High-Dimensional Graphical Model Selection

Anima Anandkumar

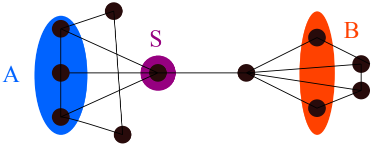
U.C. Irvine

Joint work with Vincent Tan (U. Wisc.) and Alan Willsky (MIT).

Graphical Models: Definition

Conditional Independence

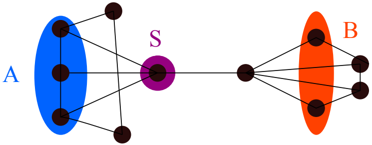
$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$



Graphical Models: Definition

Conditional Independence

$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$



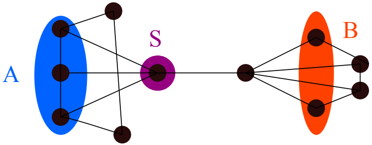
Factorization

$$P(\mathbf{x}) \propto \exp\left[\sum_{(i,j)\in G} \Psi_{i,j}(x_i, x_j)\right].$$

Graphical Models: Definition

Conditional Independence

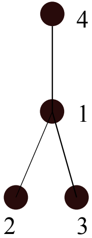
$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$



Factorization

$$P(\mathbf{x}) \propto \exp\left[\sum_{(i,j)\in G} \Psi_{i,j}(x_i, x_j)\right].$$

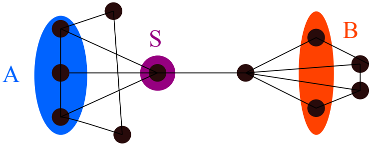
Tree-Structured Graphical Models



Graphical Models: Definition

Conditional Independence

$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$



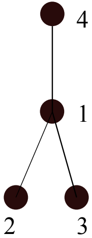
Factorization

$$P(\mathbf{x}) \propto \exp\left[\sum_{(i,j)\in G} \Psi_{i,j}(x_i, x_j)\right].$$

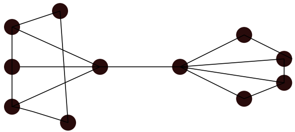
Tree-Structured Graphical Models

$$P(\mathbf{x}) = \prod_{i\in V} P_i(x_i) \prod_{(i,j)\in E} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i)P_j(x_j)}$$

$$= P_1(x_1)P_{2|1}(x_2|x_1)P_{3|1}(x_3|x_1)P_{4|1}(x_4|x_1).$$



Structure Learning of Graphical Models

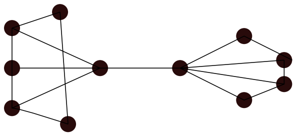


- Graphical model on p nodes
- n i.i.d. samples from multivariate distribution
- Output estimated structure $\widehat{G}^n$

Structural Consistency: $\lim_{n \to \infty} P\left[\widehat{G}^n \neq G\right] = 0.$

Structure Learning of Graphical Models

- Graphical model on p nodes
- n i.i.d. samples from multivariate distribution
- Output estimated structure $\widehat{G}^n$



Structural Consistency: $\lim_{n\to\infty} P\left[\widehat{G}^n \neq G\right] = 0.$

Challenge: High Dimensionality ("Data-Poor" Regime)

- Large p, small n regime $(p \gg n)$
- Sample Complexity: Required # of samples to achieve consistency

Challenge: Computational Complexity

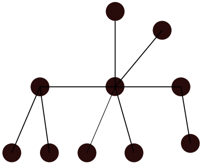
Goal: Address above challenges and provide provable guarantees

Tree Graphical Models: Tractable Learning

Maximum likelihood learning of tree structure

- Proposed by Chow and Liu (68)
- Max. weight spanning tree

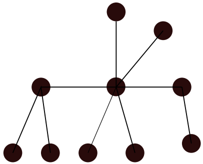
$$\hat{T}_{\mathrm{ML}} = \arg\max_{T} \sum_{k=1}^{n} \log P(\mathbf{x}_V).$$



