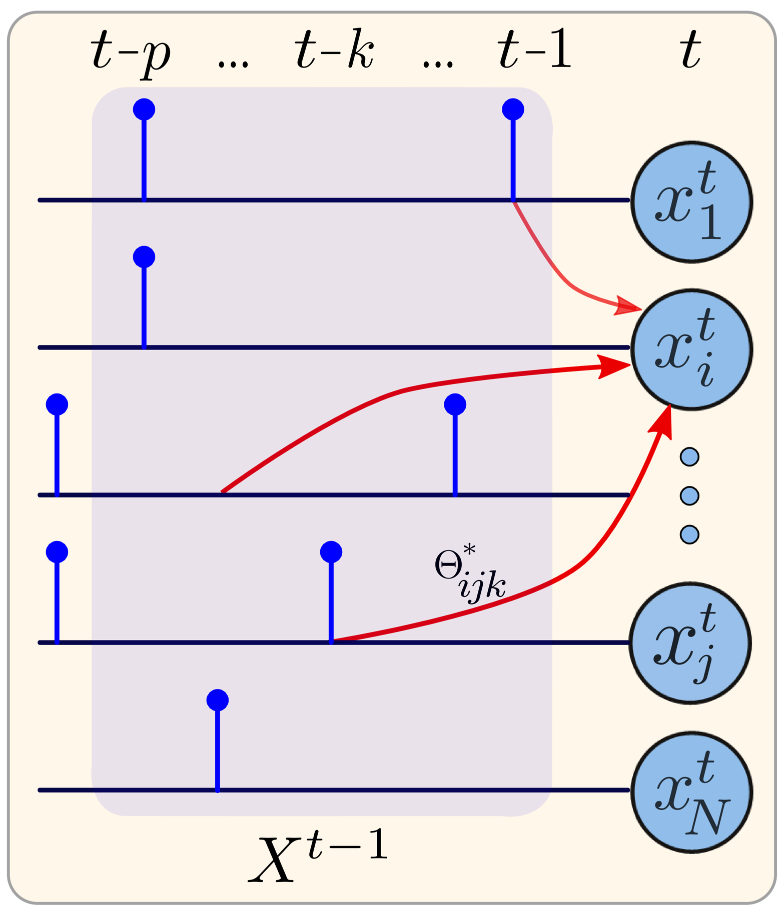
Sparse Multivariate Bernoulli Processes in High Dimensions

Parthe Pandit, Mojtaba Sahraee, Arash A. Amini, Sundeep Rangan, Alyson K. Fletcher

UCLA NYU

Bernoulli Autoregressive Model



$x_i^t \mid X^{t-1} \sim \text{Bernoulli}(z_i^t)$,

$x_i^t \mid X^{t-1} \perp\!\!\!\perp x_j^t \mid X^{t-1}$

$z_i^t = f\left(\langle \Theta^*_{i..}, X^{t-1}\rangle\right)$

$X^{t-1} = \left[x^{t-1} \mid \ldots \mid x^{t-p}\right]$

Parameters:

Θ^*_{ijk} : Describes how neuron i depends on neuron j from k lags ago

$\Theta^* \in \mathbb{R}^{N\times N\times p}$

Problem: Estimate approximately s-sparse Θ^* from $\{x_i^t\}_{t=1-p}^n$

Regularized Maximum Likelihood Estimator

$$\widehat{\Theta}_{\lambda_n} = \underset{\Theta\in\mathbb{R}^{N\times N\times p}}{\operatorname{argmin}} \mathcal{L}(\Theta; \{x^t\}_{t=1-p}^n) + \lambda_n \|\Theta\|_1$$

$$\mathcal{L}(\Theta; \{x^t\}_{t=1-p}^n) := \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{N} -x_i^t \log(f(\langle \Theta_{i..}, X^{t-1}\rangle)) - (1-x_i^t)\log(1-f(\langle\Theta_{i..}, X^{t-1}\rangle))$$

Motivation, Goal and Challenges

- **Motivation**: Multivariate Bernoulli Processes can model
  - Spike trains from an ensemble of neurons
  - Networks of dynamical systems with binary states
    - Trends in stock prices,
    - Activity in social networks
    - Crime, medical emergencies in a metropolitan area
    - Climate dynamics: atmospheric circulation patterns
    - Physiological systems, biological signaling networks
- **Goal**: Infer the structural interconnections between these dynamical systems from binary-valued observations
- **Challenges**: Non-i.i.d., non-Gaussian, nonlinear feedback, long-term dependencies (non-Markovian) data

