Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

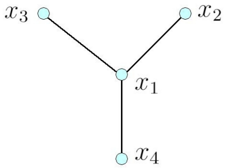
Vincent Tan
Animashree Anandkumar, Alan Willsky

Stochastic Systems Group,
Laboratory for Information and Decision Systems,
Massachusetts Institute of Technology

Allerton Conference (Sep 30, 2009)

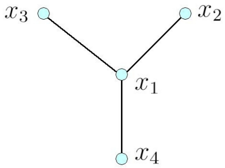
Motivation

Given a set of i.i.d. samples drawn from p, a Gaussian tree model.

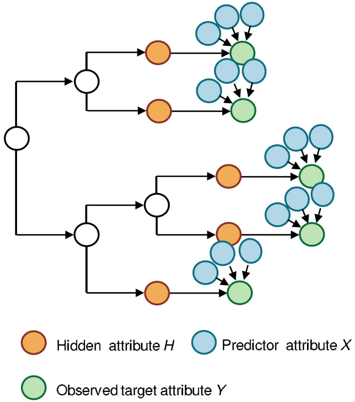


Motivation

Given a set of i.i.d. samples drawn from p, a Gaussian tree model.



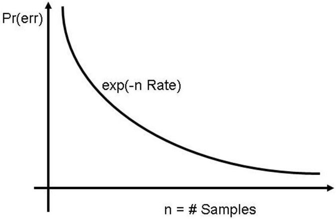
Inferring structure of Phylogenetic Trees from observed data.



Carlson et al. 2008, PLoS Comp. Bio.

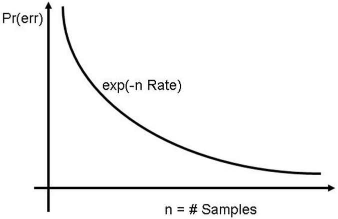
More motivation

- What is the exact rate of decay of the probability of error?



More motivation

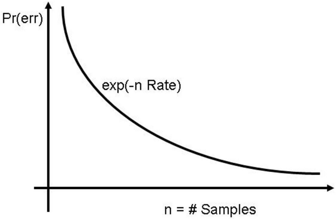
- What is the exact rate of decay of the probability of error?



- How do the structure and parameters of the model influence the error exponent (rate of decay)?

More motivation

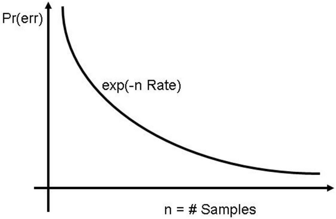
- What is the exact rate of decay of the probability of error?



- How do the structure and parameters of the model influence the error exponent (rate of decay)?
- What are extremal tree distributions for learning?

More motivation

- What is the exact rate of decay of the probability of error?



- How do the structure and parameters of the model influence the error exponent (rate of decay)?
- What are extremal tree distributions for learning?
- Consistency is well established (Chow and Wagner 1973).
- Error Exponent is a quantitative measure of the "goodness" of learning.

Main Contributions

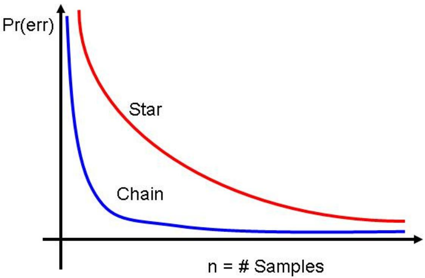
1. Provide the exact Rate of Decay for a given p.

Main Contributions

1. Provide the exact Rate of Decay for a given p.
2. Rate of decay $\approx$ SNR for learning.

Main Contributions

1. Provide the exact Rate of Decay for a given p.
2. Rate of decay $\approx$ SNR for learning.
3. Characterized the extremal trees structures for learning, i.e., stars and Markov chains.



Stars have the slowest rate. Chains have the fastest rate.

Notation and Background

$p = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: d-dimensional Gaussian tree model.

Notation and Background

$p = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: d-dimensional Gaussian tree model.

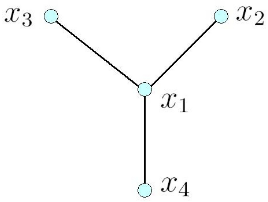
Samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn i.i.d. from p.