Tree Graphical Models: Tractable Learning

Maximum likelihood learning of tree structure

- Proposed by Chow and Liu (68)
- Max. weight spanning tree

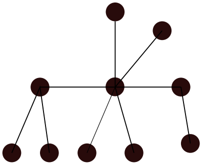


$$\hat{T}_{\mathrm{ML}} = \arg\max_T \sum_{k=1}^{n} \log P(\mathbf{x}_V).$$

$$\hat{T}_{\mathrm{ML}} = \arg\max_T \sum_{(i,j)\in T} \widehat{I}^n(X_i; X_j).$$

Tree Graphical Models: Tractable Learning

Maximum likelihood learning of tree structure



- Proposed by Chow and Liu (68)
- Max. weight spanning tree

$$\hat{T}_{\mathrm{ML}} = \arg\max_{T} \sum_{k=1}^{n} \log P(\mathbf{x}_V).$$

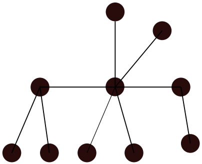
$$\hat{T}_{\mathrm{ML}} = \arg\max_{T} \sum_{(i,j)\in T} \widehat{I}^n(X_i; X_j).$$

- Pairwise statistics suffice for ML

Tree Graphical Models: Tractable Learning

Maximum likelihood learning of tree structure

- Proposed by Chow and Liu (68)
- Max. weight spanning tree



$$\hat{T}_{\mathrm{ML}} = \arg\max_T \sum_{k=1}^{n} \log P(\mathbf{x}_V).$$

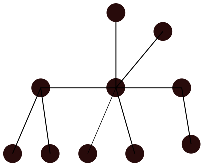
$$\hat{T}_{\mathrm{ML}} = \arg\max_T \sum_{(i,j)\in T} \widehat{I}^n(X_i; X_j).$$

- Pairwise statistics suffice for ML
- n samples and p nodes: Sample complexity: $\frac{\log p}{n} = O(1).$

Tree Graphical Models: Tractable Learning

Maximum likelihood learning of tree structure

- Proposed by Chow and Liu (68)
- Max. weight spanning tree



$$\hat{T}_{\mathrm{ML}} = \arg\max_{T} \sum_{k=1}^{n} \log P(\mathbf{x}_V).$$

$$\hat{T}_{\mathrm{ML}} = \arg\max_{T} \sum_{(i,j)\in T} \widehat{I}^{n}(X_i; X_j).$$

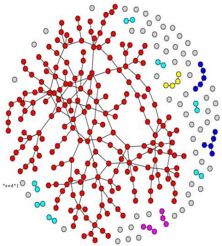
- Pairwise statistics suffice for ML
- n samples and p nodes: Sample complexity: $\dfrac{\log p}{n} = O(1).$

What other classes of graphical models are tractable for learning?

Learning Graphical Models Beyond Trees

Challenges

- Presence of cycles
  - Pairwise statistics no longer suffice
  - Likelihood function not tractable



$$P(\mathbf{x}) = \frac{1}{Z} \exp \left[\sum_{(i,j) \in G} \Psi_{i,j}(x_i, x_j) \right].$$

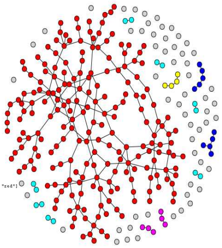
Learning Graphical Models Beyond Trees

Challenges

- Presence of cycles
  - Pairwise statistics no longer suffice
  - Likelihood function not tractable

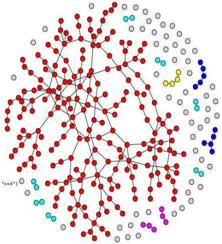
$$P(\mathbf{x}) = \frac{1}{Z} \exp \left[\sum_{(i,j) \in G} \Psi_{i,j}(x_i, x_j) \right].$$

