Sparse Multivariate Bernoulli Processes in High Dimensions

Parthe Pandit[†] **Mojtaba Sahraee**[†] **Arash A. Amini**[§] **Sundeep Rangan**[‡] **Alyson K. Fletcher**[§†]
[†]Dept. ECE, UCLA [§]Dept. Statistics, UCLA [‡]Dept. ECE, NYU

Abstract

We consider the problem of estimating the parameters of a multivariate Bernoulli process with auto-regressive feedback in the high-dimensional setting where the number of samples available is much less than the number of parameters. This problem arises in learning interconnections of networks of dynamical systems with spiking or binary valued data. We also allow the process to depend on its past up to a lag p, for a general $p \geq 1$, allowing for more realistic modeling in many applications. We propose and analyze an ℓ_1-regularized maximum likelihood (ML) estimator under the assumption that the parameter tensor is approximately sparse. Rigorous analysis of such estimators is made challenging by the dependent and non-Gaussian nature of the process as well as the presence of the nonlinearities and multi-level feedback. We derive precise upper bounds on the mean-squared estimation error in terms of the number of samples, dimensions of the process, the lag p and other key statistical properties of the model. The ideas presented can be used in the rigorous high-dimensional analysis of regularized M-estimators for other sparse nonlinear and non-Gaussian processes with long-range dependence.

1 INTRODUCTION

In many signal processing applications, the underlying time series may be modeled as a multivariate Bernoulli process (MBP). For example, the spike trains from an ensemble of neurons can be modeled as a collection of Bernoulli variables where for each neuron, at each time instant, the probability of spiking could depend on the history of the spikes from the ensemble. Similarly several other signals such as activity in social networks, trends of stock prices in financial markets [26, 27], crime occurrences in a metropolitan area [14], medical emergency call forecasting [15], climate dynamics [6] and certain physiological [12] and biological processes [8] can all be encoded as multivariate autoregressive Bernoulli processes where the history of the process affects the present outcome. We are interested in developing generalized linear models (GLM) to capture the behavior of such temporally-dependent MBPs. Such models would allow one to not only make predictions about the future of the process, but also infer the relations among the coordinates of the process (i.e., the individual time series) by a proper estimation of the parameters of the model. For example, in the neural spike train example, one could reveal a latent network among the neurons (i.e., who influences whose firing) just from observations of patterns of neural activity, a task which is of significant interest in neuroscience [19, 25, 3]. Similarly, in the context of social networks, one might be interested in who is influencing whom [20].

In a GLM, one models an invertible link function of the conditional mean of the observed variables as a linear function of the covariates. In the context of time series analysis, the role of the covariates is played by the history of the evolving time series. For Gaussian random variables with the identity link function, this leads to the classical Gaussian autoregressive (AR) process. Indeed, much of the focus in time series analysis has been on VAR (vector AR) processes [2, 5, 16, 17, 1], which provide significant richness as a rudimentary model. However, these models fail to capture higher order correlations among the variables and become harder to interpret for variables from discrete spaces.

Another fundamental class of processes is the MBP, for which very few results are known as compared to the VAR processes. For such MBPs, we are interested in understanding the dependence of each variable on the history of the process. We consider a time series of N

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Bernoulli variables where each variable depends on at most p lags of the process, resulting in N^2p possible interaction parameters. These N^2p interactions can be arranged in an $N \times N \times p$ tensor Θ, where $\Theta_{ij\ell}$ captures the effect of the response of variable j from ℓ lags ago on variable i.

Collectively, each $N \times N$ *slice* given by $\Theta_{**\ell}$ for $\ell \in \{1, 2, \ldots, p\}$ indicates the coupling between all the pairs of variables lag ℓ apart. On the other hand, each *fiber* of length p along the third dimension Θ_{ij*}, for $i, j \in \{1, 2, \ldots, N\}$, can be thought of as a *filter* that modulates the behavior of variable i on the past responses of variable j. Delineating the parameters in this way helps with the interpretation of the dynamics of the process. For example, in neural signal processing, these filters Θ_{ij*} encode properties of the spiking behavior of the neurons, as demonstrated by [30] where they provide a characterization of the filter coefficients that incite activities such as bursting, tonic spiking, phasic spiking and several others. Similarly, a slice $\Theta_{**\ell}$ can be thought of as an adjacency matrix for the *lag* ℓ *influence network* of the neurons, where $\Theta_{ij\ell} = 0$ for all ℓ if neurons i and j are not connected. In the cases where the neurons are connected however, it is possible to have different patterns of influence at different time lags, and hence the influence networks could potentially be different for each ℓ.

In order to reveal the structure of these influence networks, one needs to estimate the tensor parameter Θ, given a sample of observations from the process. In many scenarios, the parameter tensor Θ is known to be sparse. For example, in the neural setting, each neuron, or the functional unit of the cortex, has limited direct connections to other units. To exploit the sparsity assumption in forming an estimate of Θ, we propose and provide a rigorous analysis for an ℓ_1 penalized maximum likelihood estimator. For Gaussian random variables, this takes the form of a regularized least-squares (LS) tensor regression. However, for the Bernoulli GLM, maximizing the likelihood differs from the ordinary least-squares. Although one can still use a regularized LS estimator, incorporating the Bernoulli likelihood allows one to capture more information, i.e., achieve smaller variance. The resulting M-estimator in the Bernoulli case resembles a regularized logistic regression problem which is in the class of convex problems that can be solved efficiently.

