Los Alamos
NATIONAL LABORATORY
EST. 1943

Learning Discrete Graphical Models with Neural Networks

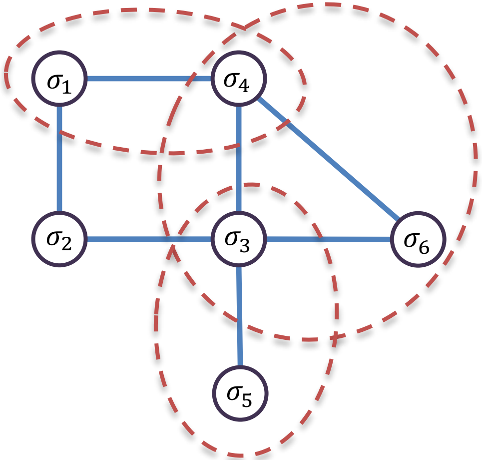
Andrey Lokhov

joint work with Abhijith Jayakumar, Sidhant Misra, Marc Vuffray

Graphical Models

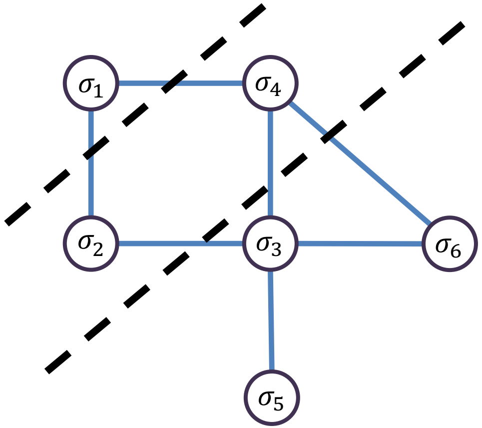
Probability distribution $\mu(\underline{\sigma})$ has conditional dependency structure according to a given graph

Factorization property



$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{c \in cliques} f_c(\sigma_c)\right)$$

Separation property



$\sigma_1 | (\sigma_2, \sigma_4)$ is independent of $(\sigma_3, \sigma_5, \sigma_6)$

Graphical Model Learning Informally

Unsupervised learning task

- Observe draws of random vectors $\underline{\sigma}$
- Learn structure and parameters of a positive distribution $\mu(\underline{\sigma}) > 0$

Dimensions of the problem

- Number of samples: n
- Number of variables: p
- Alphabet size: q $(\sigma_i \in \{1, \dots, q\})$

Prior work in computationally efficient learning

Mutual Information based **greedy** methods

Bresler (**2015**)

Hamilton, Koehler, Moitra (**2017**)

Convex optimization based methods

Vuffray, Misra, Lokhov (**2016**, **2018**)

Klivans, Meka (**2017**)

Wu, Sanghavi, Dimakis (**2019**)

Setting of Graphical Model Learning

The model has a **parametric form:**

$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{k\in K} \theta_k^* \, g_k(\underline{\sigma}_k)\right)$$

- Observe **random** draws of $\underline{\sigma}$
- **Recover parameters**

$$\left\|\underline{\hat{\theta}} - \underline{\theta}^*\right\| \leq \frac{\epsilon}{2}$$

Basis functions are centered:

$$\sum_{\sigma_i} g_k(\underline{\sigma}_k) = 0, \qquad i \in k$$

Prior ℓ_1-bound on parameters:

$$\left\|\underline{\theta}_i^*\right\|_1 = \sum_{k \ni i} |\theta_k^*| \leq \hat{\gamma}$$

Method for Solving the Inverse problem: GRISE

Arbitrary **parametric form**

$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{k\in K} \theta_k^* \, g_k(\underline{\sigma}_k)\right)$$

Generalized Regularized Interaction Screening (GRISE)

$$\underline{\hat{\theta}}_i = \arg\min_{\underline{\theta}_i} \frac{1}{n}\sum_{t=1}^{n} \exp\left(-\sum_{k\in K_i} \theta_k \, g_k(\underline{\sigma}_k^t)\right)$$

$$\text{s.t. } \left\|\underline{\theta}_i\right\|_1 \leq \hat{\gamma}$$

Local Reconstruction (one neighborhood at a time)