- Presence of high-degree nodes
  - Brute-force search not tractable



Learning Graphical Models Beyond Trees

Challenges



- Presence of cycles
  - Pairwise statistics no longer suffice
  - Likelihood function not tractable

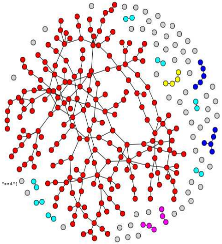
$$P(\mathbf{x}) = \frac{1}{Z} \exp \left[\sum_{(i,j) \in G} \Psi_{i,j}(x_i, x_j) \right].$$

- Presence of high-degree nodes
  - Brute-force search not tractable

Can we provide learning guarantees under above conditions?

Learning Graphical Models Beyond Trees

Challenges



- Presence of cycles
  - Pairwise statistics no longer suffice
  - Likelihood function not tractable

$$P(\mathbf{x}) = \frac{1}{Z} \exp \left[\sum_{(i,j) \in G} \Psi_{i,j}(x_i, x_j) \right].$$

- Presence of high-degree nodes
  - Brute-force search not tractable

Can we provide learning guarantees under above conditions?

Our Perspective: Tractable Graph Families

- Characterize the class of tractable families
- Incorporate all the above challenges
- Relevant for real datasets, e.g., social-network data

Related Work in Structure Learning

Algorithms for Structure Learning

- Chow and Liu (68)
- Meinshausen and Buehlmann (06)
- Bresler, Mossel and Sly (09)
- Ravikumar, Wainwright and Lafferty (10) ...

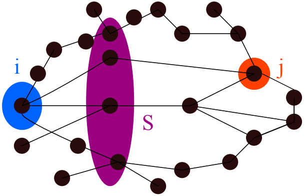
Approaches Employed

- EM/Search approaches
- Combinatorial/Greedy approach
- Convex relaxation, ...

Outline

Intuitions: Conditional Mutual Information Test

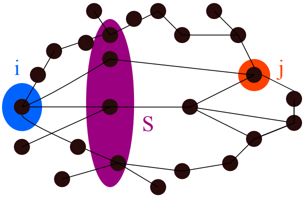
Separators in Graphical Models



$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Intuitions: Conditional Mutual Information Test

Separators in Graphical Models



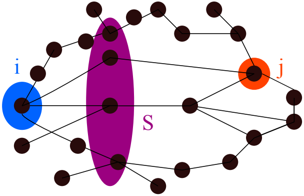
$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Observations

- Δ-separator for graphs with maximum degree Δ
  - Brute-force search for the separator: $\underset{|S| \leq \Delta}{\text{argmin}} \; I(X_i; X_j | \mathbf{X}_S)$
  - Computational complexity scales as $O(p^{\Delta})$

Intuitions: Conditional Mutual Information Test

Separators in Graphical Models



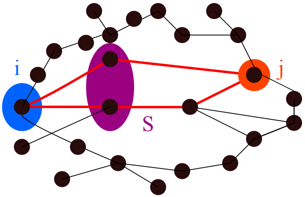
$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Observations

- Δ-separator for graphs with maximum degree Δ
  - Brute-force search for the separator: $\underset{|S| \leq \Delta}{\operatorname{argmin}} \; I(X_i; X_j | \mathbf{X}_S)$
  - Computational complexity scales as $O(p^\Delta)$
- Approximate separators in general graphs?

Intuitions: Conditional Mutual Information Test

Separators in Graphical Models



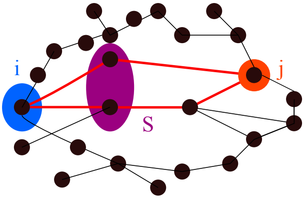
$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Observations

- Δ-separator for graphs with maximum degree Δ
  - Brute-force search for the separator: $\underset{|S| \leq \Delta}{\text{argmin}} \; I(X_i; X_j | \mathbf{X}_S)$
  - Computational complexity scales as $O(p^{\Delta})$
- Approximate separators in general graphs?