While traditional statistical methods work under the assumption that the number of available samples significantly exceeds the number of parameters, i.e., $n \gg N^2p$, the assumption is often not true in several applications. Going back to the neuroscience application, in most spiking neural data from in vivo measurements, a limited sample of spike trains are available for an experiment, owing to constraints such as subject fatigue, changes in sensor characteristics, and sensitivity to exogenous laboratory conditions. In addition, synaptic connectivity is time-varying and hence there is a limited period over which the model parameters can be assumed constant. In such a scenario, statistically sound methods are desired that provide guarantees on the fidelity of the model trained over a limited sample of observations, especially with $n \ll N^2p$. However, in spite of the model identification problem being ill-posed in this case, it is often possible to perform reliable estimation even with $n = \mathcal{O}(\log(N^2p))$, by constraining the parameter Θ to lie in a low-dimensional subspace, with s degrees of freedom, where $s \ll N^2p$. Several such low-dimensional subspaces have been investigated in the literature on compressive sensing and high-dimensional statistics. Examples include the (elementwise) *sparsity* where approximately s of the N^2p entries are assumed non-zero, or the *low-rank* assumption on the parameter tensor where it is assumed to be the sum of a few rank-1 tensors. These assumptions are often practically valid, enhance the interpretability of the model, and make the problem well-posed when $n = \mathsf{poly}(s, \log(N^2p))$. Moreover, the estimation is computationally tractable due to convex optimization based estimators for these, even for the tensor models [23].

Our focus in this work is on the "approximately sparse" model, as motivated by the network-of-fibers structure in the neural spike train application. The parameter, Θ, is assumed to be well-approximated by an s-sparse tensor. Such a constraint is incorporated by regularizing the (convex) maximum likelihood problem using the elementwise ℓ_1 penalty. Our main result in Theorem 3.1 establishes the consistency of this regularized MLE in the high-dimensional regime of $n = \mathsf{poly}(s, \log(N^2p))$ under some regularity conditions. Despite the fact that the techniques for establishing such high-dimensional guarantees on regularized M-estimators are by now fairly mature [4], significant challenges remain in analyzing dependent non-Gaussian processes.

1.1 Key contributions

A major theoretical contribution of our work is to establish the so-called restricted strong convexity (RSC) [18] for the log-likelihood of a dependent non-Gaussian process. This requires a restricted eigenvalue condition for the MBP, which is nontrivial due to the non-Gaussian and highly correlated entries of the resulting design matrix. What makes the problem more challenging is the existence of feedback from more than just the immediate past (the case $p > 1$).

We establish the RSC for general $p \geq 1$ using the novel

approach of viewing the p-block version of the process as a Markov chain. The problem becomes significantly challenging when going from $p = 1$ to even $p = 2$. The difficulty with this *higher order* Markov chain is that its contraction coefficient is trivially 1. We develop techniques to get around this issue which could be of independent interest (cf. appendix C). Our techniques hold for all $p \geq 1$.

Much of the previous work towards proving the RSC condition of the log-likelihood function has either focused on the independent sub-Gaussian case [22, 31] or the dependent Gaussian case [2, 23] for which powerful Gaussian concentration results such as the Hanson–Wright inequality [24] are still available.

Our approach is to use concentration results for Lipschitz functions of discrete Markov chains, and strengthen them to uniform results using metric entropy arguments. In doing so, we circumvent the use of empirical processes which require additional assumptions for MBP estimation [21]. Moreover, our approach allows us to identify key assumptions on the decay rate of the *fiber*s of the parameter, for sample-efficient estimation.

Although MBP time series are often modeled using the logit link function, our analysis allows for any Lipschitz continuous, log-convex link function. The analysis brings out crucial properties of the link function, and the role they play in determining the estimation error and sample complexity.

1.2 Previous work

Estimating the parameters of a multivariate time series has been of interest in recent years. Much of the work, however, focuses on Gaussian VAR(p) processes with linear feedback [2, 5, 16, 17, 1]. For these models, a restricted eigenvalue condition can be established fairly easily, by reducing the problem, even in the time-correlated setting, to the concentration of quadratic functionals of Gaussian vectors for which powerful inequalities exist [24]. These techniques do not extend to non-Gaussian setups.

In the non-Gaussian setting, Hall et al. [7, 32] recently considered a multivariate time series evolving as a GLM driven by the history of the process. The autoregressive MBP with $p = 1$ lags is a special case of this model. They provide statistical guarantees on the error rate for the ℓ_1 regularized estimator, although their assumptions on the parameter space are restrictive when applied to the MBP. More importantly, their results are restricted to the case $p = 1$ which does not allow the explicit encoding of long-term dependencies as observed in crucial neuronal phenomena such as periodic spiking. More recently, Mark et al. [14, 13] considered a similar model for multivariate Poisson processes with lags $p > 1$, albeit either through predetermined basis functions or by restricting to lags $p = 1$ or $p = 2$. A key contribution of us is to bring out the explicit dependence on p in multilag MBP models, allowing for a general $p \geq 1$. We show how the scaling of the sample complexity and the error rate with p can be controlled by the properties of the link function and a certain norm of the parameter tensor.

Our results improve upon those in [7, 14] when applied to the MBP. Due to the key observation that a p-lag MBP can be viewed as a discrete Markov chain, our analysis relaxes several assumptions made by [7, 14]. In doing so, we achieve better sample complexities with explicit dependence on p. Our analysis borrows from martingale-based concentration inequalities for Lipschitz functions of Markov chains [11].

The univariate Bernoulli process for $p \geq 1$ was considered by Kazemipour et. al [9, 10] where they analyzed a multilag Bernoulli process for a single neuron. Their analysis does not extend to $N > 1$ case. Even for $N = 1$, their analysis is restricted to the biased process with $\mathbb{P}(x_i^t = 1|X^{t-1}) < \frac{1}{2}$ for all t. Mixing times of the MBP have been considered in [8]. However, their discussion is again limited to $p = 1$.