Notation and Background

$p = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: d-dimensional Gaussian tree model.

Samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn i.i.d. from p.

- p: Markov on $T_p = (\mathcal{V}, \mathcal{E}_p)$, a tree.
- p: Factorizes according to T_p.

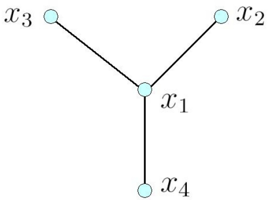


Notation and Background

$p = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: d-dimensional Gaussian tree model.

Samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn i.i.d. from p.

- p: Markov on $T_p = (\mathcal{V}, \mathcal{E}_p)$, a tree.
- p: Factorizes according to T_p.



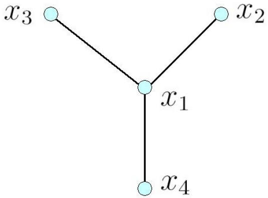
$$p(\mathbf{x}) = p_1(x_1) \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)} \frac{p_{1,3}(x_1, x_3)}{p_1(x_1)} \frac{p_{1,4}(x_1, x_4)}{p_1(x_1)},$$

Notation and Background

$p = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: d-dimensional Gaussian tree model.

Samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn i.i.d. from p.

- p: Markov on $T_p = (\mathcal{V}, \mathcal{E}_p)$, a tree.
- p: Factorizes according to T_p.



$$p(\mathbf{x}) = p_1(x_1)\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)}\frac{p_{1,3}(x_1, x_3)}{p_1(x_1)}\frac{p_{1,4}(x_1, x_4)}{p_1(x_1)}, \qquad \mathbf{\Sigma}^{-1} = \begin{pmatrix} \spadesuit & \clubsuit & \clubsuit & \clubsuit \\ \clubsuit & \spadesuit & 0 & 0 \\ \clubsuit & 0 & \spadesuit & 0 \\ \clubsuit & 0 & 0 & \spadesuit \end{pmatrix}$$

Max-Likelihood Learning of Tree Distributions (Chow-Liu)

- Denote $\widehat{p} = \widehat{p}_{\mathbf{x}^n}$ as the empirical distribution of $\mathbf{x}^n$, i.e.,

$$\widehat{p}(\mathbf{x}) := \mathcal{N}(\mathbf{x}; \mathbf{0}, \widehat{\mathbf{\Sigma}})$$

where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance matrix of $\mathbf{x}^n$.

Max-Likelihood Learning of Tree Distributions (Chow-Liu)

- Denote $\widehat{p} = \widehat{p}_{\mathbf{x}^n}$ as the empirical distribution of $\mathbf{x}^n$, i.e.,

$$\widehat{p}(\mathbf{x}) := \mathcal{N}(\mathbf{x}; \mathbf{0}, \widehat{\mathbf{\Sigma}})$$

  where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance matrix of $\mathbf{x}^n$.

- $\widehat{p}_e$: Empirical on edge e.

Max-Likelihood Learning of Tree Distributions (Chow-Liu)

- Denote $\widehat{p} = \widehat{p}_{\mathbf{x}^n}$ as the empirical distribution of $\mathbf{x}^n$, i.e.,

$$\widehat{p}(\mathbf{x}) := \mathcal{N}(\mathbf{x}; \mathbf{0}, \widehat{\mathbf{\Sigma}})$$

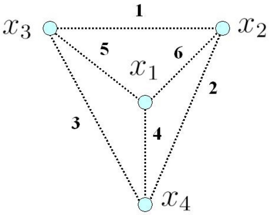
  where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance matrix of $\mathbf{x}^n$.

- $\widehat{p}_e$: Empirical on edge e.