Intuitions: Conditional Mutual Information Test

Separators in Graphical Models

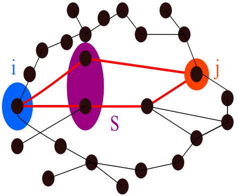


$$X_i \not\perp X_j | \mathbf{X}_S \overset{?}{\Longrightarrow} I(X_i; X_j | \mathbf{X}_S) \approx 0$$

Observations

- Δ-separator for graphs with maximum degree Δ
  - Brute-force search for the separator: $\underset{|S| \leq \Delta}{\text{argmin}} \; I(X_i; X_j | \mathbf{X}_S)$
  - Computational complexity scales as $O(p^{\Delta})$
- Approximate separators in general graphs?

Tractable Graph Families: Local Separation



γ-Local Separator $S_\gamma(i, j)$

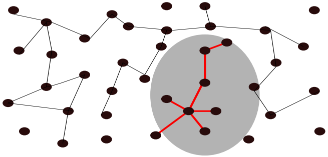
Minimal vertex separator with respect to paths of length less than γ

(η, γ)-Local Separation Property for Graph G

$$|S_\gamma(i, j)| \le \eta \text{ for all } (i, j) \notin G$$

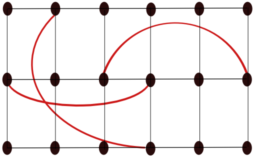
Locally tree-like

- Erdős-Rényi graphs
- Power-law/scale-free graphs



Small-world Graphs

- Watts-Strogatz model
- Hybrid/augmented graphs



Outline

Setup: Ising and Gaussian Graphical Models

- n i.i.d. samples available for structure estimation

Setup: Ising and Gaussian Graphical Models

- n i.i.d. samples available for structure estimation
- Ising and Gaussian Graphical Models

$$P(\mathbf{x}) \propto \exp\left[\frac{1}{2}\mathbf{x}^T \mathbf{J}_G \mathbf{x} + \mathbf{h}^T \mathbf{x}\right], \quad \mathbf{x} \in \{-1, 1\}^p.$$

$$f(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T \mathbf{J}_G \mathbf{x} + \mathbf{h}^T \mathbf{x}\right], \quad \mathbf{x} \in \mathbb{R}^p.$$

Setup: Ising and Gaussian Graphical Models

- n i.i.d. samples available for structure estimation
- Ising and Gaussian Graphical Models

$$P(\mathbf{x}) \propto \exp\left[\frac{1}{2}\mathbf{x}^T\mathbf{J}_G\mathbf{x} + \mathbf{h}^T\mathbf{x}\right], \quad \mathbf{x} \in \{-1,1\}^p.$$

$$f(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{J}_G\mathbf{x} + \mathbf{h}^T\mathbf{x}\right], \quad \mathbf{x} \in \mathbb{R}^p.$$

- For $(i,j) \in G$, $J_{\min} \leq |J_{i,j}| \leq J_{\max}$

Setup: Ising and Gaussian Graphical Models

- n i.i.d. samples available for structure estimation
- Ising and Gaussian Graphical Models

$$P(\mathbf{x}) \propto \exp\left[\frac{1}{2}\mathbf{x}^T \mathbf{J}_G \mathbf{x} + \mathbf{h}^T \mathbf{x}\right], \quad \mathbf{x} \in \{-1, 1\}^p.$$

$$f(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T \mathbf{J}_G \mathbf{x} + \mathbf{h}^T \mathbf{x}\right], \quad \mathbf{x} \in \mathbb{R}^p.$$

