Title

Non-asymptotic Analysis of Learning Long-range Autoregressive Generalized Linear Models for Discrete High-dimensional Data

Author
Pandit, Parthe

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Non-asymptotic Analysis of Learning Long-range Autoregressive Generalized Linear Models for Discrete High-dimensional Data

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Parthe Pandit

2021

ABSTRACT OF THE THESIS

Non-asymptotic Analysis of Learning
Long-range Autoregressive Generalized Linear Models
for Discrete High-dimensional Data

by

Parthe Pandit

Master of Science in Statistics
University of California, Los Angeles, 2021
Professor Alyson Fletcher, Chair

Fitting multivariate autoregressive (AR) models is fundamental for analysis of time-series data in a wide range of applications in science, engineering, econometrics, signal processing, and data-science. This dissertation considers the problem of learning a p-lag multivariate AR generalized linear model (GLM). In this model, the state of the time-series at each time step, conditioned on the history, is drawn from an exponential family distribution with the mean parameter depending on a linear combination of the last p states. The problem is to learn the linear connectivity tensor from a single observed trajectory of the time-series. We provide non-asymptotic error bounds on the regularized Maximum Likelihood estimator in high dimensions.

We focus on the sparse tensor setting, which arises in applications where there exists a limited number of direct connections between variables. For such problems, ℓ_1-regularized maximum likelihood estimation (or M-estimation more generally) is often straightforward to apply and works well in practice. The M-estimator can be posed as a convex optimization problem and hence can also be solved efficiently.

However, the statistical analysis of such methods is difficult due to the feedback in the

state dynamics and the presence of a non-linear link function, especially when the underlying process is non-Gaussian. Our main result in Chapter 3 provides a bound on the mean-squared error of the estimated connectivity tensor as a function of the sparsity and the number of samples, for a class of discrete multivariate AR(p) GLMs, in the high-dimensional regime. Importantly, the bound indicates that, with sufficient sparsity, consistent estimation is possible in cases where the number of samples is significantly less than the total number of free parameters.

Towards proving the main result, we present a general framework to establish the Restricted Strong Convexity (RSC) property for time-averaged loss functions often seen in time-series analysis. We also derive new concentration inequalities of functions of discrete non-Markovian random variables. These intermediate results may be of independent interest to the reader.

The thesis of Parthe Pandit is approved.

Arash Ali Amini

Lin Yang

Alyson Fletcher, Committee Chair

University of California, Los Angeles

2021

To my parents

Contents

List of Figures

List of Tables

Acknowledgements

This dissertation has been possible in large part due to the endless support from my advisors, mentors, teachers, colleagues, family, and friends.

To start off, I would like to thank my advisor Professor Alyson Fletcher, who gave me the opportunity to pursue graduate studies in machine learning at a vibrant place such as UCLA. I enjoyed a lot of intellectual autonomy working as her student, and her advise always kept me from getting stuck in local minimas. Without her awareness and encouragement, this articulated MS degree in Statistics would not have been possible.

It is hard to overstate the influence Professor Arash Amini has had in my training as a statistician. His acute attention to detail always provided me with clarity of thought. Conversations with him have been the most intellectually engaging during my time at UCLA.

I would also like to thank Professor Lin Yang for agreeing to be on my committee and for providing his valuable input. I am also very grateful to have received advice from Professors Raghu Meka and Alexander Sherstov who provided a sandbox for learning advanced concepts in theoretical computer science. Professor Suhas Diggavi's class on Information Theory helped me overcome my fear of probability, which nudged me to take up statistics and machine learning. I would also like to thank Professor Phillip Schniter for being an amazing collaborator and for making NeurIPS 2018 a far less daunting experience.

It was my utmost fortune to be guided by Professor Sundeep Rangan, whose technical prowess never ceases to amaze me. I hope that one day I can start seeing patterns in seemingly unrelated problems like him. His ability to make bold predictions and then prove them using the most elementary mathematical tools makes research look so simple.

I was lucky to have as mentors Professors Sam Coogan and Ankur Kulkarni who gave me the confidence to take up problems that I found interesting and make them into a cohesive and coherent story. The fundamental concepts and visualization techniques I learned from them has made it so much easier to understand and interpret a broad range ideas I encounter.

UCLA has an amazing Mathematics department which I greatly benefited from. Math 275 and 273 series taught by Professors Jun Yin and Wotao Yin respectively were crucial in my understanding of the landscape of problems considered in machine learning. Attending the Level Set Collective meetings every Tuesday afternoon at IPAM was, in hindsight, the best passive investment of my time, and introduced me to optimal control, optimal transport, and operator splitting. The reading group run by Professors Liza Rebrova and Palina Salanevich exposed me to Graph Signal Processing. I thank them all.

My internship experiences were amazing due to amazing mentors Sumeet Katariya, Nikhil Rao, Sheng Zha, He He, and Hua Zheng. Their guidance enriched my training as a researcher.

I acknowledge the financial support from the Gurukrupa foundation, the J. N. Tata endowment, and the K. C. Mahindra foundation.

My colleagues, coauthors, and compadres: Mojtaba Sahraee-Ardakan and Melikasadat Emami have provided a great deal of support through what could have been a much lonelier experience. I have learned a lot by being in their presence. It was also great to be among PhD students like Michael, Amber, Sundar, Vijay, Srini, Tsang-Kai, Arsalan, and Vignesh, to have long-drawn discussions over several trips to Kerchoff, Synapse, and Southern Lights.

I was lucky to have brilliant roommates like Pratik Sathe and Mihir Laghate, who provided me the best home-away-from-home. My friends Navjot, Akshay, Shantanu, Janaki, and Pratik Chaudhari made LA more like Mumbai. I am extremely fortunate to have kind friends like Arpit, Niladri, Shantanu, Saurabh, and Vighnesh. I am greatly indebted to Chaphy and Vidya-aaji, whose unbounded affection always made me feel secure in Southern California. Rosie, who I inherited from Vidya-aaji, made life in LA far more enjoyable.

CAPS and the Ashe Center helped cope with the struggles of life as an international graduate student. I shall always remember the impact that the staff, program volunteers, and the group members I interacted with, had on my health and well being.

My partner Harini Alladi has been a constant source of inspiration and joy. I have grown a lot personally and professionally because of her. I am grateful that she is a part of my life.

Finally, I would like to thank my parents, and my sister for their unwavering support.

Chapter 1

Introduction

Statistical analysis of high-dimensional time series data has several applications in science and engineering. Linear parametric models such as Autoregressive (AR) models provide a simple, interpretable and yet, powerful baseline for such applications. This dissertation provides new results regarding the sample complexity of learning the parameter of high-dimensional AR models with long-term dependencies.

1.1 Forecasting problem

AR models generalize linear models to the context of time series analysis. In the standard linear model commonly studied in statistics a response is postulated to be an affine function of the covariates. In a time series we are often interested in the following forecasting problem:

Given the history of a time series $(\boldsymbol{x}^t, \boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \ldots)$*, predict the next state* $\boldsymbol{x}^{t+1}$.

Here, the covariates are the history of the time series whereas the response is the next-state. The AR model predicts the next state as an affine function of the history of the time series.

The results presented in this dissertation appeared in [34, 31, 33]. The papers [31, 33] presented the case of learning a sparse Bernoulli AR(p) process, whereas [34] extended these results to general GLMs over discrete valued variables with the presence of a dictionary. These works are coauthored with Mojtaba Sahraee-Ardakan, Arash A. Amini, Sundeep Rangan and Alyson Fletcher.

While this may be suited for some Gaussian-like continuous valued signals, for more structured signals, a natural extension to the linear model is to consider a Generalized Linear Model (GLM), i.e., the prediction is a nonlinear function of an affine function of the covariates. These models are also well studied under the name Linear Predictive Coding (LPC) and have been influential in the development of early speech processing and communication technology.

An immediate question is how much of the history of the state is relevant to predict the next state of the time series. The AR(p) model predicts $\boldsymbol{x}^t$ using $(\boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \ldots, \boldsymbol{x}^{t-p})$. This creates a hierarchy of models, and number of lags p needs to be selected appropriately[1]. The hyperparameter p depends on the sampling resolution of the time series as well. In this work, we assume p is known and fixed, and we are interested in understanding the sample complexity of estimating the model parameter as p scales.

The learning problem is then to estimate the weights of the affine function of the GLM described above. For an appropriately chosen GLM, this can be posed as an unconstrained convex optimization problem and hence can be solved efficiently using off-the-shelf algorithms. This dissertation concerns the consistency of this M-estimator for a large class of multivariate AR(p)-GLMs.

Non-asymptotic analysis for the M-estimators of linear models and GLMs is by now a well studied problem in high dimensional statistics. Most consistency results in this area rely on the concentration phenomena of empirical processes. M-estimators in the context of time-series however, have loss functions which consist of empirical processes with temporal dependence between samples. This makes analysis extremely challenging since concentration phenomena are either unknown or hard to establish. This is far more complicated in the non-Markovian setting which we consider in this dissertation.

[1] A higher p leads to a more powerful model that is costlier to estimate - both statistically and computationally.

1.2 Organization of the dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents the general model that we will work with. Chapter 3 presents the main result regarding learning this model, followed by a discussion about its consequences, and a sketch of its proof. In Chapter 4 we present the technique to establish the restricted strong convexity (RSC) condition for loss functions which are time-averages of an empirical process, commonly seen in M-estimators in time-series analysis; whereas in Chapter 5 we present new techniques for deriving concentration inequalities for dependent multivariate processes. Chapter 6 provides numerical simulations that corroborate our theoretical predictions. Finally, Chapter 7 concludes the dissertation and lays out some open questions.

1.3 Notation

For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write either of $a_n \gtrsim b_n$ or $b_n \lesssim a_n$ or $b_n = O(a_n)$ or $a_n = \Omega(b_n)$, to mean that there is a constant $C > 0$ such that $a_n \geq Cb_n$ for all n. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$. We write either of $a_n \gg b_n$ or $b_n \ll a_n$ or $b_n = o(a_n)$, if $b_n/a_n \to 0$ as $n \to \infty$. Table 1.1 provides a list of all other notations used in the dissertation.

Table 1.1: List of notations used in the dissertation.

$[m]$	the set $\{1, 2, \ldots, m\}$	defined for any $m \in \mathbb{N}$
n	number of samples	
N	number of variables	
p	number of lags	
L	number of filters	
s_i	in-degree of variable i	var i depends on at most s_i vars
i, j	(index) of variable $\in [N]$	
k	(index) of lag $\in [p]$	
ℓ	(index) of filter $\in [L]$	
t	(index) of sample/time $\in \mathbb{Z}$	
$\mathcal{X}_i$	discrete subset of $\mathbb{R}$	examples: $\{0, 1\}, [m], \mathbb{N}, \mathbb{Z}, \mathbb{Q}$
$\mathcal{X}_i^{\times p}$	discrete subset of $\mathbb{R}^{1 \times p}$	$\left\{ \begin{bmatrix} x_1 & \ldots & x_p \end{bmatrix} \mid x_k \in \mathcal{X}_i, \forall\, k \in [p] \right\}$
$\mathcal{X}$	discrete subset of $\mathbb{R}^N$	$\prod_{i \in [N]} \mathcal{X}_i$
$\mathcal{X}^{\times p}$	discrete subset of $\mathbb{R}^{N \times p}$	$\prod_{i \in [N]} \mathcal{X}_i^{\times p}$
x_i^t	(scalar) $\in \mathcal{X}_i$, variable i at time t	
$\boldsymbol{x}^t$	(vector) $\in \mathcal{X}$, state at time t	$\begin{bmatrix} x_1^t & x_2^t & \ldots & x_N^t \end{bmatrix}^\top \in \mathcal{X}$
$\mathbf{x}_i^{t-*}$	(vector) $\in \mathcal{X}_i^{\times p}$, history of variable i	$\begin{bmatrix} x_i^{t-1} & x_i^{t-2} & \ldots & x_i^{t-p} \end{bmatrix} \in \mathcal{X}_i^{\times p}$
$\mathbf{X}^t$	(matrix) $\in \mathcal{X}^{\times p}$, p-lag history at time t	$\begin{bmatrix} \boldsymbol{x}^t & \boldsymbol{x}^{t-1} & \ldots & \boldsymbol{x}^{t-p+1} \end{bmatrix} \in \mathcal{X}^{\times p}$
$\boldsymbol{d}_\ell$	(vector) $\in \mathbb{R}^p$, filter with p lags	
$\mathbf{D}$	(matrix) $\in \mathbb{R}^{p \times L}$, dictionary of L filters	$\begin{bmatrix} \boldsymbol{d}_1 & \boldsymbol{d}_2 & \ldots & \boldsymbol{d}_L \end{bmatrix}$
Θ	(tensor) parameter $\in \mathbb{R}^{N \times N \times L}$	same as $\Theta_{[N]}$ (c.f. $\Theta_{\mathcal{U}}$ below)
Θ_i	(matrix) $\in \mathbb{R}^{N \times L}$	parameter for variable i
$\mathcal{U}$	subset of variables $\subseteq [N]$	
$\Theta_{\mathcal{U}}$	($\lvert\mathcal{U}\rvert \times N \times L$ tensor) $(\Theta_i)_{i \in \mathcal{U}}$	
$\\|\boldsymbol{M}\\|$	(norm) Frobenius norm of tensor $\boldsymbol{M}$	`sum(M**2)**0.5` in Python
$\\|M\\|_{q,r}$	(norm) of $M \in \mathbb{R}^{B \times C}$, $q, r \geq 1$	$\left(\sum_{b=1}^{B} \left\{ \sum_{c=1}^{C} \lvert M_{bc}\rvert^r \right\}^{\frac{q}{r}} \right)^{\frac{1}{q}}$
$\\|\boldsymbol{M}\\|_{p,q,r}$	(norm) of $\boldsymbol{M} \in \mathbb{R}^{A \times B \times C}$, $p, q, r \geq 1$	$\left(\sum_{a=1}^{A} \\|M_a\\|_{q,r}^p \right)^{\frac{1}{p}}$

Chapter 2

Multivariate Autoregressive Generalized Linear Models

We consider the problem of learning a p-lag AR(p) generalized linear model (GLM) for a multivariate time series involving N-variables: $\boldsymbol{x}^t = (x_i^t) \in \mathbb{R}^N$, where $x_i^t \in \mathcal{X}_i \subseteq \mathbb{R}$ for all $i \in [N]$, $t \in \mathbb{Z}$. A particular case of the model we consider is of the form,

$$x_i^t \mid z_i^t \sim \mathbb{Q}_i(z_i^t), \quad z_i^t = f_i\left(\langle \Theta_i^*, \mathbf{X}^{t-1} \rangle\right), \tag{2.1}$$

where the inner product corresponds to $\mathbb{R}^{N \times p}$, for $t = 1, 2, \ldots$ and $i = 1, 2, \ldots, N$ where $\mathbf{X}^{t-1} = [\boldsymbol{x}^{t-1} \; \boldsymbol{x}^{t-2} \; \ldots \; \boldsymbol{x}^{t-p}] \in \mathbb{R}^{N \times p}$ is the p-lag history of the process up to time $t-1$, and $\mathbb{Q}_i(z_i^t)$ is a probabilistic link function. The problem is to estimate, for $i = 1, 2, \ldots, N$, the unknown parameter $\Theta_i^* \in \mathbb{R}^{N \times p}$, which governs the influence of the process over the next state of variable i, from an observation of n time steps $\boldsymbol{x}^t$, $t = 1, \ldots, n$. The conditional distributions $\mathbb{Q}_i(\cdot)$ and link functions f_i are assumed to be known.

Modeling problems of this form appear in a wide-range of applications with time-series data. For example, in neural modeling, $\boldsymbol{x}^t$ can represent a vector of spike counts or some other measure of activity from N neurons or brain regions, as modelled by [19, 47]. In this case, estimation of the tensor Θ^* in equation (2.1) can provide insight into the neural connectivity. Other applications include genomics, econometrics [10], data science, sociology, business management, financial markets [44, 11] and natural language processing.

A key challenge in estimating the multivariate AR(p) models is the large number of unknown parameters, particularly as the dimension of the process, N, and number of time lags, p, grows. However, in many cases, one can assume some sparsity constraint in the connectivity tensor Θ^*. For example, in neural modeling, there are physically limited numbers of direct connections between brain regions. Under a sparsity assumption, it is common to estimate Θ^* via an ℓ_1-regularized M-estimator of the form,

$$\widehat{\Theta} := \operatorname*{argmin}_{\Theta \in \mathbb{R}^{N \times N \times p}} \frac{1}{n} \sum_{i=1}^{N} \sum_{t=1}^{n} \mathcal{L}_{it}\left(x_i^t\,;\, \langle \Theta_i, \mathbf{X}^{t-1} \rangle\right) + \lambda_n \|\Theta\|_{1,1,1}, \tag{2.2}$$

where $\mathcal{L}_{it} : \mathcal{X}_i \times \mathbb{R} \to \mathbb{R}$ are loss functions and $\lambda_n \|\Theta\|_{1,1,1}$ is an ℓ_1 regularizer (precise definitions will be given in the Section 2.3 below). The broad goal of this dissertation is to analyze the sample complexity of such ℓ_1-regularized M-estimators. That is, given a sparsity constraint on Θ^*, and the number of measurements, n, how well can we estimate Θ^*?

2.1 Key contributions

We consider the case where $\{\mathcal{X}_i\}_{i=1}^N$ are bounded countable subsets of $\mathbb{R}$. We analyze the ℓ_1-regularized M-estimator equation (2.2) when the loss functions $v \mapsto \mathcal{L}_{it}(u; v)$ are strongly convex, for all $u \in \mathcal{X}_i$. We assume that the connectivity tensor can be approximated by a sparse tensor with at most $s_{\max}$ non-zero values in each slice Θ_i^*. Under these assumptions, our main result in Theorem 1, stated later in Chapter 3, establishes the consistency of the regularized M-estimator equation (2.2) in the high-dimensional regime of $n = \mathsf{poly}\big(s_{\max} \log(N^2 p)\big)$ under some regularity conditions.

In proving our main result, we establish the so-called restricted strong convexity (RSC) [29] for a large class of loss functions, for a dependent non-Gaussian discrete-valued multivariate process. Our proof of the RSC property requires showing a restricted eigenvalue condition, which is nontrivial due to the non-Gaussian and highly-correlated nature of the 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 more challenging when going from $p = 1$ to even $p = 2$. The difficulty with this *higher-order* Markov chain is that its *Dobrushin contraction coefficient* is trivially 1. We develop techniques to get around this issue which could be of independent interest (cf. Section 5). Our techniques hold for all $p \geq 1$.

Much of the previous work towards proving the RSC condition has either focused on the independent sub-Gaussian case [37, 48] or the dependent Gaussian case [2, 38] for which powerful Gaussian concentration results such as the Hanson–Wright inequality [40] are still available. Our approach is to use concentration results for Lipschitz functions of Markov chains over countable spaces, 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 estimation [36]. Moreover, our approach allows us to identify key properties of the model that allow for sample-efficient estimation.

Although discrete time series are often modeled using the specific link functions such as `logit` or `softmax`, our result allows more flexibility to choose the link functions. For example in the Bernoulli AR(p) and Truncated-Poisson AR(p) cases discussed in Section 3.5, any Lipschitz continuous, log-convex link function can be used. The analysis also brings out crucial properties of the link function, and the role it plays in determining the estimation error and sample complexity.

Our model also allows for each individual time series x_i^t to lie in distinct spaces $\mathcal{X}_i$ which is desirable in practical applications with heterogeneous types of data.

2.2 Previous work

There is a vast literature on recovering sparse vectors in under-sampled settings [7, 6, 12, 13]. The generic results show that if a vector θ is s-sparse in a p-dimensions, it can be estimated in $n = \Omega(s \log(p))$ measurements. However, these results typically do not have feedback as in the AR process considered here.

The estimation of sparse Gaussian VAR(p) processes with linear feedback has been considered only more recently [2, 5, 27, 28, 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 [40]. These techniques do not extend to non-Gaussian setups.

In the non-Gaussian setting, Hall et al. [16, 49] recently considered a multivariate time series evolving as a GLM driven by the history of the process similar to our model. The Bernoulli AR(1) and Poisson AR(1) with $p = 1$ lags were considered as special cases of this model. They provide statistical guarantees on the error rate for the ℓ_1 regularized estimator. More importantly, their results are restricted to the case $p = 1$ which does not allow the explicit encoding of long-term dependencies. More recently, Mark et al. [25, 24] considered a model closer to ours for multivariate AR(p) processes with lags $p = 1$ or $p = 2$.

A key contribution of ours is to bring out the explicit dependence on p in the AR(p) models, allowing for a general $p \geq 1$. In the special cases we consider: the Bernoulli AR(p) and the Truncated-Poisson AR(p), 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 f_i and a certain norm of the parameter tensor.

Our results improve upon those in [16, 25] when applied to the Bernoulli AR(p) and Truncated-Poisson AR(p). Due to the key observation that an AR(p) over a countable space can be viewed as a higher order Markov chain, our analysis relaxes several assumptions made by [16, 25]. 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 [21].

The univariate Bernoulli AR(p) process for $p \geq 1$ was considered by Kazemipour et. al. [18, 19] where they analyzed a multilag Bernoulli process for a single neuron. Their analysis does not extend to the $N > 1$ case. Even for $N = 1$, their analysis is restricted to the biased process with $\mathbb{P}(x_1^t = 1 | \mathbf{X}^{t-1}) < \frac{1}{2}$ for all t. Mixing times of the Bernoulli AR(1) have been considered in [17]. However, their discussion is again limited to $p = 1$.

