A Large-Deviation Analysis for the Maximum-Likelihood Learning of Tree Structures

Vincent Tan †
Animashree Anandkumar †∗, Lang Tong ∗ and Alan Willsky †

†Stochastic Systems Group, LIDS,
Massachusetts Institute of Technology

∗School of ECE,
Cornell University

ISIT (Jun 30, 2009)

Motivation

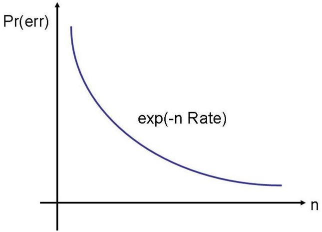
- Given a set of n i.i.d. samples drawn from P, a tree distribution.
- Maximum-Likelihood (ML) learning of the structure.

Motivation

- Given a set of n i.i.d. samples drawn from P, a tree distribution.
- Maximum-Likelihood (ML) learning of the structure.
- Large-Deviation analysis of the error in learning of the set of edges or the structure.

Motivation

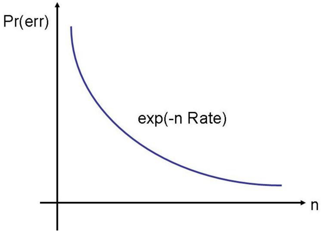
- Given a set of n i.i.d. samples drawn from P, a tree distribution.
- Maximum-Likelihood (ML) learning of the structure.
- Large-Deviation analysis of the error in learning of the set of edges or the structure.
- Questions:



1. When does the error probability decay exponentially with n?

Motivation

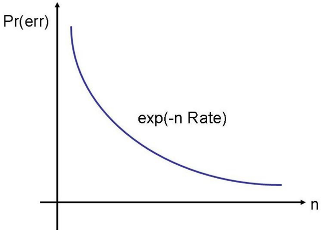
- Given a set of n i.i.d. samples drawn from P, a tree distribution.
- Maximum-Likelihood (ML) learning of the structure.
- Large-Deviation analysis of the error in learning of the set of edges or the structure.
- Questions:



1. When does the error probability decay exponentially with n?
2. What is the exact rate of decay of the probability of error?

Motivation

- Given a set of n i.i.d. samples drawn from P, a tree distribution.
- Maximum-Likelihood (ML) learning of the structure.
- Large-Deviation analysis of the error in learning of the set of edges or the structure.
- Questions:



1. When does the error probability decay exponentially with n?
2. What is the exact rate of decay of the probability of error?
3. How does the rate (error exponent) depend on the parameters of the distribution?

Main Results

- Almost every (true) tree distribution results in exponential decay.
- Quantify the exact rate of decay for a given P.

Main Results

- Almost every (true) tree distribution results in exponential decay.
- Quantify the exact rate of decay for a given P.
- Rate of decay $\approx$ SNR for learning (Intuitively appealing).

Notation and Background

- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.

Notation and Background

- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$.

Notation and Background

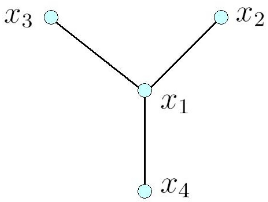
- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$.
- Vertex set: $\mathcal{V} := \{1, \ldots, d\}$. Edge set: $\mathcal{E}_P \subset \binom{\mathcal{V}}{2}$.

Notation and Background

- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$.
- Vertex set: $\mathcal{V} := \{1, \ldots, d\}$. Edge set: $\mathcal{E}_P \subset \binom{\mathcal{V}}{2}$.
  - $P(\mathbf{x})$ is Markov on $T_P = (\mathcal{V}, \mathcal{E}_P)$, a tree.
  - $P(\mathbf{x})$ factorizes according to T_P.
  - Example for P with $d = 4$.

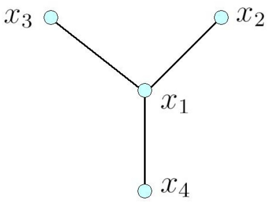
Notation and Background

- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$.
- Vertex set: $\mathcal{V} := \{1, \ldots, d\}$. Edge set: $\mathcal{E}_P \subset \binom{\mathcal{V}}{2}$.
  - $P(\mathbf{x})$ is Markov on $T_P = (\mathcal{V}, \mathcal{E}_P)$, a tree.
  - $P(\mathbf{x})$ factorizes according to T_P.
  - Example for P with $d = 4$.