2 PROBLEM FORMULATION

Consider an N-dimensional time series $\{x^t\}_{t=1}^n$, where t denotes time and each $x^t = (x_i^t) \in \mathbb{R}^N$. A general framework for analyzing $\{x^t\}$ is to use a GLM for modeling the conditional mean of the *present* given the *past*, i.e., $p_\Theta(x_i^t \mid X^{t-1})$ is such that

$$\mathbb{E}[x_i^t \mid X^{t-1}] = f\Big(\langle \Theta_{i**}, X^{t-1} \rangle\Big)$$

independently across coordinates $i \in [N]$. Here, $X^{t-1} = [x^{t-1}\ x^{t-2}\ \cdots\ x^{t-p}] \in \mathbb{R}^{N\times p}$ is the "p-lag" history at time t, and $\Theta_{i**} \in \mathbb{R}^{N\times p}$ is the i^{th} slice of the parameter tensor $\Theta \in \Omega \subseteq \mathbb{R}^{N\times N\times p}$ along the first dimension. f is the inverse link function. The notation $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product between two matrices, i.e., $\langle \Theta_{i**}, X^{t-1} \rangle$ equals

$$\sum_{j,\ell} \Theta_{ij\ell} X_{j\ell}^{t-1} = \sum_{j,\ell} \Theta_{ij\ell}\, x_j^{t-\ell} = \sum_{\ell} \langle \Theta_{i*\ell}, x^{t-\ell} \rangle$$

where we note the useful identity $X_{j\ell}^{t-1} = x_j^{t-\ell}$. (The inner product in the last equality is the usual vector inner product.) The entry $\Theta_{ij\ell}$ captures how much the j^{th} dimension of the process at time lag ℓ affects the distribution of the i^{th} dimension, at each time instant. The model relates the distribution of x^t to the history of the process over p time lags X^{t-1}.

In other words, for the MBP model, we have

$$\begin{aligned} x_i^t \mid X^{t-1} &\sim \text{Ber}\left(z_i^t\right), \\ z_i^t := z_i^t(\Theta) &:= f\left(\langle \Theta_{i**}, X^{t-1} \rangle\right), \end{aligned} \tag{1}$$

where $f : \mathbb{R} \to [\varepsilon, 1-\varepsilon]$ for some $\varepsilon \in (0, \frac{1}{2})$, is the inverse link function. Note that $z_i^t = \mathbb{P}(x_i^t = 1 \mid X^{t-1})$ represents the conditional probability of spiking for neuron i at time t, given the p lag history of the ensemble.

We are interested in estimating the "true" parameter tensor Θ from n samples $\{x^t\}_{-p+1}^n$ of a process. (We assume that the first p samples corresponding to $i = -p+1$ to $i = 0$ are given for "free" since they do not contribute any observations to the regression during parameter estimation.)

Although the parameter space Ω has ambient dimension N^2p, in real-world applications, Θ often resides in or is well-approximated by a low-dimensional subspace of Ω, allowing reliable estimation even if $n \ll N^2p$, as is desired in several applications. In the literature on high-dimensional statistics, this is often the assumption and several such low dimensional subspaces are now well-studied and ubiquitous in analyses, some examples include tensors being low-rank, exactly sparse, approximately sparse and sparse with a specific structure. Here we focus on parameter estimation for the approximately sparse parameter model where the true parameter has a "good enough" sparse approximation.

Let the process be generated by a true parameter Θ^*. We assume Θ^* to be approximately s-sparse. More precisely, we assume that the following quantity

$$\sigma_s(\Theta^*) := \min_{|S| \le s} \|\Theta^*_{S^c}\|_{1,1,1}, \tag{2}$$

decays fast as a function of s, where $\Theta^*_{S^c}$ is the tensor Θ^* with support restricted to S^c, the complement of $S \subseteq [N] \times [N] \times [p]$. This quantity $\sigma_s(\Theta^*)$ captures the $\ell_{1,1,1}$ approximation error when Θ^* is approximated by an s-sparse tensor. For an exactly s-sparse tensor Θ^*, we have $\sigma_s(\Theta^*) = 0$. In general, we do not impose any constraint on $\sigma_s(\Theta^*)$ and state a general result involving this parameter. We denote by S^* the optimal set that solves (2). For future reference, we also define

$$\tau_s^2(\Theta^*) := \frac{\sigma_s^2(\Theta^*)}{s}. \tag{3}$$

Notation. Here and in what follows $\|\Theta\|_{p,q,r}$ denotes the norm on tensors obtained by collapsing the dimensions from right to left by applying ℓ_r, ℓ_q and ℓ_p norms in that sequence. Thus $\|\cdot\|_{1,1,1}$ is the elementwise ℓ_1 norm of the tensor, i.e., the sum of the absolute values of all its entries, and $\|\cdot\|_{\infty,\infty,\infty}$ is the absolute value of the entry of the tensor with largest magnitude, which is the dual norm for $\ell_{1,1,1}$ norm. $\|\Theta\|_0$ denotes the number of non-zero entries in Θ.

3 MAIN RESULT

We study a regularized maximum likelihood estimator (R-MLE) for Θ^*. The (normalized) negative log-likelihood of the Bernoulli process (1) is given by,

$$\mathcal{L}(\Theta) = = -\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{N} \ell_{it}\left(\langle \Theta_{i**}, X^{t-1} \rangle\right) \tag{4}$$

where $\ell_{it}(u) := x_i^t \log(f(u)) + (1 - x_i^t)\log(1 - f(u))$. In order to incorporate the sparse approximability of Θ^* during estimation, we penalize the likelihood by an elementwise ℓ_1 regularizer, leading to the R-MLE

$$\widehat{\Theta}_{\text{ML}} = \underset{\Theta \in \mathbb{R}^{N\times N\times p}}{\text{argmin}} \quad \mathcal{L}(\Theta) + \lambda_n \|\Theta\|_{1,1,1}. \tag{5}$$

Here, $\|\Theta\|_{1,1,1}$ is the elementwise ℓ_1 norm of the 3-tensor Θ, encouraging a sparse solution to the optimization problem. Consider a sample of size n from the multivariate Bernoulli process generated according to (1) with parameter $\Theta = \Theta^*$. We denote this sample by $\mathbb{X} := \{x^i\}_{i=-p+1}^n$. We further assume that the process satisfies the the following regularity conditions:

(A1) The sample $\mathbb{X} = \{x^t\}_{t=-p+1}^n$ is drawn from a stable wide-sense stationary process, with power spectral density matrix

$$\mathcal{X}(\omega) := \sum_{\ell=-\infty}^{\infty} \text{Cov}(x^t, x^{t+\ell}) e^{-j\omega\ell} \in \mathbb{C}^{N\times N},$$

for angular frequencies $\omega \in [-\pi, \pi)$, whose minimum eigenvalues are uniformly bounded below,

$$\min_{\omega\in[-\pi,\pi)} 2\pi\Lambda_{\min}(\mathcal{X}(\omega)) = c_\ell^2 > 0.$$

(A2) The inverse link function $f : \mathbb{R} \to [\varepsilon, 1-\varepsilon]$, for some $\varepsilon \in (0, \frac{1}{2})$, and is twice differentiable with Lipschitz constant L_f, i.e., $|f(u) - f(v)| \le L_f|u - v|$. In addition, both $\log f$ and $\log(1-f)$ are strongly convex with curvature bounded below by $c_f > 0$.

The following quantities will be key in stating the error bounds in our main result:

$$g_f^2(\Theta^*) := \frac{3L_f^2}{2\varepsilon}\sum_{\ell=1}^{p}\sum_{i=1}^{N}\Big(\sum_{j=1}^{N}\sum_{k=\ell}^{p}|\Theta^*_{ijk}|\Big)^2, \tag{6}$$

$$G_f(\Theta^*) := 8c_f^2\Big[1 + \frac{p^2}{\left(\frac{1}{g_f(\Theta^*)} - 1\right)^2}\Big]. \tag{7}$$

Theorem 3.1 *Let $\{x^t\}_{t=-p+1}^{n}$ be a process generated by (1) with parameter Θ^* and assume that it satisfies (A1) and (A2). Then, there exist positive universal constants c, c_1 and c_2 such that for*

$$n \geq c_1 \frac{G_f(\Theta^*)}{c_f^2\, c_\ell^6}\, s^3 \log(N^2 p), \tag{8}$$

any solution $\widehat{\Theta}_{\mathrm{ML}}$ to (5) with $\lambda_n = c_2 \frac{L_f}{\varepsilon}\sqrt{\frac{\log(N^2p)}{n}}$, satisfies

$$\|\widehat{\Theta}_{\mathrm{ML}} - \Theta^*\|_F^2 \leq C\Big[\frac{s\log(N^2p)}{n} + \widetilde{\tau}^2(\Theta^*)\sqrt{\frac{\log(N^2p)}{n}}\Big],$$

with probability at least $1 - n^{-c} - 2(N^2p)^{-c_0 s}$, where $\widetilde{\tau}_s^2(\Theta^*) := \tau_s^2(\Theta^*) + \sigma_s(\Theta^*)$. The constant $c_0 = \mathcal{O}(c_\ell^{-2})$ only depends on c_ℓ and $C = \mathcal{O}(\max\{\frac{L_f}{\varepsilon c_f c_\ell^2}, 1\}^2)$ only depends on the stated constants. □

A sketch of the proof is presented in Section 4. The details are deferred to Appendix A.

3.1 Remarks on Theorem 3.1

The two terms in the error bound correspond to the estimation and approximation errors, respectively. The estimation error, in general, scales at the so-called *fast rate* $s\log(N^2p)/n$ in our setting, while the approximation error scales with the slower rate $\widetilde{\tau}^2(\Theta^*)\sqrt{\log(N^2p)/n}$. For the exact sparsity model, where $\sigma_s(\Theta^*) = 0$, the approximation error vanishes (since $\widetilde{\tau}^2(\Theta^*) = 0$) and we achieve the fast rate.

The overall sample complexity for consistent estimation (ignoring constants) is thus

$$n \gtrsim G_f(\Theta^*)\max\{s^3, \widetilde{\tau}_s^4(\Theta^*)\}\,\log(N^2p)\,. \tag{9}$$

Scaling with s. According to (9), the scaling of n in the sparsity parameter "s" is at best $O(s^3)$, corresponding to the the case of hard sparsity where $\widetilde{\tau}_s^4(\Theta^*) = 0$. While an $O(s^3)$ dependence is not ideal, it is not clear if it can be improved significantly without imposing restrictive assumptions. It is clear that one cannot do better than $O(s)$, the optimal scaling in the linear independent settings. In our proof, the additional s^2 factor comes from concentration inequality (14) in Lemma A.5. There, if one were able to show sub-Gaussian concentration for deviations of the order of $\|\Delta\|_F^2$ instead of $\|\Delta\|_{2,1,1}^2$, then the additional s^2 can be removed. It remains open whether such concentration is possible and under what additional assumptions. Figure 1c in Section 5 suggests a superlinear dependence on s, hinting that the situation may not be as simple as the i.i.d. case.

In comparison, for $p \leq 2$, a sample complexity of $\rho^3\log(N)$ was reported in [7, Cor. 1], whereas [14, Thm 4.4] requires $U^4 s\log(N)$ samples where ρ and U are parameters defined in their respective models, both of which can potentially grow as $\Omega(s)$ unless assumed otherwise.

Scaling with p. For $N = 1$, our result is the first to provide a sample complexity logarithmic in p which holds for all N. In contrast, [10, Thm. 1] requires $s^{2/3}p^{2/3}\log(p)$ samples and relies heavily on $N = 1$.

Our bound scales with p through the quantity $G_f(\Theta^*)$. The scaling depends on the behavior of the tail of $\ell \mapsto |\Theta_{ij\ell}|$, that is, how fast the "*influence from the past*" dies down. For different regimes of the influence decay, the scaling of $G_f(\Theta^*)$ is summarized in Table 1.

Table 1: Scaling of $G_f(\Theta^*)$ with p.