2.3 Models and methods

To state our results in their full generality, we consider a slightly more general model than equation (2.1). We assume that the multivariate time series $\boldsymbol{x}^t = (x_i^t) \in \mathcal{X} \subset \mathbb{R}^N$ evolves as,

$$x_i^t \mid z_i^t \sim \mathbb{Q}_i(\,\cdot \mid z_i^t) \tag{2.3a}$$

$$z_i^t = f_i\big(\, \langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \rangle_{\mathbb{R}^{N\times L}} \,\big) \tag{2.3b}$$

$$x_i^t \perp\!\!\!\perp x_j^t \mid \boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \ldots \tag{2.3c}$$

for $t = 1, 2, \ldots$ and $i = 1, 2, \ldots, N$. The key difference here is that we have added a matrix $\mathbf{D} = [\boldsymbol{d}_1 \; \boldsymbol{d}_2 \; \ldots \; \boldsymbol{d}_L] \in \mathbb{R}^{p \times L}$, a known dictionary of filters $\{\boldsymbol{d}_\ell\}_{\ell=1}^L$. When $\mathbf{D} = I_{p\times p}$, we obtain the special case equation (2.1). The role of this dictionary will be explained below. To model the discrete-valued nature of the states, we assume that $\boldsymbol{x}^t \in \mathcal{X} := \prod_{i=1}^N \mathcal{X}_i$ where each $\mathcal{X}_i$ is a bounded countable subset of $\mathbb{R}$. The matrix $\mathbf{X}^{t-1} = [\boldsymbol{x}^{t-1} \; \boldsymbol{x}^{t-2} \; \ldots \; \boldsymbol{x}^{t-p}] \in \mathbb{R}^{N\times p}$ is the p-lag history of the process up to time $t-1$, and $\mathbb{Q}_i(\,\cdot \mid z)$ is a distribution on $\mathcal{X}_i$ parameterized by z. For example an exponential family distribution with mean parameter z. The matrices $\Theta_i^* \in \mathbb{R}^{N\times L}$, $i \in [N]$ are the (unknown) model parameters and $\langle \cdot, \cdot \rangle_{\mathbb{R}^{N\times L}}$ is the inner product. A process of this form will be denoted GVAR(p).

The distribution $\mathbb{Q}_i(\,\cdot \mid z_i^t)$ represents the conditional distribution of x_i^t given the past $\boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \ldots$. Functions $f_i : \mathbb{R} \to \mathbb{R}$ are similar to the inverse-link functions in GLMs, and can be nonlinear in general. It is worth noting that $\mathcal{X}_i$ and $\mathbb{Q}_i$ can vary for every variable $i \in [N]$ making the model extremely flexible to include heterogeneous types of discrete data.

The inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^{N \times L}}$ in equation (2.3) is the Hilbert-Schmidt inner product on $\mathbb{R}^{N \times L}$, and can be expanded as:

$$\left\langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \right\rangle_{\mathbb{R}^{N \times L}} = \sum_{j=1}^{N} \sum_{\ell=1}^{L} \Theta_{ij\ell}^* \left\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \right\rangle_{\mathbb{R}^p} \tag{2.4}$$

where $\mathbf{x}_j^{t-*} := [x_j^{t-1} \; x_j^{t-2} \; \ldots \; x_j^{t-p}]$ is the p-lag history of variable j up to time $t-1$, i.e., the j^{th} row of $\mathbf{X}^{t-1}$. Note that $(\mathbf{X}^{t-1}\mathbf{D})_{j\ell} = \left\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \right\rangle_{\mathbb{R}^p}$. The parameter $(\Theta_i^*)_{j\ell} = \Theta_{ij\ell}^* \in \mathbb{R}$ captures the dependence of variable x_i^t on the past activity of variable j, via $\mathbf{x}_j^{t-*}$. The vectors $\boldsymbol{d}_\ell \in \mathbb{R}^p$ act as filters that modulate the mean of variable x_i^t based on the past activity of all the variables, that is, x_j^k for $j \in [N]$, and $t - p \leq k < t$.

2.4 Dictionary and Network interpretations

The filters $\{\boldsymbol{d}_\ell\}$ serve two main purposes: (i) interpretability and (ii) dimension reduction. For example, in neuroscience applications where the types of spiking behaviors are limited, the presence of a dictionary causes the model to favor specific forms of interactions between the spiking activities of two neurons. We refer to [47] which explores these filters for various interactive behaviors among neurons such as bursting, tonic spiking, phasic spiking, etc. The dictionary increases the interpretability of the parameter Θ_i^*—one interprets $(\Theta_i^*)_{j\ell}$ as measuring the effect of the activity of neuron i on neuron j, as explained by interaction type ℓ. Thus, the sparsity of Θ_i^* is more meaningful in the presence of a dictionary. An earlier version of this work [32] considered modeling the interaction with the past as $\langle \Theta_i^*, \mathbf{X}^{t-1} \rangle$ where Θ_i^* lies in $\mathbb{R}^{N \times p}$, corresponding to taking $\mathbf{D} = I_{p \times p}$, the identity matrix, in equation (2.3c). The formulation with a general dictionary $\mathbf{D}$ has the added advantage of potentially reducing the number of free parameters from Np to NL. When $L \ll p$, this leads to a massive dimension reduction. The bilinear term $\langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \rangle_{\mathbb{R}^{N \times L}} = \left\langle \Theta_i^* \mathbf{D}^\top, \mathbf{X}^{t-1} \right\rangle_{\mathbb{R}^{N \times p}}$ can also be thought of as a low-rank approximation to the parameter, forcing one factor to be fixed by $\mathbf{D}$. By adding pre-existing knowledge of temporal interactions between variables, the dictionary allows for a rich model with fewer parameters, leading to more (sample) efficient estimators for Θ^*.

The parameter Θ^* can be interpreted as representing a network among variables x_i^t, $i \in [N]$. A slice $\Theta_{**\ell}$ can be thought of as an adjacency matrix for the *influence network* explained by coupling behaviour ℓ. If neurons i and j are not connected, then $\Theta_{ij\ell} = 0$ for all $\ell \in [L]$. For example, in the neural spike train application, one can reveal a latent network among the neurons (i.e., who influences whose firing) just from the observations of patterns of neural activity, a task which is of significant interest in neuroscience [30, 43, 4]. Similarly, in the context of social networks, one might be interested in who is influencing whom [35].

2.5 Example Processes

The GVAR(p) process of the form equation (2.3) can be applied in a wide range of applications. For example, letting $\mathbb{Q}_i(\,\cdot\,|\,z) = \mathrm{Ber}(z)$ and $f_i(u) = (1 + e^{-u})^{-1}$ recovers the Bernoulli autoregressive process in [32]. Similarly, $\mathbb{Q}_i(\,\cdot\,|\,z) = \mathrm{Binomial}(K_i, z)$ and $f_i(u) = (1 + e^{-u})^{-1}$ models a Binomial process with K_i trials (for coordinate i) and success probability z. Such a model can be suitable for modeling count data. Another common model for point processes in neuroscience [43] is the Truncated-Poisson autoregressive process given by $\mathbb{Q}_i(\,\cdot\,|\,z) = \mathbb{P}(\min(M_i, Z) \in \,\cdot\,)$ where $Z \sim \mathrm{Poi}(z)$, and $f_i(u) = \exp(u)$ or $f_i(u) = \log(1 + e^u)$ for some integer M_i [16, 25]. Although we focus on single-parameter discrete distributions in this dissertation, the ideas can be easily extended to distributions with multiple parameters. For example, one can construct a categorical or multinomial process, by allowing z_i^t to be vector-valued and taking f_i to be the `softmax` function.

Chapter 3

Learning Sparse AR-GLMs in High dimensions

We are primarily interested in parameter estimation in the high-dimensional regime where $n \ll N$, i.e., when the number of samples n are far fewer than the number of variables N. To make the estimation feasible, we assume that variable i depends on the past values of only a few number of variables, $s_i \ll N$. We refer to s_i as the *in-degree* of variable i. Our main result provides sufficient conditions under which the parameter Θ^* can be estimated in the high-dimensional setting where $n = \mathsf{poly}(\{s_i\}_{i=1}^N, \log(NLp))$. Recall that n is the number of samples, p is the number of lags in the AR model, L is the number of filters in the dictionary, and N is the total number of variables.

Chapter Organization: Section 3.1 formally describes the regularized M-estimator. Section 3.3 provides the result which holds under the assumptions stated in Section 3.2. A few remarks on the several assumptions are given in Section 3.4, whereas special cases of Theorem 1, such as Binomial and Truncated-Poisson processes, are provided in Section 3.5. Finally, Section 3.6 outlines a sketch of the proof of Theorem 1.

3.1 Regularized M-Estimation

We are interested in learning the parameter Θ^* of the multivariate AR-GLM from equation (2.3). Given a collection of loss functions $\mathcal{L}_{it} : \mathcal{X}_i \times \mathbb{R} \to \mathbb{R}$, for $i \in [N]$ and $t \in \mathbb{Z}$, we consider the following ℓ_1-regularized M-estimator

$$\begin{aligned} \widehat{\Theta} &:= \underset{\Theta \in \mathbb{R}^{N \times N \times L}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_i(\Theta_i) + \lambda_n \|\Theta\|_{1,1,1}, \\ \mathcal{L}_i(\Theta_i) &:= \tfrac{1}{n} \sum_{t=1}^{n} \mathcal{L}_{it}\left(x_i^t\,;\, \langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D} \rangle\right). \end{aligned} \tag{3.1}$$

Since both the loss function and the ℓ_1 penalty are decomposable, we can solve each of the N problems in equation (3.1) indexed by i separately,

$$\widehat{\Theta}_i := \underset{\Theta_i \in \mathbb{R}^{N \times L}}{\operatorname{argmin}}\ \mathcal{L}_i(\Theta_i) + \lambda_n \|\Theta_i\|_{1,1} \quad \forall\, i \in [N]. \tag{3.2}$$

The possible dependence of $\mathcal{L}_{it}$ on t in the M-estimator equation (3.1) allows for the incorporation of time-discounting factors such as γ^t for some $\gamma < 1$. We consider a large class of loss functions later stated explicitly in assumptions (A2) and (A3). This class always includes the negative-log likelihood function for exponential family distributions $\mathbb{Q}_i(\,\cdot \mid f_i(v))$ with log-concave link f_i, and pseudo-likelihood functions in some cases. When $\mathcal{L}_{it}$ are chosen to be convex, the whole problem equation (3.1) is unconstrained, convex, with a coercive objective function, whereby the solution $\widehat{\Theta}$ is unique. Furthermore, the estimator equation (3.1) can be solved efficiently using any non-smooth convex optimization solver, such as the subgradient methods or proximal gradient descent methods [3]. An implementation for the general problem in equation (3.1) is available at [41] which implements both the subgradient method as well as the proximal gradient method.

Each iteration of both of these methods involve computation of the gradient of the loss function followed by finding the sub-gradient or proximal mapping for the regularization.

Computing the gradient of the loss is the most expensive step. The gradient of the loss is

$$\nabla\mathcal{L}(\Theta_i) = \frac{1}{n}\sum_{t=1}^{n}\mathcal{L}'_{it}\left(x_i^t\,;\,\langle\Theta_i,\mathbf{X}^{t-1}\mathbf{D}\rangle\right)\mathbf{X}^{t-1}\mathbf{D}, \tag{3.3}$$

where in $\mathcal{L}'_{it}(\cdot\,;\,\cdot)$ the derivative is with respect to the second argument. To compute the gradient, $\mathbf{X}^{t-1}\mathbf{D}$ can be precomputed once by multiplying $\mathbb{X} := \{\boldsymbol{x}^t\}_{t=-p+1}^{n}$ and $\mathbf{D}$. Hence, the complexity of obtaining the gradient $\nabla\mathcal{L}(\Theta_i)$ at each iteration is dominated by that of computing $\langle\Theta_i,\mathbf{X}^{t-1}\mathbf{D}\rangle$ for all i, that is, $O(nNL)$. To solve the optimization problem, one can then use the subgradient method with a provable convergence rate of $1/\sqrt{k}$ after k steps. This relatively slow rate is due to the non-smoothness of the objective function. Alternatively, we can use the proximal gradient method that converges at a rate of $1/k$. Then, the overall computational complexity of obtaining an ε-optimal solution is $O(nNL/\varepsilon)$. The parallel implementation in equation (3.2) allows for massive speed-ups in computation when using GPUs. The main result of this dissertation concerns the statistical complexity of the estimator and is agnostic to the choice of the optimization solver.

Our main result establishes the statistical properties of estimator equation (3.1) such as consistency, sample complexity and error rate. Our analysis also highlights desirable properties of the loss functions $\mathcal{L}_{it}$ and the nonlinearities f_i for achieving consistency. The result also shows the effect of the dictionary $\mathbf{D}$ in increasing the sample-efficiency of the estimator.

3.2 Assumptions for the Main Result

Our main result concerns the estimation error of the parameters $\{\widehat{\Theta}_i\}_{i=1}^{N}$, obtained by solving equation (3.2). We implicitly assume Θ_i^* to be approximately s_i-sparse. This assumption is encoded via the ℓ_1-approximation errors

$$\omega_i := \min_{\boldsymbol{\beta}\in\mathbb{R}^{N\times L}}\{\|\boldsymbol{\beta}-\Theta_i^*\|_1 \mid \|\boldsymbol{\beta}\|_{0,0}\leq s_i\}. \tag{3.4}$$

We also impose the following assumptions:

(A1) The process is wide-sense stationary and stable, i.e., the power spectral density matrix exists:

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

$$\min_{\omega\in[-\pi,\pi)} \lambda_{\min}\left(\mathcal{X}(\omega)\right) \geq C_{\mathcal{X}}^2 > 0.$$

(A2) The loss function $v \mapsto \mathcal{L}_{it}(u,v)$ is twice differentiable and strongly convex for all u, with curvature $\kappa_i > 0$, i.e., $\partial_v^2 \mathcal{L}_{it}(u;v) \geq \kappa_i$ for all $u \in \mathcal{X}_i, v \in \mathbb{R}, i \in [N], t \in \mathbb{N}_+$.

(A3) $|\partial_v \mathcal{L}_{it}(u,v)| \leq C_{\mathcal{L}}$, and for all $v \in \mathbb{R}$, $i \in [N], t \in \mathbb{N}_+$ we have

$$U \sim \mathbb{Q}_i(\,\cdot \mid f_i(v)) \implies \mathbb{E}\big[\partial_v \mathcal{L}_{it}(U;v)\big] = 0.$$

Assumption (A3) guarantees that Θ^* is the minimizer of the population loss, and is necessary for the consistency of the M-estimator. The second half of the assumption is generally satisfied if the loss is taken to be the log-likelihood function. The next example verifies this for single-parameter exponential families.

Example 1. Assume that $Q_i(\,\cdot \mid z)$ is an exponential family with density $x \mapsto \exp(xz - \phi(z))$, for all i. Here, z is the so-called natural parameter of the family and ϕ is the log-partion function. Let $U \sim Q(\,\cdot \mid f_i(v))$ and take $\mathcal{L}_{it}(x,v)$ to be the log-likelihood of this model, that is,

$$\mathcal{L}_{it}(x;v) = -x f_i(v) + \phi(f_i(v)).$$

This class includes Bernoulli, Poisson, and Gaussian (with known variance) AR processes among others. We have

$$\partial_v \mathcal{L}_{it}(U;v) = -U f_i'(v) + \phi'(f_i(v)) f_i'(v).$$

By a standard property of the exponential family $\mathbb{E}[U] = \phi'(f_i(v))$, hence $\mathbb{E}[\partial_v \mathcal{L}_{it}(U; v)] = 0$ verifying the second half of (A3). If, in addition, the family has bounded support and both ϕ and f_i are Lipschitz, then the entire (A3) holds. Distributions such as Poisson and Gaussian violate the boundedness assumption. However, the truncated version of these distributions belong to the exponential family and satisfy the boundedness condition.

Example 2. Under the same exponential family distribution as in Example 1, the second half of (A3) also holds for the squared error loss

$$\mathcal{L}_{it}(x; v) = \big[x - \phi'(f_i(v))\big]^2.$$

To verify this, it is enough to observe that

$$\partial \mathcal{L}_{it}(U; v) = 2\big[U - \phi'(f_i(v))\big] \cdot \phi''(f_i(v)) f_i'(v),$$

and use $\mathbb{E}[U] = \phi'(f_i(v))$.

These two examples show that (A3) is satisfied for commonly used loss functions. As for (A2), we recall that in an exponential family with the natural parameterization, the log-partition function $\phi(\cdot)$ is convex. Assumption (A2), however, requires the map $v \mapsto \mathcal{L}_{it}(u, v)$ to be strongly convex. Extra care should be taken in choosing the loss and $f_i(\cdot)$ to ensure that this assumption is satisfied. The stability assumption (A1) is further discussed in the remarks following the main result.

Let us now define a few constants necessary to state our main result. Let

$$\begin{aligned} C_{\mathbf{D}} &:= \max_{\ell} \|\boldsymbol{d}_\ell\|_1, \\ G = G(\Theta^*) &:= 64 C_{\mathbf{D}}^4 B^4 \Big(1 + p^2 \psi\big(\tau_1(\Theta^*)\big)\Big), \end{aligned} \tag{3.5}$$

where $\psi(x) = (1 - x^{-1})^{-2}$ and

$$\begin{aligned} \tau_1(\Theta^*) &:= \sup_{\boldsymbol{z},\boldsymbol{y} \in \mathcal{X}^{\times p}} \|\mathbb{P}_{\boldsymbol{z}} - \mathbb{P}_{\boldsymbol{y}}\|_{\mathrm{TV}} < 1, \\ \mathbb{P}_{\boldsymbol{z}} &:= \mathbb{P}(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = \boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{X}^{\times p}. \end{aligned} \tag{3.6}$$

Here, $\mathcal{X}^{\times p} \subset \mathbb{R}^{N \times p}$ denotes the set of matrices consisting of p columns, each from $\mathcal{X}$. Note that $\mathbb{P}_{\boldsymbol{z}}$ is invariant to t. Now fix $\mathcal{U} \subset [N]$ and let us define the following constants needed to state the main result:

$$s_{\max} := \max_{i \in \mathcal{U}} s_i, \tag{3.7a}$$

$$s_+ := \sum_{i \in \mathcal{U}} s_i, \tag{3.7b}$$

$$\overline{\kappa} := \max_{i \in \mathcal{U}} \kappa_i \tag{3.7c}$$

$$\underline{\kappa} := \frac{C_{\mathcal{X}}^2}{8} \min_{i \in \mathcal{U}} \kappa_i, \tag{3.7d}$$

$$\widetilde{\omega}_+ := \sum_{i \in \mathcal{U}} \overline{\kappa} \frac{\omega_i^2}{s_i} + 4\omega_i, \tag{3.7e}$$

where $C_{\mathcal{X}}$ and κ_i are specified in assumptions Item (A1) and item (A2) respectively.

3.3 Main Result

We are now ready to state the main result:

Theorem 1. *Suppose that $\{\boldsymbol{x}^t\}_{t=-p+1}^n$ are samples from process given in equation (2.3), with each $\mathcal{X}_i$ being a countable subset of $[-B, B]$ for some $B > 0$, and satisfying (A1). Fix a subset $\mathcal{U} \subseteq [N]$ and let $\{\widehat{\Theta}_i\}_{i \in \mathcal{U}}$ be the solutions of the optimization problem equation (3.2) with loss functions $\mathcal{L}_{it}$ satisfying assumptions (A2) and (A3). Fix $c_1 > 2$ and let $c = c_1/2 - 1$. If*

$$\lambda_n = 2BC_{\mathcal{L}}C_{\mathbf{D}}\sqrt{c_1 \log(|\mathcal{U}|NL)/n}, \qquad \text{and} \qquad n \gtrsim \frac{G}{C_{\mathcal{X}}^6}\, s_{\max}^3 \log(NL), \tag{3.8}$$

then, with probability at least $1-(NL)^{-Cs_{\max}}-(|\mathcal{U}|NL)^{-c}$,

$$\sum_{i\in\mathcal{U}}\|\widehat{\Theta}_i-\Theta_i^*\|_F^2 \;\le\; \frac{9}{\underline{\kappa}^2}s_+\lambda_n^2+\frac{\widetilde{\omega}_+}{\underline{\kappa}}\lambda_n. \tag{3.9}$$

where $C=O(C_{\mathcal{X}}^{-2})$ *only depends on* $C_{\mathcal{X}}$.

The error bound in equation (3.9) can be written, up to constants, as:

$$\sum_{i\in\mathcal{U}}\|\widehat{\Theta}_i-\Theta_i^*\|_F^2 \lesssim \frac{s_+\log(NL)}{n}+\widetilde{\omega}_+\sqrt{\frac{\log(NL)}{n}}. \tag{3.10}$$

The two terms in the bound correspond to the estimation and approximation errors, respectively. The estimation error scales at the so-called *fast rate* $\log(NL)/n$, while the approximation error scales at the slower rate $\sqrt{\log(NL)/n}$. For the exact sparsity model, where $\omega_i=0$ for all i, the approximation error vanishes and the estimator achieves the fast rate. For simplicity, assume that $C_{\mathcal{L}},C_{\mathbf{D}}\lesssim 1\lesssim C_{\mathcal{X}}$. Then, the overall (excess) sample complexity for consistent estimation is

$$n\gg \max\left\{Gs_{\max}^3,\ s_+,\ (\widetilde{\omega}_+)^2\right\}\log(NL). \tag{3.11}$$

By consistency, we mean that the estimator converges to the true parameter when n grows to infinity, as long as the above condition holds, even when the rest of the parameters s,p,L and N grow to infinity alongside n. We discuss the meaning of the "excess" qualification for the sample complexity in the remarks below.

Bound equation (3.10) has a logarithmic dependence on N, the number of variables in the process, which is a notable feature of our work. Compared to some of the previous work [19], we overcome the $N>1$ barrier for the BAR model while allowing for $p>1$ dependence on the past. The bound also depends logarithmically on L. This means that dictionary $\mathbf{D}$ can be overcomplete, allowing for Θ^* to be sparse, for nearly no additional cost.

3.4 Remarks on the Main Result

Let us make a few comments on the various choices in Theorem 1.

3.4.1 Choice of regularization parameter λ_n

The upper bound provided in equation (3.9) depends on the choice of the regularization parameter. The result however also specifies a particular value for λ_n. This choice is governed largely by the proof technique. In short, we require that the value of λ_n be larger than the dual-regularizer-norm of the gradient of the loss function at Θ^*. More details on this are provided in Section 3.7 after outlining the sketch of the proof of Theorem 1.

3.4.2 Choice of the loss $\mathcal{L}$

Theorem 1 holds for any loss function satisfying conditions (A2) and (A3). For the Bernoulli AR process, the negative log-likelihood $\mathcal{L}_{i,t}(u,v) = -u \log f_i(v) - (1-u)\log(1-f_i(v))$ satisfies these assumptions for any log-concave f_i; see [32]. For the Truncated-Poisson AR process, the negative log-likelihood takes the form $\mathcal{L}_{it}(u,v) = f_i(v) - u\log f_i(v) + \log(u!)$ and satisfies the assumptions for $f_i(v) = \exp(v)$ or $f_i(v) = \log(1+e^v)$.