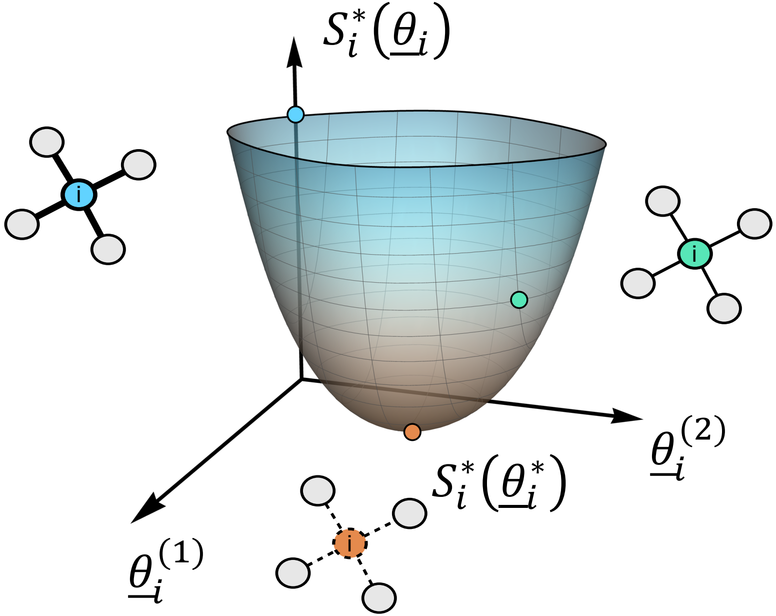
Convex Function (with low complexity minimization using entropic descent)

Intuition Behind GRISE: Infinite Sample Size Limit

$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{k\in K} \theta_k^* \, g_k(\underline{\sigma}_k)\right)$$

$$S_i(\underline{\theta}_i) \xrightarrow{n\to\infty} S_i^*(\underline{\theta}_i) = \mathbb{E}\left[\exp\left(-\sum_{k\in K_i} \theta_k \, g_k(\underline{\sigma}_k^t)\right)\right]$$

$$\boxed{\nabla_{\underline{\theta}_i} S_i^*(\underline{\theta}_i^*) = 0}$$



Theorem for Learning Gibbs Distributions with GRISE

(Informal) With high probability, **GRISE estimates**:

$$\left\| \underline{\hat{\theta}} - \underline{\theta}^* \right\| \leq \frac{\epsilon}{2}$$

with a **number of samples**:

$$n = \tilde{O}(q^{2L} \log(p) / \epsilon^4)$$

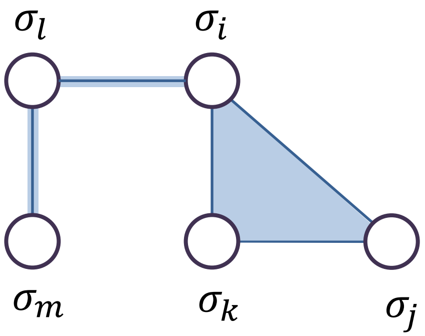
and **computational complexity**:

$$\tilde{O}(p^L)$$

Precise finite sample analysis with proofs: **arXiv:1902.00600**

Complete Basis Function Hierarchies: Monomial Basis Example

$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{i\in V}\theta_i^*\sigma_i + \sum_{(i,j)\in E_2}\theta_{ij}^*\sigma_i\,\sigma_i + \sum_{(i,j,k)\in E_3}\theta_{ijk}^*\sigma_i\,\sigma_j\sigma_k + \cdots\right)$$

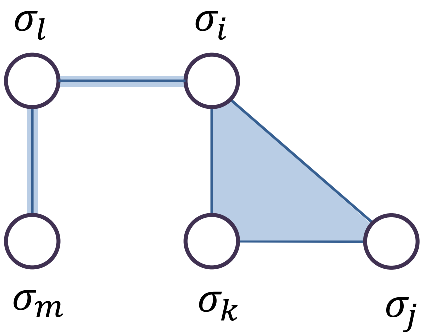


Binary alphabet $\underline{\sigma} \in \{-1,+1\}^p$

Monomial basis functions $g_k(\underline{\sigma}_k) \in \{\sigma_i, \sigma_i\sigma_j, \sigma_i\sigma_j\sigma_k, \ldots\}$

Complete Basis Function Hierarchies: Monomial Basis Example

$$\mu(\underline{\sigma}) \propto \exp\left(\sum_{i\in V}\theta_i^*\sigma_i + \sum_{(i,j)\in E_2}\theta_{ij}^*\sigma_i\,\sigma_i + \sum_{(i,j,k)\in E_3}\theta_{ijk}^*\sigma_i\,\sigma_j\sigma_k + \cdots\right)$$



Interaction Screening Loss:

$$\underline{\hat{\theta}_i} = \arg\min_{\underline{\theta_i}} \frac{1}{n}\sum_{t=1}^{n}\exp\left(-\sigma_i\left(\theta_i + \sum_{j}\theta_{ij}\,\sigma_i + \sum_{j,k}\theta_{ijk}\,\sigma_j\sigma_k + \cdots\right)\right)$$

For L-wise models, the computational complexity of GRISE is $\widetilde{O}(p^L)$.