$\|\Theta_{ij\ell}\|$ \ L_f	$\mathcal{O}(1)$	$\mathcal{O}(p^{-1})$	$\mathcal{O}(p^{-2})$
$\mathcal{O}(1)$	$\mathcal{O}(p^5)$	$\mathcal{O}(p^3)$	$\mathcal{O}(1)$
$\mathcal{O}(\ell^{-\alpha})$ or $\mathcal{O}(e^{-\beta\ell})$	$\mathcal{O}(p^2)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$

In the worst case, without any assumptions on Θ and with $L_f = \mathcal{O}(1)$, $G_f(\Theta^*)$ could scale as p^5. Although this is not ideal, it is analogous to constant U^4 in the sample complexity of [14], which is derived under more assumptions on Θ^*. (The dependence of U on p in that result is also somewhat complicated.)

On the other hand, under mild assumptions of polynomial or exponential tail decay, the dependence on p is much better: If $|\Theta_{ij\ell}|$ decays polynomially in lag ℓ, i.e., $|\Theta_{ij\ell}| = \mathcal{O}(\ell^{-\alpha})$, uniformly in i, j, or exponentially $|\Theta_{ij\ell}| = \mathcal{O}(e^{-\beta\ell})$, for any $\alpha > 3/2$ or $\beta > 0$, then the sample complexity reduces by a factor of p^3. Furthermore, if the Lipschitz constant L_f is allowed to drop with p, more reduction in $G_f(\Theta^*)$ and hence the sample complexity is possible, as illustrated in Table 1.

Appendix D provides derivations for Table 1. Better bounds on G_f and hence the sample complexity can be obtained by imposing suitable structural assumptions on Θ^*.

Scaling with N. Our results all have a logarithmic dependence on N, the number of neurons in the context of neuronal ensembles, which is a notable feature of our work. We overcome the $N > 1$ barrier for the MBP model, while also allowing $p \geq 1$.

Assumptions. We use Assumption (A1) to guarantee that the restricted strong convexity (RSC) property holds at the population level. The RSC is key in guaranteeing that any parameter tensor $\widehat{\Theta}$ that maximizes the regularized likelihood cannot stray too far away from the true parameter. For the $N = 1$ case,

it implies that the process does not have zeros on the unit circle in the spectral domain. Assumption (A1) is by now standard in estimating time-series [23, 2]. It relates to the "flatness" of the power spectral density (PSD) [2]. Controlling the scaling of c_ℓ in terms of Θ is a non-trivial research question, even for Gaussian AR(p) processes, since the relation between the PSD and the parameter is via the Z-transform. While there could be pathological Θ for which $c_\ell = o(1)$, the set of parameters for which $c_\ell = \Omega(1)$ is largely believed to be non-trivial. A line of work [2, 7, 14] obtains weak bounds on c_ℓ which could decrease with s, hence the authors need to further assume $s = O(1)$.

Assumption (A2) is not too restrictive. For example, for the `logit` link, $f(u) = f_\alpha(u) = 1/(1+e^{\alpha u})$, the sigmoid function, we have

$$c_f = \min_u \alpha^2 f_\alpha(u)(1 - f_\alpha(u)) = \alpha^2 \varepsilon(1-\varepsilon)$$

assuming that $f \in [\varepsilon, 1-\varepsilon]$, that is, $\alpha|u| \le \log \frac{1-\varepsilon}{\varepsilon}$. The Lipschitz constant in this case satisfies $L_f \le \frac{\alpha}{4}$.

4 PROOF SKETCH

The main challenge, when $n \ll N^2 p$, is that the empirical Hessian $\nabla^2 \mathcal{L}$ is rank-deficient and hence the likelihood cannot be strongly convex (i.e., have a positive curvature) in all directions around the true parameter Θ^*. This means that even though a candidate solution $\widehat{\Theta}$ is a stationary point of (5), it could be far away from Θ^* if the error vector $\widehat{\Delta} := \widehat{\Theta} - \Theta^*$ lies in the null-space of $\nabla^2 \mathcal{L}$. However, for sufficiently large values of λ_n, one can guarantee that $\widehat{\Delta}$ lies in a small "cone-like" subset. Hence, it suffices that $\mathcal{L}_n$ be strongly convex only over this subset (i.e., $\nabla^2 \mathcal{L}$ be uniformly quadratically bounded from below on this set) . This observation is by now standard in the high-dimensional analysis of estimators. For example, we refer to [18] whose general framework we use to establish the result.

Let us briefly sketch the approach of [18]; see Section A.1 in the supplement for more details. Let

$$R\mathcal{L}(\Delta; \Theta^*) := \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle$$

be the remainder of the first-order Taylor expansion of the loss function around Θ^*. The main ingredient of the proof is showing that $R\mathcal{L}(\Delta; \Theta^*)$ satisfies restricted strong convexity as defined below.

Definition The loss function $\mathcal{L}$ satisfies restricted strong convexity (RSC) relative to Θ^* and $S \subseteq [N]^2 \times [p]$ with curvature $\kappa > 0$ and tolerance τ^2 if

$$R\mathcal{L}(\Delta; \Theta^*) \ge \kappa \|\Delta\|_F^2 - \tau^2 \tag{10}$$

for any $\Delta \in \mathbb{R}^{N \times N \times p}$ such that

$$\|\Delta_{S^c}\|_{1,1,1} \le 3\|\Delta_S\|_{1,1,1} + 4\|\Theta^*_{S^c}\|_{1,1,1}. \tag{11}$$

The set of Δ in (11) is denoted as $\mathbb{C} = \mathbb{C}(S; \Theta^*)$.