3.4.3 Choice of $\mathcal{U}$

The result in Theorem 1 has been stated for a general $\mathcal{U} \subseteq [N]$. Taking $\mathcal{U} = [N]$, gives a bound on the Frobenius norm of the entire tensor $\|\widehat{\Theta} - \Theta^*\|_F^2$. On the other extreme, we can take $\mathcal{U} = \{i\}$ to obtain bounds on each slice of the tensor with better scaling with sparsity. For example, in the exact sparsity setting, we obtain $\|\widehat{\Theta}_i - \Theta_i^*\|_F^2 \lesssim s_i \log(NL)/n$, avoiding the extra price of $(\sum_{j\neq i} s_j)\log(NL)/n$ that we pay for the entire tensor.

3.4.4 Scaling with sparsity

Considering the exact sparsity setting, the scaling of the sample complexity equation (3.11) with sparsity is $n = \Omega(s_+ \vee s_{\max}^3)$. In the worst case, $s_+ = s_{\max}$ and we get a cubic dependence on sparsity which is not ideal. However, when $s_+ \gtrsim s_{\max}^3$, Theorem 1 requires $n = \Omega(s_+)$ which is the optimal scaling with sparsity. (This can be seen by noting that in the linear independent setting, one cannot do better than $n = \Omega(s_+)$.) Our result also holds for the more general case of $\omega_i \neq 0$. For example, for the ℓ_q ball sparsity with $q \in (0,1)$, we have $\omega_i = O(s_i^{1-1/q})$ hence $\omega_i^2/s_i + w_i = O(\omega_i) = O(s_i)$ and $\widetilde{\omega}_+ = O(s_+)$ and the same sample complexity as the exact sparsity case holds.

It is not clear if the worst-case cubic dependence on the sparsity can be improved without imposing restrictive assumptions. It is worth noting that in our proof, the additional s_i^2 factor comes from concentration inequality equation (4.4) in Lemma 10. This additional factor can be removed if one were able to show sub-Gaussian concentration for deviations of the order of $\|\boldsymbol{\beta}\|_F^2$ instead of $\|\boldsymbol{\beta}\|_{1,1}^2$, in Lemma 10. It remains open whether such concentration is possible and under what additional assumptions. Section 5 provides a more detailed discussion on this concentration inequality. Figure 6.4a in Section 6 suggests a superlinear dependence on s, hinting that the situation may not be as simple as the i.i.d. case.

For $p = 1$, a sample complexity of $\rho^3 \log(N)$ was reported in [16, Cor. 1]. One can verify that ρ in their model is equal to $s_{\max}$ in ours, hence they obtain the same $s_{\max}^3$ dependence on sparsity. Similarly for $p = 2$, the result in [25, Thm 4.4] requires $(s/r_\rho^2)\log(N)$ samples where s and r_ρ are sparsity parameters defined therein and r_ρ is inversely related to $s_{\max}$ in the worst case, yielding a similar cubic dependence on sparsity as ours. Furthermore, it appears that their analysis only holds for $s_{\max} = \mathcal{O}(1)$, whereas we make no such assumption. In short, to our knowledge, no prior work has broken the $s_{\max}^3$ barrier in the non-Gaussian AR setting.

3.4.5 Scaling with lag p

Our result is the first to provide sufficient conditions for a sample complexity logarithmic in p in the case of the identity dictionary, for any value of N. As will be discussed in Section 3.5, the dependence of the (excess) sample size n on p could be as good as $O(\log L)$ for a general dictionary, under certain tail and normalization conditions. In these cases, one could obtain an $O(1)$ growth of n as function of p in the best case (when $L = O(1)$) and an $O(\log p)$ growth in the worse case (the identity dictionary). In contrast, [19, Thm. 1] requires $s^{2/3}p^{2/3}\log(p)$ samples, for the identity dictionary, and their proof relies heavily on $N = 1$.

Our bound scales with p through G which is defined in terms of the contraction coefficient $\tau_1(\Theta^*)$ in equation (3.6). The contraction coefficient only depends on Θ^* and is always less than 1. Intuitively, if Θ^* is too large, then for two different initializations $\boldsymbol{z}$ and $\boldsymbol{y}$, the distributions $\mathbb{P}(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = \boldsymbol{y})$ and $\mathbb{P}(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = \boldsymbol{z})$ may significantly differ. A clear sufficient condition for $G = O(1)$ is to have $\tau_1(\Theta^*) = O(p^{-1})$ as well as $C_{\mathbf{D}} \lesssim 1$. The challenge is to control $\tau_1(\Theta^*)$ in terms of the size of Θ^*. Section 3.5 further discusses sufficient conditions under which $G = O(1)$. There, we show that for certain exponential families, the scaling depends on the behavior of the tail of $k \mapsto |(\boldsymbol{d}_\ell)_k|$, that is, how fast the *influence from the past* dies down in the filters $\{\boldsymbol{d}_\ell\}$.

A subtle point worth noting here, which does not arise in ordinary M-estimation with i.i.d. measurements, is that n is in fact the *excess* sample-size one needs beyond the p initial samples. It is clear that at least p initial samples are needed for estimating a p-lag process. Examples discussed in Section 3.5 provide conditions that guarantee that the excess sample size, n, needed for consistent estimation is $O(\log L)$ as p grows, the smallest order one could hope for.

3.4.6 Stability assumption (A1)

We use assumption (A1) to guarantee that the strong convexity holds for the population loss $\Theta \mapsto \mathbb{E}\,\mathcal{L}(\Theta)$. This is key in guaranteeing that any parameter tensor $\widehat{\Theta}$ that maximizes the regularized loss function in equation (3.1) does not deviate far from the true parameter Θ^*.

Assumption (A1) is by now standard in time-series estimation literature [38, 2, 23]. The quantity $C_{\mathcal{X}}$ is fundamental to multivariate time-series analysis, however, its behavior as a function of the parameters of the model is not yet fully understood. Intuitively, $C_{\mathcal{X}}$ is related to the *flatness* of the power spectral density (PSD) $\mathcal{X}$, and the stability of the process. For the $N = 1$ case, $C_{\mathcal{X}} > 0$ implies that the process does not have zeros on the unit circle in the spectral domain.

In general, $C_{\mathcal{X}}$ could potentially depend on N, indirectly via Θ^*. In subsequent discussions of Theorem 1, we have assumed that $C_{\mathcal{X}}$ stays uniformly bounded away from zero as N grows. This assumption is explicitly stated as $C_{\mathcal{X}} \gtrsim 1$. Our main result (Theorem 1), however, holds for all positive values of $C_{\mathcal{X}}$, regardless of its growth rate. Even if $C_{\mathcal{X}} = o(1)$ with respect to N, Theorem 1 still gives a consistency result, albeit with a worse dependence on N.

The dependence of $C_{\mathcal{X}}$ on N occurs through the scaling of the true parameter Θ^*. That $C_{\mathcal{X}}$ is in general bounded below by a constant (or has a slow decay as a function of N) is part of the folklore of the time series literature. It is reasonable to assume that this holds for certain structured Θ^*. However, obtaining exact conditions on Θ^* for $C_{\mathcal{X}} \gtrsim 1$ to hold is, in general, a non-trivial open problem, even for univariate Gaussian AR(p) processes. The main difficulty is that the relation between the power spectral density of the process and its parameter is indirect and via the Z-transform. Nevertheless, conditions are known in special cases. See for example the discussion surrounding Proposition 2.2 in [2], where explicit conditions are given on the parameter matrix of a VAR(1) Gaussian process, for $C_{\mathcal{X}}$ to stay bounded away from zero.

3.5 Special Cases of the Main Result

Let us now look at the applications of Theorem 1 to two special cases often considered in discrete-valued time series modeling — Binomial and Poisson AR processes. We take

$\mathcal{U} = [N]$ throughout this section. To apply the theorem, we need to upper-bound $G(\Theta^*)$ in each case. Since the ψ function in equation (3.5) is non-decreasing on $[0, 1)$, it is enough to control $\tau_1(\Theta^*)$. In fact, a sufficient condition for $G(\Theta^*) = O(1)$ is to have $\tau_1(\Theta^*) = O(\frac{1}{p})$ and $C_{\mathbf{D}} = O(1)$.

The quantity $\tau_1(\Theta^*)$ is the maximum total variation distance between the p-step conditional distributions of the process, starting from two initial states $\boldsymbol{y}$ and $\boldsymbol{z}$. The Pinsker's inequality [9, p. 44] can be used to further control the total variation distance by the KL divergence, which is the natural choice for comparing two exponential family distributions with independent coordinates.

Recall $\mathcal{X} = \prod_{i=1}^N \mathcal{X}_i \subset [-B, B]^N$ and the notation $\mathbb{P}_{\boldsymbol{z}}$ from equation (3.6). Pinsker's inequality yields

$$\tau_1^2(\Theta^*) \leq \sup_{\boldsymbol{z},\boldsymbol{y} \in \mathcal{X}^{\times p}} \tfrac{1}{2} D_{\text{KL}}(\mathbb{P}_{\boldsymbol{z}} \| \mathbb{P}_{\boldsymbol{y}}), \tag{3.12}$$

where $D_{\text{KL}}(\cdot\|\cdot)$ is the KL-divergence. We now state upper bounds on $D_{\text{KL}}(\mathbb{P}_{\boldsymbol{z}} \| \mathbb{P}_{\boldsymbol{y}})$ for the two cases of the Binomial and Poisson processes. A quantity of interest is the tail decay of the dictionary elements $\{\boldsymbol{d}_\ell\}_{\ell=1}^L$, measured by

$$\gamma_{t\ell} := \sum_{m=t}^{p} |(\boldsymbol{d}_\ell)_m|. \tag{3.13}$$

Let us define the following norm on Θ,

$$\|\Theta\|_\star := \Big(\sum_{i,t} L_i^2 \Big[\sum_{j,\ell} \gamma_{t\ell} |\Theta_{ij\ell}| \Big]^2 \Big)^{1/2}$$

where L_i is the Lipschitz constant of the link function f_i, and the summations run over $(i, t, j, \ell) \in [N] \times [p] \times [N] \times [L]$. One can often establish a bound of the form

$$D_{\text{KL}}(\mathbb{P}_{\boldsymbol{z}} \| \mathbb{P}_{\boldsymbol{y}}) \leq C_f B^2 \|\Theta^*\|_\star^2 \tag{3.14}$$

where C_f depends on $\{f_i\}$ and Θ^* is the true parameter generating the samples.

Lemma 2. *Consider a Binomial AR process given by equation* (2.3) *with* $\mathcal{X}_i = \{0, 1, \dots, K_i\}$, *where* $K_i \leq B$, *and* $\mathbb{Q}_i(\cdot \mid z) = \text{Bin}(K_i, z)$. *Assume that* f_i *is* L_i*-Lipschitz, and for some* $\varepsilon \in (0, \frac{1}{2})$, $f_i : \mathbb{R} \to [\varepsilon, 1 - \varepsilon]$ *for all* i. *Then, equation* (3.14) *holds with* $C_f = 6/\varepsilon$.

The case of $B = 1$ recovers the result for the Bernoulli Autoregressive Process in [32]. The proof is provided in the Appendix.

Lemma 3. *Consider a Truncated Poisson AR process given by equation* (2.3) *with* $\mathcal{X}_i = \{0, 1, \dots, K_i\}$ *and* $\mathbb{Q}_i(\cdot \mid z) = \mathbb{P}(\min(K_i, Z) \in \cdot)$ *where* $Z \sim \text{Poi}(z)$ *and* $K_i \leq B$. *Assume that* f_i *is* L_i*-Lipschitz, and for some* $\varepsilon > 0$, $f_i : \mathbb{R} \to [\varepsilon, \infty)$ *for all* i. *Then, equation* (3.14) *holds with* $C_f = 4/\varepsilon$.

Combining with equation (3.12), we have the following corollary. The proof is provided in the Appendix.

Corollary 4. *Under the assumptions of Lemma 2 or 3,*

$$\tau_1(\Theta^*) \lesssim \frac{B}{\sqrt{\varepsilon}} \|\Theta^*\|_\star.$$

In particular, if $C_{\mathcal{L}}, C_{\mathbf{D}} \lesssim 1 \lesssim C_{\mathcal{X}}$ *and* $\|\Theta^*\|_\star = O(1/p)$, *then* $G = O(1)$ *and the following is sufficient for consistency:*

$$n \gg \max\left\{s_{\max}^3,\ s_+,\ (\widetilde{\omega}_+)^2\right\} \log(NL).$$

In other words, Corollary 4 provides conditions under which consistent estimation is possible with (excess) sample complexity that grows at most logarithmically in L.

Let us consider some examples for which $\|\Theta^*\|_\star = O(1/p)$. For the purpose of illustration, let us separate the tail decay of Θ^*, along the lag dimension, by assuming that

$$|\Theta^*_{ij\ell}| \leq R_{ij} h_\ell, \quad \forall\, (i, j, \ell) \in [N] \times [N] \times [L].$$

for some sequence $\{h_\ell\}_{\ell=1}^\infty$ such that $\sum_{\ell=1}^\infty h_\ell < \infty$ and a matrix $R = (R_{ij})$. Assume that $\Theta^*_{ij\ell}$ is normalized so that $\|R\|_{2,1} = O(1)$. Moreover, assume that $\max_i L_i = O(1/p)$. Since in model equation (2.3), the input to each f_i involves terms $\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \rangle_{\mathbb{R}^p}$, each of which is essentially a sum of p terms (cf. equation (2.4)), the aforementioned assumption on the Lipschitz constant is a natural normalization that prevents the saturation of the nonlinearities f_i as p grows. Equivalently, we can make this condition more explicit by replacing $f_i(\cdot)$ in the definition of model equation (2.3) with $\tilde{f}_i(\frac{1}{p}\cdot)$ and assuming that $\tilde{f}_i$ have Lipschitz constants uniformly bounded by a constant.

Under the above modeling assumptions, consider the following two dictionaries:

Case (a): The identity dictionary, where $L = p$ and $(\boldsymbol{d}_\ell)_m = 1\{m = \ell\}$. In this case, $\gamma_{t\ell} = 1\{t \le \ell\}$. Then,

$$\|\Theta\|_\star \lesssim \frac{1}{p}\|R\|_{2,1}\Big[\sum_{t=1}^p \Big(\sum_{\ell=t}^p h_\ell\Big)^2\Big]^{1/2} = O\Big(\frac{1}{p}\Big)$$

assuming that $\sum_{t=1}^\infty (\sum_{\ell=t}^\infty h_\ell)^2 < \infty$ which holds, for example, if h_ℓ decays at least as fast as $\ell^{-1-\alpha/2}$ for some $\alpha > 1$. Note that in this case $C_{\mathbf{D}} \asymp 1$ is trivially satisfied.

Case (b): A general dictionary, with filters satisfying the decay rate $\max_\ell |(\boldsymbol{d}_\ell)_m| \lesssim m^{-\alpha-1}$ for some $\alpha > 1$. Then, $\max_\ell \gamma_{t\ell} \lesssim t^{-\alpha}$ and

$$\|\Theta\|_\star \lesssim \frac{1}{p}\|R\|_{2,1}\Big(\sum_{t=1}^p t^{-2\alpha}\Big)^{1/2} \sum_{\ell=1}^p h_\ell = O\Big(\frac{1}{p}\Big)$$

using $\sum_{t=1}^\infty t^{-2\alpha} < \infty$ and $\sum_{\ell=1}^\infty h_\ell < \infty$. Moreover, since we have $C_{\mathbf{D}} \lesssim \sum_{m=1}^p m^{-\alpha-1}$, it follows that $C_{\mathbf{D}} = O(1)$ as p grows.

Thus in both cases, Corollary 4 guarantees that the excess sample size n needed for consistency grows at most logarithmically in L. This translates to an $O(\log p)$ growth in the case the identity dictionary but could be as low as $O(1)$ for a dictionary with the number of filters L not growing with p. Note that the summability condition on h_ℓ in case (b) is milder than that in case (a), showing the trade-off between the tail decay of Θ (along the lag dimension) and the tail decay of the dictionary filters. Having fast decaying filters relaxes the decay requirement on the tails of Θ.

3.6 Proof Sketch of the Main Result

We now outline the proof of Theorem 1. Our analysis applies the framework of Negahban et al. [29]. Let

$$\mathcal{L}_i(\boldsymbol{\beta}) := \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{it,}(x_i^t; \langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\rangle), \quad \boldsymbol{\beta} \in \mathbb{R}^{N\times L}.$$

Fix $\mathcal{U} \subseteq [N]$ and set $\Theta_{\mathcal{U}} := (\Theta_i)_{i\in\mathcal{U}}$ and similarly $\Theta^*_{\mathcal{U}} := (\Theta^*_i)_{i\in\mathcal{U}}$ and $\widehat{\Theta}_{\mathcal{U}} := (\widehat{\Theta}_i)_{i\in\mathcal{U}}$, all tensors in $\mathbb{R}^{|\mathcal{U}|\times N\times L}$. We also write $\mathcal{L}_{\mathcal{U}}(\Theta_{\mathcal{U}}) = \sum_{i\in\mathcal{U}} \mathcal{L}_i(\Theta_i)$. Now we have,

$$\widehat{\Theta}_{\mathcal{U}} = \underset{\Theta_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times N\times L}}{\operatorname{argmin}} \mathcal{L}_{\mathcal{U}}(\Theta_{\mathcal{U}}) + \|\Theta_{\mathcal{U}}\|_{1,1,1}. \tag{3.15}$$

In the sequel, $\nabla\mathcal{L}_{\mathcal{U}}$ and $\nabla^2\mathcal{L}_{\mathcal{U}}$ are the gradient and Hessian of $\mathcal{L}_{\mathcal{U}}$ with respect to variable $\Theta_{\mathcal{U}}$. When $n \ll |\mathcal{U}|NL$, the empirical Hessian, $\nabla^2\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})$, is rank-deficient, hence the loss function is flat in many directions around $\Theta^*_{\mathcal{U}}$. The approach of Negahban et al. [29] is to guarantee that $\mathcal{L}_{\mathcal{U}}$ is positively curved in certain directions, including $\widehat{\Delta}_{\mathcal{U}} := \widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}}$.

In particular, if the regularization parameter λ_n is large enough, specifically

$$\lambda_n \geq 2\|\nabla\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})\|_{\infty,\infty,\infty}, \tag{3.16}$$

then, the error tensor $\widehat{\Delta}_{\mathcal{U}}$ lies in a small *cone-like* subset $\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$—to be defined below—and on this set, $\mathcal{L}_{\mathcal{U}}$ is "nearly" strongly convex, i.e., $\nabla^2\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})$ is uniformly quadratically bounded below.

For a set $S \subseteq [N]\times[L]$, let $\boldsymbol{\beta}_S$ denote the projection of $\boldsymbol{\beta}$ on the subspace of matrices with support S. For $\boldsymbol{\beta}^*$ define:

$$\mathbb{C}(S;\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_{1,1} \leq 3\|\boldsymbol{\beta}_S\|_{1,1} + 4\|\boldsymbol{\beta}^*_{S^c}\|_{1,1}\}. \tag{3.17}$$

Note that this is a *cone-like* subset of $\mathbb{R}^{N\times L}$ around $\boldsymbol{\beta}^*$. See [29] for a visualization. Let $\mathcal{S} := \bigcup_{i\in\mathcal{U}}\{i\}\times S_i$ where $S_i \subseteq [N]\times[L]$ for $i\in\mathcal{U}$. Equivalently, $\mathcal{S} = \bigsqcup_{i\in\mathcal{U}} S_i$ using the notation of *disjoint union*. With some abuse of notation, we write $\mathcal{S}^c := \bigcup_{i\in\mathcal{U}}\{i\}\times S_i^c$. The cone-like set $\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$ is defined as follows:

$$\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}}) := \left\{(\Delta_i)_{i\in\mathcal{U}} : \Delta_i \in \mathbb{C}(S_i;\Theta^*_i),\, \forall i\in\mathcal{U}\right\}. \tag{3.18}$$

For loss functions $\mathcal{L}_i$, $i\in\mathcal{U}$, and for $\boldsymbol{\delta},\boldsymbol{\beta}^*\in\mathbb{R}^{N\times L}$, let

$$R\mathcal{L}_i(\boldsymbol{\delta};\boldsymbol{\beta}^*) := \mathcal{L}_i(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - \mathcal{L}_i(\boldsymbol{\beta}^*) - \langle\nabla\mathcal{L}_i(\boldsymbol{\beta}^*),\boldsymbol{\delta}\rangle\,, \tag{3.19}$$

be the remainder of the first-order Taylor expansion of $\mathcal{L}_i$ around $\boldsymbol{\beta}^*$. Following [29], we say that $\mathcal{L}_{\mathcal{U}}$ satisfies restricted strong convexity (RSC) at $\Theta^*_{\mathcal{U}}$ with curvature $\kappa > 0$ and tolerance τ^2 if for all $\Delta\in\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$, we have,

$$\sum_{i\in\mathcal{U}} R\mathcal{L}_i(\Delta_i;\Theta^*_i) \geq \kappa\sum_{i\in\mathcal{U}}\|\Delta_i\|_F^2 - \tau^2. \tag{3.20}$$

The left-hand side is the remainder of the first-order Taylor expansion of $\mathcal{L}_{\mathcal{U}}$ around $\Theta^*_{\mathcal{U}}$, that is, $R\mathcal{L}_{\mathcal{U}}(\Delta_{\mathcal{U}};\Theta^*_{\mathcal{U}})$—defined similar to equation (3.19).

Now, assume that equation (3.16) and equation (3.20) hold. Then, [29, Theorem 1] implies that $\widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}} \in \mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$, and that

$$\|\widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}}\|_F^2 \leq \tfrac{9\lambda_n^2}{\kappa^2}|\mathcal{S}| + \tfrac{\lambda_n}{\kappa}(2\tau^2 + 4\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^c}\|_{1,1,1}). \tag{3.21}$$

The above inequality provides a family of bounds, one for each choice of $\mathcal{S} = \bigsqcup_{i\in\mathcal{U}} S_i$.