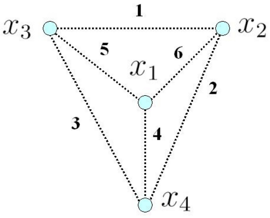
- Reduces to a max-weight spanning tree problem (Chow-Liu 1968)

$$\widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{argmax}_{\mathcal{E}_q \,:\, q \in \mathrm{Trees}} \sum_{e \in \mathcal{E}_q} I(\widehat{p}_e). \qquad I(\widehat{p}_e) := \widehat{I}(X_i; X_j).$$

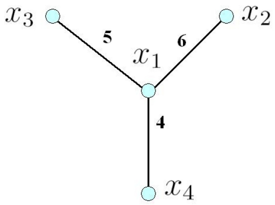
Max-Likelihood Learning of Tree Distributions



Max-Likelihood Learning of Tree Distributions

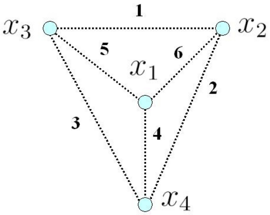


True MIs $\{I(p_e)\}$

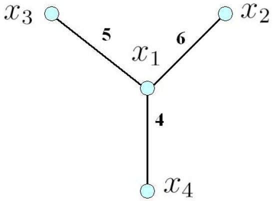


Max-weight spanning tree $\mathcal{E}_p$

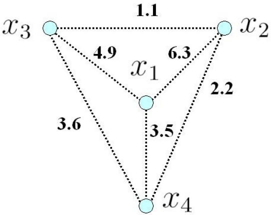
Max-Likelihood Learning of Tree Distributions



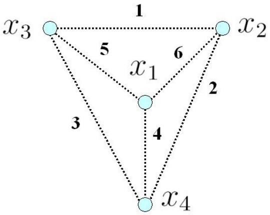
True MIs $\{I(p_e)\}$



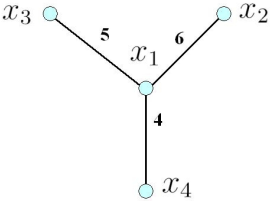
Max-weight spanning tree $\mathcal{E}_p$



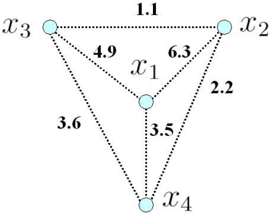
Max-Likelihood Learning of Tree Distributions



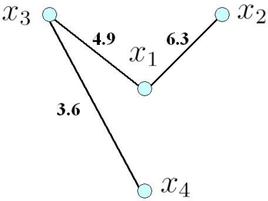
True MIs $\{I(p_e)\}$



Max-weight spanning tree $\mathcal{E}_p$



Empirical MIs $\{I(\widehat{p}_e)\}$ from $\mathbf{x}^n$



Max-weight spanning tree $\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p$

Problem Statement

- The estimated edge set is $\widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n)$

Problem Statement

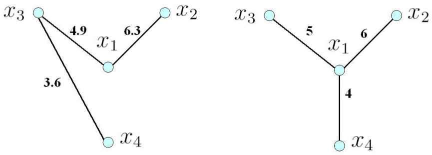
- The estimated edge set is $\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n)$ and the error event is

$$\left\{\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\right\}.$$

Problem Statement

- The estimated edge set is $\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n)$ and the error event is

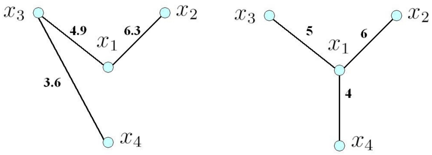
$$\left\{ \widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p \right\}.$$



Problem Statement

- The estimated edge set is $\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n)$ and the error event is

$$\left\{\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\right\}.$$



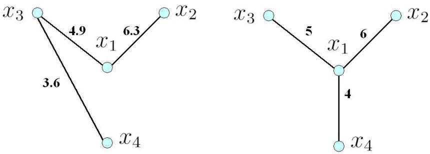
- Find and analyze the error exponent K_p:

$$K_p := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\left\{\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\right\}\right).$$

Problem Statement

- The estimated edge set is $\widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n)$ and the error event is

$$\left\{ \widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p \right\}.$$



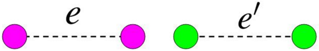
- Find and analyze the error exponent K_p:

$$K_p := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\left\{ \widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p \right\} \right).$$

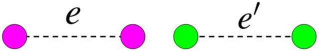
- Alternatively,

$$\mathbb{P}\left(\left\{ \widehat{\mathcal{E}}_{\mathrm{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p \right\} \right) \doteq \exp(-nK_p).$$

