$,

$$\mathbb{E}\left[\sup_{\Delta_m\in\mathcal{B}'_m(\delta_m)}\{Y_T - \|\Delta_m\|_T^2\}\right] \leq \frac{\omega\delta_m^2}{4}$$

with probability $1 - (MT)^{-2}$.

Overall this tells us that on the set $\mathcal{B}'_m(\delta_m)$ we have that $\|\Delta_m\|_T^2 \geq \frac{3\omega}{4}\|\Delta_m\|_2^2$ with high probability. Now we return to the main proof. After considering all three cases that can follow from 6, we have

$$\max(\|\Delta_m\|_2^2, \|\Delta_m\|_T^2) \leq \frac{144}{\sigma^2\omega^2\xi^2}\rho_m\lambda^2$$

with probability at least $1 - \exp(\frac{c'\rho_m}{\omega^2}\log(2M) - \frac{c\omega^2 T}{\rho_m^2})$, which bounds the error accrued on any single row, as a function of the sparsity of the true row. Combining, to get an overall error yields,

$$\|\widehat{A} - A^*\|_F^2 \leq \frac{144}{\sigma^2\omega^2\xi^2}\lambda^2\sum_{m=1}^{M}\rho_m = \frac{144}{\sigma^2\omega^2\xi^2}\lambda^2 s$$

with probability at least

$$1 - \exp\left(\log(M) + \frac{c'\rho}{\omega^2}\log(2M) - \frac{c\omega^2 T}{\rho^2}\right) \tag{9}$$

. □

4.2 Proof of Theorem 2

4.2.1 Part 1

Proof. The matrix Γ_t can be expanded as

$$\mathbb{E}[X_t X_t^\top | X_{t-1}] = \mathbb{E}[X_t|X_{t-1}]\mathbb{E}[X_t|X_{t-1}]^\top + \mathrm{Diag}(\mathrm{Var}(X_t|X_{t-1}))$$

Thus Γ_t has two parts, one is the outer product of a vector with itself, and the second is a diagonal matrix. Therefore, the smallest eigenvalue will be lower bounded by the smallest element of the diagonal matrix, because the outer product matrix will always be positive semi-definite with smallest eigenvalue equal to 0. Using properties of the Bernoulli distribution, the conditional variance is explicitly given as $(2+\exp(\nu + A^* X_{t-1}) + \exp(-\nu - A^* X_{t-1}))^{-1}$ and therefore the smallest eigenvalue of Γ_t is lower bounded by $(3 + \exp(\tilde{\nu} + \rho\tilde{a}))^{-1}$. □

4.2.2 Part 2

Proof. In order to prove this part of the Theorem, we use of Markov's inequality and Lemma 5 in the case of the Bernoulli autoregressive process. Define the sequence $(Y_n, n \in \mathbb{N})$ as

$$Y_n \triangleq \frac{1}{T}\sum_{t=0}^{n-1} X_{t,m}(X_{t+1,\ell} - \mathbb{E}[X_{t+1,\ell}|X_t]).$$

Notice the following values:

$$\begin{aligned} Y_n - Y_{n-1} =& \frac{X_{n-1,m}}{T}\left(X_{n,\ell} - \mathbb{E}[X_{n,\ell}|X_{n-1}]\right) \\ M_n^k =& \sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{X_{i-1,m}}{T}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^k | X_1, \ldots, X_{i-1}\right]. \end{aligned}$$

The first value shows that $\mathbb{E}[Y_n - Y_{n-1}|X_1, \ldots, X_{n-1}] = 0$ and therefore Y_n (and the negative of the sequence, $-Y_n$) is a martingale. Additionally, we know $|Y_n - Y_{n-1}| \leq \frac{1}{T} \triangleq B$ and

$$\begin{aligned} M_n^2 =& \sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{X_{i-1,m}}{T}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^2 | X_1, \ldots, X_{i-1}\right] \\ =& \frac{1}{T^2}\sum_{i=1}^{n} X_{i-1,m}^2\mathbb{E}[(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])^2|X_{i-1}] \leq \frac{n}{4T^2} \triangleq \widehat{M}_n^2 \end{aligned}$$

where the last step follows because Bernoulli random variables are bounded by one, and the variance is bounded by $1/4$. We also need to bound M_n^k as follows:

$$\begin{aligned}
M_n^k =& \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^k |X_1, \ldots, X_{i-1}\right] \\
=& \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^2 \left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^{k-2} |X_{i-1}\right] \\
\leq& B^{k-2} M_n^2
\end{aligned}$$

We use these values to get a bound on the summation term used in Lemma 5.

$$\begin{aligned}
D_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!} M_n^k \leq \sum_{k\geq 2} \frac{\eta^k B^{k-2} M_n^2}{k!} \leq \frac{\widehat{M}_n^2}{B^2} \sum_{k\geq 2} \frac{(\eta B)^k}{k!} \triangleq \widehat{D}_n \\
\widetilde{D}_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!} (-1)^k M_n^k \leq \widehat{D}_n.
\end{aligned}$$

In the above $\widetilde{D}_n$ corresponds to the sum corresponding to the negative sequence $-Y_0, -Y_1, \ldots$ which we also need to obtain the desired bound. Now we use a variant of Markov's inequality to get a bound on the desired quantity.

$$\begin{aligned}
\mathbb{P}(|Y_n| \geq y) =& \mathbb{P}(Y_n \geq y) + \mathbb{P}(-Y_n \geq y) \leq \mathbb{E}[e^{\eta Y_n}] e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n)}] e^{-\eta y} \\
=& \mathbb{E}[e^{\eta Y_n - D_n + D_n}] e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n + \widetilde{D}_n}] e^{-\eta y} \\
\leq& \mathbb{E}[e^{\eta Y_n - D_n}] e^{\widehat{D}_n - \eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n}] e^{\widehat{D}_n - \eta y} \leq 2e^{\widehat{D}_n - \eta y}.
\end{aligned}$$

The final inequality comes from the use of Lemma 5, which states that the given terms are supermartingales with initial term equal to 1, so the entire expectation is less than or equal to 1. The final step of the proof is to find the optimal value of η to minimize this upper bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp(\widehat{D}_n - \eta y) = 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(e^{\eta B} - 1 - \eta B\right) - \eta y\right)$$

Setting $\eta = \frac{1}{B}\log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)$ yields the lowest such bound, giving

$$\begin{aligned}
\mathbb{P}(|Y_n| \geq y) \leq& 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(\frac{yB}{\widehat{M}_n^2} - \log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)\right) - \frac{y}{B}\log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)\right) \\
=& 2\exp\left(-\frac{\widehat{M}_n^2}{B^2} H\left(\frac{yB}{\widehat{M}_n^2}\right)\right)
\end{aligned}$$

where $H(x) = (1+x)\log(1+x) - x$. We use the fact that $H(x) \geq \frac{3x^2}{2(x+3)}$ for $x \geq 0$ to further simplify the bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp\left(\frac{-3y^2}{2yB + 6\widehat{M}_n^2}\right) = 2\exp\left(-\frac{6y^2T^2}{4yT + 3n}\right)$$

To complete the proof, we set $n = T$ and take a union bound over all indices because Y_T considered specific indices m and ℓ, which gives the bound

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le i,j\le M}\frac{1}{T}\left|\sum_{t=0}^{T-1}X_{t-1,i}(X_{t,j}-\mathbb{E}[X_{t,j}|X_{t-1}])\right|\ge 3\frac{\log(MT)}{\sqrt{T}}\Big)&\\ \le\exp\left(\log(2M^2)-\frac{54\log(MT)}{12/\sqrt{T}+3}\right)&\\ \le\frac{1}{MT}.&\end{aligned}$$

Here we have additionally assumed that $T \ge 2$ and that $\log(MT) \ge 1$. □

4.3 Proof of Theorem 3

4.3.1 Part 1

Proof. We start with the following observation:

$$\mathbb{E}[X_{\mathcal{T}_t}X_{\mathcal{T}_t}^\top|X_{\mathcal{T}_{t-1}}] = \mathbb{E}[X_{\mathcal{T}_t}|X_{\mathcal{T}_{t-1}}]\mathbb{E}[X_{\mathcal{T}_t}|X_{\mathcal{T}_{t-1}}]^\top + \mathrm{Diag}(\mathrm{Var}(X_{\mathcal{T}_t}|X_{\mathcal{T}_{t-1}}))$$