Decreasing $|\mathcal{S}|$ reduces the first term, but potentially increases $\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^c}\|_{1,1,1}$. We choose $\mathcal{S}$ to balance the two. Let $S^*_i \subset [N] \times [L]$ be the support of the minimizer in equation (3.4), so that $|S^*_i| = s_i$. We take $\mathcal{S} = \mathcal{S}^* = \bigsqcup_{i \in \mathcal{U}} S^*_i$. Consequently, $|\mathcal{S}^*| = \sum_{i \in \mathcal{U}} s_i$ and $\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^{*c}}\|_{1,1,1} = \sum_{i \in \mathcal{U}} \omega_i$. For this choice of $\mathcal{S}$, Proposition 5 below shows that equation (3.20) holds, with high probability. To state the concentration inequality, recall the definitions equation (3.7).

Proposition 5 (Restricted Strong Convexity)**.** Under assumptions (A1) and (A2), if we have,

$$n \gtrsim \frac{G}{C_{\mathcal{X}}^6} s_{\max}^3 \log(NL) \tag{3.22}$$

then, the RSC given in equation (3.20) for $\mathcal{S} = \mathcal{S}^*$ holds with curvature $\kappa = \underline{\kappa}$ and tolerance $\tau^2 = \frac{\overline{\kappa}}{2} \sum_{i \in \mathcal{U}} \omega_i^2 / s_i$, with probability at least $1 - (NL)^{-Cs_{\max}}$ where $C = O(C_{\mathcal{X}}^{-2})$.

Lemma 6 below shows that $\Theta^*_{\mathcal{U}}$ is in fact the minimizer of the expected loss $\mathbb{E}\mathcal{L}_{\mathcal{U}}(\cdot)$. Later, Lemma 7 shows that taking $\lambda_n = O(\sqrt{\log(|\mathcal{U}|NL)/n})$ is enough for equation (3.16) to hold with high probability. Putting the pieces together proves Theorem 1. Chapter 4 is dedicated to proving Proposition 5.

Lemma 6. *Under assumptions (A1)–(A3), we have* $\Theta^*_i \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \mathbb{E}\,\mathcal{L}_i(\boldsymbol{\beta})$.

Proof. This is a direct consequence of Lemma 9 and assumption (A3). Notice that from Lemma 8 we have

$$\mathcal{L}_i(\Theta^*_i + \Delta_i) \geq \mathcal{L}_i(\Theta^*) + \langle \nabla \mathcal{L}_i(\Theta^*_i), \Delta_i \rangle + \mathcal{E}(\Delta_i; \mathbb{X}).$$

Taking expectations on both sides, and applying lemma 9, we get

$$\mathbb{E}\mathcal{L}_i(\Theta^*_i + \Delta_i) \geq \mathbb{E}\mathcal{L}_i(\Theta^*) + \langle \mathbb{E}\nabla \mathcal{L}_i(\Theta^*_i), \Delta_i \rangle + C_{\mathcal{X}}^2 \|\Delta_i\|_F^2.$$

It follows from Assumption (A3) that $\mathbb{E}\nabla\mathcal{L}_i(\Theta^*) = 0$. Thus we get

$$\mathbb{E}\mathcal{L}_i(\Theta_i^* + \Delta_i) \geq \mathbb{E}\mathcal{L}_i(\Theta^*)$$

for all $\Delta_i \in \mathbb{R}^{N \times L}$, which proves the claim. ∎

3.7 Choice of regularization hyperparameter

Lemma 7. *For any constant $c_1 > 2$,*

$$\|\nabla\mathcal{L}_{\mathcal{U}}(\Theta_{\mathcal{U}}^*)\|_{\infty,\infty,\infty} \leq BC_{\mathcal{L}}C_{\mathbf{D}}\sqrt{c_1 \log(|\mathcal{U}|NL)/n} \tag{3.23}$$

with probability at least $1 - (|\mathcal{U}|NL)^{-c}$, where $c = c_1/2 - 1$.

Proof. Fix $i, j \in [N]$ and $\ell \in [L]$. Then we have,

$$\frac{\partial\mathcal{L}_i(\Theta_i)}{\partial\Theta_{ij\ell}} = \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}'_{it}(x_i^t, \langle\Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\,(\mathbf{X}^{t-1}\mathbf{D})_{j\ell} = \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}'_{it}(x_i^t, \langle\Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\left\langle \boldsymbol{x}_j^{t-*}, \boldsymbol{d}_\ell \right\rangle$$

where $\mathcal{L}'_{it}(u, v) := \partial_v\mathcal{L}_{it}(u, v)$ It follows that

$$\frac{\partial\mathcal{L}_i(\Theta_i^*)}{\partial\Theta_{ij\ell}} = \frac{1}{n}\sum_{t=1}^{n} D_{ij\ell}^t \quad \text{where} \quad D_{ij\ell}^t := \mathcal{L}'_{it}(x_i^t, \left\langle\Theta_i^*, \mathbf{X}^{t-1}\mathbf{D}\right\rangle)\left\langle \boldsymbol{x}_j^{t-*}, \boldsymbol{d}_\ell \right\rangle.$$

Let $\mathcal{F}^{t-1} = \sigma(\boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \dots)$ be the σ-field generated by the past observations of the process. From assumption (A3), we have $\mathbb{E}[\mathcal{L}'_{it}(x_i^t, \langle\Theta_i^*, \mathbf{X}^{t-1}\mathbf{D}\rangle) \mid \mathcal{F}^{t-1}] = 0$, hence

$$\mathbb{E}[D_{ij\ell}^t \mid \mathcal{F}^{t-1}] = 0.$$

That is, $\{D_{ij\ell}^t\}_t$ is a *martingale difference sequence.* Similarly, by assumption (A3), we get $\|\mathcal{L}'_{it}\|_\infty \leq C_{\mathcal{L}}$. If follows that $\{D_{ij\ell}^t\}_t$ is also bounded, i.e., $|D_{ij\ell}^t| \leq C_{\mathcal{L}} \cdot C_D$. By the

Azuma–Hoeffding inequality for martingale differences [45],

$$\mathbb{P}\left(\Big|\frac{\partial\mathcal{L}(\Theta^*)}{\partial\Theta_{ij\ell}}\Big| > t\right) = \mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} D_{ij\ell}^{t}\Big| > t\right) \leq 2\exp\Big(-\frac{nt^2}{2C_{\mathcal{L}}^2 C_D^2}\Big), \quad t > 0.$$

Writing $\|\nabla \sum_i \mathcal{L}(\Theta_i^*)\|_{\infty,\infty,\infty} = \sup_{ij\ell} |\frac{\partial\mathcal{L}_i(\Theta_i^*)}{\partial\Theta_{ij\ell}}|$, by the union bound we have,

$$\mathbb{P}\Big(\|\nabla \sum_{i\in\mathcal{U}} \mathcal{L}_i(\Theta_i^*)\|_{\infty,\infty,\infty} > t\Big) \leq 2|\mathcal{U}|NL \cdot \exp\Big(-\frac{nt^2}{2C_{\mathcal{L}}^2 \cdot C_D^2}\Big) \leq \delta, \quad t > 0.$$

Taking $t = C_{\mathcal{L}} \cdot C_D \sqrt{2\log(|\mathcal{U}|NL/\delta)/n}$ with $\delta = (|\mathcal{U}|NL)^{-c}$ establishes the result. ∎

Chapter 4

Restricted Strong Convexity of Time-averaged losses

The restricted strong convexity (RSC) property for the loss function defining the estimator is crucial in proving the main result. The formal result of Proposition 5 is restated below.

Proposition 5 (Restricted Strong Convexity)**.** Under assumptions (A1) and (A2), if we have,

$$n \gtrsim \frac{G}{C_{\mathcal{X}}^{6}} s_{\max}^{3} \log(NL) \tag{3.22}$$

then, the RSC given in equation (3.20) for $\mathcal{S} = \mathcal{S}^*$ holds with curvature $\kappa = \underline{\kappa}$ and tolerance $\tau^2 = \frac{\overline{\kappa}}{2} \sum_{i \in \mathcal{U}} \omega_i^2 / s_i$, with probability at least $1 - (NL)^{-C s_{\max}}$ where $C = O(C_{\mathcal{X}}^{-2})$.

In this chapter we will outline the proof technique for showing this property and provide details on each step.

Showing the RSC property given in equation (3.20) for a particular choice of $\mathcal{S}$ is a major contribution of this dissertation. This is a nontrivial task since it involves uniformly controlling a dependent non-Gaussian empirical process. Even for i.i.d. samples, the task is challenging since the quantity to be controlled, $\Delta \mapsto R\mathcal{L}(\Delta; \Theta^*)$, is a *random function* that needs to be uniformly bounded below. Controlling the behavior of this function becomes significantly harder without the independence assumption.

4.1 Proof Sketch for showing Restricted Strong Convexity

We proceed by a establishing a series of intermediate lemmas which are proved in Section **??**. First, we show that $\boldsymbol{\beta} \mapsto R\mathcal{L}_i(\boldsymbol{\beta}; \Theta_i^*)$ is lower-bounded by the following quadratic form:

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) := \frac{1}{n} \sum_{t=1}^{n} \langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D} \rangle^2, \tag{4.1}$$

where $\mathbb{X} := \{\boldsymbol{x}^t\}_{t=-p+1}^{n}$.

Lemma 8 (Quadratic lower bound)**.** *Under assumption (A2),*

$$R\mathcal{L}_i(\boldsymbol{\beta}; \Theta_i^*) \;\geq\; \frac{\kappa_i}{2}\, \mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \tag{4.2}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^{N \times L}$ and $i \in [N]$.

Notice that $\boldsymbol{\beta} \mapsto \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ is a random function due to the randomness in $\mathbb{X}$. Importantly, $\mathcal{E}(\,\cdot\,; \mathbb{X})$ does not depend on the choice of i. The following set of results establish some important properties of the random function $\mathcal{E}(\,\cdot\,; \mathbb{X})$.

Lemma 9 (Strong convexity at the population level)**.** *Under assumption (A1),*

$$\mathbb{E}\, \mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq C_{\mathbb{X}}^2 \, \|\boldsymbol{\beta}\|_F^2, \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{N \times L}. \tag{4.3}$$

Next, we show that for a fixed $\boldsymbol{\beta}$, the quantity $\mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ concentrates around its mean. Section 5 provides a sketch of the proof of the following concentration inequality:

Lemma 10 (Concentration inequality)**.** *For any* $\boldsymbol{\beta} \in \mathbb{R}^{N \times L}$*, if* $\mathbb{X}$ *is generated as equation* (2.3)*, then with probability at least* $1 - 2\exp(-nt^2/G)$*, we have*

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) > \mathbb{E}\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) - t\|\boldsymbol{\beta}\|_{1,1}^2. \tag{4.4}$$

Finally, for a fixed $i \in [N]$ we use the structural properties of set $\mathbb{C}(S_i^*; \Theta_i^*)$ along with Lemmas 9 and 10 to give a uniform quadratic lower bound on $\mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$, which holds with high probability:

Lemma 11. *Fix* $i \in \mathcal{U}$*. For constants* $C_1, C_2 > 0$*, if* $s_i \geq \frac{C_{\mathbb{X}}^2}{C_1}$*, then with probability* $\geq 1 - \exp(\frac{C_2}{C_{\mathbb{X}}^2} s_i \log(NL) - \frac{nC_{\mathbb{X}}^4}{16Gs_i^2})$,

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq \tfrac{C_{\mathbb{X}}^2}{4}\|\boldsymbol{\beta}\|^2 - \omega_i^2/s_i, \quad \forall \boldsymbol{\beta} \in \mathbb{C}(S_i^*; \Theta_i^*).$$

The proof of Lemma 11 (cf. Appendix 4.4) makes use of a discretization argument. Proving uniform laws are challenging when the parameter space is not finite. The discretization of the set $\mathbb{C}(S^*; \Theta^*)$ uses estimates of the *entropy numbers* for absolute convex hulls of collections of points (Lemma 13). These estimates are well-known in approximation theory and have been previously adapted to the analysis of regression problems in [37]. The following technical lemma allows us to put the above results together:

Lemma 12. *For all* $i \in \mathcal{U}$*, let* a_i, b_i, d_i, p_i *be positive constants, and consider random variables* $X_i, Y_i \in \mathbb{R}$ *which satisfy* $Y_i \geq a_i X_i$*, and* $\mathbb{P}(X_i < b_i - d_i) \leq p_i$ *for all* $i \in \mathcal{U}$*. Then with probability at least* $1 - |\mathcal{U}| \max_{i \in \mathcal{U}} p_i$*, we have,*

$$\sum_{i \in \mathcal{U}} Y_i > (\min_{i \in \mathcal{U}} a_i) \sum_{i \in \mathcal{U}} b_i - (\max_{i \in \mathcal{U}} a_i) \sum_{i \in \mathcal{U}} d_i$$

Proof of Lemma 12. We start by stating a general result that for sets $A, B, \{A_i\}_{i=1}^N, \{B_i\}_{i=1}^N$

from a σ-algebra such that (i) $\bigcap_i A_i \subseteq A \subseteq B$, and (ii) $B_i \subseteq A_i$ for all i, then

$$\mathbb{P}(B) \geq \mathbb{P}(A) \geq \mathbb{P}\Big(\bigcap_i A_i\Big) \geq 1 - \sum_{i=1}^{N} \mathbb{P}(A_i^c) \geq 1 - N \max_i \mathbb{P}(A_i^c) \geq 1 - N \max_i \mathbb{P}(B_i^c). \quad (4.5)$$

The first two inequalities follows from (i), the third inequality is the union bound to $\mathbb{P}(\bigcap_i A_i) = 1 - \mathbb{P}(\cup_i A_i^c)$. The last inequality follows from (ii).

Recall that $Y_i > a_i X_i$, and consider the set definitions $B_i = \{X_i > b_i - d_i\}$, $A_i = \{a_i X_i > (\min_i a_i) b_i - (\max_i a_i) d_i\}$, $A = \{\sum_i a_i X_i > (\min_i a_i)\sum_i b_i - (\max_i a_i)\sum_i d_i\}$ and $B = \{\sum_i Y_i > (\min_i a_i)\sum_i b_i - (\max_i a_i)\sum_i d_i\}$ which satisfy the above inclusion for $a_i, b_i, d_i > 0$. The lemma follows immediately from equation (4.5). ∎

The RSC property, or Proposition 5 follows by applying these choices in Lemma 12: $Y_i = R\mathcal{L}_i(\Delta_i; \Theta_i^*)$, $X_i = \mathcal{E}(\Delta_i, \mathbb{X})$, $a_i = \frac{\kappa_i}{2}$, $b_i = \frac{C_{\mathbb{X}}^2}{4}\|\Delta_i\|_F^2$, and $d_i = \omega_i^2/s_i$.

4.2 Quadratic lower bound on Remainder terms: $\mathcal{E}(\beta, \mathbb{X})$

Fix $i \in \mathcal{U}$. Recall that the loss $\mathcal{L}_i$ can be written as

$$\mathcal{L}_i(\Theta_i) = \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{i,t}(x_i^t, \langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)$$

We have

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_i(\Theta_i)}{\partial \Theta_{iab}\partial \Theta_{ik\ell}} &= \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}''_{i,t}(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\,(\mathbf{X}^{t-1}\mathbf{D})_{ab}(\mathbf{X}^{t-1}\mathbf{D})_{k\ell} \\ &= \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}''_{i,t}(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\left\langle \mathbf{x}_a^{t-*}, \boldsymbol{d}_b\right\rangle \left\langle \mathbf{x}_k^{t-*}, \boldsymbol{d}_\ell\right\rangle.\end{aligned}$$

Let $\nabla^2 \mathcal{L}_i(\Theta_i) \in \mathbb{R}^{(N\times L)\times(N\times L)}$ denote the Hessian matrix of $\mathcal{L}_i$, i.e.

$$\nabla^2 \mathcal{L}_i(\Theta_i) = \left[\frac{\partial^2 \mathcal{L}_i(\Theta_i)}{\partial \Theta_{iab}\partial \Theta_{ik\ell}}\right], \quad (a,b) \in [N]\times[L], (k,\ell) \in [N]\times[L],$$

and define the vector $\mathbf{h}^t := [\langle \mathbf{x}_a^{t-*}, \boldsymbol{d}_b \rangle] \in \mathbb{R}^{N\times L}$. Then we have

$$\nabla^2 \mathcal{L}_i(\Theta_i) = \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}''_{i,t}(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\mathbf{h}^t\mathbf{h}^{t\top}. \tag{4.6}$$

Hence, for all $\Theta_i, \boldsymbol{\beta} \in \mathbb{R}^{N\times L}$, the quadratic form of the Hessian of $\mathcal{L}_i$ satisfies

$$\begin{aligned}\langle \boldsymbol{\beta}\nabla^2\mathcal{L}_i(\Theta_i), \boldsymbol{\beta}\rangle &= \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}''_{i,t}(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\,\mathrm{vec}(\boldsymbol{\beta})^\top \mathbf{h}^t\mathbf{h}^{t\top}\,\mathrm{vec}(\boldsymbol{\beta}) \\ &= \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}''_{i,t}(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)\,\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\rangle^2 \\ &\overset{(i)}{\geq} \frac{\kappa_i}{n}\sum_{t=1}^{n} \langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\rangle^2 := \kappa_i \mathcal{E}(\boldsymbol{\beta}; \mathbb{X}), \end{aligned} \tag{4.7}$$

where vec(β) represents the vectorized form of the matrix $\boldsymbol{\beta}$ (in the same order as rows/columns of $\nabla^2\mathcal{L}_i$), and inequality (i) follows from $\mathcal{L}''_{i,t}(x_i^t, \cdot) \geq \kappa_i > 0$, which holds by Assumption (A2).

Next, consider the function $f(t) := \mathcal{L}(\Theta_i^* + t\boldsymbol{\beta})$. By Taylor's Theorem we have

$$f(1) - f(0) - f'(0) = \frac{1}{2}f''(\xi), \quad \text{for some } \xi \in [0,1].$$

Therefore, there exist a $\xi \in [0,1]$ such that

$$R\mathcal{L}_i(\boldsymbol{\beta}; \Theta_i^*) = \frac{1}{2}\langle \boldsymbol{\beta}\nabla^2\mathcal{L}_i(\Theta_i^* + \xi\boldsymbol{\beta}), \boldsymbol{\beta}\rangle \geq \frac{\kappa_i}{2}\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}),$$

where the last inequality follows from equation (4.7). This completes the proof. □

4.3 Uniform lower bound on $\mathbb{E}\mathcal{E}(\beta; \mathbb{X})$

In this section we provide the proof of Lemma 9.

Using the notation in equation (4.13) and equation (4.14), equation (4.15) implies

$$\mathbb{E}\mathcal{E}(\Delta; \mathbb{X}) = \mathbb{E}\|\mathsf{X}_{t*}\mathsf{S}(\boldsymbol{\beta})\|_2^2 \quad \text{for all } t,$$

since by assumption the process is wide-sense stationary (i.e., the second moments of the distribution of X_{t*} is the same for all t). Recall the stacking operator $\mathtt{S}(\boldsymbol{\beta}) \in \mathbb{R}^{NL}$ defined in equation (4.14), and let $\mathbf{R} := \mathbb{E}\mathsf{X}_{t*}^\top \mathsf{X}_{t*} \in \mathbb{R}^{NL \times NL}$ be the population autocorrelation matrix, again independent of t by stationarity. Then,

$$\mathbb{E}\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) = \mathbb{E}\|\mathsf{X}_{t*}\mathtt{S}(\boldsymbol{\beta})\|_2^2 = \mathbb{E}\,\mathrm{tr}\left(\mathsf{X}_{t*}^\top \mathsf{X}_{t*}\mathtt{S}(\boldsymbol{\beta})\mathtt{S}(\boldsymbol{\beta})^\top\right) = \mathrm{tr}\left(\mathbf{R}\mathtt{S}(\boldsymbol{\beta})\mathtt{S}(\boldsymbol{\beta})^\top\right).$$

Since $\mathbf{R} - \lambda_{\min}(\mathbf{R})I \succeq 0$, we have that

$$\mathbb{E}\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \;\geq\; \lambda_{\min}(\mathbf{R})\|\mathtt{S}(\boldsymbol{\beta})\|_2^2. \tag{4.8}$$

We note that $\mathbf{R}$ is a block symmetric matrix with blocks $\mathbf{R}_{ij} := \mathbb{E}[\boldsymbol{x}^{t-i}(\boldsymbol{x}^{t-j})^\top] \in \mathbb{R}^{N \times N}$. We also note that due to the stationarity, $\mathbf{R}_{ij}$ only depends on $i - j$, so with some abuse of notation we write $\mathbf{R}_{ij} = \mathbf{R}_{i-j}$, i.e., $\mathbf{R}$ is block Toeplitz. Let $\mathbf{C}_{i-j}$ denote the centered autocorrelation matrix $\mathbb{E}[(\boldsymbol{x}^{t-i} - \mathbb{E}\boldsymbol{x}^t)(\boldsymbol{x}^{t-j} - \mathbb{E}\boldsymbol{x}^t)^\top]$, whereby $\mathbf{R}_{i-j} = \mathbf{C}_{i-j} + \mathbb{E}\boldsymbol{x}^t(\mathbb{E}\boldsymbol{x}^t)^\top$. Define $\mathbf{C}$ similarly as a block Toeplitz matrix with $\mathbf{C}_{ij} = \mathbf{C}_{i-j}$. Consequently $\lambda_{\min}(\mathbf{R}) \geq \lambda_{\min}(\mathbf{C})$.