The Crossover Rate I



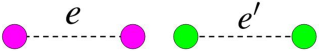
The Crossover Rate I



Two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $p_{e,e'}$, s.t.

$$I(p_e) > I(p_{e'}).$$

The Crossover Rate I



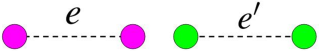
Two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $p_{e,e'}$, s.t.

$$I(p_e) > I(p_{e'}).$$

Consider the crossover event:

$$\{I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\}.$$

The Crossover Rate I



Two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $p_{e,e'}$, s.t.

$$I(p_e) > I(p_{e'}).$$

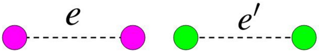
Consider the crossover event:

$$\{I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\}.$$

Definition: Crossover Rate

$$J_{e,e'} \;:=\; \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\}\right).$$

The Crossover Rate I



Two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $p_{e,e'}$, s.t.

$$I(p_e) > I(p_{e'}).$$

Consider the crossover event:

$$\{I(\widehat{p}_e) \le I(\widehat{p}_{e'})\}.$$

Definition: Crossover Rate

$$J_{e,e'} \;:=\; \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{I(\widehat{p}_e) \le I(\widehat{p}_{e'})\}\right).$$

This event may potentially lead to an error in structure learning. Why?

The Crossover Rate II

Theorem

The *crossover rate* *is*

$$J_{e,e'} = \inf_{q \in \text{Gaussians}} \left\{ D(q \,||\, p_{e,e'}) : I(q_{e'}) = I(q_e) \right\}.$$

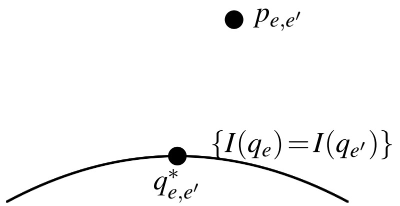
The Crossover Rate II

Theorem

The crossover rate is

$$J_{e,e'} = \inf_{q \in \text{Gaussians}} \left\{ D(q \,||\, p_{e,e'}) : I(q_{e'}) = I(q_e) \right\}.$$

By assumption $I(p_e) > I(p_{e'})$.



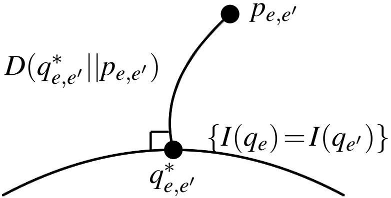
The Crossover Rate II

Theorem

The crossover rate is

$$J_{e,e'} = \inf_{q \in \text{Gaussians}} \left\{ D(q \,||\, p_{e,e'}) : I(q_{e'}) = I(q_e) \right\}.$$

By assumption $I(p_e) > I(p_{e'})$.



Error Exponent for Structure Learning II

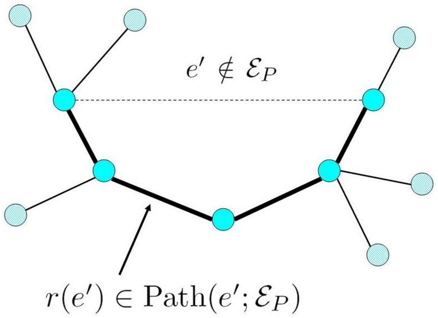
$$\mathbb{P}\left(\left\{\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\right\}\right) \doteq \exp(-nK_p).$$

Error Exponent for Structure Learning II

$$\mathbb{P}\left(\left\{\widehat{\mathcal{E}}_{\mathsf{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\right\}\right) \doteq \exp(-nK_p).$$

Theorem (First Result)

$$K_p = \min_{e' \notin \mathcal{E}_p} \left(\min_{e \in \mathrm{Path}(e';\mathcal{E}_p)} J_{e,e'} \right).$$



Approximating the Crossover Rate I

Definition: $p_{e,e'}$ satisfies the very noisy learning condition if

$$||\rho_e| - |\rho_{e'}|| < \epsilon \qquad \Rightarrow \qquad I(p_e) \approx I(p_{e'}).$$

Euclidean Information Theory (Borade, Zheng 2007).