Notation and Background

- Assume we have a set of samples $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ each drawn i.i.d. from $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{X}$ is a finite set.
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$.
- Vertex set: $\mathcal{V} := \{1, \ldots, d\}$. Edge set: $\mathcal{E}_P \subset \binom{\mathcal{V}}{2}$.
  - $P(\mathbf{x})$ is Markov on $T_P = (\mathcal{V}, \mathcal{E}_P)$, a tree.
  - $P(\mathbf{x})$ factorizes according to T_P.
  - Example for P with $d = 4$.



$$P(\mathbf{x}) = P_1(x_1) \times \frac{P_{1,2}(x_1, x_2)}{P_1(x_1)} \times \frac{P_{1,3}(x_1, x_3)}{P_1(x_1)} \times \frac{P_{1,4}(x_1, x_4)}{P_1(x_1)}.$$

ML Learning of Tree Distribution (Chow-Liu) I

ML Learning of Tree Distribution (Chow-Liu) I

- Solve the ML problem given $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

$$P_{\mathrm{ML}} = \operatorname*{argmax}_{Q \in \mathrm{Trees}} \log Q^n(\mathbf{x}^n).$$

ML Learning of Tree Distribution (Chow-Liu) I

- Solve the ML problem given $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

$$P_{\mathsf{ML}} = \operatorname*{argmax}_{Q \in \mathrm{Trees}} \log Q^n(\mathbf{x}^n).$$

- $\widehat{P}(\mathbf{x}) = \widehat{P}_{\mathbf{x}^n}(\mathbf{x})$: the empirical distribution of $\mathbf{x}^n$.

ML Learning of Tree Distribution (Chow-Liu) I

- Solve the ML problem given $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

$$P_{\text{ML}} = \operatorname*{argmax}_{Q \in \text{Trees}} \log Q^n(\mathbf{x}^n).$$

- $\widehat{P}(\mathbf{x}) = \widehat{P}_{\mathbf{x}^n}(\mathbf{x})$: the empirical distribution of $\mathbf{x}^n$.
- $\widehat{P}_e$: the pairwise marginal of $\widehat{P}$ on edge $e = (i, j)$.

ML Learning of Tree Distribution (Chow-Liu) I

- Solve the ML problem given $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

$$P_{\mathsf{ML}} = \operatorname*{argmax}_{Q \in \text{Trees}} \log Q^n(\mathbf{x}^n).$$

- $\widehat{P}(\mathbf{x}) = \widehat{P}_{\mathbf{x}^n}(\mathbf{x})$: the empirical distribution of $\mathbf{x}^n$.
- $\widehat{P}_e$: the pairwise marginal of $\widehat{P}$ on edge $e = (i, j)$.
- Reduces to a max-weight spanning tree problem (Chow-Liu 1968)

$$\mathcal{E}_{\mathsf{ML}} = \operatorname*{argmax}_{\mathcal{E}_Q \,:\, Q \in \text{Trees}} \sum_{e \in \mathcal{E}_Q} I(\widehat{P}_e),$$

ML Learning of Tree Distribution (Chow-Liu) I

- Solve the ML problem given $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

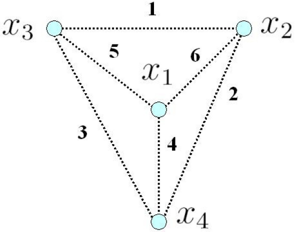
$$P_{\mathrm{ML}} = \operatorname*{argmax}_{Q \in \mathrm{Trees}} \log Q^n(\mathbf{x}^n).$$

- $\widehat{P}(\mathbf{x}) = \widehat{P}_{\mathbf{x}^n}(\mathbf{x})$: the empirical distribution of $\mathbf{x}^n$.
- $\widehat{P}_e$: the pairwise marginal of $\widehat{P}$ on edge $e = (i, j)$.
- Reduces to a max-weight spanning tree problem (Chow-Liu 1968)

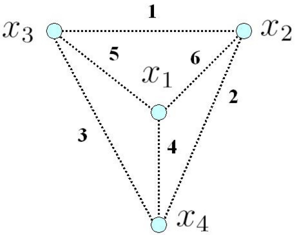
$$\mathcal{E}_{\mathrm{ML}} = \operatorname*{argmax}_{\mathcal{E}_Q \,:\, Q \in \mathrm{Trees}} \sum_{e \in \mathcal{E}_Q} I(\widehat{P}_e), \qquad I(\widehat{P}_e) := \sum_{x_i, x_j} \widehat{P}_{i,j}(x_i, x_j) \log \frac{\widehat{P}_{i,j}(x_i, x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}.$$

- Samples $\Rightarrow \{I(\widehat{P}_e) : e \in \binom{\mathcal{V}}{2}\} \Rightarrow$ max-weight spanning tree.