Let $\mathcal{X}(\omega) \in \mathbb{C}^{N \times N}$ be the power spectrum matrix of the process as in assumption (A1) so that

$$\mathbf{C}_\ell := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{X}(\omega)\, e^{j\omega\ell} d\omega, \tag{4.9}$$

Also, recall from assumption (A1) that

$$C_{\mathcal{X}}^2 := \min_{\omega \in [-\pi, \pi)} \lambda_{\min}(\mathcal{X}(\omega)) > 0. \tag{4.10}$$

It is well-known that $\lambda_{\min}(\mathbf{C}) \geq C_{\mathcal{X}}^2$. See for example [2, Proposition 2.3] or [15, Lemma 4.1]. For completeness, we prove this assertion below. This together with equation (4.8) and $\|\mathtt{S}(\boldsymbol{\beta})\|_2^2 = \|\boldsymbol{\beta}\|_F^2$ proves Lemma 9. □

4.3.1 Proof of $\lambda_{\min}(\mathbf{C}) \geq C_{\mathcal{X}}^2$

Fix $\mathbf{u}^\top = \begin{bmatrix} u_0^\top \ u_1^\top \ \dots \ u_{p-1}^\top \end{bmatrix}$, where $u_i \in \mathbb{R}^N$ and set $G(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{r=0}^{p-1} u_r e^{-jr\omega}$. Then, $\mathbf{u}^\top \mathbf{C} \mathbf{u}$ equals,

$$\sum_{r,s=0}^{p-1} u_r^\top \mathbf{C}_{r-s} u_s = \sum_{r,s=0}^{p-1} u_r^\top \Big[\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{X}(\omega) e^{j(r-s)\omega} d\omega\Big] u_s = \int_{-\pi}^{\pi} G^{\mathrm{H}}(\omega) \mathcal{X}(\omega) G(\omega) d\omega. \tag{4.11}$$

Since $\mathcal{X}(\omega)$ is a Hermitian matrix, $G^{\mathrm{H}}(\omega)\mathcal{X}(\omega)G(\omega)$ is always a real matrix. Moreover, we have that

$$G^{\mathrm{H}}(\omega)\mathcal{X}(\omega)G(\omega) \geq \lambda_{\min}(\mathcal{X}(\omega))\, G^{\mathrm{H}}(\omega)G(\omega) \geq C_{\mathcal{X}}^2\, G^{\mathrm{H}}(\omega)G(\omega)$$

hence

$$\mathbf{u}^\top \mathbf{C} \mathbf{u} \geq C_{\mathcal{X}}^2 \int_{-\pi}^{\pi} G^{\mathrm{H}}(\omega) G(\omega) d\omega = C_{\mathcal{X}}^2 \sum_{r,s=0}^{p-1} u_r^\top (\delta_{r-s} I_N) u_s = C_{\mathcal{X}}^2 \|\mathbf{u}\|_2^2,$$

by Parseval's theorem. (Alternatively, reverse the operation in equation (4.11) with $\mathcal{X}(\omega) = 1 \cdot I_N$ and recall that the inverse of a flat spectrum is the delta function). Here, $\delta_x = 1\{x = 0\}$. Taking the minimum over $\|\mathbf{u}\|_2 = 1$ completes the proof. □

4.4 Uniform law for $\mathcal{E}(\beta; \mathbb{X})$

In this section we provide the Proof of Lemma 11.

For the current proof, we have fixed $i \in [N]$. We also use the notation $\|\boldsymbol{\beta}\|_q := \|\boldsymbol{\beta}\|_{q,q}$ for the ℓ_q norm of a matrix $\boldsymbol{\beta} \in \mathbb{R}^{N \times L}$. Note that $\|\boldsymbol{\beta}\|_2 = \|\boldsymbol{\beta}\|_F$. We also use the following notation.

$$\begin{aligned}
&\mathbb{B}_1(r) := \{\boldsymbol{\beta} \in \mathbb{R}^{N\times L} : \|\boldsymbol{\beta}\|_1 \leq r\}, \quad \partial\mathbb{B}_2(r) := \{\boldsymbol{\beta} \in \mathbb{R}^{N\times L} : \|\boldsymbol{\beta}\|_2 = r\}, \\
&\mathbb{B}_p^d(u) := \{D \in \mathbb{R}^d : \|D\|_p \leq u\}. \\
&\quad \omega_i := \omega_{s_i}(\Theta_i^*) = \min_{\boldsymbol{\beta} \in \mathbb{R}^{N\times L}} \{\|\boldsymbol{\beta} - \Theta_i^*\|_1 \mid \|\boldsymbol{\beta}\|_0 \leq s_i\}. \qquad (4.12) \\
&\quad \mathbb{C}_i^* := \mathbb{C}(S_i^*, \Theta_i^*) = \big\{\boldsymbol{\beta} \in \mathbb{R}^{N\times L} : \ \|\boldsymbol{\beta}_{S_i^{*c}}\|_1 \leq 3\|\boldsymbol{\beta}_{S_i^*}\|_1 + 4\omega_i\big\}.
\end{aligned}$$

where S_i^* is the support of the best ℓ_1 approximator of Θ_i^* that has cardinality s_i, i.e., the support of the optimal solution to equation (4.12). One can then show that $\|\Theta^*_{S_i^{*c}}\|_1 = \omega_i$.

We want to show the following inequality,

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq \frac{1}{4} C_{\mathbb{X}}^2 \|\boldsymbol{\beta}\|_F^2 - \tau_i^2, \qquad \forall\, \boldsymbol{\beta} \in \mathbb{C}_i^*.$$

We show this inequality by breaking $\mathbb{C}^*$ into the sets

$$\{\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)\} \ \cup \ \{\mathbb{C}_i^* \cap (\mathbb{B}_F(r_1))^c\} \ \cup \ \{\mathbb{C}_i^* \cap \mathbb{B}_F(\omega_i^2/\sqrt{s_i})\}.$$

For the first two sets of these, the inequality can be shown without any tolerance ($\tau_i^2 = 0$). We need to allow for some tolerance $\tau_i^2 = \omega_i^2/s_i$ when $\omega_i > 0$.

4.4.1 Fixed ℓ_2 norm

Consider the set $\mathbb{C}_i^* \cap \partial\mathbb{B}_2(r_1)$, where $r_1^2 = (\omega_i^2)/s_i + \mathbf{1}_{\{\omega_i = 0\}}$.

Note that for any $\boldsymbol{\beta} \in \mathbb{C}_i^*$, we have $\boldsymbol{\beta} = \boldsymbol{\beta}_{S_i^*} + \boldsymbol{\beta}_{S_i^{*c}}$, and hence

$$\|\boldsymbol{\beta}\|_1 = \ \|\boldsymbol{\beta}_{S_i^*}\|_1 + \|\boldsymbol{\beta}_{S_i^{*c}}\|_1 \ \leq \ 4\|\boldsymbol{\beta}_{S_i^*}\|_1 + 4\|\Theta_{S_i^{*c}}\|_1 \ \leq \ 4\big(\sqrt{s}\|\boldsymbol{\beta}\|_F + \omega_i\big) \qquad \forall\, \boldsymbol{\beta} \in \mathbb{C}_i^*$$

using $\|\boldsymbol{\beta}_{S_i^*}\|_1 \leq \sqrt{s_i}\|\boldsymbol{\beta}_{S_i^*}\|_F$ and $\|\Theta_{S_i^{*c}}\|_1 \leq \omega_i$. It follows that for any $r_1 > 0$,

$$\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1) \ \subseteq \ \mathbb{B}_1\big(r_2\big), \qquad \text{where} \quad r_2 := 4\big(r_1\sqrt{s_i} + \omega_i\big)$$

Next we consider covering $\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)$ by finding a minimum ε-cover of $\mathbb{B}_1(r_2)$. For a metric space (T, ρ), let $\mathcal{N}$ be a minimum ε-cover of T in ρ, i.e., the smallest set $\mathcal{N}$ which satisfies

$$\forall\, \boldsymbol{\beta} \in T, \quad \exists\, \boldsymbol{\beta}' \in \mathcal{N}, \qquad \text{such that} \qquad \rho(\boldsymbol{\beta}, \boldsymbol{\beta}') \leq \varepsilon.$$

The quantity $\mathcal{N}(\varepsilon, T, \rho) := \log |\mathcal{N}|$ for a minimum ε-cover $\mathcal{N}$ is called the metric entropy. The following is an adaptation of a result of [37, Lemma 3, case $q = 1, p = 2$]:

Lemma 13. *Let $\mathsf{X} \in \mathbb{R}^{n\times d}$ be a matrix with column normalization $\|\mathsf{X}_{*j}\|_2 \le \sqrt{n}$ for all j. Consider the following (pseudo) metric in the space $\mathbb{R}^d$, $\rho(D_1, D_2) := \frac{1}{\sqrt{n}}\|\mathsf{X}(D_1 - D_2)\|_2$ on $\mathbb{R}^d$. Then, for a sufficiently small constant $C_1 > 0$, the metric entropy of $\mathbb{B}_1(u)$ in ρ is bounded as*

$$\mathcal{N}\big(\varepsilon, \mathbb{B}_1^d(u), \rho\big) \le \widetilde{C}_2 \frac{u^2}{\varepsilon^2}\log(d), \quad \forall \varepsilon \le \widetilde{C}_1 u.$$

Now, consider a design matrix $\mathsf{X} \in \mathbb{R}^{n\times NL}$ defined as,

$$\mathsf{X}_{t*} := [(\mathbf{X}^{t-1}\boldsymbol{d}_1)^\top \ (\mathbf{X}^{t-1}\boldsymbol{d}_2)^\top \ldots (\mathbf{X}^{t-1}\boldsymbol{d}_L)^\top] \in \mathbb{R}^{1\times NL}, \quad t = 1, 2, \ldots n \tag{4.13}$$

Note that X satisfies the column normalization property $\|\mathsf{X}_{*j}\|_2 \le C_{\mathbb{X}}\sqrt{n}$ for all ℓ since $\mathsf{X}_{tj} \in [-C_{\mathbb{X}}, C_{\mathbb{X}}]$ for all $t \in [n]$ and $j \in [NL]$. Fix $\varepsilon \in (0, 2\widetilde{C}_1 r_2/r_1)$ for sufficiently small $\widetilde{C}_1 > 0$. It follows that there exists an $(r_1\varepsilon/2)$-cover, denoted by $\mathcal{N}_i''$, of $\mathbb{B}_1^{NL}(r_2)$ in the metric defined in Lemma 13 with cardinality bounded as

$$\log |\mathcal{N}_i''| \ \le \widetilde{C}_2 \ \frac{r_2^2}{r_1^2\varepsilon^2}\log(NL).$$

Define a *stacking operator* $\mathtt{S} : \mathbb{R}^{N\times L} \to \mathbb{R}^{NL}$ that flattens a matrix into a vector columnwise:

$$\mathtt{S}(\boldsymbol{\beta}) := \begin{bmatrix} \boldsymbol{\beta}_{*1} \\ \vdots \\ \boldsymbol{\beta}_{*L} \end{bmatrix} \in \mathbb{R}^{NL}. \tag{4.14}$$

Also denote for a set A denote by $\mathtt{S}(A) = \{\mathtt{S}(a) \mid a \in A\}$. Then we have

$$\mathtt{S}(\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)) \;\subseteq\; \mathtt{S}(\mathbb{B}_1\big(r_2\big)) = \mathbb{B}_1^{NL}(r_2).$$

Define a (pseudo) metric on the matrix space $\mathbb{R}^{N\times L}$ as $\overline{\rho}(\boldsymbol{\beta}, \boldsymbol{\beta}') := \rho(\mathtt{S}(\boldsymbol{\beta}), \mathtt{S}(\boldsymbol{\beta}'))$. Since $\mathtt{S}$ is a bijection, it follows that there is an exterior $(r_1\varepsilon/2)$-covering of $\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)$ in metric $\overline{\rho}$ with the same cardinality as $\mathcal{N}_i''$; call it $\mathcal{N}_i'$. (Here, the exterior covering means that the elements need not belong the set they cover. Elements of $\mathcal{N}_i'$ are matrices in $\mathbb{B}_1(r_2)$ but not necessarily in $\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)$.)

We can pass from $\mathcal{N}_i'$ to an $(r_1\varepsilon)$-cover of $\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)$, denoted by $\mathcal{N}_i$ such that $|\mathcal{N}_i| \leq |\mathcal{N}_i'|$ (see Exercise 4.2.9 in [46, p.75]). In particular, we have $\mathcal{N}_i \subseteq \mathbb{C}_i^* \cap \mathbb{B}_F(r_1)$.

Using the following equality which is proved in Appendix 4.4.4,

$$\mathcal{E}(\Delta; \mathbb{X}) = \frac{1}{n}\|\mathsf{X}\,\mathtt{S}(\boldsymbol{\beta})\|_2^2, \tag{4.15}$$

by the triangle inequality $\big||a| - |b|\big| \leq |a - b|$, we get,

$$|\sqrt{\mathcal{E}(\boldsymbol{\beta}; \mathbb{X})} - \sqrt{\mathcal{E}(\boldsymbol{\beta}'; \mathbb{X})}| \leq \overline{\rho}(\boldsymbol{\beta}, \boldsymbol{\beta}'), \qquad \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^{N\times L}$$

for any two matrices $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Using $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$, with $b = \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$, and $a = \mathcal{E}(\boldsymbol{\beta}'; \mathbb{X})$ we have

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq \frac{1}{2}\mathcal{E}(\boldsymbol{\beta}'; \mathbb{X}) - \overline{\rho}^2(\boldsymbol{\beta}, \boldsymbol{\beta}').$$

If follows that

$$\inf_{\boldsymbol{\beta} \in \mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)} \mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \;\geq\; \frac{1}{2} \inf_{\boldsymbol{\beta} \in \mathcal{N}_i} \mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) - (r_1\varepsilon)^2$$

By Lemma 10 and the union bound, with probability at least $1-|\mathcal{N}_i|\exp(\frac{-nt^2}{G})$, we have

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X})-\mathbb{E}\mathcal{E}(\boldsymbol{\beta};\mathbb{X})\geq -t\|\boldsymbol{\beta}\|_{1,1}^2, \quad \forall\, \boldsymbol{\beta}\in\mathcal{N}_i.$$

Since $\mathcal{N}_i\subseteq\mathbb{C}_i^*\cap\mathbb{B}_F(r_1)$, for any $\boldsymbol{\beta}\in\mathcal{N}_i$ we have $\|\boldsymbol{\beta}\|_{1,1}^2\leq s_i\|\boldsymbol{\beta}\|_F^2$ and $\|\boldsymbol{\beta}\|_F=r_1$. It follows that with the same probability $1-|\mathcal{N}_i|\exp(\frac{-nt^2}{G})$,

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) \;\geq\; \mathbb{E}\mathcal{E}(\boldsymbol{\beta};\mathbb{X})-t\,sr_1^2 \;\geq\; (C_{\mathbb{X}}^2-ts)\,r_1^2, \quad \forall \boldsymbol{\beta}\in\mathcal{N}_i$$

where we have used Lemma 9 in the second inequality. It follows that with the same probability

$$\inf_{\boldsymbol{\beta}\,\in\,\mathbb{C}^*\,\cap\,\partial\mathbb{B}_F(r_1)}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) \;\geq\; \Big(\frac{1}{2}C_{\mathbb{X}}^2-\frac{1}{2}ts-\varepsilon^2\Big)\,r_1^2. \tag{4.16}$$

Taking $r_1=(\omega_i+\mathbf{1}_{\{\omega_i=0\}})/\sqrt{s_i}$, we can balance the two terms in r_2. We obtain

$$4\sqrt{s_i}\;\leq r_2/r_1\;\leq 8\sqrt{s_i}.$$

The constraint on ε is $\varepsilon\leq 2\widetilde{C}_1(r_2/r_1)$. It is enough to require $\varepsilon\leq 8\widetilde{C}_1\sqrt{s_i}$. Taking $\varepsilon^2=\frac{1}{8}C_{\mathbb{X}}^2$ and assuming that $s_i\geq\frac{C_{\mathbb{X}}^2}{512\widetilde{C}_1^2}=:C_{\mathbb{X}}^2/C_1$ satisfies the constraint. Also, taking $t=\frac{1}{4}C_{\mathbb{X}}^2/s_i$, we obtain

$$\mathbb{P}\Big(\inf_{\boldsymbol{\beta}\,\in\,\mathbb{C}_i^*\,\cap\,\partial\mathbb{B}_F(r_1)}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) \;\geq\; \big(\frac{1}{4}C_{\mathbb{X}}^2\big)\,r_1^2\Big)\geq 1-\exp\Big(\log|\mathcal{N}_i|-C_{\mathbb{X}}^4\frac{n}{16s_i^2G}\Big)=:P_i \tag{4.17}$$

Noting that

$$\log|\mathcal{N}_i|\leq\widetilde{C}_2\big(8\sqrt{s_i}\big)^2\Big(\frac{8}{C_{\mathbb{X}}^2}\Big)\log(NL),$$

the probability is further bounded as

$$1 - P_1 \;\leq\; \exp\Big(\frac{C_2}{C_{\mathbb{X}}^2} s_i \log(NL) - C_{\mathbb{X}}^4 \frac{n}{16 s_i^2 G}\Big),$$

where $C_2 := 512\widetilde{C}_2$. Thus, we have established RSC with high probability for matrices in $\mathbb{C}_i^* \cap \partial\mathbb{B}_F(r_1)$ with curvature $\kappa = \frac{1}{4}C_{\mathbb{X}}^2$ and tolerance $\tau^2 = 0$, as shown in equation (4.17).

Note that when $\omega_i = 0$ (i.e., the case of hard sparsity), $\mathbb{C}_i^*$ is a cone hence the above extends immediately to all $\boldsymbol{\beta} \in \mathbb{C}_i^*$, since $\mathcal{E}(c\boldsymbol{\beta}; \mathbb{X}) = c^2\mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ for all $c > 0$, thus completing the proof. Let us assume $\omega_i > 0$ in the rest of the proof.

4.4.2 Extending to the complement of the ℓ_2 norm ball

For $\omega_i > 0$, since $\mathbb{C}_i^*$ is not a cone, we cannot use a scale-invariance argument to extend to general matrices. However, we have the following:

Lemma 14. *Assume that RSC holds for $\mathcal{E}$ in the sense of $\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq \kappa\|\boldsymbol{\beta}\|_F^2$, for all $\boldsymbol{\beta} \in \mathbb{C}_i^* \cap \partial\mathbb{B}_F(r)$. Then, RSC holds in the same sense for all $\boldsymbol{\beta} \in \mathbb{C}_i^* \cap \{\boldsymbol{\beta} :\ \|\boldsymbol{\beta}\|_F \geq r\}$.*

We skip the proof since it can be verified without much difficulty. The lemma establishes the RSC of the previous step for all of $\mathbb{C}_i^* \cap \{\boldsymbol{\beta} :\ \|\boldsymbol{\beta}\|_F \geq r_1\}$. The proof is straightforward and follows from the observation that $\mathcal{E}(\cdot\,; \mathbb{X})$ satisfies $\mathcal{E}(c\boldsymbol{\beta}; \mathbb{X}) = c^2\mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$, for $c \geq 1$.

4.4.3 Extending to small radii

It remains to extend the result to $\boldsymbol{\beta} \in \mathbb{C}^* \cap \{\boldsymbol{\beta} :\ \|\boldsymbol{\beta}\|_F < r_1\}$. In this case, we simply take $\tau^2 := r_1^2 = \omega_i^2/s_i$ (since $\omega_i > 0$ by assumption) so that

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq 0 \geq \|\boldsymbol{\beta}\|_F^2 - \tau^2$$

so that the RSC holds with curvature $= 1$ and tolerance τ^2. Putting the pieces together, we have the RSC for all $\boldsymbol{\beta} \in \mathbb{C}_i$ with the probability given in Step 1, curvature $\kappa_i = \min\{\frac{1}{4}C_{\mathbb{X}}^2, 1\}$

and tolerance $\tau_i^2 = \omega_i^2/s_i$. This concludes the proof. □

4.4.4 Proof of equality equation (4.15)

The right hand side is

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}(\mathtt{X}_{t*}\mathtt{S}(\boldsymbol{\beta}))^2 &= \frac{1}{n}\sum_{t=1}^{n}\sum_{\ell=1}^{L}(\mathbf{X}^{t-1}\boldsymbol{d}_\ell)^\top\boldsymbol{\beta}_{*\ell} \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{\ell=1}^{L}\boldsymbol{\beta}_{*\ell}^\top(\mathbf{X}^{t-1}\boldsymbol{d}_\ell) \\
&= \frac{1}{n}\text{trace}(\boldsymbol{\beta}^\top\mathbf{X}^{t-1}\mathbf{D}) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left\langle\boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\right\rangle
\end{aligned}$$

This proves the claim. □

Chapter 5

Concentration of Empirical Processes under Long-range Dependence

In this chapter, we sketch the proof of Lemma 10 which is a concentration inequality for $\boldsymbol{\beta} \mapsto \mathcal{E}(\boldsymbol{\beta};\mathbb{X})$, a quadratic empirical process based on dependent non-Gaussian variables with long-term dependence. We restate Lemma 10 below. Recall the definition of $\mathcal{E}(\beta;\mathbb{X})$:

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) := \frac{1}{n}\sum_{t=1}^{n}\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\rangle^2, \tag{5.1}$$

where $\mathbb{X} := \{\boldsymbol{x}^t\}_{t=-p+1}^{n}$ are observations of a trajectory of the time-series.

Lemma 10 (Concentration inequality)**.** *For any $\boldsymbol{\beta} \in \mathbb{R}^{N\times L}$, if $\mathbb{X}$ is generated as equation* (2.3)*, then with probability at least* $1 - 2\exp\left(-nt^2/G\right)$*, we have*

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) > \mathbb{E}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) - t\|\boldsymbol{\beta}\|_{1,1}^2. \tag{4.4}$$

For independent sub-Gaussian variables $\{\mathbf{X}^{t-1}\}$, such a concentration result is often called the Hanson–Wright inequality [40, Thm. 1]. Providing similar inequalities for dependent random variables is significantly more challenging. For dependent Gaussian variables, the machinery of the Hanson–Wright inequality can still be adapted to derive the desired result [2, Prop. 2.4]. However, these arguments do not extend easily to non-Gaussian dependent variables and hence other techniques are needed to provide such concentration inequalities.

Recent results [14, 8] on the concentration of empirical processes derived from Markov chains could provide improvements on the results we present here. However, since we are dealing with a non-Markovian process (when $p > 1$), such results are not directly applicable. To that end we derive some new results regarding the mixing properties of the $p-$Markov chains that elicit conditions under which we can show concentration of running averages to their mean. We note that this is an active area of research in theoretical statistics. The concentration inequality in statement of the lemma is an application of the Azuma-Hoeffding inequality for bounded martingale difference sequences. See [45]. A key observation, discussed in Section 5.3, is that process equation (2.3) can be represented as a discrete-space *p-Markov chain*. This allows us to use concentration results for dependent processes in countable metric spaces. There are several results for such processes; see [21, 26, 42] and [20] for a review. Here, we apply that of Kontorovich et. al. [21]. These concentration inequalities are stated in terms of various mixing and contraction coefficients of the underlying process. The challenge is to control the contraction coefficients in terms of the process parameter Θ^*, which in our case is done using quantities $\tau_1(\Theta^*)$ and $G(\Theta^*)$. Some results developed in this section hold more generally for any $p-$Markov process, even those outside the current autoregressive framework.