Thus Γ_t has two parts, one is the outer product of a vector with itself, and the second is a diagonal matrix. Therefore, the smallest eigenvalue will be lower bounded by the smallest element of the diagonal matrix. In order to lower bound this variance, we must consider the two cases, one where $\mathcal{T}_{t-1} = \mathcal{T}_t - 1$ where the previous term in the sequence $\mathcal{T}$ is the previous term in the overall sequence, and the other case where $\mathcal{T}_{t-1} < \mathcal{T}_t - 1$ where the previous term is not in the sequence $\mathcal{T}$. The variance of $X_{\mathcal{T}}$ can be characterized based on these two possible situations:

$$\mathrm{Var}(X_{\mathcal{T}_{t,i}}|X_{\mathcal{T}_{t-1}}) = p\mathrm{Var}(X_{\mathcal{T}_{t,i}}|X_{\mathcal{T}_{t-1}}, \mathcal{T}_{t-1} = \mathcal{T}_t - 1) + (1-p)\mathrm{Var}(X_{\mathcal{T}_{t,i}}|X_{\mathcal{T}_{t-1}}, \mathcal{T}_{t-1} < \mathcal{T}_t - 1)$$

where p is the probability that $\mathcal{T}_{t-1} = \mathcal{T}_t - 1$. Because variances are lower bounded by 0, we can lower bound this entire term by the first part of the sum, where $\mathcal{T}_{t-1} = \mathcal{T}_t - 1$. For this term, we know that $X_{\mathcal{T}_t}$ is drawn from a Poisson distribution, with the added information that each element is bounded above by U because it is an element of the sequence $X_{\mathcal{T}_1}, X_{\mathcal{T}_2}, \ldots$. Thus using Lemma 3 we know that the variance of each value is lower bounded by $\frac{4}{5}\exp(\nu_i + a_i^{*\top}X_{\mathcal{T}_{t-1}})$ which can in turn be lower bounded by $\exp(\nu_{\min} + \rho a_{\min}U)$. Finally, since there are at least ξT elements of $1, 2, \ldots, T$ which are in the bounded set of observations, then the worst case distribution of the observations with elements greater than U is that they are never consecutive. This maximizes the number of times there is a break in the sequence $\mathcal{T}_1, \mathcal{T}_2, \ldots$, which means there would be a total of $T - \xi T$ times when there was a break. Thus the probability that consecutive elements are in the set is at least ξ, meaning that the minimum eigenvalue of $\mathbb{E}[X_{\mathcal{T}_t}X_{\mathcal{T}_t}^\top|X_{\mathcal{T}_{t-1}}]$ is lower bounded by $\frac{4\xi}{5}\exp(\nu_{\min} + \rho a_{\min}U)$.

□

4.3.2 Part 2

Proof. To prove this part of the Theorem, we use of Markov's inequality and Lemma 5 as they pertain specifically to our problem. Define the sequence $(Y_n, n \in \mathbb{N})$ as

$$Y_n \triangleq \frac{1}{T}\sum_{t=0}^{n-1}X_{t,m}(X_{t+1,\ell} - \mathbb{E}[X_{t+1,\ell}|X_t]).$$

Notice the following values:

$$\begin{aligned}Y_n - Y_{n-1} =& \frac{X_{n-1,m}}{T}\left(X_{n,\ell} - \mathbb{E}[X_{n,\ell}|X_{n-1}]\right)\\ M_n^k =& \sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{X_{i-1,m}}{T}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^k|X_1, \ldots, X_{i-1}\right].\end{aligned}$$

The first value shows that $\mathbb{E}[Y_n - Y_{n-1}|X_1, \ldots, X_{n-1}] = 0$ and therefore Y_n (and the negative of the sequence, $-Y_n$) is a martingale. Additionally, we have assumed that $|X_{m,i}| \leq C \log MT$ for $1 \leq m \leq M$ and $1 \leq i \leq T$, so it is true that $|Y_n - Y_{n-1}| \leq \frac{C^2 \log^2(MT)}{T} \triangleq B$. Additionally:

$$\begin{aligned}
M_n^2 =& \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_{i-1,m}}{T}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^2 |X_1, \ldots, X_{i-1}\right] \\
=& \frac{1}{T^2}\sum_{i=1}^{n} X_{i-1,m}^2 \mathbb{E}[(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])^2|X_{i-1}] \\
=& \frac{1}{T^2}\sum_{i=1}^{n} X_{i-1,m}^2 \exp(\nu_\ell + a_\ell^{*\top} X_{i-1}) \leq \frac{nC^2 \log^2(MT) e^{\nu_{\max}}}{T^2} \triangleq \widehat{M}_n^2
\end{aligned}$$

where the last step follows because $X_{\ell,i}|X_{i-1} \sim \text{Poisson}(\exp(\nu_\ell + a_\ell^{*\top} X_{i-1}))$ and the mean and variance of a Poisson random variable are equal. The final line uses the fact that X_t is bounded. We will also need to bound M_n^k as follows:

$$\begin{aligned}
M_n^k =& \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^k |X_1, \ldots, X_{i-1}\right] \\
=& \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^2 \left(\frac{X_{i-1,m}}{T^2}(X_{i,\ell} - \mathbb{E}[X_{i,\ell}|X_{i-1}])\right)^{k-2} |X_{i-1}\right] \\
\leq& B^{k-2} M_n^2
\end{aligned}$$

We need to use these values to get a bound on the summation term used in Lemma 5.

$$\begin{aligned}
D_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!} M_n^k \leq \sum_{k\geq 2} \frac{\eta^k B^{k-2} M_n^2}{k!} \leq \frac{\widehat{M}_n^2}{B^2} \sum_{k\geq 2} \frac{(\eta B)^k}{k!} \triangleq \widehat{D}_n \\
\widetilde{D}_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!} (-1)^k M_n^k \leq \widehat{D}_n
\end{aligned}$$

In the above $\widetilde{D}_n$ corresponds to the sum corresponding to the negative sequence $-Y_0, -Y_1, \ldots$ which we will also need to obtain the desired bound. Now we are able to use a variant of Markov's inequality to get a bound on the desired quantity.

$$\begin{aligned}
\mathbb{P}(|Y_n| \geq y) =& \mathbb{P}(Y_n \geq y) + \mathbb{P}(-Y_n \geq y) \leq \mathbb{E}[e^{\eta Y_n}] e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n)}] e^{-\eta y} \\
=& \mathbb{E}[e^{\eta Y_n - D_n + D_n}] e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n + \widetilde{D}_n}] e^{-\eta y} \\
\leq& \mathbb{E}[e^{\eta Y_n - D_n}] e^{\widehat{D}_n - \eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n}] e^{\widehat{D}_n - \eta y} \leq 2e^{\widehat{D}_n - \eta y}
\end{aligned}$$

The final inequality comes from the use of Lemma 5, which states that the given terms are supermartingales with initial term equal to 1, so the entire expectation is less than or equal to 1. The final step of the proof is to find the optimal value of η to minimize this upper bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp(\widehat{D}_n - \eta y) = 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(e^{\eta B} - 1 - \eta B\right) - \eta y\right)$$

Setting $\eta = \frac{1}{B}\log\left(\frac{yB}{\widehat{M}_n^2}+1\right)$ yields the lowest such bound, giving

$$\begin{aligned}\mathbb{P}(|Y_n| \geq y) \leq & 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(\frac{yB}{\widehat{M}_n^2} - \log\left(\frac{yB}{\widehat{M}_n^2}+1\right)\right) - \frac{y}{B}\log\left(\frac{yB}{\widehat{M}_n^2}+1\right)\right) \\ = & 2\exp\left(-\frac{\widehat{M}_n^2}{B^2}H\left(\frac{yB}{\widehat{M}_n^2}\right)\right)\end{aligned}$$

where $H(x) = (1+x)\log(1+x) - x$. We can use the fact that $H(x) \geq \frac{3x^2}{2(x+3)}$ for $x \geq 0$ to further simplify the bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp\left(\frac{-3y^2}{2yB + 6\widehat{M}_n^2}\right) = 2\exp\left(-\frac{3y^2T^2}{2C^2(Ty + 3ne^{\nu_{\max}})\log^2(MT)}\right)$$