- For $(i, j) \in G$, $J_{\min} \leq |J_{i,j}| \leq J_{\max}$
- Graph G satisfies (η, γ) local separation property

Setup: Ising and Gaussian Graphical Models

- n i.i.d. samples available for structure estimation
- Ising and Gaussian Graphical Models

$$P(\mathbf{x}) \propto \exp\left[\frac{1}{2}\mathbf{x}^T\mathbf{J}_G\mathbf{x} + \mathbf{h}^T\mathbf{x}\right], \quad \mathbf{x} \in \{-1, 1\}^p.$$

$$f(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{J}_G\mathbf{x} + \mathbf{h}^T\mathbf{x}\right], \quad \mathbf{x} \in \mathbb{R}^p.$$

- For $(i, j) \in G$, $J_{\min} \leq |J_{i,j}| \leq J_{\max}$
- Graph G satisfies (η, γ) local separation property

Tradeoff between $\eta, \gamma, J_{\min}, J_{\max}$ for tractable learning

Regime of Tractable Learning

Efficient Learning Under Approximate Separation

- Maximum edge potential $J_{\max}$ of Ising model satisfies

$$J_{\max} < J^*.$$

J^* is threshold for phase transition for conditional uniqueness.

Regime of Tractable Learning

Efficient Learning Under Approximate Separation

- Maximum edge potential $J_{\max}$ of Ising model satisfies

$$J_{\max} < J^*.$$

  J^* is threshold for phase transition for conditional uniqueness.

- Gaussian model is α-walk summable

$$\|\overline{\mathbf{R}}_G\| \leq \alpha < 1.$$

  $\overline{\mathbf{R}}_G$ is absolute partial correlation matrix.

$$\mathbf{J}_G = \mathbf{I} - \mathbf{R}_G.$$

Regime of Tractable Learning

Efficient Learning Under Approximate Separation

- Maximum edge potential $J_{\max}$ of Ising model satisfies

$$J_{\max} < J^*.$$

  J^* is threshold for phase transition for conditional uniqueness.

- Gaussian model is α-walk summable

$$\|\overline{\mathbf{R}}_G\| \leq \alpha < 1.$$

  $\overline{\mathbf{R}}_G$ is absolute partial correlation matrix.

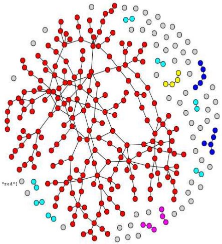
$$\mathbf{J}_G = \mathbf{I} - \mathbf{R}_G.$$

Tractable Parameter Regime for Structure Learning

Tractable Graph Families and Regimes

- Graph G satisfies (η, γ)-local separation property where

$$\eta = O(1).$$



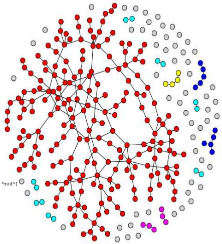
Tractable Graph Families and Regimes

- Graph G satisfies (η, γ)-local separation property where

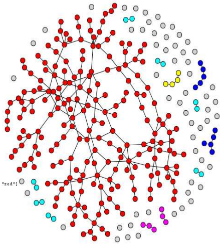
$$\eta = O(1).$$

- Maximum edge potential $J_{\max}$ satisfies

$$\alpha := \frac{\tanh J_{\max}}{\tanh J^*} < 1 \text{ or } \|\overline{\mathbf{R}}_G\| \leq \alpha < 1.$$



Tractable Graph Families and Regimes



- Graph G satisfies (η, γ)-local separation property where

$$\eta = O(1).$$

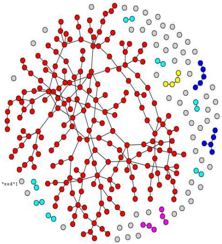
- Maximum edge potential $J_{\max}$ satisfies

$$\alpha := \frac{\tanh J_{\max}}{\tanh J^*} < 1 \text{ or } \|\overline{\mathbf{R}}_G\| \leq \alpha < 1.$$