5.1 Concentration of functions of Dependent variables

We start by stating the result of Kontorovich et. al. [21] for a process $\{X^t\}_{t\in[n]}$ consisting of (possibly dependent) random variables taking values in a countable space $\mathcal{X}$. For any $\ell \geq k \geq 1$, define the *mixing coefficient*

$$\eta_{k\ell} \triangleq \sup_{w,w',y} \left\| \mathbb{P}\left(X_\ell^n = \cdot \mid X_k = w', X_1^{k-1} = y\right) - \mathbb{P}\left(X_\ell^n = \cdot \mid X_k = w, X_1^{k-1} = y\right) \right\|_{\mathrm{TV}}, \quad (5.2)$$

where the supremum is over $w, w' \in \mathcal{X}$ and $y \in \mathcal{X}^{k-1}$. Here, $X_u^v := (X^t, u \leq t \leq v)$ is viewed either as a member of $\mathcal{X}^{\times(v-u+1)}$ (the set of a matrices with $v-u+1$ columns from $\mathcal{X}$) or simply as a vector in $\mathcal{X}^{v-u+1}$. Let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be an upper triangular matrix with entries $\eta_{k\ell}$ for $\ell \geq k$ and zero otherwise. Let $\|\mathbf{H}\|_\infty := \max_k \sum_{\ell \geq k} \eta_{k\ell}$ be the ℓ_∞ operator norm of $\mathbf{H}$.

Proposition 15. *[21, Theorem 1.1] Let $\phi : \mathcal{X}^n \to \mathbb{R}$ be an L_ϕ-Lipschitz function of $\{X^t\}_{t=1}^n$ with respect to the Hamming norm, then for all $\varepsilon > 0$, with probability at least*

$$1 - 2\exp\left(-\frac{\varepsilon^2}{2nL_\phi^2\|\mathbf{H}\|_\infty^2}\right),$$

we have

$$|\phi(\{X^t\}_{t=1}^n) - \mathbb{E}\phi(\{X^t\}_{t=1}^n)| < \varepsilon. \tag{5.3}$$

We apply the above result to $\phi = \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ by finding an upper bound for the Lipschitz constant L_ϕ of the map $\mathbb{X} \mapsto \mathcal{E}(\boldsymbol{\beta}, \mathbb{X})$ with respect to the Hamming distance over $\mathcal{X}^{\times(n+p-1)} = (\prod_{i=1}^N \mathcal{X}_i)^{\times(n+p-1)}$. Lemmas 16 and 17 in Section 5.2 shows that $L_\phi \leq (4B^2C_{\mathbf{D}}^2/n)\|\boldsymbol{\beta}\|_{1,1}^2$, and that $\|\mathbf{H}\|_\infty^2 \leq 2(1 + p^2\psi_1(\Theta^*))$, where the quantity $\psi_1(\Theta^*)$ is defined below equation (3.5). Lemma 17 is a general result that applies to any p-lag Markov chain, including the GVAR(p) processes considered in this dissertation. In Section 5.3 we also develop some tools for controlling $\|\mathbf{H}\|_\infty$ in terms of the contraction coefficient of another related Markov chain obtained via a non-standard state augmentation.

Returning to the proof of Lemma 10: applying Theorem 15 with $\varepsilon = t\|\boldsymbol{\beta}\|_{1,1}^2$, and using the upper bounds for L and $\|\mathbf{H}\|_\infty^2$ concludes the proof of Lemma 10.

5.2 Proof Sketch for Concentration inequality

In this section, we prove the following two main lemmas used in Chapter 5.

Lemma 16. *The map $\mathbb{X} \mapsto \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ is Lipschitz with respect to the Hamming distance on $\mathcal{X}^{\times(n+p-1)}$, with Lipschitz constant at most $(4B^2 C_{\mathbf{D}}^2/n)\|\boldsymbol{\beta}\|_{1,1}^2$.*

A process over a countable space $\mathcal{X}$ is referred to as a *p -Markov chain* if for some finite p,

$$\mathbb{P}(\boldsymbol{x}^t = z | \{\boldsymbol{x}^{t-k}\}_{k \in \mathbb{N}_+}) = \mathbb{P}(\boldsymbol{x}^t = z | \{\boldsymbol{x}^{t-k}\}_{k=1}^{p}), \tag{5.4}$$

for all $z \in \mathcal{X}$, for all $t \in \mathbb{Z}$. To keep the exposition simple, we assume that $\mathbb{P}$ above does not depend on t, i.e., the process is homogeneous.

Recall the notation $\tau_1(\Theta^*)$ defined in equation (3.6), whereby $\tau_1(\mathcal{K}^p) = \tau_1(\Theta^*)$ by definition. The following lemma provides an upper bound for $\|\mathbf{H}\|_\infty$ as a function of $\tau_1(\Theta^*)$.

Lemma 17. *For a p-Markov process over $\mathcal{X}$, with equivalent kernel $\mathcal{K} \in \mathbb{R}^{|\mathcal{X}|^p \times |\mathcal{X}|^p}$ given by equation* (5.13) *with $r = p$, the mixing coefficients defined in equation* (5.2) *are bounded as*

$$\eta_{k\ell} \;\le\; \tau_1(\Theta^*)^{1+\lfloor (\ell-k-1)/p) \rfloor}, \quad \ell \;\ge k. \tag{5.5}$$

In particular, for any $\tau \in [\tau_1(\Theta^*), 1)$

$$\|\mathbf{H}\|_\infty^2 := \Big(\max_{k \in [n]} \sum_{\ell \ge k} \eta_{k\ell} \Big)^2 \;\le\; 2 + \frac{2p^2}{(\tau^{-1} - 1)^2}. \tag{5.6}$$

5.2.1 Bounding Lipschitz constant: Proof of Lemma 16

It is enough to consider two sequences $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{y}^t\}$ that differ in a single time step, say at time point r, so that the state vectors can be written as $\mathbb{X} = (\boldsymbol{x}^{-p+1}, \boldsymbol{x}^{-p+2}, \dots, \boldsymbol{x}^r, \dots, \boldsymbol{x}^{n-1})$ and $\mathbb{Y} = (\boldsymbol{x}^{-p+1}, \boldsymbol{x}^{-p+2}, \dots, \boldsymbol{y}^r, \dots, \boldsymbol{x}^{n-1})$, where r will be fixed. The general case follows, via triangle inequality, since any $\widetilde{\mathbb{Y}}$ can be reached from $\mathbb{X}$ by a sequence $\mathbb{X} =: \mathbb{X}_{(0)}, \mathbb{X}_{(1)}, \dots, \mathbb{X}_{(h)} := \widetilde{\mathbb{Y}}$ such that $\mathbb{X}_{(i)}$ and $\mathbb{X}_{(i-1)}$ are Hamming distance 1 apart, for $i = 1, 2, \dots h$, where h is the hamming distance of $\mathbb{X}$ and $\widetilde{\mathbb{Y}}$ in $\mathcal{X}^{n+p-1}$.

Let $\mathbf{X}^{t-1}$ and $\mathbf{Y}^{t-1}$ be defined based on $\mathbb{X}$ and $\mathbb{Y}$ as before, i.e., the corresponding

p-lag history at time $t-1$. Note that $\mathbf{X}^{t-1}$ and $\mathbf{Y}^{t-1}$ are different only for t such that $t \in \{r+1, \ldots, r+p\}$, and for such t, we have via Hölder's inequality:

$$|\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D} - \mathbf{Y}^{t-1}\mathbf{D}\rangle| \leq 2B \|(\boldsymbol{\beta}\mathbf{D}^\top)_{*,t-r}\|_1$$
$$|\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D} + \mathbf{Y}^{t-1}\mathbf{D}\rangle| \leq 2B \|\boldsymbol{\beta}\mathbf{D}^\top\|_{1,1}.$$

where $\boldsymbol{M}_{*,i}$ is the i^{th} column of a matrix $\boldsymbol{M}$. Note the inner products above are over matrices in $\mathbb{R}^{N\times L}$. In the above inequality we have also used the fact that for any $\boldsymbol{M} \in \mathbb{R}^{N\times p}$, we have $\langle \boldsymbol{\beta}, \boldsymbol{M}\mathbf{D}\rangle = \left\langle \boldsymbol{\beta}\mathbf{D}^\top, \boldsymbol{M}\right\rangle$ where the second inner product is over $\mathbb{R}^{N\times p}$. Combining the above inequalities we obtain

$$\begin{aligned}
|\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) - \mathcal{E}(\boldsymbol{\beta}; \mathbb{Y})| &= \frac{1}{n}\Big| \sum_{t=r+1}^{r+p} \left[\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\rangle^2 - \langle \boldsymbol{\beta}, \mathbf{Y}^{t-1}\mathbf{D}\rangle^2\right] \Big| \\
&\leq \sum_{t=r+1}^{r+p} |\langle \boldsymbol{\beta}, (\mathbf{X}^{t-1} - \mathbf{Y}^{t-1})\mathbf{D}\rangle||\langle \boldsymbol{\beta}, (\mathbf{X}^{t-1} + \mathbf{Y}^{t-1})\mathbf{D}\rangle| \\
&\leq \frac{4B^2}{n} \sum_{t=r+1}^{r+p} \|(\boldsymbol{\beta}\mathbf{D}^\top)_{*,t-r}\|_1 \, \|\boldsymbol{\beta}\mathbf{D}^\top\|_{1,1} = \frac{4B^2}{n} \|\boldsymbol{\beta}\mathbf{D}^\top\|_{1,1}^2
\end{aligned}$$

Finally, $\|\boldsymbol{\beta}\mathbf{D}^\top\|_{1,1} = \|\mathbf{D}\boldsymbol{\beta}^\top\|_{1,1} = \sum_{\ell=1}^{L} \|\mathbf{D}(\boldsymbol{\beta}^\top)_{*,\ell}\|_1 \leq C_{\mathbf{D}} \|\boldsymbol{\beta}_{\ell,*}\|_1 = C_{\mathbf{D}}\|\boldsymbol{\beta}\|_{1,1}$, where we have used the fact that $C_{\mathbf{D}}$ is the 1 → 1 operator norm of the matrix $\mathbf{D}$, i.e., $C_{\mathbf{D}} = \max_{\boldsymbol{u}\neq \mathbf{0}} \frac{\|\mathbf{D}\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_1}$. This proves the claim. □

5.3 Contraction in p-Markov chains

In this section we develop the necessary background to prove Lemma 17. We start by recalling a well-known contraction quantity, the *Dobrushin ergodicity coefficient*, and relating it to the mixing coefficients of p-Markov processes.

5.3.1 Dobrushin ergodicity coefficient

For a Markov chain (or 1-Markov process) over a discrete space $\mathcal{X}$, let $P = (P_{ij}) \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ be its transition kernel. The kernel is a nonnegative stochastic matrix, i.e., each row is a probability distribution. Thus, $P \geq 0$ and $P\mathbf{1} = \mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^{|\mathcal{X}|}$ is the all-ones vector. Let

$$\mathcal{H}_1 := \big\{u \in \mathbb{R}^{|\mathcal{X}|} \;|\; \mathbf{1}^\top u = 0\big\}. \tag{5.7}$$

This subspace is invariant to every Markov kernels $P \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$, i.e., for all $u \in \mathcal{H}_1$, we have $u^\top P \in \mathcal{H}_1$. The *Dobrushin ergodicity coefficient* of P is defined as

$$\tau_1(P) := \sup_{u\in\mathcal{H}_1} \frac{\|u^\top P\|_1}{\|u\|_1}. \tag{5.8}$$

It follows from the invariance of $\mathcal{H}_1$ to P that

$$\|u^\top P^\ell\|_1 \;\leq\; \tau_1(P)^\ell \|u\|_1 \quad \forall u \in \mathcal{H}_1. \tag{5.9}$$

The following alternative characterization is well-known [39].

Lemma 18. *The Dobrushin ergodicity coefficient of P satisfies*

$$\tau_1(P) = \tfrac{1}{2} \sup_{x,y\,\in\mathcal{X}} \|(e_x - e_y)^T P\|_1 \tag{5.10}$$

where e_x is the x-th basis vector of $\mathbb{R}^{\mathcal{X}}$.

Proof. Optimization problem in equation (5.8) is scale invariant, hence,

$$\tau_1(P) = \sup_{u\,\in\,\mathcal{H}_1(1)} \|u^\top P\|_1, \tag{5.11}$$

where $\mathcal{H}_1(1) = \{u \in \mathcal{H}_1 \mid \|u\|_1 \leq 1\}$. We will show that the set $\mathcal{H}_1(1) = C := \mathrm{conv}(\{\frac{1}{2}(e_x - e_y)\})$. Using this, equation (5.11) is a maximization of a convex function $\|u^\top P\|_1$ over a polytope with extreme points $\frac{1}{2}(e_x - e_y), x, y \in \mathcal{X}$. It follows that the maximum occurs, at least, at an extreme point, which gives the desired result. The inequality in the statement of the lemma follows since the total-variation is bounded by 1.

The rest of the proof establishes $\mathcal{H}_1(1) = C$. The inclusion $C \subseteq \mathcal{H}_1(1)$ can be verified easily by checking the membership of extreme points of C in $\mathcal{H}_1(1)$, since $\mathcal{H}_1(1)$ is a convex set. We now prove the nontrivial direction $\mathcal{H}_1(1) \subseteq C$.

Let the ambient space be $\mathbb{R}^m$, Δ_m the probability simplex in $\mathbb{R}^m$, and $\partial\mathbb{B}_1 := \{u \in \mathbb{R}^m : \|u\|_1 = 1\}$ the boundary of ℓ_1 ball. We have $C = \frac{1}{2}\Delta_m + \frac{1}{2}(-\Delta_m)$, which is a Minkowski sum. This follows since taking the Minkowski sum and taking the convex hull commute [22, Theorem 3]. Hence, it suffices to show that for any vector $u \in \mathcal{H}_1(1)$, there exists a pair of probability vectors $\pi_1, \pi_2 \in \Delta_m$ such that $u = \frac{1}{2}(\pi_1 - \pi_2)$. Since $0 \in C$, and $\mathcal{H}_1(1) = \operatorname{conv}(0, \partial\mathbb{B}_1 \cap \mathcal{H}_1)$, it is enough to consider $u \in \partial\mathbb{B}_1 \cap \mathcal{H}_1$.

Let $u \in \partial\mathbb{B}_1 \cap \mathcal{H}_1$, and let u_+ and u_- be the positive and negative parts of u, that is, $(u_+)_i = \max(u_i, 0)$ and $(u_-)_i = -\min(u_i, 0)$. Taking $\pi_1 = 2u_+$ and $\pi_2 = 2u_-$, we have $u = \frac{1}{2}(\pi_1 - \pi_2)$. Also, due to $u \in \partial\mathbb{B}_1$, $1 = \|u\|_1 = \frac{1}{2}\|\pi_1\|_1 + \frac{1}{2}\|\pi_2\|_1$ whereas due to $u \in \mathcal{H}_1$, $0 = \mathbf{1}^\top u = \frac{1}{2}\|\pi\|_1 - \|\pi_2\|_1$. It follows that $\|\pi_1\|_1 = \|\pi_2\|_1 = 1$, that is, $\pi_1, \pi_2 \in \Delta_m$. This concludes the proof. ∎

Recall that $\|\pi_1 - \pi_2\|_{\mathrm{TV}}$ denotes the *total variation* distance between probability distributions π_1 and π_2. For discrete distributions we have, $\|\pi_1 - \pi_2\|_{\mathrm{TV}} = \frac{1}{2}\|\pi_1 - \pi_2\|_1 \leq\ 1$, with equality if and only if π_1 and π_2 are orthogonal, i.e., have completely mismatched supports. Consequently, for any stochastic matrix P, we have $\tau_1(P) \leq 1$. Furthermore, the inequality is strict if and only if no two rows of P are orthogonal. Markov kernels with $\tau_1(\cdot) < 1$ are said to be *scrambling.* A sufficient condition for $\tau_1(P) < 1$ is P having at least one column with all entries positive.

5.3.2 The p-step chain

A p-Markov process can be equivalently represented by a Markov kernel $\mathcal{K} \in [0,1]^{|\mathcal{X}|^p \times |\mathcal{X}|^p}$ that gives transition probabilities for consecutive blocks of size p. For any $t \in \mathbb{Z}$,

$$\mathcal{K}_{\boldsymbol{i}\boldsymbol{j}} = \mathbb{P}\Big((\boldsymbol{x}^{t+1-k})_{k=1}^{p} = \boldsymbol{j} \Big| \, (\boldsymbol{x}^{t-k})_{k=1}^{p} = \boldsymbol{i} \Big), \tag{5.12}$$

for all $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{X}^{\times p}$. Kernel matrix $\mathcal{K}$ is constrained since $\mathcal{K}_{\boldsymbol{i}\boldsymbol{j}}$ can be nonzero only if $(j_2, j_3, \ldots, j_p) = (i_1, i_2, \ldots, i_{p-1})$. The r-step chain associated with $\mathcal{K}$ has kernel $\mathcal{K}^r$. In general, for all $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{X}^{\times p}$ and for $r \geq 1$

$$(\mathcal{K}^r)_{\boldsymbol{i}\boldsymbol{j}} = \mathbb{P}\Big((\boldsymbol{x}^{t+r-k})_{k=1}^{p} = \boldsymbol{j} \;\Big|\; (\boldsymbol{x}^{t-k})_{k=1}^{p} = \boldsymbol{i} \Big). \tag{5.13}$$

Similarly, $(\mathcal{K}^r)_{\boldsymbol{i}\boldsymbol{j}}$ can be nonzero only if $(j_{r+1}, j_{r+2}, \ldots j_p) = (i_1, i_2, \ldots, i_{p-r})$, for $r < p$. However, no such constraint applies for $r \geq p$. Moreover, one can verify that for $r < p$, a pair of rows $(\mathcal{K}^r)_{\boldsymbol{i}*}$ and $(\mathcal{K}^r)_{\boldsymbol{i}'*}$ are always orthogonal for $\boldsymbol{i}, \boldsymbol{i}' \in \mathcal{X}^{\times p}$ such that $i_1 \neq i_1'$. Consequently, $\tau_1(\mathcal{K}^r) = 1$ for all $r < p$.

Fortunately for $r = p$, one can show that $\tau_1(\mathcal{K}^p) < 1$, under the mild assumption that

$$\mathbb{P}\big(\boldsymbol{x}^t = z | (\boldsymbol{x}^{t-k})_{k=1}^{p} = \boldsymbol{j}\big) > 0 \;\text{ for all } z \in \mathcal{X} \text{ and } \boldsymbol{j} \in \mathcal{X}^{\times p},$$

since this implies that $\mathcal{K}^p$ is a positive matrix and hence *scrambling.* Note that the above condition always holds for the process defined in equation (2.3).

5.4 Bounding Dobrushin ergodicity coefficient

In this section we provide a proof of equation (5.5). We assume that the reader has familiarized themself with the relevant background explained in Section 5.3.

Recall that $\mathbb{X}_{-p+1}^{n} := \{\boldsymbol{x}^n, \boldsymbol{x}^{n-1}, \ldots, \boldsymbol{x}^{-p+1}\}$ together make n steps of the p-Markov process.

Fix $k \geq 1$ and take $w \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{X}^{p-1}$, and $\boldsymbol{z} \in \mathcal{X}^{k-1}$. We use the shorthand $\mathbb{X}_{-p+1}^{k} = w\boldsymbol{y}\boldsymbol{z}$, to denote $\boldsymbol{x}^k = w, \mathbb{X}_{k-p+1}^{k-1} = \boldsymbol{y}$ and $\mathbb{X}_{-p+1}^{k-p} = \boldsymbol{z}$ and define the law

$$\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z}) := \mathbb{P}\big(\mathbb{X}_\ell^n = \cdot \mid \mathbb{X}_{-p+1}^{k} = w\boldsymbol{y}\boldsymbol{z}\big) = \mathbb{P}\big(\mathbb{X}_\ell^n = \cdot \mid \mathbb{X}_{k-p+1}^{k} = w\boldsymbol{y}\big) =: \mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y})$$

using the p-Markov property, showing that $\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z})$ does not depend on $\boldsymbol{z}$. Thus, we also write $\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y})$ for $\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z})$.