To prove the proof, we set $n = T$ and take a union bound over all indices because Y_T considered specific indices m and ℓ, which gives the bound

$$\begin{aligned}&\mathbb{P}\Big(\max_{i,j}\frac{1}{T}\left|\sum_{t=0}^{T-1} X_{t-1,i}(X_{t,j} - \mathbb{E}[X_{t,j}|X_{t-1}])\right| \geq 4C^2e^{\nu_{\max}}\frac{\log^3(MT)}{\sqrt{T}}\Big) \\ &\quad\leq \exp\left(\log(2M^2) - \frac{48C^4\exp^{2\nu_{\max}}\log^4(MT)}{8C^4e^{\nu_{\max}}\log^3(MT)/\sqrt{T} + 6C_1^2e_{\max}^{\nu}}\right) \\ &\quad\leq \exp\left(2\log(MT) - \frac{24C^2e_{\max}^{\nu}\log(MT)}{4C^2/\sqrt{T}+3}\right) \\ &\quad\leq \exp\left(-c\log(MT)\right)\end{aligned}$$

where $c = \frac{24C^2e^{\nu_{\max}} - 8C^2 - 6}{4C^2+3}$ which is positive for sufficiently large C. Here we have additionally assumed that $T \geq 2$ and that $\log(MT) \geq 1$. □

5 Discussion

Corollaries 1 and 2 provide several important facts about the inference process. Primarily, if ρ is fixed as a constant for increasing M (suggesting that the maximum degree of a node does not increase with the number of nodes in a network), then the error scales inversely with T, linearly with the sparsity level s and only logarithmically with the dimension M in order to estimate M^2 parameters. These parameters will dictate how much data needs to be collected to achieve a desired accuracy level. This rate illustrates the idea that doing inference in sparse settings can greatly reduce the needed amount of sensing time, especially when $s \ll M^2$. Another quantity to notice is that we require $T \geq \omega^{-4}\rho^3\log(M)$. If ρ is fixed as a constant for increasing M, this tells us that T needs to be on the order of $\log(M)$, which is significantly less than the total M^2 parameters which are being estimated, and therefore including the sparsity assumption has lead to a significant gain. One final observation from the risk bound is that it provides guidance in the setting of the regularization parameter. We see that we would like to set λ generally as small as possible, since the error scales approximately like λ^2, but we also require λ at least as large as $\tilde{O}(T^{-1/2})$ for the bounds to hold. The balance between setting λ small enough to have low error, while maintaining that it's large enough is an equivalent argument to needing to set λ large enough for it to take effect, but not too large to cause over smoothing.

5.1 Dense rows of A^*

The exponential scaling in Corollaries 1 and 2 with the maximum number of non-zeros in a row, ρ, at first seems unsatisfying. However, we can imagine a worst-case scenario where a large ρ relative to s and M would actually lead to very poor estimation. Consider the case of a large star-shaped network, where every node in the network influences and is influenced by a single node, and there are no other edges in the network. This would correspond to a matrix

with a single, dense row and corresponding column. Therefore, we would have $\rho = M$ and $s = 2M - 1$. In the Poisson setting, this network would have $M - 1$ independently and identically distributed Poisson random variables at every time with mean ν, but the central node of the network would be constantly inhibited, almost completely. In a large network, it would be very difficult to know if this inhibition was coming from a few strong connections or from the cumulative effect of all the inhibitions. Additionally, since the central node would almost never have a positive count, it would also be difficult to learn about the influence that node has on the rest of the network. Because of networks like this, it is important that not only is the overall network sparse, but each row also needs to be sparse. This requirement might seem restrictive, but it has been shown in many real world networks that the degree of a node in the network follows a power-law which is independent of the overall size of the network [63], and ρ would grow slowly with growing M.

5.2 Bounded observations and higher-order autoregressive processes

Recall that Assumption 1 ensures that most observations are bounded. Bounded observations are important to our analysis because we use martingale concentration inequalities [64] which depend on bounded conditional means and conditional variances, the latter condition being equivalent to Z being strongly convex. Since the conditional means and variances are data-dependent, bounded data (at least with high probability) is a sufficient condition for bounded conditional means and conditional variances. In some settings (*e.g.,* Bernoulli), bounded observations are natural and in Assumption 1 $\xi = 1$. In other settings (*e.g.,* Poisson) there is no constant U independent of T that is an upper bound for *all* observations with high probability. Furthermore, if we allow U to increase with T in violation of Assumption 1, we derive a bound on $\|\widehat{A} - A^*\|_F^2$ that *increases* polynomially with T. To avoid this and get the far better bound in Theorem 1, our proof focuses on characterizing the error on the set $\mathcal{T}$ defined in Assumption 1.

Thus far we have focused on the case where $X_{t+1,m}|X_t \sim p(\nu + a_m^{*\top} X_t)$, a first order autoregressive process. However, we could imagine a simple, higher-order version where $X_{t+1,m}|X_{t-q+1}, \ldots, X_t \sim p(\nu + a_m^{*\top} \sum_{i=0}^{q-1} \alpha_i X_{t-i})$ for some known sequence α_i. This process could be reformulated as a process $X_{t+1,m}|X_{t-q+1}, \ldots, X_t \sim p(\nu + a_m^{*\top} \tilde{X}_t)$ where $\tilde{X}_t \triangleq \sum_{i=0}^{q-1} \alpha_i X_{t-1}$, and much of the same proof techniques would still hold, especially in the case of the Bernoulli autoregressive process, where $\mathcal{T}$ is easily defined. However, in the more general GLAR case finding the right analogy to $\mathcal{T}$ in the higher space is not an obvious extension. A true order-q autoregressive process where $X_{t+1,m}|X_{t-q+1}, \ldots, X_t \sim p(\nu + \sum_{i=0}^{q-1} a_{m,i}^{*\top} X_{t-i})$ could also be formulated as an order-1 process by properly stacking vectors and matrices, however, in this case proving the key lemmas and showing that the process obeys Assumption 1 is also an open question.

5.3 Stationarity

As stated in the problem formulation, we restrict our attention to bounded matrices $A^* \in [a_{\min}, a_{\max}]^{M \times M}$; in the specific context of the log-linear Poisson autoregressive model, we use $a_{\max} = 0$, corresponding to a model that only accounts for inhibitory interactions. One might ask whether these constraints could be relaxed and whether the Poisson model could also account for stimulatory interactions.

These boundedness constraints are sufficient to ensure that the observed process has a stationary distribution. The stationarity of processes is heavily studied; once a process has reached its stationary distribution, then data can be approximated as independent samples from this distribution and temporal dependencies can be can be ignored. While stationarity does not play an explicit role in our analysis, we can identify several sufficient conditions to ensure the vector GLAR processes of interest are stationary. In particular we assume that $A^* = A^{*\top}$ which ensures reversibility of the Markov chain described by the process defined by $X_{t+1,m}|X_t \sim p(\nu_m + a_m^{*\top} X_t)$. We derive the stationary distribution $\pi(x)$, and then establish bounds on the mixing time. Note that this is a Markov chain with transition kernel:

$$P(x, y) = \mathbb{P}(X_{t+1} = y | X_t = x) = \exp\left(\nu^\top y + y^\top A^* x - \sum_{i=m}^{M} Z(\nu_m + a_m^{*\top} X)\right) \prod_{m=1}^{M} h(y_m).$$

If we further assume that the entries of X_t take on values on a countable domain to ensure a countable Markov chain, we can derive bounds on the mixing time.