It is shown in [18] that if (5) is solved with a regularization parameter satisfying

$$\lambda_n \ge 2\|\nabla \mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty}, \tag{12}$$

then for any $S \subseteq [N]^2 \times [p]$, the error vector $\widehat{\Delta}$ belongs to the cone-like set $\mathbb{C}(S; \Theta^*)$. Combined with the RSC property (11), Thm. 1 in [18] implies the error bound:

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \le 9\frac{\lambda_n^2 |S|}{\kappa^2} + \frac{\lambda_n}{\kappa}\left(2\tau^2 + 4\|\Theta^*_{S^c}\|_1\right). \tag{13}$$

The above gives a family of bounds, one for each choice of S. Decreasing $|S|$ potentially increases $\|\Theta_{S^c}\|_1 := \|\Theta_{S^c}\|_{1,1,1}$ and hence presents a trade-off. We choose an S that balances all the terms in the bound.

Specifically, we show that taking $\lambda_n = \mathcal{O}(\sqrt{\log(N^2 p)/n})$ is enough for (12) to hold with high probability. We then choose an S^* that solves (2), such that $|S^*| = s$ and $\|\Theta^*_{S^{*c}}\|_{1,1,1} = \sigma_s(\Theta^*)$; for which we can show (10) holds over $\mathbb{C}(S^*; \Theta^*)$ with high probability for $\kappa = \Omega(1)$, $\tau^2 = \sigma_s^2(\Theta^*)/s$. Putting these together proves Theorem 3.1.

4.1 Restricted Strong Convexity of $\mathcal{L}$

Proving RSC property (10) for a particular choice of S is a major contribution of our work. This is a nontrivial result since it involves uniformly controlling a dependent non-Gaussian empirical process. Proving the RSC is challenging even for i.i.d. samples, due the fact that the quantity to be controlled, namely, $\Delta \mapsto R\mathcal{L}(\Delta; \Theta^*)$, is a *random function* that needs to be uniformly controlled from below. Controlling the behavior of this function becomes significantly harder without the independence assumption. We establish the RSC property for the log-likelihood loss (4) under the MBP model (1) in the following:

Proposition 4.1 *Let $\sigma_s^2 = \sigma_s^2(\Theta^*)$. There exists a numerical constant $c_1 > 0$ such that if (8) holds, then, the RSC (10) holds with $\kappa = \min\left\{\frac{1}{4}c_f c_\ell^2, 1\right\}$, and $\tau^2 = \sigma_s^2/s$ for all tensors $\Delta \in \mathbb{C}(S^*; \Theta^*)$, with probability at least $1 - 2(N^2 p)^{-c_0 s}$. The constant $c_0 = \mathcal{O}(c_\ell^{-2})$ only depends on c_ℓ.*

To establish RSC in Proposition 4.1, we proceed as follows: **(i)** First we show that $R\mathcal{L}(\Delta; \Theta^*)$ is lower bounded by a quadratic function of the error tensor Δ

of the form $\mathcal{E}(\Delta;\mathbb{X}) := \frac{c_f}{n}\sum_{t=1}^{n}\sum_{k=1}^{N}\langle \Delta_{k**}, X^{t-1}\rangle^2$. (Lem. A.2). **(ii)** We show using Assumption (A1) that the population mean of $\mathcal{E}(\Delta;\mathbb{X})$ is strongly convex (Lem. A.4). **(iii)** In Lem. A.5, we show that for a fixed Δ, the random function $\mathcal{E}(\Delta;\mathbb{X})$ concentrates as,

$$\mathbb{P}\Big(\big|\mathcal{E}(\Delta;\mathbb{X}) - \mathbb{E}\mathcal{E}(\Delta;\mathbb{X})\big| > t\|\Delta\|_{2,1,1}^2\Big) \le 2e^{\frac{-nt^2}{G_f}} \quad (14)$$

by relying on the concentration result of [11] for the Lipschitz functions of discrete Markov chains (Prop. A.7). **(iv)** Finally, using a discretization argument, we uniformly bound $\mathcal{E}(\Delta;\mathbb{X})$ over all $\Delta \in \mathbb{C}$ (Lem. A.6). Proving uniform bounds of this form are challenging when the parameter space is not finite. The discretization step uses the estimates of the entropy numbers for absolute convex hulls of finite sets (Lem. A.8). These estimates are well-known in approximation theory and have been previously adapted to the analysis of regression problems in [22].

4.2 Upper bound on $\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty}$

To set λ_n such that (12) holds, we need to find an upper bound on $\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty}$. Since λ_n affects the error bound directly, we would like to choose the smallest λ_n that satisfies (12). In general, one would like to have a vanishing λ_n (as $n \to \infty$) to guarantee consistency. Our next result provides the necessary bound on the gradient of the loss, leading to a suitable choice for the regularization parameter λ_n:

Lemma 4.2 *For universal constants $c_2, c > 0$,*

$$\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty} \le \frac{c_2 L_f}{2\varepsilon}\sqrt{\frac{\log(N^2p)}{n}}. \quad (15)$$

with probability at least $1 - n^{-c}$. □

The proof uses Azuma–Hoeffding concentration inequality for martingale difference sequences [28] and can be found in Appendix A.2.

5 SIMULATIONS

We evaluate the performance of the estimator in (5) using simulated data. We use two different metrics of performance: (1) the estimation error in Frobenius norm, and (2) support recovery, i.e., assuming that the true parameter tensor is exactly s-sparse, how does the support estimated from $\widehat{\Theta}_{\mathrm{ML}}$ compare to the support of Θ^*. To do so, we need to estimate the support from $\widehat{\Theta}_{\mathrm{ML}}$. If we know the sparsity, we can estimate the support by taking the indices corresponding to the s largest entries of $\widehat{\Theta}_{\mathrm{ML}}$ in magnitude. If we do not know the sparsity in advance, we can estimate the support based on a threshold chosen by cross-validation. Given a threshold γ, the estimated support would be $\widehat{\mathrm{supp}}(\Theta) := \{(j,k,\ell) : |\widehat{\Theta}_{\mathrm{ML}_{jk\ell}}| \ge \gamma\}$. Note that our theoretical results do not give any guarantees for support recovery. In order to guarantee support recovery, a stronger result bounding the error uniformly for each entry of $\widehat{\Theta}_{\mathrm{ML}}$ is required, i.e., we need $|\hat{\Theta}_{\mathrm{ML}_{jk\ell}} - \Theta^*_{jk\ell}| \le \delta$ for all j,k, and ℓ with high probability. More work is needed to obtain theoretical guarantees for support recovery. Nevertheless, our simulations show that the estimator is able to recover the support very well.