Case 1. Assuming $\ell + p \leq n$, we have

$$\begin{aligned}
&\mathbb{P}\big(X_\ell^n = x_\ell^n \mid X_{k-p+1}^{k} = w\boldsymbol{y}\big) \\
&= \mathbb{P}\big(\mathbb{X}_{\ell+p}^n = x_{\ell+p}^n \mid \mathbb{X}_\ell^{\ell+p-1} = x_\ell^{\ell+p-1}\big) \cdot \mathbb{P}\big(\mathbb{X}_\ell^{\ell+p-1} = x_\ell^{\ell+p-1} \mid \mathbb{X}_{k-p+1}^{k} = w\boldsymbol{y}\big) \\
&= \phi\big(x_{\ell+p}^n \mid x_\ell^{\ell+p-1}\big) \cdot \psi_{w\boldsymbol{y}}(x_\ell^{\ell+p-1})
\end{aligned}$$

where we have defined $\phi(u \mid v) := \mathbb{P}\big(\mathbb{X}_{\ell+p}^n = u \mid \mathbb{X}_\ell^{\ell+p-1} = v\big)$ and

$$\psi_{w\boldsymbol{y}}(\beta) := \mathbb{P}\big(\mathbb{X}_\ell^{\ell+p-1} = \beta \mid \mathbb{X}_{k-p+1}^{k} = w\boldsymbol{y}\big)$$

We note that $\psi_{w\boldsymbol{y}}(\cdot)$ is the $w\boldsymbol{y}$-th row of $\mathcal{K}^{\ell+p-k-1}$ which follows by comparing the definition of $\psi_{w\boldsymbol{y}}$ with equation (5.13) applied with $t = k+1$ and $r = \ell + p - k - 1$. Letting e_i denote the i^{th} row of identity in $\mathbb{R}^{|\mathcal{X}|^p \times |\mathcal{X}|^p}$, we have

$$\psi_{w\boldsymbol{y}} = e_{w\boldsymbol{y}}^{\top}\, \mathcal{K}^{\ell+p-k-1}.$$

Now, we have

$$
\begin{aligned}
2\|\mathcal{L}^{(\ell \to n)}_{w\boldsymbol{y}\boldsymbol{z}} - \mathcal{L}^{(\ell \to n)}_{w'\boldsymbol{y}\boldsymbol{z}}\|_{\mathrm{TV}} &= \sum_{x_\ell^n} \left|\mathbb{P}\big(\mathbb{X}_\ell^n = x_\ell^n \mid \mathbb{X}^k_{k-p+1} = w\boldsymbol{y}\big) - \mathbb{P}\big(\mathbb{X}_\ell^n = x_\ell^n \mid \mathbb{X}^k_{k-p+1} = w'\boldsymbol{y}\big)\right| \\
&= \sum_{x_\ell^{\ell+p-1}} \sum_{x^n_{\ell+p}} \phi\big(x^n_{\ell+p} \mid x_\ell^{\ell+p-1}\big) \left|\psi_{w\boldsymbol{y}}(x_\ell^{\ell+p-1}) - \psi_{w'\boldsymbol{y}}(x_\ell^{\ell+p-1})\right| \\
&= \sum_{x_\ell^{\ell+p-1}} \left|\psi_{w\boldsymbol{y}}(x_\ell^{\ell+p-1}) - \psi_{w'\boldsymbol{y}}(x_\ell^{\ell+p-1})\right| \\
&= \|\psi_{w\boldsymbol{y}} - \psi_{w'\boldsymbol{y}}\|_1 = 2\|\mathcal{L}^{(\ell \to \ell+p-1)}_{w\boldsymbol{y}} - \mathcal{L}^{(\ell \to \ell-p+1)}_{w'\boldsymbol{y}}\|_{\mathrm{TV}}. \qquad (5.14)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\eta_{k\ell} &= \sup_{w,w',\boldsymbol{y},\boldsymbol{z}} \|\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z}) - \mathcal{L}_k^{(\ell \to n)}(w'\boldsymbol{y}\boldsymbol{z})\|_{\mathrm{TV}} \\
&= \frac{1}{2} \sup_{w,w',y} \|\psi_{w\boldsymbol{y}} - \psi_{w'\boldsymbol{y}}\|_1 = \frac{1}{2} \sup_{w,w',y} \|(e_{w\boldsymbol{y}} - e_{w'\boldsymbol{y}})^\top \mathcal{K}^{\ell+p-1-k}\|_1.
\end{aligned}
$$

Let $m = \ell - k - 1$. Writing $m = p\lfloor m/p \rfloor + (m \bmod p)$ and using $\frac{1}{2}(e_{w\boldsymbol{y}} - e_{w'\boldsymbol{y}}) \in \mathcal{H}_1$ (see Definition equation (5.7)), we get

$$
\begin{aligned}
\eta_{k\ell} &\leq \sup_{v \in \mathcal{H}_1} \|v^\top \mathcal{K}^{p+p\lfloor m/p \rfloor + (m \bmod p)}\|_1 && (5.15a) \\
&\overset{(a)}{\leq} \sup_{v \in \mathcal{H}_1} \tau_1\left(\mathcal{K}^{(m \bmod p)}\right) \|v^\top \mathcal{K}^{p+p\lfloor m/p \rfloor}\|_1 && (5.15b) \\
&\overset{(b)}{\leq} \sup_{v \in \mathcal{H}_1} \|v^\top (\mathcal{K}^p)^{1+\lfloor m/p \rfloor}\|_1 \;\leq\; \tau_1(\mathcal{K}^p)^{1+\lfloor m/p \rfloor}, && (5.15c)
\end{aligned}
$$

where (a) follows from equation (5.9) applied for $u^\top = v^\top \mathcal{K}^{p+p\lfloor m/p \rfloor}$ which also belongs to $\mathcal{H}_1$, while (b) follows from the inequality in Lemma 18 and the last inequality follows from inequality equation (5.9) applied for $u = v$. This is the desired result which holds for $\ell + p \leq n$.

Case 2. When $\ell + p > n$, the reduction in equation (5.14) is unnecessary, i.e., there are fewer than p variables between ℓ and n. We cannot write the difference of the two underlying laws in terms of rows of $\mathcal{K}^r$ for some integer r. But, we can augment and consider $\mathcal{L}_k^{(\ell \to n+u)}(w\boldsymbol{y}\boldsymbol{z})$ where $u = \ell+p-n$ and then get $\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z})$ by marginalization. We have for any $w, w' \in \mathcal{X}$,

$$\|\mathcal{L}_k^{(\ell \to n)}(w\boldsymbol{y}\boldsymbol{z}) - \mathcal{L}_k^{(\ell \to n)}(w'\boldsymbol{y}\boldsymbol{z})\|_{\mathrm{TV}} \;\leq\; \|\mathcal{L}_k^{(\ell \to n+u)}(w\boldsymbol{y}\boldsymbol{z}) - \mathcal{L}_k^{(\ell \to n+u)}(w'\boldsymbol{y}\boldsymbol{z})\|_{\mathrm{TV}}$$

since marginalization does not increase the total variation distance. This follows from the triangle inequality: Assuming $p(\cdot,\cdot)$ and $q(\cdot,\cdot)$ to be some probability mass functions,

$$\sum_x \Big|p(x) - q(x)\Big| = \sum_x \Big|\sum_y p(x,y) - \sum_y q(x,y)\Big| \leq \sum_x \sum_y |p(x,y) - q(x,y)|.$$

Since $\ell+p = n+u$, the proof in this case reduces to that of Case 1. The proof of equation (5.5) is complete.

5.4.1 Sum of mixing coefficients: Proof of equation (5.6)

It is enough to prove the inequality for $\tau = \tau_1(\mathcal{K}^p)$. Then, the result follows since $\frac{1}{(\frac{1}{x}-1)^2}$ is increasing on $[\tau_1(\mathcal{K}^p), 1)$. For this τ, we have for any fixed k (recalling $\eta_{kk} = 1$),

$$\sum_{\ell \geq k} \eta_{k\ell} \;\leq\; 1 + \sum_{\ell > k} \tau^{1+\lfloor(\ell-k-1)/p\rfloor} \;\leq 1 + \sum_{m \geq 1} \sum_{\ell=(m-1)p+k+1}^{mp+k} \tau^m = 1 + \frac{p\tau}{1-\tau}.$$

It follows that

$$\|\mathbf{H}\|_\infty^2 := \Big(\max_k \sum_{\ell \geq k} \eta_{k\ell}\Big)^2 \leq \Big(1 + \frac{p\tau}{1-\tau}\Big)^2 \leq 2 + 2\frac{p^2\tau^2}{(1-\tau)^2}$$

which is the desired result. □

Chapter 6

Numerical Experiments

In this section, we evaluate the performance of the estimator in equation (3.1) using simulated data. We generate the data using the model in equation (2.3). In all the examples, we first randomly generate Θ^* and $\mathbf{D}$. To generate Θ^*, we select the support of Θ_i^* for each i uniformly at random based on the sparsity s_i. We then fill the support with i.i.d. draws of the normal distribution, and finally normalize such that $\|\Theta_i^*\|_{1,1}$ is a constant.

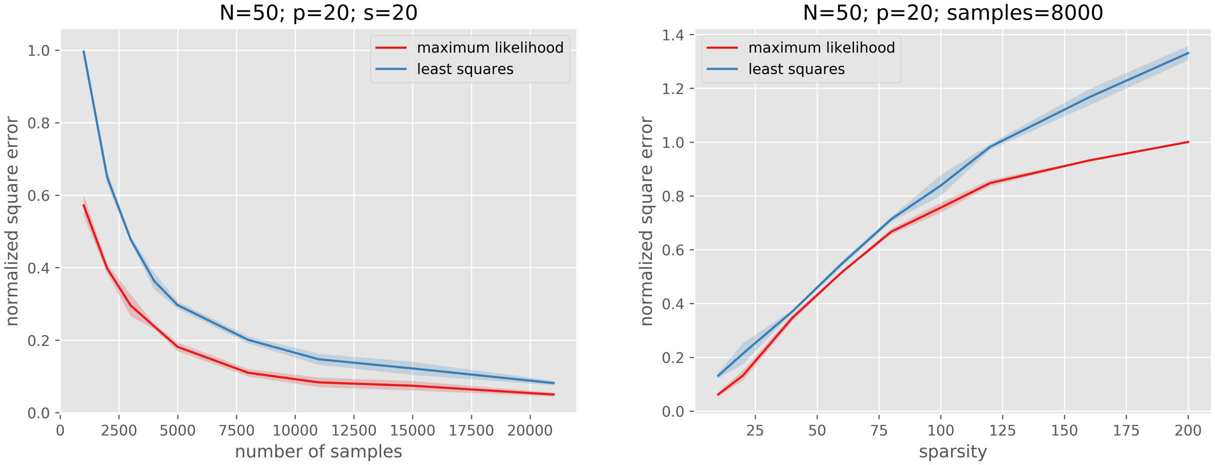
To report the performance of equation (3.1), we use the metric normalized squared error (NSE) defined as:

$$\mathrm{NSE}(\Theta^*, \widehat{\Theta}) = \frac{\|\Theta^* - \widehat{\Theta}\|_F^2}{\|\Theta^*\|_F^2}. \tag{6.1}$$

to normalize variations in the size of the parameter across independent instances of Θ^*. An implementation is provided at [41]. We consider the following 3 processes:

6.1 Poisson AR(p) process without dictionary

We evaluate the performance of the regularized maximum likelihood and the regularized least-squares estimators on a Poisson process with no dictionary, i.e., $\mathbf{D} = \boldsymbol{I}_p$. For the Poisson process, we use the inverse link function $f_i(z) = \log(1 + e^z)$. Then, these estimators have the



(a) NSE vs. sample size for a Poisson process without dictionary.

(b) NSE vs. sparsity for a Poisson process without dictionary.

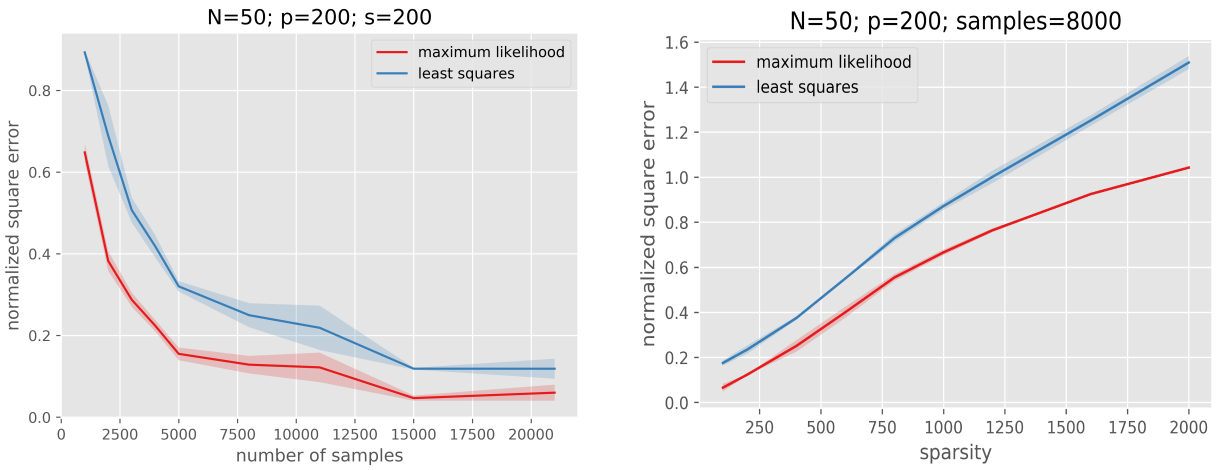
Figure 6.1: Poisson AR(p) process without a dictionary (i.e., $\mathbf{D} = \boldsymbol{I}_p$).

form of equation (3.1) with

$$\mathcal{L}_{it}^{\mathrm{ML}}\left(x_i^t\,;\,z_i^t\right) = z_i^t - x_i^t \log(z_i^t), \tag{6.2a}$$

$$\mathcal{L}_{it}^{\mathrm{LS}}\left(x_i^t\,;\,z_i^t\right) = (x_i^t - z_i^t)^2, \tag{6.2b}$$

where $z_i^t = f(\langle \Theta_i^*, \mathbf{X}^{t-1}\rangle)$, since $\mathbf{D} = \boldsymbol{I}_p$. Note that the M-estimation problem in equation (3.1) corresponding to equation (6.2a) is convex, whereas it is non-convex for equation (6.2b) (we report a local minimum). Here, we generate the ground truth parameters as mentioned before with $N = 50$ and $p = 20$ and we use $\lambda_n = 0.05/\sqrt{n}$. When comparing NSE v/s n, each Θ_i has sparsity 20. The results are shown in Figure 6.1. The error shades correspond to one standard deviation over 5 independent instances of $(\Theta^*, \widehat{\Theta})$. With the NSE metric, the regularized maximum likelihood estimator appears to perform better for the Poisson AR(p) process, for the random ensemble of problems generated in these examples.



(a) NSE vs. sample size for a Poisson process with dictionary.

(b) NSE vs. sparsity for a Poisson process with dictionary.

Figure 6.2: Poisson AR(p) process with dictionary of size $L = 20$.

6.2 Poisson AR(p) process with dictionary

We choose $\mathbf{D}$ to be entrywise i.i.d. Gaussian with standard deviation σ/p for a constant σ, so that the ℓ_1-norm of all columns of $\mathbf{D}$ are close to a constant for large p (the constant being the mean of a folded normal distribution). The process is generated as in the previous example using equation (2.3). We take $N = 50, p = 200$, and $L = 20$ such that the process has very long range dependencies. We again consider the two regularized M-estimators: the regularized maximum likelihood and the regularized least-squares with the inverse link function $f(z) = \log(1 + e^z)$. These estimators are identical to the ones in equation (6.2a) and equation (6.2b), except that $z_i^t = f(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)$ with $\mathbf{D} \neq \boldsymbol{I}_p$.

The results are shown in Figure 6.2. They are very similar to Figure 6.1. In accordance with our theoretical results, these figures suggest that for an AR processes with very long range dependencies, estimating the parameter is easier in the presence of a dictionary.

6.3 Bernoulli AR(p) process without dictionary

Finally, we look at a Bernoulli autoregressive process. We use the sigmoid function, $f(z) = 1/(1+e^{-z})$, as the inverse link function. We compare the performance of regularized maximum likelihood estimator to regularized least-squares estimator. Both of these estimators have the form of equation (3.1) with

$$\mathcal{L}_{it}^{\text{ML}}\left(x_i^t\,;\,z_i^t\right) = -z_i^t \log(x_i^t) - (1-x_i^t)\log(1-z_i^t) \tag{6.3a}$$

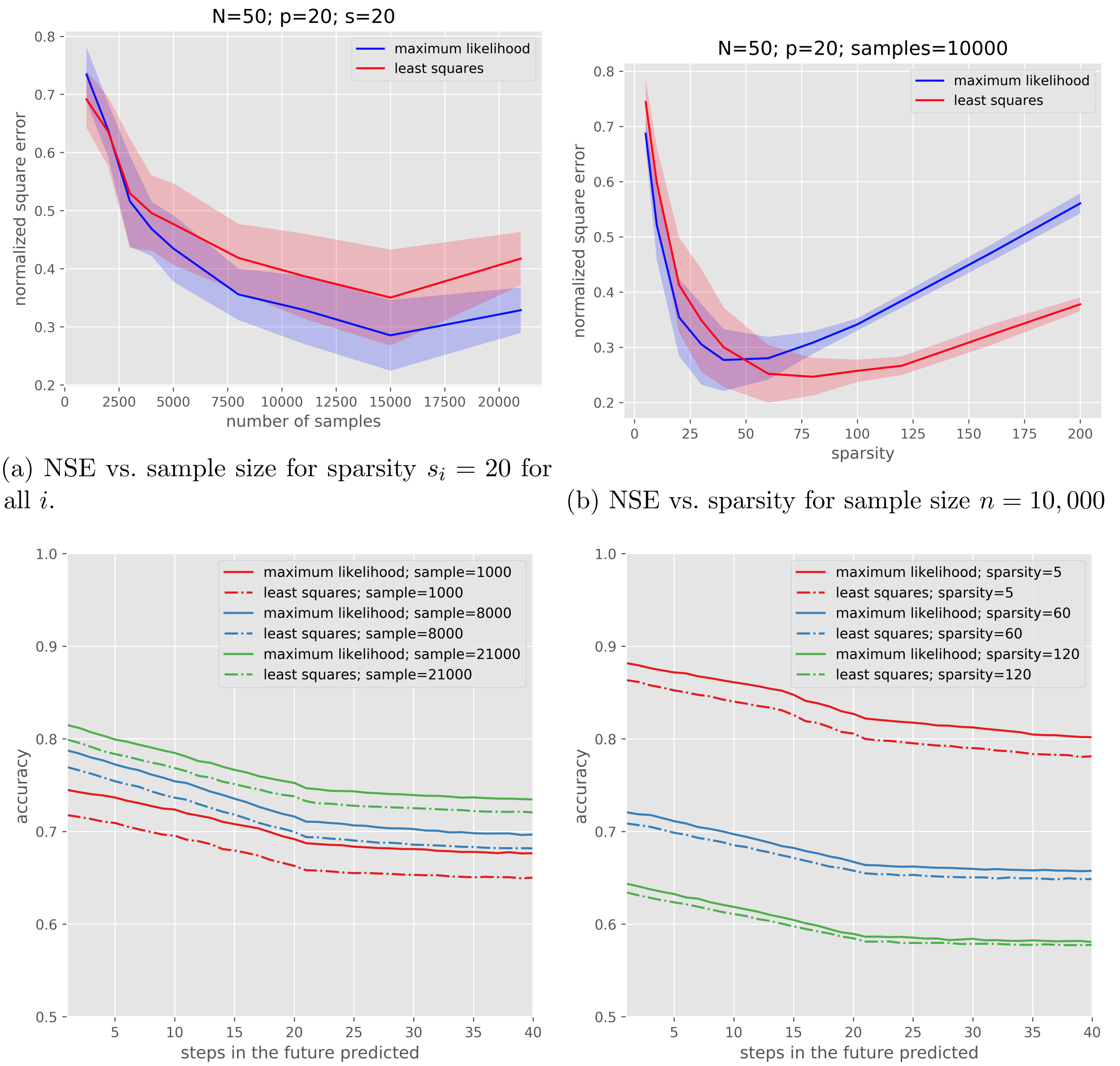
$$\mathcal{L}_{it}^{\text{LS}}\left(x_i^t\,;\,z_i^t\right) = (x_i^t - z_i^t)^2, \tag{6.3b}$$

where $z_i^t = f(\langle\Theta_i, \mathbf{X}^{t-1}\rangle)$ is the mean parameter of the dimension i of the Bernoulli process at time t. Note that due to inverse link function, despite convexity of square loss with respect to z_i^t, the optimization problem corresponding to least square estimator is non-convex and our results do not apply to it. Nevertheless, we observe that its performance is similar to maximum likelihood estimator.

Figure 6.3 shows different measures of performance of the regularized maximum likelihood estimator. We have set $N = 50, p = 20$ and $\lambda_n = 0.05/\sqrt{n}$ as recommended by Theorem 1, in these examples. Figure 6.3a shows how the normalized estimation error changes with respect to the number of training samples.

The sparsity is 20 for each Θ_i. Note that we are using the same regularization parameter for both estimators and not the optimal λ_n, i.e.without any cross-validation. The error shades correspond to one standard deviation. Figure 6.3b shows the normalized square error for different sparsity levels. For small values of sparsity, the denominator Θ^* has a small norm which causes high normalized error, however for higher values of sparsity, we see the linear dependence on sparsity as predicted by Theorem 1.

The next two figures correspond to generalization error as opposed to estimation error in the first two figures. Here, we use the estimated parameters $\widehat{\Theta}$ to predict the process in the future and calculate the accuracy of prediction. We use 5 MCMC runs of the process to



(a) NSE vs. sample size for sparsity $s_i = 20$ for all i.

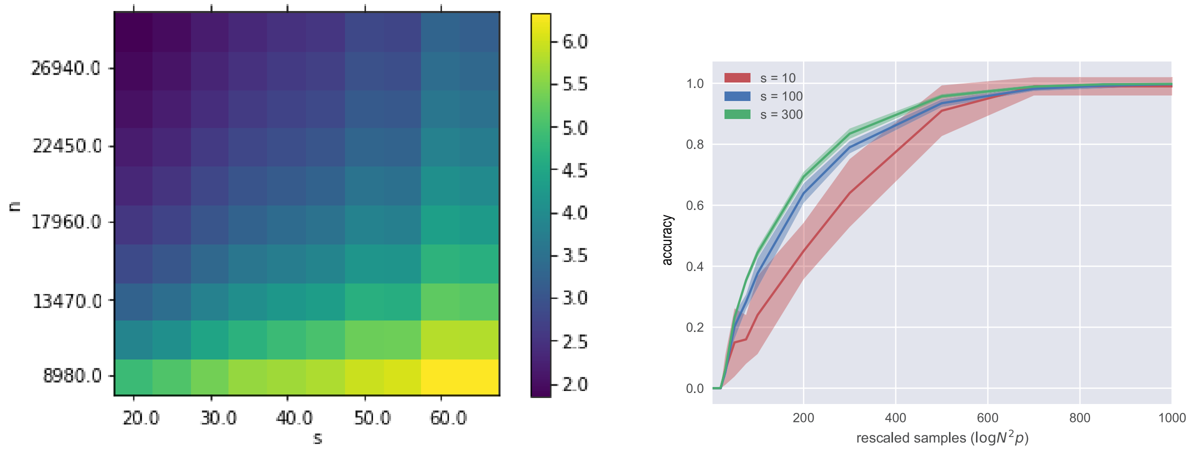
(b) NSE vs. sparsity for sample size $n = 10,000$

(c) Accuracy vs. steps predicted in the future for different n.

(d) Accuracy vs. steps predicted in the future for different s.