Lemma 2. *Assume $A^* = A^{*\top}$, then the Markov chain $X_{t+1,m} \sim p(\nu_m + a_m^{*\top} X_t)$ is a reversible Markov chain with stationary distribution:*

$$\pi(x) = C_{\nu,A^*} \exp\left(\nu^\top x + \sum_{m=1}^{M} Z(\nu_m + a_m^{*\top} x)\right) \prod_{m=1}^{M} h(x_m)$$

for $C_{\nu,A^*} = \int_{x_1}\int_{x_2}\dots\int_{x_M} \exp\left(\nu^\top x + \sum_{m=1}^{M} Z(\nu_m + a_m^{*\top} x)\right) \prod_{m=1}^{M} h(x_m) dx_M \dots dx_2 dx_1$. Further, if $X_t \in \mathbb{Z}_+^M$, $a_{\max} = 0$ and $Z(\cdot)$ is an increasing function, then for any $y \in \mathbb{Z}_+^M$, if $\nu_m \le \nu_{\max} < \infty$ for all $1 \le m \le M$ and $a_{\min} \le 0$ we have that

$$\|P^t(y,.) - \pi(.)\|_{TV} \le \left(1 - h(0)^{-2M} e^{-2MZ(\nu_{\max})}\right)^t.$$

Notice that for large M, the chain will mix very slowly, and additionally this bound has no dependence on the sparsity of the true matrix A^*. Conversely, our results require T to be greater than a value that scales roughly like $\rho^3 \log(M)$, which has a much milder dependence on M, and varies based on the sparsity of A^*. What we can conclude from these observations is that while the RMLE needs a certain amount of observations to yield good results, we do not necessarily need enough data to reach the stationary distribution. Additionally, under conditions where mixing time guarantees are not given (i.e. non-symmetric A^*, uncountable domain), we still have guarantees on the performance of the RMLE.

6 Conclusions

Instances of the generalized linear autoregressive process has been used successfully in many settings to learn network structure. However, this model is often used without rigorous non-asymptotic guarantees of accuracy. In this paper we have shown important properties of the Regularized Maximum Likelihood Estimator of the GLAR process under a sparsity assumption. We have proven bounds on the error of the estimator as a function of sparsity, maximum degree of a node, ambient dimension and time, and shown how these bounds look for the specific examples of the Bernoulli and Poisson autoregressive proceses. In order to prove this risk bound, we have incorporated many recently developed tools of statistical learning, including concentration bounds for dependent random variables. Our results show that by incorporating sparsity the amount of data needed is on the order of $\rho^3 \log(M)$ for bounded degree networks, which is a significant gain compared to the M^2 parameters being estimated.

While this paper has focused on generalized linear models, we believe that the extension of these ideas to other models is possible. Specifically, for modeling firing rates of neurons in the brain, we are interested in settings in which we observe

$$X_{t+1,m} | X_t \sim \text{Poisson}(g(a_m^{*\top} X_t + \nu))$$

and exploring possible functions g beyond the exponential function considered here. Such analysis would allow our results to apply to stimulatory effects in addition to inhibitory effects, but key challenges include ensuring that the process is stable and, with high probability, bounded. Another direction would be settings where the counts are drawn from more complicated higher-order or autoregressive moving average (ARMA) models which would better model real-world point processes.

7 Appendix

7.1 Supplementary Lemmas

First we present supplementary Lemmas which we use throughout the proofs of the main Theorems.

Lemma 3. *Let X be a Poisson random variable, with the following probability density function:*

$$p(X = k|\lambda) = \frac{\lambda^k e^{-k}}{k!}$$

and let X' be a random variable defined by the following pdf:

$$q(k|\lambda) = \begin{cases} \frac{c}{k!}\lambda^k e^{-\lambda} \text{ if } k \leq U \\ 0 \text{ otherwise} \end{cases}$$

where $c = \frac{1}{1-\mathbb{P}(X>U)} > 1$. Roughly speaking, X' is generated by taking a Poisson pdf, and removing the tail probability, and scaling the remaining density so that it is a valid pdf. For this random variable, assuming $U \geq \max(6, 1.5e\lambda, \lambda + 5)$ then

$$\text{Var}(X') \geq \frac{4}{5}\text{Var}(X) = \frac{4\lambda}{5}$$

Proof. Define the error terms $\epsilon_1 \triangleq \mathbb{E}[X]^2 - \mathbb{E}[X']^2$ and $\epsilon_2 \triangleq \mathbb{E}[X^2] - \mathbb{E}[X'^2]$. We know

$$\begin{aligned} \text{Var}(X') &= \mathbb{E}[X'^2] - \mathbb{E}[X']^2 = (\mathbb{E}[X^2] - \epsilon_2) - (\mathbb{E}[X]^2 - \epsilon_1) \\ &\geq \underbrace{(\mathbb{E}[X^2] - \mathbb{E}[X]^2)}_{\text{Var}(X)} - (|\epsilon_1| + |\epsilon_2|) = \lambda - (|\epsilon_1| + |\epsilon_2|) \end{aligned} \tag{10}$$

Our strategy will be to show ϵ_1, ϵ_2 are small relative to λ, which will tell us $\text{Var}(X') \approx \text{Var}(X) = \lambda$. Intuitively, the error terms should be small relative to λ because X' differs from X only by cutting off the extreme edge of the pdf, given the assumptions on the size of U relative to λ.

First, we bound ϵ_1. We have

$$\epsilon_1 = \mathbb{E}[X]^2 - \mathbb{E}[X']^2 = (\mathbb{E}[X] + \mathbb{E}[X'])(\mathbb{E}[X] - \mathbb{E}[X'])$$

Since $\mathbb{E}[X'] \leq \mathbb{E}[X]$, the first term is bounded by $2\mathbb{E}[X] = 2\lambda$. To bound the second term, we note that the pdf for X' is given explicitly as

$$q(k|\lambda) = \begin{cases} \frac{c}{k!}\lambda^k e^{-\lambda} \text{ if } k \leq U \\ 0 \text{ otherwise} \end{cases}$$

where $c = \frac{1}{1-\mathbb{P}(X>U)} > 1$. And therefore

$$\mathbb{E}[X'] = c\sum_{k=1}^{U} \frac{\lambda^k e^{-\lambda}}{(k-1)!} \geq \sum_{k=1}^{U} \frac{\lambda^k e^{-\lambda}}{(k-1)!}$$

Using this fact to bound $\mathbb{E}[X] - \mathbb{E}[X']$ gives us

$$\mathbb{E}[X] - \mathbb{E}[X'] \leq \mathbb{E}[X] - \sum_{k=1}^{U} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \sum_{k=U+1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \frac{\lambda}{e^\lambda}\sum_{k=U}^{\infty} \frac{\lambda^k}{k!}$$

Note $\sum_{k=U}^{\infty} \frac{\lambda^k}{k!}$ is the remainder term of the degree $U-1$ Taylor Polynomial for e^λ. We can bound this using Taylor's Remainder theorem:

$$\sum_{k=U}^{\infty} \frac{\lambda^k}{k!} \leq \frac{e^\lambda \lambda^U}{U!}$$

and so

$$\mathbb{E}[X] - \mathbb{E}[X'] \leq \lambda\frac{\lambda^U}{U!} \leq \frac{\lambda}{1.5^U}\frac{(\frac{U}{e})^U}{U!}$$

where the second inequality comes from the assumption that $U \geq 1.5e\lambda$. Here, the second fraction is small by Sterling's approximation formula. Formally, Sterling tells us

$$\frac{(\frac{U}{e})^U}{U!} \leq \frac{1}{\sqrt{2\pi U}}$$

and therefore

$$\mathbb{E}[X] - \mathbb{E}[X'] \leq \frac{\lambda}{1.5^U\sqrt{2\pi U}}.$$

Combining the two terms tells us

$$|\epsilon_1| \leq 2\lambda\frac{\lambda}{1.5^U\sqrt{2\pi U}} \leq \frac{\lambda}{10}$$

since $U \geq 6$.

Next we bound $\epsilon_2 = \mathbb{E}[X^2] - \mathbb{E}[X'^2]$. We have

$$\mathbb{E}[X'^2] = c\sum_{k=1}^{U}\frac{k\lambda^k e^{-\lambda}}{(k-1)!} \geq \sum_{k=1}^{U}\frac{k\lambda^k e^{-\lambda}}{(k-1)!}$$

and therefore

$$\epsilon_2 \leq \mathbb{E}[X^2] - \sum_{k=1}^{U}\frac{k\lambda^k e^{-\lambda}}{(k-1)!} = \sum_{k=U+1}^{\infty}\frac{k\lambda^k e^{-\lambda}}{(k-1)!} \leq \frac{(U+1)\lambda^2}{Ue^{\lambda}}\sum_{k=U-1}^{\infty}\frac{\lambda^k}{k!}$$

where the last inequality is due to the fact that $\frac{k}{k-1} \leq \frac{U+1}{U}$ for all $k \geq U+1$. Here $\sum_{k=U-1}^{\infty}\frac{\lambda^k}{k!}$ is the remainder term for the degree $U-2$ Taylor Polynomial approximation to e^{λ}. By the Taylor's remainder formula, we can bound this by

$$\frac{e^{\lambda}\lambda^{U-1}}{(U-1)!}$$

and so

$$|\epsilon_1| \leq \lambda(U+1)\frac{\lambda^U}{U!}$$

and since $\lambda \leq \frac{U}{1.5e}$, it follows from Sterling's approximation that

$$|\epsilon_1| \leq \lambda\frac{U+1}{1.5^U\sqrt{2\pi U}} \leq \frac{\lambda}{10}$$

since $U \geq 6$.