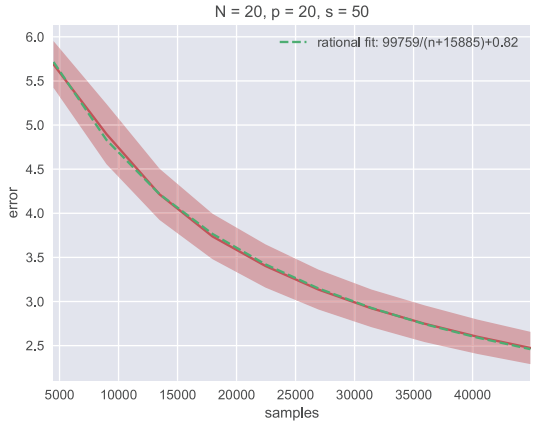
We first simulate a network with $N = 20$ units and $p = 20$ lags, i.e., the parameter space has dimension $N^2p = 8000$. We generate a uniformly random sparsity pattern over this space, and then generate the data using the parameter tensor. Next, we use this data to obtain $\widehat{\Theta}_{\mathrm{ML}}$ and compute the estimation error in Frobenius norm $\|\widehat{\Theta}_{\mathrm{ML}} - \Theta^*\|_F$. This process is repeated for 20 independent runs.

Figure 1a shows how the error changes with the sample size. The shaded area represents one standard deviation of estimation error over 20 runs. For comparison, a rational function fit of the form $a + b/(n + c)$ is also plotted on top of the error. In Figure 1b, we plot the error vs. sparsity for a fixed sample size. The error grows almost linearly with sparsity. A linear fit of the error is also shown for comparison. Figure 1c shows the average error over the runs for different sparsity levels and different sample sizes. As expected, the error goes down as Θ becomes sparser, or sample size n increases.

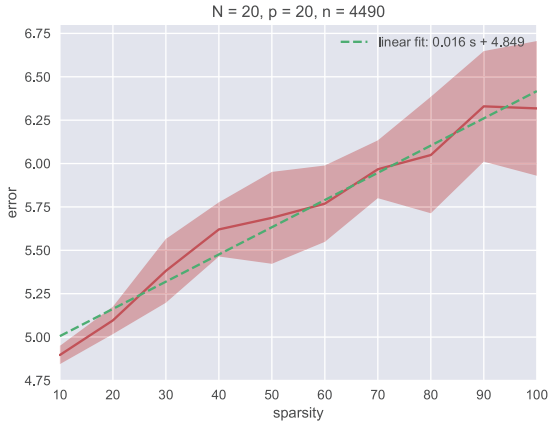
Finally, the support recovery performance is shown in Figure 1d. In this experiment, the network has $N = 100$ units but only $p = 1$ lag for $s = 10, 100, 300$. For recovering the support, we assumed that the sparsity s is known, and took the indices corresponding to the s largest entries of $\widehat{\Theta}_{\mathrm{ML}}$ as the recovered support. The fraction of the correctly recovered indices is plotted against the sample size. The figure shows that if the sample size is below some threshold, no entries of the support are recovered, while above the threshold, the recovered fraction gradually increases to 1.

6 DISCUSSION

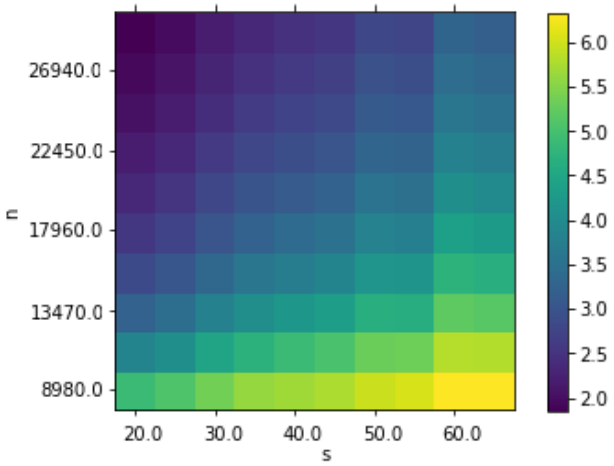
We analyzed a sparse multivariate Bernoulli process which evolves as an autoregressive GLM. This model provides a framework for identifying the spiking characteristics and the underlying network of an ensemble of neurons from a finite sample of their spike trains. The conditional probability of spiking for each neuron is modeled as a GLM with a link function, based on the history of the multivariate process. The work extends the model considered by [10] to the multi-



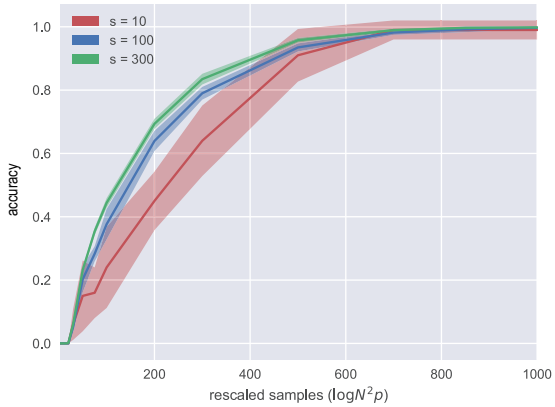
(a) Error vs. sample size for sparsity $s = 50$.



(b) Error vs. sparsity for sample size $n = 4490$



(c) Average Frobenius norm error over 20 runs for $N = 20$ and $p = 20$ lags. Each pixel corresponds to a sample size n and a sparsity level s.



(d) Fraction of support size recovered using s largest entries of $\hat{\Theta}$ as the estimator of support, for $N = 100$, $p = 1$.