- Minimum edge potential $J_{\min}$ is sufficiently strong

$$\frac{J_{\min}}{\alpha^\gamma} = \widetilde{\omega}(1).$$

Tractable Graph Families and Regimes



- Graph G satisfies (η, γ)-local separation property where

$$\eta = O(1).$$

- Maximum edge potential $J_{\max}$ satisfies

$$\alpha := \frac{\tanh J_{\max}}{\tanh J^*} < 1 \text{ or } \|\overline{\mathbf{R}}_G\| \leq \alpha < 1.$$

- Minimum edge potential $J_{\min}$ is sufficiently strong

$$\frac{J_{\min}}{\alpha^{\gamma}} = \widetilde{\omega}(1).$$

- Edge potentials are generic.

Example: girth g, maximum degree Δ

- Structural criteria: (η, γ)-local separation property is satisfied

$$\eta = 1, \quad \gamma = g.$$

- Parameter criteria: The maximum edge potential satisfies

$$J_{\max} < J^* = \operatorname{atanh}(\Delta^{-1}), \quad \alpha := \frac{\tanh J_{\max}}{\tanh J^*}.$$

- Tradeoff: The minimum edge potential satisfies

$$J_{\min} \alpha^g = \omega(1).$$

For example, when

$$J_{\min} = \Theta(\Delta^{-1}) \;\Rightarrow\; \Delta \alpha^g = o(1).$$

Learnability regime involves a tradeoff between degree and girth.

Outline

Algorithm for Structure Learning

Conditional Mutual Information Thresholding (CMIT)

- Empirical Conditional Mutual Information from samples
- Attempt to search for approx. separator of size η

$$\boxed{(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}}$$

Algorithm for Structure Learning

Conditional Mutual Information Thresholding (CMIT)

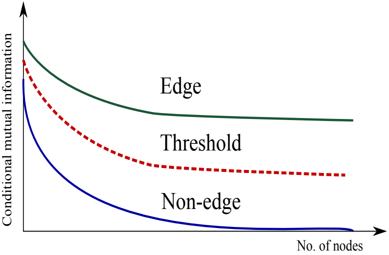
- Empirical Conditional Mutual Information from samples
- Attempt to search for approx. separator of size η

$$\boxed{(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}}$$

Threshold $\xi_{n,p}$

- Depends only on # of samples n and # of nodes p

$$\xi_{n,p} = O(J^2_{\min}) \cap \omega(\alpha^{2\gamma}) \cap \Omega\left(\frac{\log p}{n}\right).$$



Algorithm for Structure Learning

Conditional Mutual Information Thresholding (CMIT)

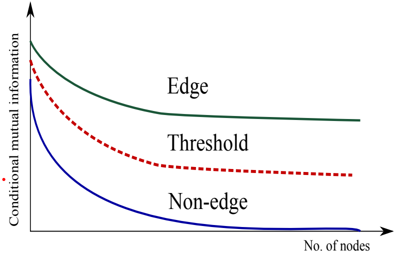
- Empirical Conditional Mutual Information from samples
- Attempt to search for approx. separator of size η

$$(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}$$

Threshold $\xi_{n,p}$

- Depends only on # of samples n and # of nodes p

$$\xi_{n,p} = O(J_{\min}^2) \cap \omega(\alpha^{2\gamma}) \cap \Omega\left(\frac{\log p}{n}\right).$$



Local Test Using Low-order Statistics

Guarantees on Conditional Mutual Information Test

$$(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}$$

- Ising/Gaussian graphical model on p nodes
- No. of samples n such that

$$n = \Omega(J_{\min}^{-4} \log p).$$

Theorem

CMIT is structurally consistent

$$\lim_{\substack{p,n \to \infty \\ n = \Omega(J_{\min}^{-4} \log p)}} P\left[\widehat{G}_p^n \neq G_p\right] = 0.$$