Putting the bounds for ϵ_1 and ϵ_2 back into Equation 10 to get the final form of the Lemma

$$\mathrm{Var}(X') \geq \frac{4}{5}\mathrm{Var}(X) = \frac{4}{5}\lambda$$

□

We next present a one-sided concentration bound for Poisson random variables due to Bobkov and Ledoux [65]: random variables.

Lemma 4 (Proposition 10 in [65]). *If $X \sim$ Poisson(λ):*

$$\mathbb{P}\big(X - \lambda > t\big) \leq \exp(-\frac{t}{4}\log(1 + \frac{t}{2\lambda})).$$

Lemma 5 (Lemma 3.3 in [64]). *Let $(Y_n, n \in \mathbb{N})$ be a martingale. For all $k \geq 2$, let*

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

Then for all integers $n \geq 1$ and for all η such that for all $i \leq n$, $\mathbb{E}[\exp(|\eta(Y_i - Y_{i-1})|)] \leq \infty$,

$$\varepsilon_n \triangleq \exp\left(\eta Y_n - \sum_{k\geq 2}\frac{\eta^k}{k!}M_n^k\right)$$

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

Lemma 6. *Let $(\epsilon_t)_{t=0}^T$ be i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\epsilon_t = +1) = \mathbb{P}(\epsilon_t = -1) = -1$ and $\{X_t\}_{t=0}^T$ are a sequence of random variables, where $X_t \in [0, U]^M$, $X_t(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})$ is a function of $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})$. Then*

$$\sup_{X_1,\ldots,X_T} \|\frac{1}{T}\sum_{t=1}^{T} X_t(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})\epsilon_t\|_\infty \leq 2U\frac{\log(MT)}{\sqrt{T}},$$

with probability at least $1 - \frac{1}{(MT)^2}$.

Proof. To prove this Lemma, we once again use Markov's inequality and Lemma 5. For a fixed $m \in \{1, \ldots, M\}$, define the sequence $(Y_n, n \in \mathbb{N})$ as

$$Y_n \triangleq \frac{1}{T}\sum_{t=1}^{n} X_{t,m}\epsilon_t.$$

Notice the following values:

$$\begin{aligned} Y_n - Y_{n-1} =& \frac{1}{T}\epsilon_n X_{n,m} \\ M_n^k =& \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{1}{T}X_{t,m}\epsilon_t\right)^k |\epsilon_1, \ldots, \epsilon_{t-1}\right]. \end{aligned}$$

The first value shows that $\mathbb{E}[Y_n - Y_{n-1}|\epsilon_1, \ldots, \epsilon_{n-1}] = 0$ and therefore Y_n (and the negative of the sequence, $-Y_n$) is a martingale. Additionally, we have assumed that $0 \leq X_{m,i} \leq U$ for $1 \leq m \leq M$ and $1 \leq i \leq T$, so it is true that $|Y_n - Y_{n-1}| \leq \frac{2U}{T} \triangleq B$. Additionally:

$$\begin{aligned} M_n^2 =& \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{1}{T}X_{t,m}\epsilon_t\right)^2 |\epsilon_1, \ldots, \epsilon_{t-1}\right] \\ =& \frac{1}{T^2}\sum_{t=1}^{n} \epsilon_t^2 \mathbb{E}\left[X_{t,m}^2|\epsilon_1, \ldots, \epsilon_{t-1}\right] \\ \leq& \frac{4nU^2}{T^2} \triangleq \widehat{M}_n^2 \end{aligned}$$

We will also need to bound M_n^k as follows:

$$\begin{aligned} M_n^k =& \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{\epsilon_t}{T^2}X_{t,m}\right)^k |\epsilon_1, \ldots, \epsilon_{t-1}\right] \\ =& \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{\epsilon_t}{T^2}X_{t,m}\right)^2 \left(\frac{\epsilon_t}{T^2}X_{t,m}\right)^{k-2} |\epsilon_1, \ldots, \epsilon_{t-1}\right] \\ \leq& B^{k-2}M_n^2 \end{aligned}$$

We need to use these values to get a bound on the summation term used in Lemma 5.

$$\begin{aligned} D_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!} M_n^k \leq \sum_{k\geq 2} \frac{\eta^k B^{k-2} M_n^2}{k!} \leq \frac{\widehat{M}_n^2}{B^2}\sum_{k\geq 2}\frac{(\eta B)^k}{k!} \triangleq \widehat{D}_n \\ \widetilde{D}_n \triangleq& \sum_{k\geq 2} \frac{\eta^k}{k!}(-1)^k M_n^k \leq \widehat{D}_n \end{aligned}$$

In the above $\widetilde{D}_n$ corresponds to the sum corresponding to the negative sequence $-Y_0, -Y_1, \ldots$ which we will also need to obtain the desired bound. Now we are able to use a variant of Markov's inequality to get a bound on the

desired quantity.

$$\begin{aligned}\mathbb{P}(|Y_n| \geq y) =& \mathbb{P}(Y_n \geq y) + \mathbb{P}(-Y_n \geq y) \leq \mathbb{E}[e^{\eta Y_n}]e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n)}]e^{-\eta y} \\ =& \mathbb{E}[e^{\eta Y_n - D_n + D_n}]e^{-\eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n + \widetilde{D}_n}]e^{-\eta y} \\ \leq& \mathbb{E}[e^{\eta Y_n - D_n}]e^{\widehat{D}_n - \eta y} + \mathbb{E}[e^{\eta(-Y_n) - \widetilde{D}_n}]e^{\widehat{D}_n - \eta y} \leq 2e^{\widehat{D}_n - \eta y}\end{aligned}$$

The final inequality comes from the use of Lemma 5, which states that the given terms are supermartingales with initial term equal to 1, so the entire expectation is less than or equal to 1. The final step of the proof is to find the optimal value of η to minimize this upper bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp(\widehat{D}_n - \eta y) = 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(e^{\eta B} - 1 - \eta B\right) - \eta y\right)$$

Setting $\eta = \frac{1}{B}\log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)$ yields the lowest such bound, giving

$$\begin{aligned}\mathbb{P}(|Y_n| \geq y) \leq& 2\exp\left(\frac{\widehat{M}_n^2}{B^2}\left(\frac{yB}{\widehat{M}_n^2} - \log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)\right) - \frac{y}{B}\log\left(\frac{yB}{\widehat{M}_n^2} + 1\right)\right) \\ =& 2\exp\left(-\frac{\widehat{M}_n^2}{B^2}H\left(\frac{yB}{\widehat{M}_n^2}\right)\right)\end{aligned}$$

where $H(x) = (1+x)\log(1+x) - x$. We can use the fact that $H(x) \geq \frac{3x^2}{2(x+3)}$ for $x \geq 0$ to further simplify the bound.

$$\mathbb{P}(|Y_n| \geq y) \leq 2\exp\left(\frac{-3y^2}{2yB + 6\widehat{M}_n^2}\right) = 2\exp\left(-\frac{3y^2T^2}{8(Ty + 3n)}\right)$$

To complete the proof, we set $n = T$ and take a union bound over all indices because Y_T considered specific indices m, which gives the bound

$$\begin{aligned}&\mathbb{P}\Big(\max_m \frac{1}{T}\left|\sum_{t=1}^{T} X_{t,m}\epsilon_t\right| \geq 2U\frac{\log(MT)}{\sqrt{T}}\Big) \\ &\quad\leq \exp\left(\log(M) - \frac{12U^2T\log^2(MT)}{4U^2(\sqrt{T}\log(MT) + 3T)}\right) \\ &\quad\leq \exp\left(\log(MT) - \frac{3\log(MT)}{1/\sqrt{T} + 3/\log(MT)}\right) \\ &\quad\leq \exp\left(-2\log(MT)\right).\end{aligned}$$