Approximating the Crossover Rate II

Theorem (Second Result)

The approximate crossover rate is:

$$\widetilde{J}_{e,e'} = \frac{(I(p_{e'}) - I(p_e))^2}{2\,\mathrm{Var}(s_{e'} - s_e)}$$

Approximating the Crossover Rate II

Theorem (Second Result)

The approximate crossover rate is:

$$\widetilde{J}_{e,e'} = \frac{(I(p_{e'}) - I(p_e))^2}{2\,\mathrm{Var}(s_{e'} - s_e)}$$

where s_e is the information density:

$$s_e(x_i, x_j) = \log \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}$$

Approximating the Crossover Rate II

Theorem (Second Result)

The approximate crossover rate is:

$$\widetilde{J}_{e,e'} = \frac{(I(p_{e'}) - I(p_e))^2}{2\,\mathrm{Var}(s_{e'} - s_e)}$$

where s_e is the information density:

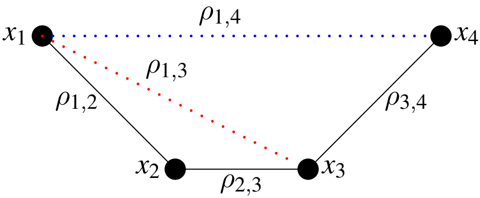
$$s_e(x_i, x_j) = \log \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}$$

The approximate error exponent is

$$\widetilde{K}_p = \min_{e' \in \mathcal{E}_p} \left(\min_{e \in \mathrm{Path}(e';\mathcal{E}_p)} \widetilde{J}_{e,e'} \right).$$

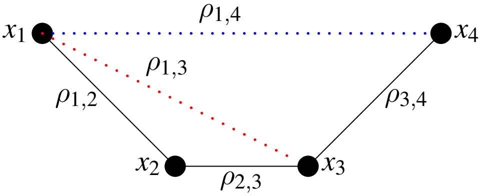
Correlation Decay

Correlation Decay



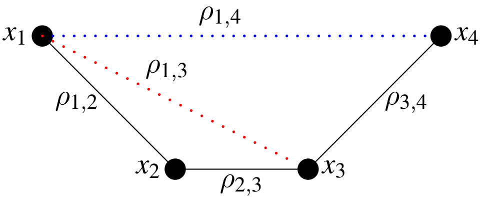
- $\rho_{i,j} = \mathbb{E}[x_i\, x_j]$.

Correlation Decay



- $\rho_{i,j} = \mathbb{E}[x_i x_j]$.
- Markov property $\Rightarrow \quad \rho_{1,3} = \rho_{1,2} \times \rho_{2,3}$.
- Correlation decay $\Rightarrow \quad |\rho_{1,4}| \leq |\rho_{1,3}|$.

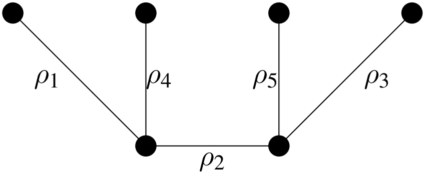
Correlation Decay



- $\rho_{i,j} = \mathbb{E}[x_i x_j]$.
- Markov property $\Rightarrow$ $\rho_{1,3} = \rho_{1,2} \times \rho_{2,3}$.
- Correlation decay $\Rightarrow$ $|\rho_{1,4}| \leq |\rho_{1,3}|$.
- $(1,4)$ is not likely to be mistaken as a true edge.
- Only need to consider triangles in the true tree.

Extremal Structures I

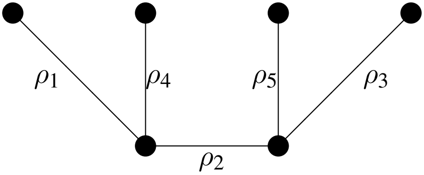
Fix $\boldsymbol{\rho}$, a vector of correlation coefficients on the tree, e.g.



$\boldsymbol{\rho} := [\rho_1, \rho_2, \rho_3, \rho_4, \rho_5]$.

Extremal Structures I

Fix $\boldsymbol{\rho}$, a vector of correlation coefficients on the tree, e.g.



$\boldsymbol{\rho} := [\rho_1, \rho_2, \rho_3, \rho_4, \rho_5]$.