Assumptions

1.	Nonlinearity	$f : \mathbb{R} \to [\epsilon, 1-\epsilon]$, L_f-Lipschitz, f, $1-f$ strongly log concave
2.	Stability	$\min_{\omega\in[-\pi,\pi)} \lambda_{\min}(\mathcal{X}(\omega)) = c_\ell^2 > 0$; $\mathcal{X}(\omega) := \sum_\ell \text{Cov}(x^t, x^{t+\ell})e^{-j\omega\ell} \in \mathbb{C}^{N\times N}$
3.	Size of parameter	$\frac{3L_f^2}{2\epsilon}\sum_{\ell=1}^{p}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\sum_{k=\ell}^{N}\lvert\Theta^*_{ijk}\rvert\right)^2 < 1$
4.	Approx. sparsity	$\sigma_s := \min_{\|\Theta\|_0\le s}\|\Theta^* - \Theta\|_1$

Contact me

References

[1] Negahban, Sahand N., et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." Statistical Science 27.4 (2012): 538-557.

[2] Kontorovich, Leonid Aryeh, and Kavita Ramanan. "Concentration inequalities for dependent random variables via the martingale method." The Annals of Probability 36.6 (2008): 2126-2158.

Main Result

If number of samples n, regularization parameter λ_n satisfy

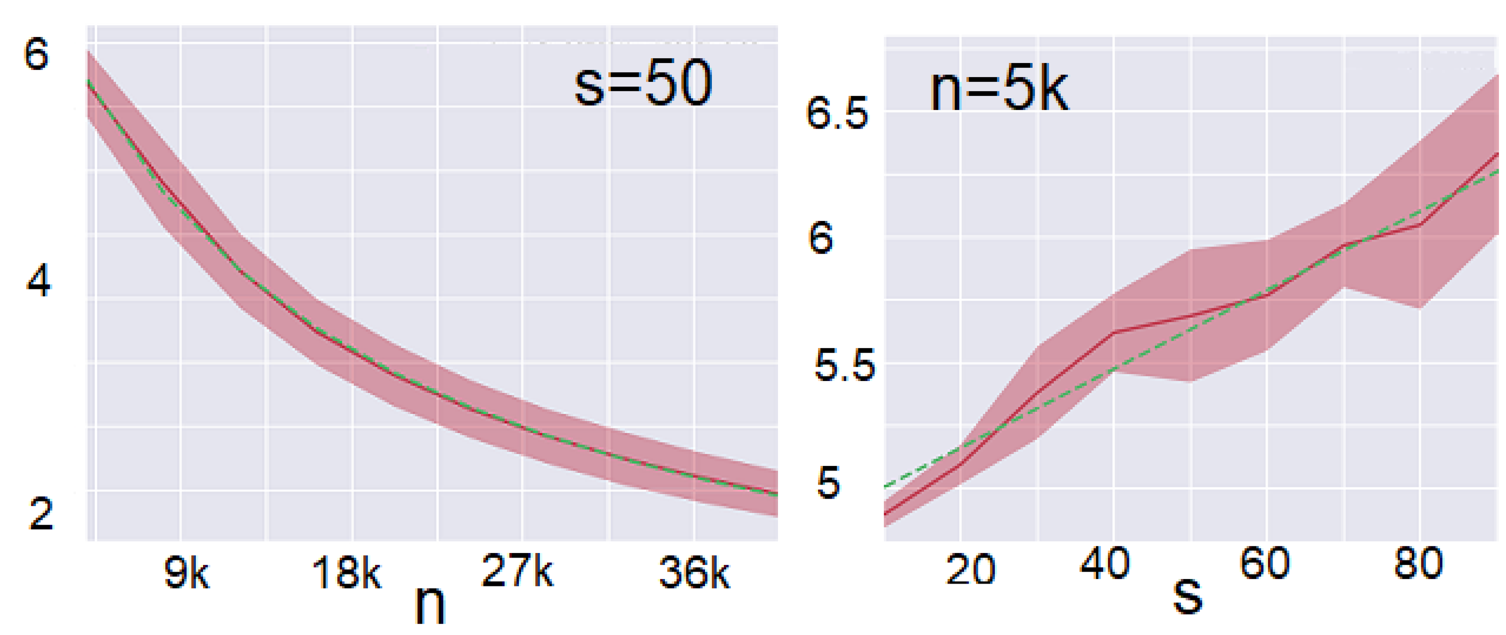
$$n = \Omega\left(G_f \cdot s^3 \log(N^2 p)\right), \quad \lambda_n = \Omega\left(\sqrt{\log(N^2p)/n}\right).$$