□

7.2 Proof of Lemma 1

7.2.1 Part 1

Proof. For all $1 \leq t \leq T$ and $1 \leq m \leq M$, $X_{t,m}|X_{t-1}$ is drawn from a Poisson distribution with mean $e^{\nu_m + a_m^{*\top}X_{t-1}}$ for some $a_m^* \in [a_{\min}, 0]^M$. Because of the range of values a_m^* can take, we know that $e^{\nu_m + a_m^{*\top}X_{t-1}} \leq e^{\nu_{\max}}$ where $\nu_m \leq \nu_{\max}$ for some $\nu_{\max} < \infty$ for all m. Therefore, we know that

$$\mathbb{P}(X_{t,m} \geq \eta + e^{\nu_{\max}}|X_{t-1}) \leq \mathbb{P}(Y \geq \eta + e^{\nu_{\max}})$$

where Y is a Poisson random variable with mean $e^{\nu_{\max}}$. To bound this quantity we use the result of Lemma 4,

$$\mathbb{P}(Y > \eta + e^{\nu_{\max}}) \leq \exp\left(-\frac{\eta}{4}\log\left(1 + \frac{\eta}{2e^{\nu_{\max}}}\right)\right).$$

Setting $\eta = C\log MT - e^{\nu_{\max}}$,

$$\begin{aligned}\mathbb{P}(Y > C\log MT) &\leq \exp\left(-\frac{C\log MT - e^{\nu_{\max}}}{4}\log\left(1 + \frac{C\log MT - e^{\nu_{\max}}}{2e^{\nu_{\max}}}\right)\right)\\ &\leq \exp\left(-\frac{C\log MT - e^{\nu_{\max}}}{4}\right).\end{aligned}$$

Here, we have assumed that $C \geq e^{\nu_{\max}}(2e-1)$ and $\log MT \geq 1$. This upper bound is not dependent on the value of X_{t-1}, so this quantity is also an upper bound for the unconditional probability of $X_{t,m} \geq C\log MT$. Using this for a single index t, m of our data X, and taking a union bound over all possible indices $1 \leq m \leq M, 1 \leq t \leq T$ gives

$$\mathbb{P}\left(\max_{1\leq m\leq M, 1\leq t\leq T} X_{t,m} > C\log MT\right) \leq \exp\left(\log MT - \frac{C\log MT - e^{\nu_{\max}}}{4}\right) \leq \exp(-c\log MT) \tag{11}$$

for $c \leq \frac{C - e^{\nu_{\max}}}{4} - 1$. Thus if $C > \max(e^{\nu_{\max}}(2e-1), 4 + e^{\nu_{\max}})$, then $c > 0$, and the bound is valid. □

7.2.2 Part 2

Proof. We are interested in bounding the number of observations $X_{t,m}$ for $1 \leq m \leq M$ and $1 \leq t \leq T$ that are above the value U. Saying at least $j \triangleq \alpha MT$ observations are less than a certain value, is equivalent to saying that the j^{th} smallest observation is less than that value. Therefore,

$$\begin{aligned}\mathbb{P}(j^{th}\text{smallest observation } X_{t,m} > U) &= \mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} Y_{t,m} \leq j-1\right)\\ &= \sum_{\ell=0}^{j-1}\mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} Y_{t,m} = \ell\right) \leq \sum_{\ell=0}^{j}\sum_{y\in\mathcal{Y}^\ell}\mathbb{P}(Y=y).\end{aligned}$$

Here we define $Y_{t,m} \triangleq \mathbf{1}\{X_{t,m} \leq U\}$, and $\mathcal{Y}^\ell = \{y \in \{0,1\}^{M\times T} | \sum_{t=1}^{T}\sum_{m=1}^{M} y_{t,m} = \ell\}$. We then condition the values of Y_t on all previous values of Y and then understand this as a marginal of the joint distribution over Y_t and X_{t-1}. Below we use the notation $Y_{1:t}$ to denote all the time indices of Y from 1 to t, and similarly for y.

$$\begin{aligned}\mathbb{P}(Y=y) &= \prod_{t=1}^{T}\mathbb{P}(Y_t = y_t | Y_{1:t-1} = y_{1:t-1})\\ &= \prod_{t=1}^{T}\sum_{x_{t-1}}(\mathbb{P}(Y_t = y_t | Y_{1:t-1} = y_{1:t-1}, X_{t-1} = x_{t-1})\mathbb{P}(X_{t-1} = x_{t-1} | Y_{1:t-1} = y_{1:t-1}))\\ &= \prod_{t=1}^{T}\sum_{x_{t-1}}\left(\left(\prod_{m=1}^{M}\mathbb{P}(Y_{t,m} = y_{t,m} | X_{t-1} = x_{t-1})\right)\mathbb{P}(X_{t-1} = x_{t-1} | Y_{1:t-1} = y_{1:t-1})\right)\end{aligned}$$

In the last line we use the fact that conditioned on X_{t-1}, Y_t is independent across dimensions m, and independent of previous values $Y_{1:t-1}$. We now make the observation that $\mathbb{P}(X_{t,m} > U | X_{t-1} = x_{t-1})$ is exactly the probability that a Poisson random variable with rate $\exp(\nu_m + a_m^{*\top} x_{t-1})$ is greater than U, which can be upper-bounded by the probability that a Poisson random variable with rate $\exp(\nu_{\max})$ is greater than U because we have assumed all values of a_m^* are non-positive. Call this probability $p_{\nu_{\max}}$. Thus we have $\mathbb{P}(Y=y) \leq p_{\nu_{\max}}^{MT - \sum_{t=1}^{T}\sum_{m=1}^{m} y_{t,m}}$ and therefore,

$$\mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} Y_{t,m} \leq j-1\right) \leq \sum_{\ell=0}^{j}\binom{MT}{\ell} p_{\nu_{\max}}^{MT-\ell} = (1+p_{\nu_{\max}})^{MT} - \sum_{\ell=0}^{MT-j-1}\binom{MT}{\ell} p_{\nu_{\max}}^{\ell}$$
$$\leq \binom{MT}{MT-j}(1+p_{\nu_{\max}})^{j} p_{\nu_{\max}}^{MT-j} \leq \left(\frac{MTe}{MT-j}\right)^{MT-j}(1+p_{\nu_{\max}})^{j} p_{\nu_{\max}}^{MT-j}.$$

The second inequality is from the application of Taylor's Remainder Theorem, and the third is from the fact that $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^{k}$. Now use the fact that $j = \alpha MT$ as stated in the Lemma, to give

$$\mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} Y_{t,m} \leq j-1\right) \leq \left(\frac{p_{\nu_{\max}} e}{1-\alpha}\right)^{(1-\alpha)MT}(1+p_{\nu_{\max}})^{\alpha MT} \leq \left[\left(\frac{p_{\nu_{\max}} e}{1-\alpha}\right)^{1-\alpha} 2^{\alpha}\right]^{MT}.$$

By using Lemma 4 in a similar way as was used in the proof of Lemma 1 part 1, $p_{\nu_{\max}}$ can be controlled by U in the following way,

$$p_{\nu_{\max}} = \mathbb{P}(X > U) \leq \exp\left(-\frac{U-e^{\nu_{\max}}}{4}\log\left(1+\frac{U-e^{\nu_{\max}}}{2e^{\nu_{\max}}}\right)\right) \leq \exp\left(-\frac{U-e^{\nu_{\max}}}{4}\right),$$

when $U \geq e^{\nu_{\max}}(2e-1)$. Plugging the result back into the bound gives

$$\mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} Y_{t,m} \leq j-1\right) \leq \left[\left(\frac{\exp(1-(U-e^{\nu_{\max}})/4)}{1-\alpha}\right)^{1-\alpha} 2^{\alpha}\right]^{MT}.$$

When $U > 4 + e^{\nu_{\max}} + \frac{4\alpha\log(2)}{1-\alpha} - 4\log(1-\alpha)$ and additionally greater than $e^{\nu_{\max}}(2e-1)$ the condition from above, then the probability of this event is decaying in M and T. Therefore, for $c = -\left(1 - \frac{U-e^{\nu_{\max}}}{4} - \log(1-\alpha)\right)(1-\alpha) - \alpha\log(2)$, we have the inequality

$$\mathbb{P}(\text{at least } \alpha MT \text{ observations } X_{t,m} \leq U) \geq 1 - e^{-cMT}$$