Neural Net Parametrization of the Partial Energy Function

Interaction Screening Loss:

$$\underline{\hat{\theta}_i} = \arg\min_{\underline{\theta_i}} \frac{1}{n}\sum_{t=1}^{n} \exp\left(-\sigma_i\left(\theta_i + \sum_j \theta_{ij}\,\sigma_i + \sum_{j,k} \theta_{ijk}\,\sigma_j\sigma_k + \cdots\right)\right)$$

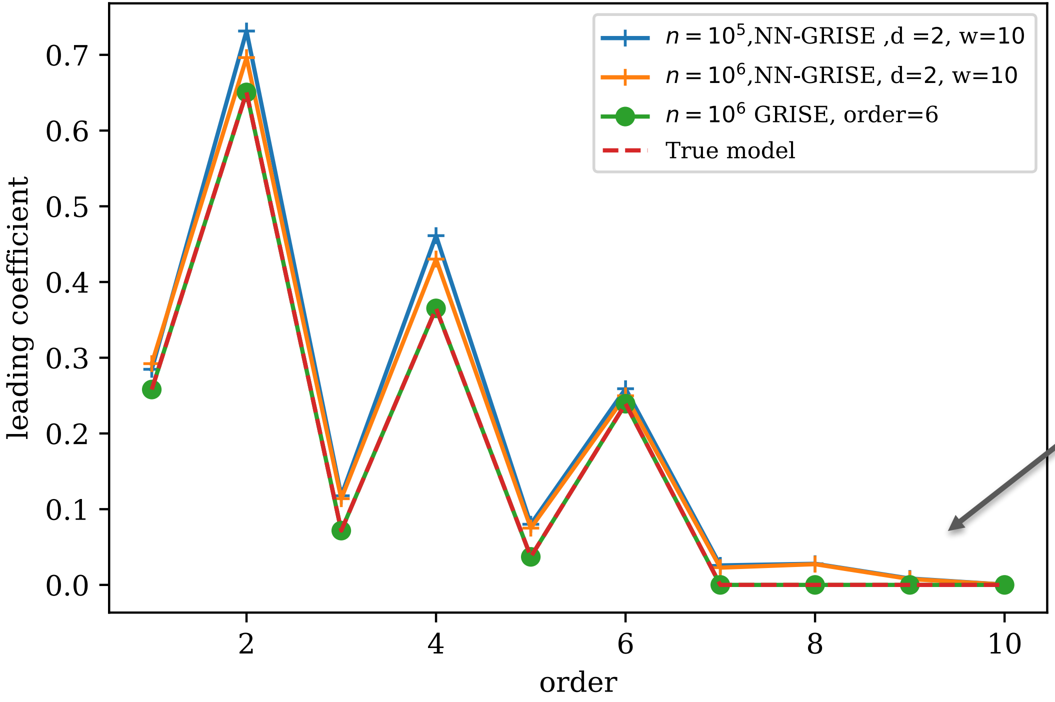
Neural Net Interaction Screening Loss:

$$\underline{\widehat{w}_i} = \arg\min_{\underline{w_i}} \frac{1}{n}\sum_{t=1}^{n} \exp\left(-\sigma_i\,\mathrm{NN}(\underline{\sigma}\backslash\sigma_i;\underline{w_i})\right)$$

If Neural Net is expressive enough, the global minima of **NN-GRISE** loss are interaction screening minima corresponding to recovered local energy

Illustration on a small ($p = 10$) tractable model of order $\mathrm{L} = 6$

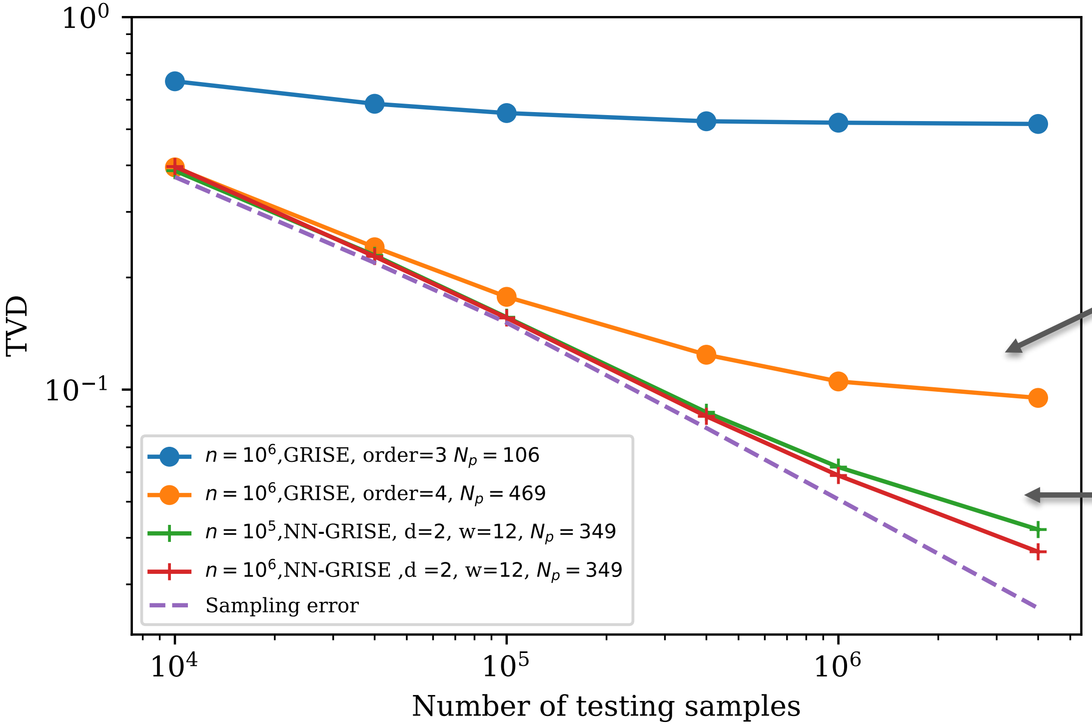
NN-GRISE hierarchy contains higher-order polynomials in its hypothesis space