Figure 6.3: Bernoulli AR(p) process without dictionary.

estimate the accuracy. The plot shows average accuracy over all N variables of the process. Figure 6.3c shows the accuracy vs. steps in the future for different training sample sizes and Figure 6.3d shows it for different levels of sparsity. There is a prominent change in in the accuracy plots at 21 steps. This corresponds to $p = 20$ where the future of the process is being estimated purely based on simulated samples using the estimated parameter. Prior



(a) Average Frobenius norm of the error over 20 runs with $N = 20$, $p = 20$. Each pixel corresponds to a pair (s, n) for Θ^*.

(b) Fraction of support recovered by taking the largest s entries of $\hat{\Theta}$ as the estimator of support. Here $N = 100$, $p = 1$.

Figure 6.4: Simulation results for Bernoulli AR(p) process.

to this point, parts of the samples being used to make the predictions are True values and not estimated ones. As expected, the accuracies improve as the number of training samples increase with sparsity fixed, and they decrease as sparsity level increases with number of training samples fixed. Figure 6.4a shows the estimation error for different sample sizes and sparsity levels.

Finally, we also use the regularized maximum likelihood estimator to perform support recovery, i.e. assuming that the true parameter tensor is exactly s-sparse, how does the support estimated from $\widehat{\Theta}$ compare to the support of Θ^*? To do so, we need to estimate the support from $\widehat{\Theta}$. If we know the sparsity s, we can estimate the support by taking the indices corresponding to the s largest entries of $\widehat{\Theta}$ 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

$$\widehat{\text{supp}}(\Theta) := \{(j, k, \ell) : |\widehat{\Theta}_{jk\ell}| \geq \gamma\}.$$

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 $\widehat{\Theta}$ is required, i.e., we need to control $\|\widehat{\Theta} - \Theta^*\|_{\infty,\infty,\infty}$ with high probability. Therefore, 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. Figure 6.4b shows the results for a process with $p = 1, N = 100$ and three different sparsities. For recovering the support, we assumed that the sparsity s is known, and took the indices corresponding to the s largest entries of $\widehat{\Theta}$ as the recovered support. The fraction of the correctly recovered indices is plotted against the sample size. Figure 6.4b 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.

Chapter 7

Conclusion

Fitting autoregressive AR(p) models with multiple lags is of broad interest in multivariate time series analysis. We consider a large class of multivariate discrete-valued AR(p) processes with nonlinear feedback. We study statistical properties of a general ℓ_1 regularized M-estimator for this model, and provide sufficient conditions on the model hyperparameters under which consistent estimation is possible. Under assumptions of approximate sparsity, our result shows that a sample complexity $\Omega(\mathsf{poly}(s), \log(Np))$ is achievable. Our experiments validate the theoretical results on simulated data. Commonly occurring special cases of discrete-valued processes such as Bernoulli AR(p) and Truncated-Poisson AR(p) are explored in detail. The proof technique develops concentration inequalities and identifies mixing properties of higher order Markov chains which may be of independent interest. These techniques were previously unknown to the best of our knowledge.

Several open questions remain to be uncovered for the general AR(p) model. For example the current model explores the case of bounded, discrete valued data. Getting around this assumption requires finding concentration inequalities for random averages of the form in Lemma 10 for real-valued random processes. Also, it remains unknown whether the dependence on the sparsity hyperparameter s is optimal, since there is a small gap between our upper bound and the naive lower bound. Finally, it would be interesting to study parameter estimation, and potentially even controls, for the case of partial observability, i.e., when we observe $g(\boldsymbol{x}^t)$ and not $\boldsymbol{x}^t$ fully, akin to partially-observed Markov decision processes (POMDPs).

Appendix

We provide a proof for lemmas 2 and 3 restated below.

Lemma 2. *Consider a Binomial AR process given by equation* (2.3) *with* $\mathcal{X}_i = \{0, 1, \dots, K_i\}$*, where* $K_i \le B$*, and* $\mathbb{Q}_i(\,\cdot \,|\, z) = \mathrm{Bin}(K_i, z)$*. Assume that* f_i *is* L_i*-Lipschitz, and for some* $\varepsilon \in (0, \frac{1}{2})$*,* $f_i : \mathbb{R} \to [\varepsilon, 1 - \varepsilon]$ *for all* i*. Then, equation* (3.14) *holds with* $C_f = 6/\varepsilon$*.*

Lemma 3. *Consider a Truncated Poisson AR process given by equation* (2.3) *with* $\mathcal{X}_i = \{0, 1, \dots, K_i\}$ *and* $\mathbb{Q}_i(\,\cdot \,|\, z) = \mathbb{P}(\min(K_i, Z) \in \,\cdot\,)$ *where* $Z \sim \mathrm{Poi}(z)$ *and* $K_i \le B$*. Assume that* f_i *is* L_i*-Lipschitz, and for some* $\varepsilon > 0$*,* $f_i : \mathbb{R} \to [\varepsilon, \infty)$ *for all* i*. Then, equation* (3.14) *holds with* $C_f = 4/\varepsilon$*.*

Proof of Lemmas 2 and 3

We start by defining some notation. Recall that for $z \in \mathcal{X}^{\times p}$,

$$\mathbb{P}_{\boldsymbol{z}} := \mathbb{P}(\mathbb{X}_t^{t+p-1} = \cdot \mid \mathbb{X}_{t-p}^{t-1}) = \mathbb{P}(\mathbb{X}_1^p = \cdot \mid \mathbb{X}_{1-p}^0),$$

using the invariance of the conditional distribution to time shifts. We also write $p_{\boldsymbol{z}}(\cdot)$ for the probability mass function of $\mathbb{P}_{\boldsymbol{z}}$, i.e.,

$$p_{\boldsymbol{z}}(\boldsymbol{a}) := \mathbb{P}(\mathbb{X}_t^{t+p-1} = \boldsymbol{a} \mid \mathbb{X}_{t-p}^{t-1} = z) = \mathbb{P}(\mathbb{X}_1^p = \boldsymbol{a} \mid \mathbb{X}_{1-p}^0 = \boldsymbol{z}), \qquad \forall \boldsymbol{a} \in \mathcal{X}^{\times p}.$$

We also let $q(\xi \mid \boldsymbol{a}) := \mathbb{P}\big(\boldsymbol{x}^t = \xi \mid \mathbb{X}_{t-p}^{t-1} = \boldsymbol{a}\big)$ for $\xi \in \mathcal{X}$, $\boldsymbol{a} \in \mathcal{X}^{\times p}$, and define

$$d_K(\boldsymbol{a};\boldsymbol{a}') := D_{\mathrm{KL}}\Big(q(\cdot \mid \boldsymbol{a}) \parallel q(\cdot \mid \boldsymbol{a}')\Big), \qquad \boldsymbol{a},\boldsymbol{a}' \in \mathcal{X}^{\times p},$$

where D_{KL} denotes the KL divergence. The following lemma gives a decomposition for the KL divergence between two samples of a p-Markov process. Lemmas 19, 20 and 21 are proved later in this Appendix.

Lemma 19. *Assume that the process is p-Markov in the sense of equation* (5.4)*. Then,*

$$D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{z}} \parallel \mathbb{P}_{\boldsymbol{y}}) = \sum_{t=1}^{p} \mathbb{E}_{\boldsymbol{z}}\left[d_K\Big((\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0)\ ;\ (\mathbb{X}_1^{t-1}, \boldsymbol{y}_{t-p}^0)\Big)\right].$$

Here, $\mathbb{E}_{\boldsymbol{z}}$ denotes the expectation assuming that $\mathbb{X}_t^{t-1}$ is distributed as $\mathbb{P}_{\boldsymbol{z}}$. The notation $(\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0) \in \mathcal{X}^{\times p}$ denotes an $N \times p$ matrix with columns in $\mathcal{X}$, partitioned across columns into $N \times (t-1)$ matrix $\mathbb{X}_1^{t-1}$ and $N \times (p-t+1)$ matrix $\boldsymbol{z}_{t-p}^0$.

We also note the following bounds on the KL divergences between Bernoulli random variables and Poisson random variables to be used in proving Lemmas 2 and 3 respectively.

Lemma 20. *Let* $U \sim \mathrm{Ber}(p)$, *and* $V \sim \mathrm{Ber}(q)$ *for* $p, q \in [\varepsilon, 1-\varepsilon]$ *for some* $\varepsilon \in (0, \frac{1}{2})$*. Then,*

$$D_{\mathrm{KL}}(U\|V) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \ \leq\ \frac{3}{4\varepsilon(1-\varepsilon)}(p-q)^2.$$

Lemma 21. *Let* $U = \min\{M, \mathrm{Poisson}(p)\}$, *and* $V = \min\{M, \mathrm{Poisson}(q)\}$ *for* $p, q > \varepsilon > 0$ *for some* ε*. Then,*

$$D_{\mathrm{KL}}(U\|V) \leq p\log\frac{p}{q} + (q-p)\ \leq\ \frac{1}{q}(p-q)^2 \leq \frac{1}{\varepsilon}(p-q)^2$$

Proof of Lemma 2. Continuing with the proof of Lemma 2, recall that $\mathcal{S} = \{0,1\}^N$, and

$$\boldsymbol{x}^t \mid \mathbb{X}_{t-p}^{t-1} \ \sim\ \prod_{i=1}^{N} \mathrm{Ber}\left(f_i\left(\langle \Theta_i, \mathbb{X}_{t-p}^{t-1}\mathbf{D}\rangle\right)\right).$$

Let $\alpha_i^t = \big\langle \Theta_i^*, (\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0)\mathbf{D} \big\rangle$ and $\beta_i^t = \big\langle \Theta_i, (\mathbb{X}_1^{t-1}, \boldsymbol{y}_{t-p}^0)\mathbf{D} \big\rangle$. Then using the decomposability of the KL divergence for product measures,

$$\begin{aligned} d_K\big((\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0) \,\|\, (\mathbb{X}_1^{t-1}, \boldsymbol{y}_{t-p}^0)\big) &= \sum_{i=1}^N D_{\mathrm{KL}}\left(\mathrm{Ber}\big(f_i(\alpha_i^t)\big) \,\big\|\, \mathrm{Ber}\big(f_i(\beta_i^t)\big)\right), \\ &\le \frac{3}{4\varepsilon(1-\varepsilon)} \sum_{i=1}^N \big[f_i(\alpha_i^t) - f_i(\beta_i^t)\big]^2. \end{aligned}$$

By the Lipschitz assumption, $[f_i(\alpha_i^t) - f_i(\beta_i^t)]^2 \le L_i^2\big(\alpha_i^t - \beta_i^t\big)^2$. Using $\varepsilon < 1/2$, it follows that

$$D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{z}} \,\|\, \mathbb{P}_{\boldsymbol{y}}) \le \frac{3}{2\varepsilon} \sum_{i=1}^N L_i^2 \sum_{t=1}^p \mathbb{E}_{\boldsymbol{z}}\big(\alpha_i^t - \beta_i^t\big)^2.$$

Let $d_{m\ell} = (\boldsymbol{d}_\ell)_m$ be the (m,ℓ)th entry of $\mathbf{D}$. Let $z_j^{t-m}, m = t, \dots, p$ denote entries on the jth row of $\boldsymbol{z}_{t-p}^0$ and similarly for $\boldsymbol{y}_{t-p}^0$. We have

$$\alpha_i^t - \beta_i^t = \big\langle \Theta_i^*, (\mathbf{0}_{N\times(t-1)}, \boldsymbol{z}_{t-p}^0 - \boldsymbol{y}_{t-p}^0)\mathbf{D} \big\rangle = \sum_{j\ell} \Theta_{ij\ell}^* \sum_{m=t}^p (z_j^{t-m} - y_j^{t-m}) d_{m\ell},$$

where $\mathbf{0}_{N\times(t-1)}$ is the $N \times (t-1)$ zero matrix. Assuming that $\mathcal{X}_i \subset [-B_i, B_i]$, we have

$$\begin{aligned} |\alpha_i^t - \beta_i^t| &\le \sum_{j\ell} |\Theta_{ij\ell}| \sum_{m=t}^p \big(|z_j^{t-m}| + |y_j^{t-m}|\big)|d_{m\ell}| \\ &\le 2B \sum_{j\ell} |\Theta_{ij\ell}| \sum_{m=t}^p |d_{m\ell}|. \end{aligned}$$

Putting the pieces together finishes the proof. ▮

Proof of Lemma 3

The proof of Lemma 3, proceeds almost identically to that of 2. In this case however $\mathcal{S} = \mathbb{N}^N$, and

$$\boldsymbol{x}^t \mid \mathbb{X}_{t-p}^{t-1} \sim \prod_{i=1}^{N} \text{Poisson}\left(f_i\left(\left\langle \Theta_i^*, \mathbb{X}_{t-p}^{t-1}\mathbf{D}\right\rangle\right)\right).$$

Let $\alpha_i^t = \big\langle \Theta_i^*, (\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0)\mathbf{D}\big\rangle$ and $\beta_i^t = \big\langle \Theta_i^*, (\mathbb{X}_1^{t-1}, \boldsymbol{y}_{t-p}^0)\mathbf{D}\big\rangle$. Then using the decomposability of the KL divergence for product measures,

$$\begin{aligned} d_K\big((\mathbb{X}_1^{t-1}, \boldsymbol{z}_{t-p}^0) \parallel (\mathbb{X}_1^{t-1}, \boldsymbol{y}_{t-p}^0)\big) &= \sum_{i=1}^{N} D_{\text{KL}}\left(\text{Poisson}\big(f_i(\alpha_i^t)\big) \,\big\|\, \text{Poisson}\big(f_i(\beta_i^t)\big)\right), \\ &\le \frac{1}{\varepsilon}\sum_{i=1}^{N}\big[f_i(\alpha_i^t) - f_i(\beta_i^t)\big]^2 \;\le\; \frac{1}{\varepsilon}\sum_{i=1}^{N} L_i^2\big(\alpha_i^t - \beta_i^t\big)^2, \end{aligned}$$

where the first inequality is using Lemma 21 and the second by the Lipschitz assumption on f_i. The rest follows identically as in the proof of Lemma 2. □

Proof of Lemma 19

Recall the notation $X_1^p = (x_p, \ldots, x_1)$. Similarly, let $a = (a_p, \ldots, a_1) \in \mathcal{X}^{\times p}$ so that $X_1^p = a$ is the same as $X_u = a_u$ for all $u \in [p]$. We also write $a_1^{t-1} = (a_{t-1}, \ldots, a_1)$ and so on for elements of $\mathcal{X}^{\times p}$. For any $a, z \in \mathcal{X}^{\times p}$, we have

$$\begin{aligned} p_z(a) &= \mathbb{P}(X_1^p = a \mid X_{1-p}^0 = z) \\ &= \prod_{t=1}^{p} \mathbb{P}(x_t = a_t \mid X_1^{t-1} = a_1^{t-1},\, X_{t-p}^0 = z_{t-p}^0) \\ &= \prod_{t=1}^{p} \mathbb{P}\big(x_t = a_t \;\big|\; X_{t-p}^{t-1} = (a_1^{t-1}, z_{t-p}^0)\big) = \prod_{t=1}^{p} q(a_t \mid (a_1^{t-1}, z_{t-p}^0)) \end{aligned}$$

where the second line is by the Markov property. Replacing a with a random variable $X_1^p \in \mathcal{X}^{\times p}$,

$$p_z(X_1^p) = \prod_{t=1}^{p} q\big(x_t \mid (X_1^{t-1}, z_{t-p}^0)\big).$$

Letting $\mathbb{E}_z$ denote the expectation assuming $X_1^p \sim \mathbb{P}_z$, we have

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_z \,\|\, \mathbb{P}_y) &= \mathbb{E}_z \, \log \frac{p_z(X_1^p)}{p_y(X_1^p)} \\
&= \sum_{t=1}^{p} \mathbb{E}_z \, \log \frac{q\big(x_t \mid (X_1^{t-1}, z_{t-p}^0)\big)}{q\big(x_t \mid (X_1^{t-1}, y_{t-p}^0)\big)} \\
&= \sum_{t=1}^{p} \mathbb{E}_z \mathbb{E}_z \Big[\log \frac{q\big(x_t \mid (X_1^{t-1}, z_{t-p}^0)\big)}{q\big(x_t \mid (X_1^{t-1}, y_{t-p}^0)\big)} \;\Big|\; X_1^{t-1} \Big] \\
&= \sum_{t=1}^{p} \mathbb{E}_z \, d_K\big((X_1^{t-1}, z_{t-p}^0) \,\|\, (X_1^{t-1}, y_{t-p}^0)\big)
\end{aligned}$$

where the last line follows by noting that under $X_1^p \sim \mathbb{P}_z$, further conditioning on X_1^{t-1} is equivalent to conditioning on X_1^{t-1} and $X_{t-p}^0 = z_{t-p}^0$, i.e., x_t will have distribution $q(\,\cdot \mid (X_1^{t-1}, z_{t-p}^0))$ under this conditioning. □

Proof of Lemma 20

It is enough to prove for the case $q \geq p$ (the other case follows by applying the proven case to $1-p$ and $1-q$). The second claim follows from the decomposition of the KL divergence for product distributions. Let $\delta := \varepsilon(1-\varepsilon)$. Fix p and consider the function

$$f(q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - \frac{1}{4\delta}(p-q)^2,$$

over $q \in [p, 1-\varepsilon]$. We have

$$f'(q) = (q-p)\Big(\frac{1}{q(1-q)} - \frac{1}{2\delta}\Big).$$

We have $f(q) = f(p) + f'(\tilde{q})(q - p)$ for some $\tilde{q} \in [p, q]$. Note that $f(p) = 0$ and

$$f'(\tilde{q}) \leq (\tilde{q} - p)\Big(\frac{1}{\delta} - \frac{1}{2\delta}\Big) \leq \frac{1}{2\delta}(q - p)$$

using the fact that $(\tilde{q}(1 - \tilde{q}))^{-1} \in [4, \delta^{-1}]$. Thus, we have $f(q) \leq (q - p)^2/(2\delta)$. □

Proof of Lemma 21

The KL divergence between two Poisson distributions with parameters p and q is given by

$$p \log \frac{p}{q} + (q - p) - \frac{(q - p)^2}{q} = p(\log \frac{p}{q} + 1 - \frac{p}{q})$$

We show that the truncation only reduces the KL divergence using Jensen's inequality for the convex function $g(u, v) = u \log(u, v)$. Let $p_i := e^{-p} \frac{p^i}{i!}$ and $q_i := e^{-q} \frac{q^i}{i!}$. Next, observe that the KL divergence for the truncated version is

$$\sum_{i<M} p_i \log \frac{p_i}{q_i} + \sum_{i \geq M} p_i \log \frac{\sum_{i \geq M} p_i}{\sum_{i \geq M} q_i}$$

Applying the Jensen's inequality to second term, we get that the quantity above is at most

$$\sum_{i<M} p_i \log \frac{p_i}{q_i} + \sum_{i \geq M} p_i \log \frac{p_i}{q_i}$$

which is the KL divergence between Poisson(p) and Poisson(q). Finally, observe that for $p, q > 0$

$$p \log \frac{p}{q} + (q - p) - \frac{(q - p)^2}{q} = p(\log \frac{p}{q} + 1 - \frac{p}{q}) \leq 0$$

where we use the inequality $\log x \leq x - 1$. □

Bibliography

[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] Dimitri P Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

[4] 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.

[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] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[7] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[8] Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. *arXiv preprint arXiv:1201.0559*, 2012.

[9] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[10] Christine De Mol, Domenico Giannone, and Lucrezia Reichlin. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328, 2008.

[11] David N DeJong and Charles H Whiteman. The temporal stability of dividends and stock prices: Evidence from the likelihood function. *The American Economic Review*, pages 600–617, 1991.

[12] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[13] Yonina C. Eldar and Gitta Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge Univ. Press, June 2012.

[14] Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's lemma for markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.

[15] Robert M Gray et al. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[16] Eric C Hall, Garvesh Raskutti, and Rebecca M Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4):2401–2422, 2018.

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

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

[19] 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.

[20] Aryeh Kontorovich. Obtaining measure concentration from markov contraction. *Markov Processes and Related Fields*, 18:613–638, 2012.

[21] 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.

[22] M Krein and V Smulian. On regularly convex sets in the space conjugate to a banach space. *Annals of Mathematics*, pages 556–583, 1940.

[23] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[24] 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.

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

[26] Katalin Marton et al. Bounding $\overline{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

[27] 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.

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

[29] 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.

[30] 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.

[31] Parthe Pandit, Mojtaba Sahraee-Ardakan, Arash Amini, Sundeep Rangan, and Alyson K. Fletcher. Sparse Multivariate Bernoulli Processes in High Dimensions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 457–466, April 2019.

[32] Parthe Pandit, Mojtaba Sahraee-Ardakan, Arash Amini, Sundeep Rangan, and Alyson K Fletcher. Sparse multivariate bernoulli processes in high dimensions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 457–466, 2019.

[33] Parthe Pandit, Mojtaba Sahraee-Ardakan, Arash A. Amini, Sundeep Rangan, and Alyson K. Fletcher. High-Dimensional Bernoulli Autoregressive Process with Long-Range Dependence. *arXiv:1903.09631 [cs, eess, math, stat]*, March 2019.

[34] Parthe Pandit, Mojtaba Sahraee-Ardakan, Arash A Amini, Sundeep Rangan, and Alyson K Fletcher. Generalized autoregressive linear models for discrete high-dimensional data. *IEEE Journal on Selected Areas in Information Theory*, 1(3):884–896, 2020.

[35] 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.

[36] 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.

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

[38] Garvesh Raskutti, Ming Yuan, Han Chen, et al. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

[39] Adolf Rhodius. On the maximum of ergodicity coefficients, the Dobrushin ergodicity coefficient, and products of stochastic matrices. *Linear algebra and its applications*, 253(1-3):141–154, 1997.

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

[41] Mojtaba Sahraee-Ardakan, Parthe Pandit, Arash Amini, Sundeep Rangan, and Alyson K Fletcher. *Multivariate Autoregressive Generalized Linear Model regression in PyTorch*, 2020. https://github.com/mojtabasah/AR_process.

[42] Paul-Marie Samson et al. Concentration of measure inequalities for markov chains and Φ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

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

[44] Allan Timmermann. Excess volatility and predictability of stock prices in autoregressive dividend models with learning. *The Review of Economic Studies*, 63(4):523–557, 1996.

[45] 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.

[46] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[47] 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.

[48] 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.

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