□

7.3 Proof of Lemma 2

Proof. To prove the form of the stationary distribution we show that

$$\pi(y) = \int_{x} \pi(x) P(x,y),$$

where

$$P(x,y) = \exp\left(\nu^{\top} y + y^{\top} A^{*} x - \sum_{m=1}^{M} Z(\nu_m + a_m^{\top} x)\right)\prod_{m=1}^{M} h(y_m).$$

Plugging in $\pi(x)$ as specified,

$$\begin{aligned}
\int_x \pi(x)P(x,y) =& C_{\nu,A^*} \int_x \exp\left(\nu^\top x + \sum_{m=1}^M Z(\nu_m + a_m^{*\top} x) + \nu^\top y + y^\top A^* x - \sum_{m=1}^M Z(\nu_m + a_m^{*\top} x)\right) \prod_{m=1}^M h(x_m)h(y_m) \\
=& C_{\nu,A^*} \exp(\nu^\top y) \prod_{m=1}^M (h(y_m)) \int_x \exp\left(\nu^\top x + y^\top A^* x\right) \prod_{m=1}^M h(x_m) \\
=& C_{\nu,A^*} \exp(\nu^\top y) \prod_{m=1}^M (h(y_m)) \int_x \exp\left(\nu^\top x + x^\top A^* y\right) \prod_{m=1}^M h(x_m) \\
=& C_{\nu,A^*} \exp(\nu^\top y) \prod_{m=1}^M \left(h(y_m) \int_{x_m} \exp\left(\nu_m x_m + x_m a_m^{*\top} y\right) h(x_m)\right) \\
=& C_{\nu,A^*} \exp\left(\nu^\top y + \sum_{m=1}^M Z(\nu_m + a_m^{*\top} y)\right) \prod_{m=1}^M h(y_m) = \pi(y)
\end{aligned}$$

The second to last equality uses the definition of Z as the log partition function, and the third uses the assumption that $A^* = A^{*\top}$.

To prove the upper bound on total variation distance for Markov chains on countable domains, we define two chains, one chain Y_t begins at the stationary distribution and the other independent chain starts at X_t begins at some arbitrary random state x, both with transition kernel P. These two chains are said to be coupled if they are run independently until the first time where the states are equal, then are equal for the rest of the trial. The notation $P^t(x,y)$ denotes the probability of transitioning from state y to state x in exactly t steps. Theorem 5.2 of [66] asserts that:

$$\|P^t(x,\cdot) - \pi(\cdot)\|_{TV} \leq \mathbb{P}_x(\tau_{couple} > t),$$

where $\tau_{couple} := \left\{\min_{t>0} : X_t = Y_t\right\}$. Note first that $\mathbb{P}(\tau_{couple} > t) \leq \prod_{\tau=0}^t (1 - \mathbb{P}(X_\tau = Y_\tau = 0))$. Since the chains are independent until τ_{couple}, $\mathbb{P}(X_\tau = Y_\tau = 0) = \mathbb{P}(X_\tau = 0)\mathbb{P}(Y_\tau = 0)$. Note also that:

$$\begin{aligned}
\mathbb{P}(X_\tau = 0 | X_{\tau-1} = x) =& h(0)^M \exp\left(-\sum_{m=1}^M Z(\nu_m + a_m^{*\top} x)\right) \\
\geq& h(0)^M \exp\left(-\sum_{m=1}^M Z(\nu_m)\right) \geq h(0)^M \exp(-MZ(\nu_{\max})),
\end{aligned}$$

where the first inequality is due to the fact that Z is an increasing function, and from the assumption that $A_{i,j} \geq 0$. Hence $\mathbb{P}(\tau_{couple} > t) \leq \prod_{\tau=0}^t (1 - h(0)^{-2M} \exp(-2MZ(\nu_{\max}))) = (1 - h(0)^{-2M} \exp(-2MZ(\nu_{\max})))^t$. □

7.4 Empirical processes for martingale sequences

To concretely define the martingale, let $(X_t)_{t\geq 1}$ be a sequence of random variables adapted to the filtration $(\mathcal{A}_t)_{t\geq 1}$. First we present a bounded difference inequality for martingales developed by van de Geer [61].

Theorem 4 (Theorem 2.6 in [61]). *Fix $T \geq 1$ and let Z_T be an $\mathcal{A}_\mathcal{T}$-measurable random variable, satisfying for each $t = 1, 2, \ldots, t$,*

$$L_t \leq \mathbb{E}[Z_T | \mathcal{A}_t] \leq U_t,$$

almost surely where $L_t < U_t$ are constants. Define $C_T^2 = \sum_{t=1}^T (U_t - L_t)^2$. Then for all $a > 0$,

$$\mathbb{P}(Z_T - \mathbb{E}[Z_T] \geq a) \leq \exp(-\frac{2a^2}{C_T^2}).$$

The second important result we need is a notion of sequential Rademacher complexity for martingales that allows us to do symmetrization, an important step in empirical process theory (see e.g. [67]). To do this we use machinery developed in [62]. Recall that $(X_t)_{t\geq 1}$ is a martingale and let χ be the range of each X_t. Let $\mathcal{F}$ be a function class where for all $f \in \mathcal{F}$, $f : \chi \to \mathbb{R}$.

To define the notion of sequential Rademacher complexity, we first let $(\epsilon_t)_{t=1}^T$ be a sequence of independent Rademacher random variables (i.e. $\mathbb{P}(\epsilon_t = +1) = \mathbb{P}(\epsilon_t = -1) = \frac{1}{2}$). Next we define a tree process as a function of these independent Rademacher random variables.

A χ-valued tree $\mathbf{x}$ of depth T is a rooted complete binary tree with nodes labelled by elements of χ. We identify the tree $\mathbf{x}$ with the sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ of labeling functions $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \chi$ which provide the labels for each node. Here $\mathbf{x}_1 \in \chi$ is the label for the root of the tree, while $\mathbf{x}_t$ for $t > 1$ is the label of the node obtained by following the path of length $t-1$ from the root, with $+1$ indicating "right" and -1 indicating "left." Based on this tree, $\mathbf{x}_t$ is a function of $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})$.

Based on this, we define the sequential Rademacher complexity of a function class $\mathcal{F}$.

Definition 1 (Definition 3 in [62])**.** The sequential Rademacher complexity of a function class $\mathcal{F}$ on a χ-valued tree $\mathbf{x}$ is defined as

$$\mathcal{R}_T(\mathcal{F}) \triangleq \sup_{\mathbf{x}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \epsilon_t f(\mathbf{x}_t(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1}))\right]$$

where the outer supremum is taken over all χ-valued trees. Importantly note that $\big(\epsilon_t f(\mathbf{x}_t(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1}))\big)_{t\geq 1}$ is a martingale. Now we are in a position to state the main result which allows us to do symmetrization for functions of martingales.

Theorem 5 (Theorem 2 in [62])**.**

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[f(X_t)|\mathcal{A}_{t-1}] - f(X_t)\right] \leq 2\mathcal{R}_T(\mathcal{F}).$$

For further details refer to [62].

References

[1] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[2] T. P. Coleman and S. Sarma, "Using convex optimization for nonparametric statistical analysis of point processes," in *Proc. ISIT*, 2007.

[3] A. C. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Computation*, vol. 15, pp. 965–991, 2003.

[4] M. Hinne, T. Heskes, and M. A. J. van Gerven, "Bayesian inference of whole-brain networks," *arXiv:1202.1696 [q-bio.NC]*, 2012.

[5] M. Ding, CE Schroeder, and X. Wen, "Analyzing coherent brain networks with Granger causality," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 5916–8.

[6] 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*, vol. 454, pp. 995–999, 2008.

[7] M. S. Masud and R. Borisyuk, "Statistical technique for analysing functional connectivity of multiple spike trains," *Journal of Neuroscience Methods*, vol. 196, no. 1, pp. 201–219, 2011.

[8] Praneeth Netrapalli and Sujay Sanghavi, "Learning the graph of epidemic cascades," in *ACM SIGMETRICS Performance Evaluation Review*. ACM, 2012, vol. 40, pp. 211–222.