Figure 1: Simulation results.

dimensional setting and provides the first result in the high-dimensional regime for this model.

Some key features of our work are as follows: We provide the first rigorous analysis of the regularized maximum likelihood estimator of the MBP model (1) with (i) an approximately sparse parameter, (ii) a Lipschitz, log-convex non-linear inverse link function, and (iii) for the general $p \geq 1$, $N \geq 1$ case. We proved that the estimator achieves the fast rate $\mathcal{O}(s \log(N^2 p)/n)$ under exact sparsity, and the slow rate $\mathcal{O}(\widetilde{\tau}_s^2(\Theta^*)\sqrt{\log(N^2 p)/n})$ under approximate sparsity, where $\widetilde{\tau}_s(\Theta^*)$ is a measure of the ℓ_1 approximation error.

Our error bounds are valid under fairly general assumptions on the stability and wide-sense stationarity of the process. We bring out key properties of the nonlinear link function that can provide control on the estimation error and the sample complexity. We also identify a new norm on the true parameter Θ^*, defined in (6), that affects the error bounds. These ideas provide potential directions for identifying structural properties of the true parameter that can enhance estimation. For example, we demonstrate that when the parameters decay with the lag, i.e., $|\Theta^*_{ij\ell}| = \mathcal{O}(\ell^{-\alpha})$ or $\mathcal{O}(e^{-\beta\ell})$, a $\mathcal{O}(p^3)$ improvement can be achieved. Our result in general requires a sample complexity of $n = \Omega(s^3 \log(N^2 p))$, allowing a high-dimensional scaling in (N, p).

Our analysis for establishing the RSC is novel and the proof techniques can be applied to any multivariate processes over discrete countable spaces with long range dependencies. Mixing properties of higher-order Markov chains are also discussed during the proof which may be of independent interest.

Acknowledgements. P. Pandit, M. Sahraee and A.K. Fletcher were supported in part by the National Science Foundation under Grants 1738285 and 1738286 and the Office of Naval Research under Grant N00014-15-1-2677. S. Rangan was supported in part by the National Science Foundation under Grants 1116589, 1302336, and 1547332, and the industrial affiliates of NYU WIRELESS.

References

[1] Daniel Felix Ahelegbey, Monica Billio, and Roberto Casarin. Sparse graphical vector autoregression: A bayesian approach. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (123/124):333–361, 2016.

[2] Sumanta Basu, George Michailidis, et al. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[3] Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456, 2004.

[4] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[5] T Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[6] Yanira Guanche, Roberto Mínguez, and Fernando J Méndez. Autoregressive logistic regression applied to atmospheric circulation patterns. *Climate dynamics*, 42(1-2):537–552, 2014.

[7] Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

[8] Dimitrios Katselis, Carolyn Beck, and R Srikant. Mixing times and structural inference for bernoulli autoregressive processes. *IEEE Transactions on Network Science and Engineering*, 2018.

[9] Abbas Kazemipour. Compressed sensing beyond the iid and static domains: Theory, algorithms and applications. *arXiv preprint arXiv:1806.11194*, 2018.

[10] Abbas Kazemipour, Min Wu, and Behtash Babadi. Robust estimation of self-exciting generalized linear models with application to neuronal modeling. *IEEE Transactions on Signal Processing*, 65(14):3733–3748, 2017.

[11] Leonid Aryeh Kontorovich, Kavita Ramanan, et al. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

[12] Rasoul Kowsar, Behrooz Keshtegar, Mohamed A Marey, and Akio Miyamoto. An autoregressive logistic model to predict the reciprocal effects of oviductal fluid components on in vitro spermophagy by neutrophils in cattle. *Scientific Reports*, 7(1):4482, 2017.

[13] Ben Mark, Garvesh Raskutti, and Rebecca Willett. Network estimation via poisson autoregressive models. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2017 IEEE 7th International Workshop on*, pages 1–5. IEEE, 2017.

[14] Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Network estimation from point process data. *arXiv preprint arXiv:1802.04838*, 2018.

[15] David S Matteson, Mathew W McLean, Dawn B Woodard, Shane G Henderson, et al. Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B):1379–1406, 2011.

[16] Timothy L McMurry, Dimitris N Politis, et al. High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9(1):753–788, 2015.

[17] Jonathan Mei and José MF Moura. Signal processing on graphs: Causal modeling of unstructured data. *IEEE Trans. Signal Processing*, 65(8):2077–2092, 2017.

[18] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, pages 538–557, 2012.

[19] Murat Okatan, Matthew A Wilson, and Emery N Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17(9):1927–1961, 2005.

[20] Maxim Raginsky, Rebecca M Willett, Corinne Horn, Jorge Silva, and Roummel F Marcia.

Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8):5544–5562, 2012.

[21] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[22] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[23] Garvesh Raskutti, Ming Yuan, and Han Chen. Convex regularization for high-dimensional multi-response tensor regression. *arXiv preprint arXiv:1512.01215*, 2015.

[24] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[25] Anne C Smith and Emery N Brown. Estimating a state-space model from point process observations. *Neural computation*, 15(5):965–991, 2003.

[26] Richard Startz. Binomial autoregressive moving average models with an application to us recessions. *Journal of business & economic statistics*, 26(1):1–8, 2008.

[27] James W Taylor and Keming Yu. Using auto-regressive logit models to forecast the exceedance probability for financial risk management. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(4):1069–1092, 2016.

[28] Sara A van de Geer. On hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Springer, 2002.

[29] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[30] Alison I Weber and Jonathan W Pillow. Capturing the dynamical repertoire of single neurons with generalized linear models. *Neural computation*, 29(12):3260–3289, 2017.

[31] Cun-Hui Zhang, Jian Huang, et al. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[32] Hao Zhou and Garvesh Raskutti. Non-parametric sparse additive auto-regressive network models. *arXiv preprint arXiv:1801.07644*, 2018.