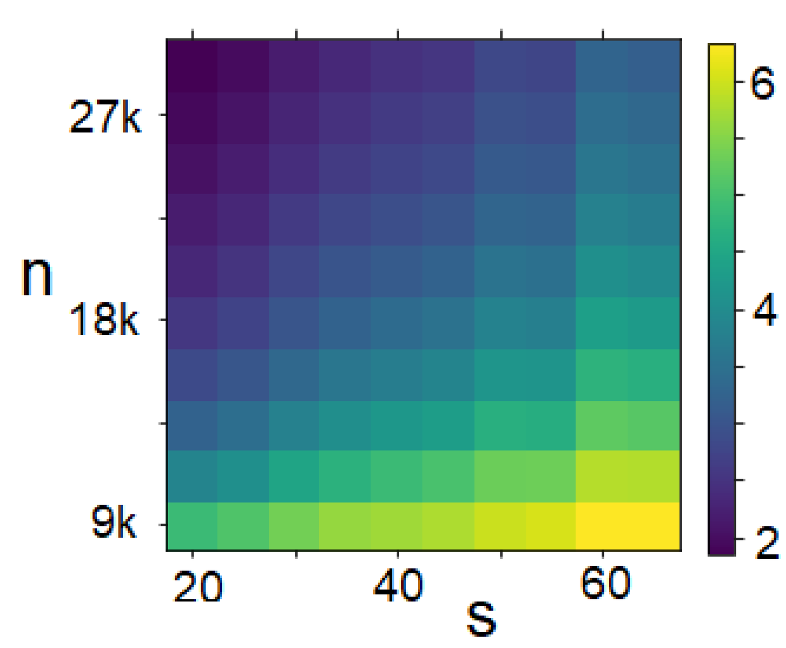
then with high probability,

$$\left\|\widehat{\Theta}_{\lambda_n} - \Theta^*\right\|_F^2 \lesssim \frac{s\log(N^2p)}{n} + (\sigma_s + \frac{\sigma_s^2}{s})\sqrt{\frac{\log(N^2p)}{n}}.$$

- **Consistency guaranteed**, rate $\mathcal{O}(n^{-\frac{1}{2}})$ for hard sparsity
- For consistent estimation, *sample complexity* n grows as
  - $\mathcal{O}(G_f \log p)$ with *lags* p; a sufficient condition for $G_f = \mathcal{O}(1)$ is $L_f = \mathcal{O}(p^{-1})$ and the tail $\Theta_{ij\ell}$ decays faster than $\mathcal{O}(\ell^{-3/2})$
  - $\mathcal{O}(s^3)$ with *sparsity* s, non-ideal but simulations suggest $s^{1+\delta}$
  - $\mathcal{O}(\log N)$ with *dimension* N, previously unknown for $p > 1$
- **Main challenge**: summation in definition of $\mathcal{L}(\Theta)$ is not i.i.d.

Simulations





- $N = 20, p = 20$
- Mean, Variance of $\|\cdot\|_F^2$ error for 20 expt.
- Notice $n = \Omega(s^{1+\delta})$ equal-error contours

Sketch of the Proof and Key insights

1. **Restricted Strong Convexity of $\mathcal{L}$ around Θ^***

$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle\nabla\mathcal{L}(\Theta^*), \Delta\rangle \ge \kappa\|\Delta\|_F^2 - \tau^2,\ \Delta \in \mathbb{C}(\Theta^*)$$

- LHS $\ge \mathcal{E}(\Delta; \{x^t\}) := \frac{c_f}{n}\sum_{t=1}^{n}\sum_{i=1}^{N}\langle\Delta_{i..}, X^{t-1}\rangle^2$
- $\mathcal{E}(\Delta; \{x^t\}) \ge$ RHS with high probability
  - Inequality holds for $\mathbb{E}\,\mathcal{E}(\Delta; \{x^t\})$ due to Assumptions 1,2
  - For a fixed Δ, $\mathcal{E}(\Delta; \{x^t\})$ concentrates near $\mathbb{E}\,\mathcal{E}(\Delta; \{x^t\})$

    $\mathbb{P}\{\lvert\mathcal{E} - \mathbb{E}\,\mathcal{E}\rvert > t\|\Delta\|_{2,1,1}^2\} \le e^{-nt^2/G_f}$

    since $\{x^t\} \mapsto \mathcal{E}(\Delta; \{x^t\})$ is $\|\Delta\|_{2,1,1}^2$-Lipschitz and the process $\{x^t\}$ is η-mixing due to Assumption 3
  - Uniform law over $\Delta \in \mathbb{C}$ using *Covering* arguments
  - For $\Delta \in \mathbb{C}$, we have $\frac{\|\Delta\|_{2,1,1}^2}{\|\Delta\|_F^2} = \mathcal{O}(s^2)$. This causes an additional s^2 factor in the sample complexity

2. **Choice of regularization parameter**: $\lambda_n \ge \|\nabla\mathcal{L}(\Theta^*)\|_\infty$

- $\nabla\mathcal{L}(\Theta^*)$ is a zero mean martingale difference sequence
- Azuma-Hoeffding's gives $\|\nabla\mathcal{L}(\Theta^*)\|_\infty = \mathcal{O}(\sqrt{\log(N^2p)/n})$