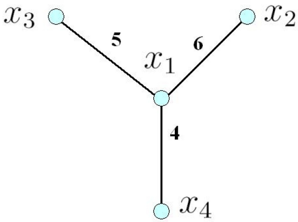
ML Learning of Tree Distribution (Chow-Liu) II



ML Learning of Tree Distribution (Chow-Liu) II

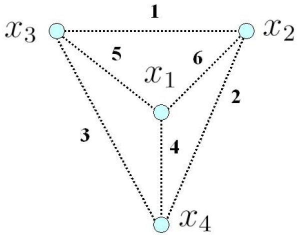


True MIs $\{I(P_e)\}$

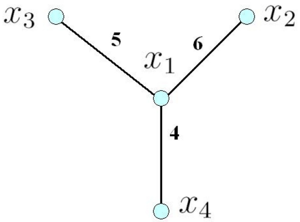


Max-weight spanning tree $\mathcal{E}_P$

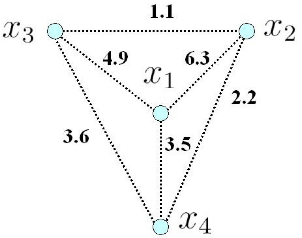
ML Learning of Tree Distribution (Chow-Liu) II



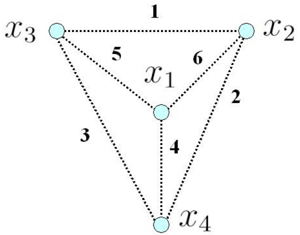
True MIs $\{I(P_e)\}$



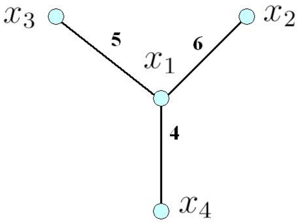
Max-weight spanning tree $\mathcal{E}_P$



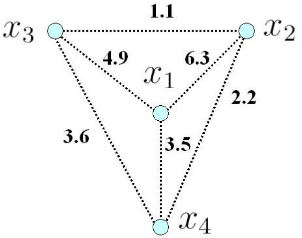
ML Learning of Tree Distribution (Chow-Liu) II



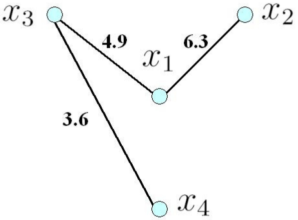
True MIs $\{I(P_e)\}$



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



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



Max-weight spanning tree $\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P$

Problem Statement

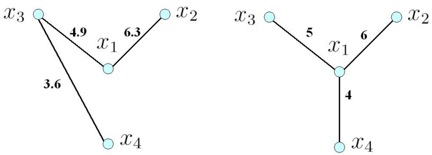
- Define $\mathcal{E}_{\mathrm{ML}}$ to be the ML edge set and the error event to be:

$$\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}.$$

Problem Statement

- Define $\mathcal{E}_{\text{ML}}$ to be the ML edge set and the error event to be:

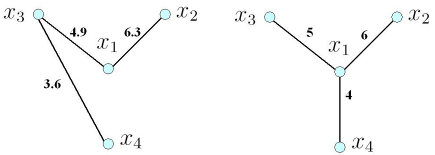
$$\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\} .$$



Problem Statement

- Define $\mathcal{E}_{\text{ML}}$ to be the ML edge set and the error event to be:

$$\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\}.$$



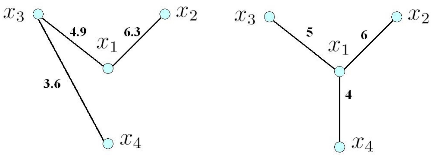
- Find the error exponent K_P:

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\}\right).$$

Problem Statement

- Define $\mathcal{E}_{\text{ML}}$ to be the ML edge set and the error event to be:

$$\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\}.$$



- Find the error exponent K_P:

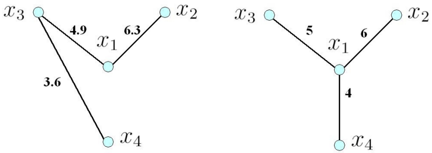
$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\}\right).$$

- Alternatively, $\mathbb{P}\left(\{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\}\right) \doteq \exp(-nK_P)$.

Problem Statement

- Define $\mathcal{E}_{\mathrm{ML}}$ to be the ML edge set and the error event to be:

$$\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}\,.$$

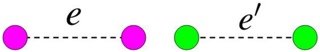


- Find the error exponent K_P:

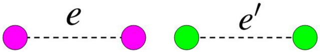
$$K_P \;:=\; \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left(\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}\right).$$

- Alternatively, $\mathbb{P}\left(\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}\right) \;\doteq\; \exp(-nK_P)$.
- Easier to consider crossover events first.