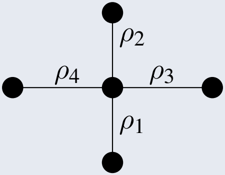
Which structures gives the highest and lowest exponents?

Extremal Structures II

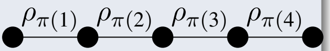
Theorem (Main Result)

- *Worst: The star minimizes $\widetilde{K}_p$.*

$$\widetilde{K}_{\text{star}} \le \widetilde{K}_p.$$



- *Best: The Markov chain maximizes $\widetilde{K}_p$.*



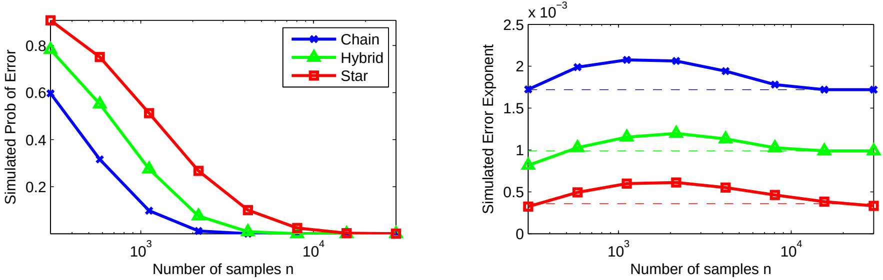
$$\widetilde{K}_{\text{chain}} \ge \widetilde{K}_p.$$

Extremal Structures III

Chain, Star and Hybrid Graphs for $d = 10$.

Extremal Structures III

Chain, Star and Hybrid Graphs for $d = 10$.



Plot of the error probability and error exponent for 3 tree graphs.

Extremal Structures IV

Remarks:

- Universal result.
- Extremal structures wrt diameter are the extremal structures for learning.

Extremal Structures IV

Remarks:

- Universal result.
- Extremal structures wrt diameter are the extremal structures for learning.
- Corroborates our intuition about correlation decay.

Extensions

- Significant reduction of complexity for computation of error exponent.

Extensions

- Significant reduction of complexity for computation of error exponent.
- Finding the best distributions for fixed ρ.

Extensions

- Significant reduction of complexity for computation of error exponent.
- Finding the best distributions for fixed ρ.
- Effect of adding and deleting nodes and edges on error exponent.

Conclusion

1. Found the rate of decay of the error probability using large deviations.

Conclusion

1. Found the rate of decay of the error probability using large deviations.

Conclusion

1. Found the rate of decay of the error probability using large deviations.
2. Used Euclidean Information Theory to obtain an SNR-like expression for crossover rate.

Conclusion

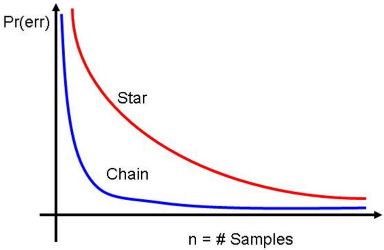
1. Found the rate of decay of the error probability using large deviations.
2. Used Euclidean Information Theory to obtain an SNR-like expression for crossover rate.
3. We can say which structures are easy and hard based on the error exponent.

   Extremal structures in terms of the tree diameter.

Conclusion

1. Found the rate of decay of the error probability using large deviations.
2. Used Euclidean Information Theory to obtain an SNR-like expression for crossover rate.
3. We can say which structures are easy and hard based on the error exponent.

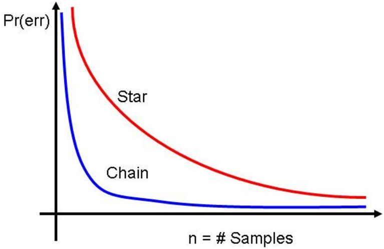
   Extremal structures in terms of the tree diameter.



Conclusion

1. Found the rate of decay of the error probability using large deviations.
2. Used Euclidean Information Theory to obtain an SNR-like expression for crossover rate.
3. We can say which structures are easy and hard based on the error exponent.

   Extremal structures in terms of the tree diameter.



Full versions can be found at

- http://arxiv.org/abs/0905.0940.
- http://arxiv.org/abs/0909.5216.