NN-GRISE explores a different basis functions hierarchy, and gets close to the true model with less parameters

Comparison of conditional distributions for a larger problem

For p=15, L=6 problem, monomial basis contains 3472 terms, and GRISE becomes intractable

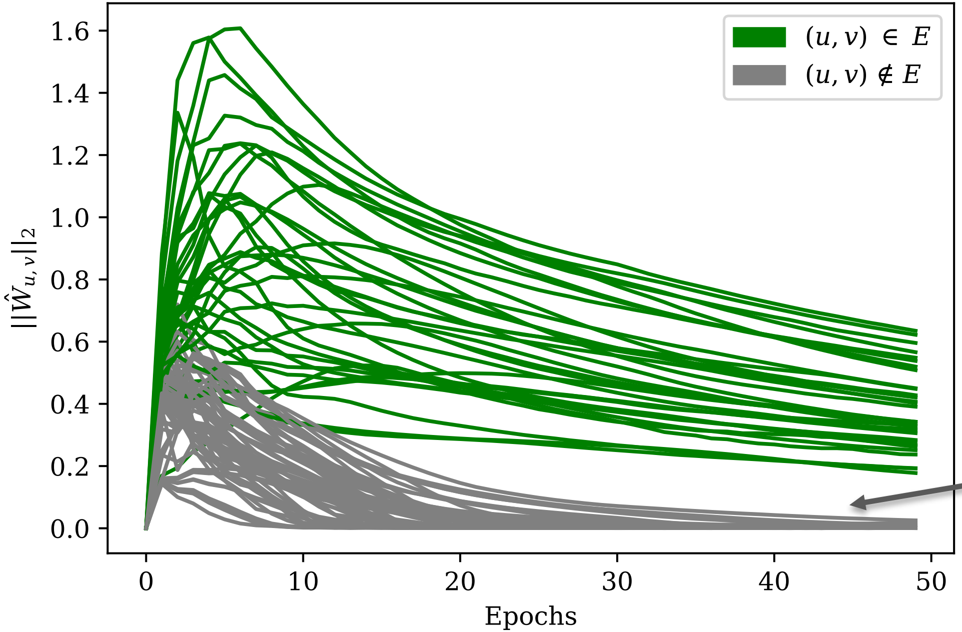


Only order L=4 is practically feasible with GRISE

NN basis has less parameters (349) and uses less training samples

Structure Learning with NN-GRISE

$$\underline{\widehat{w}_u} = \arg\min_{\underline{w_u}} \left(\frac{1}{n} \sum_{t=1}^{n} \exp\left(-\sigma_u \, \mathrm{NN}(\underline{\sigma} \backslash \sigma_u; \underline{w_u})\right) + \lambda \left\| \underline{w_u}^{(1)} \right\|_1 \right)$$



Regularization through penalty on first layer weights

Variables v outside of the neighborhood of u do not influence the output at the interaction screening minima

Summary

- **GRISE** is a convex estimator for learning arbitrary discrete graphical models with rigorous guarantees, improving upon sampling complexities of previous methods

arXiv **Efficient Learning of Discrete Graphical Models**
M. Vuffray, S. Misra, A. Y. Lokhov (2020)

- **NN-GRISE** is a computationally efficient non-convex estimator that uses the non-linear representation power of Neural Nets to exploit sparse basis hierarchies

- **NN-GRISE** can still learn the MRF structure, full energy function representation, and conditional distributions that can be used for re-sampling from the learned model

arXiv **Learning of Discrete Graphical Models with Neural Networks**
Abhijith J., A. Y. Lokhov, S. Misra, M. Vuffray (2020)

Questions?