- Probability measure on both graph and samples

Lower Bound on Sample Complexity

- Erdős-Rényi random graph $G \sim \mathcal{G}(p, c/p)$

Theorem

For any estimator $\widehat{G}_p^n$, it is necessary that

- Discrete distribution over $\mathcal{X}$: $n \geq \dfrac{c \log_2 p}{2 \log_2 |\mathcal{X}|}$
- Gaussian with α-walk summability: $n \geq \dfrac{c \log_2 p}{\log_2 \left[2\pi e \left(\frac{1}{1-\alpha} + 1\right)\right]}$

$$\lim_{n \to \infty} P\left[\widehat{G}_p^n \neq G_p\right] = 0.$$

$\Omega(c \log p)$ samples needed for random graph structure estimation.

Lower Bound on Sample Complexity

- Erdős-Rényi random graph $G \sim \mathcal{G}(p, c/p)$

Theorem

For any estimator $\widehat{G}_p^n$, it is necessary that

- Discrete distribution over $\mathcal{X}$: $n \geq \dfrac{c \log_2 p}{2 \log_2 |\mathcal{X}|}$
- Gaussian with α-walk summability: $n \geq \dfrac{c \log_2 p}{\log_2 \left[2\pi e \left(\frac{1}{1-\alpha} + 1 \right) \right]}$

$$\lim_{n \to \infty} P\left[\widehat{G}_p^n \neq G_p \right] = 0.$$

Proof Techniques

- Fano's inequality over typical graphs
- Characterize typical graphs for Erdős-Rényi ensemble

$\Omega(c \log p)$ samples needed for random graph structure estimation.

Proof Ideas

$$\boxed{(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}}$$

- Correctness of algorithm under exact statistics
- Consistency under prescribed sample complexity
  - Concentration bounds for empirical quantities

Proof Ideas

$$(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}$$

- Correctness of algorithm under exact statistics
- Consistency under prescribed sample complexity
  - Concentration bounds for empirical quantities

Analysis for non-neighbors

- Conditional mutual information upon conditioning by local separator
- Derive rate of decay for conditional mutual information

  Self-avoiding walk tree analysis for Ising models

  Walk-sum analysis for Gaussian models

Proof Ideas

$$(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}$$

- Correctness of algorithm under exact statistics
- Consistency under prescribed sample complexity
  - Concentration bounds for empirical quantities

Analysis for non-neighbors

- Conditional mutual information upon conditioning by local separator
- Derive rate of decay for conditional mutual information

  Self-avoiding walk tree analysis for Ising models

  Walk-sum analysis for Gaussian models

Analysis for neighbors

- Lower bound under generic edge potentials

Proof Ideas

$$(i,j) \in \widehat{G} \text{ if } \min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}$$

- Correctness of algorithm under exact statistics
- Consistency under prescribed sample complexity
  - Concentration bounds for empirical quantities

Analysis for non-neighbors

- Conditional mutual information upon conditioning by local separator
- Derive rate of decay for conditional mutual information

  Self-avoiding walk tree analysis for Ising models

  Walk-sum analysis for Gaussian models

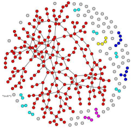
Analysis for neighbors

- Lower bound under generic edge potentials

Outline

Summary and Outlook

Summary



- Local algorithm based on low-order statistics
- Transparent assumptions
- Logarithmic sample complexity

Outlook

- Is structure learning beyond this regime hard?
- Connections with incoherence conditions
- Structure learning with latent variables

A. Anandkumar, V. Tan and Alan Willsky, "High-Dimensional Structure Learning of Ising Models: Tractable Graph Families" ArXiv 1107.1736.

A. Anandkumar, V. Tan and Alan Willsky, "High-Dimensional Gaussian Graphical Model Selection: Tractable Graph Families" ArXiv 1107.1270.