[9] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alessandro Ingrosso, and Riccardo Zecchina, "The patient-zero problem with noisy observations," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 10, pp. P10016, 2014.

[10] David Kempe, Jon Kleinberg, and Éva Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[11] M. Kuperman and G. Abramson, "Small world effect in an epidemiological model," *Physical Review Letters*, vol. 86, no. 13, pp. 2909, 2001.

[12] Per Johansson, "Speed limitation and motorway casualties: a time series count data regression approach," *Accident Analysis & Prevention*, vol. 28, no. 1, pp. 73–87, 1996.

[13] David S Matteson, Mathew W McLean, Dawn B Woodard, and Shane G Henderson, "Forecasting emergency medical service call arrival rates," *The Annals of Applied Statistics*, pp. 1379–1406, 2011.

[14] Tina Hviid Rydberg and Neil Shephard, "A modelling framework for the prices and times of trades made on the new york stock exchange," Tech. Rep., Nuffield College, 1999, Working Paper W99-14.

[15] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. A. Laeven, "Modeling financial contagion using mutually exciting jump processes," Tech. Rep., National Bureau of Economic Research, 2010.

[16] V. Chavez-Demoulin and J. A. McGill, "High-frequency financial data modeling using Hawkes processes," *Journal of Banking & Finance*, vol. 36, no. 12, pp. 3415–3426, 2012.

[17] A Colin Cameron and Pravin K Trivedi, *Regression analysis of count data*, vol. 53, Cambridge university press, 2013.

[18] 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*, vol. 58, no. 8, pp. 5544–5562, 2012.

[19] J. Silva and R. Willett, "Hypergraph-based anomaly detection in very large networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 563–569, 2009, doi:10.1109/TPAMI.2008.232.

[20] A. Stomakhin, M. B. Short, and A. Bertozzi, "Reconstruction of missing data in social networks based on temporal patterns of interactions," *Inverse Problems*, vol. 27, no. 11, 2011.

[21] C. Blundell, K. A. Heller, and J. M. Beck, "Modelling reciprocating relationships with Hawkes processes," in *Proc. NIPS*, 2012.

[22] 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.

[23] Shyh-Jier Huang and Kuang-Rong Shih, "Short-term load forecasting via arma model identification including non-Gaussian process considerations," *Power Systems, IEEE Transactions on*, vol. 18, no. 2, pp. 673–679, 2003.

[24] D. Vere-Jones and T. Ozaki, "Some examples of statistical estimation applied to earthquake data," *Ann. Inst. Statist. Math.*, vol. 34, pp. 189–207, 1982.

[25] Y. Ogata, "Seismicity analysis through point-process modeling: A review," *Pure and Applied Geophysics*, vol. 155, no. 2-4, pp. 471–507, 1999.

[26] Kurt Brännäs and Per Johansson, "Time series count data regression," *Communications in Statistics-Theory and Methods*, vol. 23, no. 10, pp. 2907–2925, 1994.

[27] Iain L MacDonald and Walter Zucchini, *Hidden Markov and other models for discrete-valued time series*, vol. 110, CRC Press, 1997.

[28] Scott L Zeger, "A regression model for time series of counts," *Biometrika*, vol. 75, no. 4, pp. 621–629, 1988.

[29] Bent Jørgensen, Soren Lundbye-Christensen, PX-K Song, and Li Sun, "A state space model for multivariate longitudinal count data," *Biometrika*, vol. 86, no. 1, pp. 169–181, 1999.

[30] Ludwig Fahrmeir and Gerhard Tutz, *Multivariate statistical modelling based on generalized linear models*, Springer Science & Business Media, 2013.

[31] Gary K Grunwald, Rob J Hyndman, Leanna Tedesco, and Richard L Tweedie, "Theory & methods: Non-Gaussian conditional linear AR (1) models," *Australian & New Zealand Journal of Statistics*, vol. 42, no. 4, pp. 479–495, 2000.

[32] Michael A Benjamin, Robert A Rigby, and D Mikis Stasinopoulos, "Generalized autoregressive moving average models," *Journal of the American Statistical association*, vol. 98, no. 461, pp. 214–223, 2003.

[33] Christian Gouriéroux and Joann Jasiak, "Autoregressive gamma processes," *Les Cahiers du CREF of HEC Montréal Working Paper*, , no. 05-03, 2005.

[34] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim, "Poisson autoregression," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1430–1439, 2009.

[35] Fukang Zhu and Dehui Wang, "Estimation and testing for a Poisson autoregressive model," *Metrika*, vol. 73, no. 2, pp. 211–230, 2011.

[36] Konstantinos Fokianos and Dag Tjøstheim, "Log-linear Poisson autoregression," *Journal of Multivariate Analysis*, vol. 102, no. 3, pp. 563–578, 2011.

[37] A. G. Hawkes, "Point spectra of some self-exciting and mutually-exciting point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 83–90, 1971.

[38] A. G. Hawkes, "Point spectra of some mutually-exciting point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 33, pp. 438–443, 1971.

[39] 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.

[40] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard, "LASSO and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, pp. 83–143, 02 2015.

[41] Muzy Bacry, Gaiffas, "A generalization error bound for sparse and low-rank multivariate hawkes processes," arXiv:1501.00725, 2015.

[42] Andréas Heinen, "Modeling time series count data: an autoregressive conditional Poisson model," *Available at SSRN 1117187*, 2003.

[43] Fukang Zhu, "A negative binomial integer-valued garch model," *Journal of Time Series Analysis*, vol. 32, no. 1, pp. 54–67, 2011.

[44] Fukang Zhu, "Modeling time series of counts with COM-poisson INGARCH models," *Mathematical and Computer Modelling*, vol. 56, no. 9, pp. 191–203, 2012.

[45] Fukang Zhu, "Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued garch models," *Journal of Mathematical Analysis and Applications*, vol. 389, no. 1, pp. 58–71, 2012.

[46] R. Willett and R. Nowak, "Multiscale Poisson intensity and density estimation," *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3171–3187, 2007, doi:10.1109/TIT.2007.903139.

[47] M. Raginsky, R. Willett, Z. Harmany, and R. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 3990–4002, 2010, arXiv:0910.5146.

[48] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank, "Performance bounds for expander-based compressed sensing in Poisson noise," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, 2011, arXiv:1007.2377.

[49] X. Jiang, R. Willett, and G. Raskutti, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Transactions on Information Theory*, vol. 61, pp. 4458–4474, 2015.

[50] S. van de Geer, "High-dimensional generalized linear models and the LASSO," *Annals of Statistics*, vol. 36, pp. 614–636, 2008.

[51] V. Koltchinskii and M. Yuan, "Sparse recovery in large ensembles of kernel machines," in *Proceedings of COLT*, 2008.

[52] L. Meier, S. van de Geer, and P. Buhlmann, "High-dimensional additive modeling," *Annals of Statistics*, vol. 37, pp. 3779–3821, 2009.

[53] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2010.

[54] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Transactions on Information Theory*, vol. 57, pp. 6976–6994, 2011.

[55] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *Journal of Machine Learning Research*, vol. 13, pp. 398–427, 2012.

[56] P. Zhao and B. Yu, "On model selection consistency of LASSO," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2567, 2006.

[57] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[58] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals of Statistics*, vol. 43, no. 4, pp. 1535–1567, 2015.

[59] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett, "A data-dependent weighted LASSO under Poisson noise," *arXiv preprint arXiv:1509.08892*, 2015.

[60] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[61] S. van de Geer, *Empirical Process Techniques for Dependent Data*, Springer-Verlag, New York, NY, 2002.

[62] A. Rakhlin, K. Sridharan, and A. Tewari, "Sequential complexities and uniform martingale laws of large numbers," *Probability Theory and Related Fields*, vol. 1, no. 161, pp. 111–153, February 2015.

[63] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[64] Christian Houdré and Patricia Reynaud-Bouret, "Exponential inequalities, with constants, for U-statistics of order two," in *Stochastic inequalities and applications*, pp. 55–69. Springer, 2003.

[65] S. G. Bobkov and M. Ledoux, "On modified logarithmic Soboloev inequalities for Bernoulli and Poisson measures," *Journal of Functional Analysis*, vol. 156, pp. 347–365, 1998.

[66] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2008.

[67] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.