The Crossover Rate I



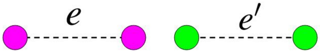
The Crossover Rate I



Given two node pairs $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, s.t.

$$I(P_e) > I(P_{e'}).$$

The Crossover Rate I



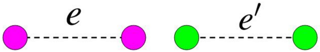
Given two node pairs $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, s.t.

$$I(P_e) > I(P_{e'}).$$

Consider the crossover event of the empirical mutual informations

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

The Crossover Rate I



Given two node pairs $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, s.t.

$$I(P_e) > I(P_{e'}).$$

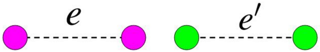
Consider the crossover event of the empirical mutual informations

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

Def: Crossover Rate

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

The Crossover Rate I



Given two node pairs $e, e' \in \binom{\mathcal{V}}{2}$ with distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, s.t.

$$I(P_e) > I(P_{e'}).$$

Consider the crossover event of the empirical mutual informations

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

Def: Crossover Rate

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

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

The Crossover Rate II

Theorem

The crossover rate for empirical mutual informations is

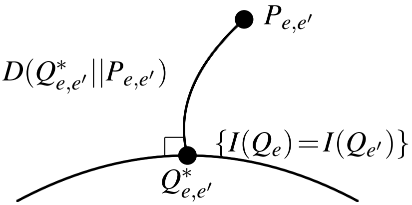
$$J_{e,e'} \;=\; \min_{Q\in\mathcal{P}(\mathcal{X}^4)} \Big\{ D(Q\,||\,P_{e,e'}) : I(Q_{e'}) = I(Q_e) \Big\}.$$

The Crossover Rate II

Theorem

The *crossover rate* *for empirical mutual informations is*

$$J_{e,e'} = \min_{Q \in \mathcal{P}(\mathcal{X}^4)} \left\{ D(Q \,||\, P_{e,e'}) : I(Q_{e'}) = I(Q_e) \right\}.$$



- Sanov's Theorem and The Contraction Principle.
- Exact but not very intuitive.
- Non-Convex.

The Crossover Rate III

- Euclidean Information Theory [Borade & Zheng '08]:

$$Q \approx P \quad \Rightarrow \quad D(Q \,||\, P) \;\approx\; \frac{1}{2}\|Q - P\|_P^2$$

The Crossover Rate III

- Euclidean Information Theory [Borade & Zheng '08]:

$$Q \approx P \quad \Rightarrow \quad D(Q\,||\,P) \;\approx\; \frac{1}{2}\|Q-P\|_P^2$$

- Def: Very noisy learning condition on $P_{e,e'}$

$$P_e \approx P_{e'} \qquad \Rightarrow$$

The Crossover Rate III

- Euclidean Information Theory [Borade & Zheng '08]:

$$Q \approx P \quad \Rightarrow \quad D(Q \,||\, P) \;\approx\; \frac{1}{2}\|Q - P\|_P^2$$

- Def: Very noisy learning condition on $P_{e,e'}$

$$P_e \approx P_{e'} \qquad \Rightarrow \qquad I(P_e) \approx I(P_{e'}).$$

The Crossover Rate III

- Euclidean Information Theory [Borade & Zheng '08]:

$$Q \approx P \quad \Rightarrow \quad D(Q \,||\, P) \;\approx\; \frac{1}{2}\|Q - P\|_P^2$$

- Def: Very noisy learning condition on $P_{e,e'}$

$$P_e \approx P_{e'} \qquad \Rightarrow \qquad I(P_e) \approx I(P_{e'}).$$

- Def: Given a $P_e = P_{i,j}$ the information density function is

$$s_e(x_i, x_j) \;:=\; \log \frac{P_{i,j}(x_i, x_j)}{P_i(x_i)P_j(x_j)}, \quad \forall\, (x_i, x_j) \in \mathcal{X}^2.$$

The Crossover Rate IV

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

The Crossover Rate IV

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

Theorem

The ***approximate*** *crossover rate is:*

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

Signal-to-noise ratio for structure learning.

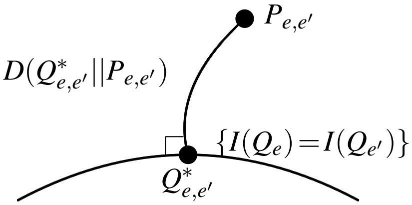
$$\mathrm{SNR} = \left(\frac{\text{mean}}{\text{stddev}}\right)^2.$$

The Crossover Rate V

Convexifying the optimization problem by linearizing the constraints.

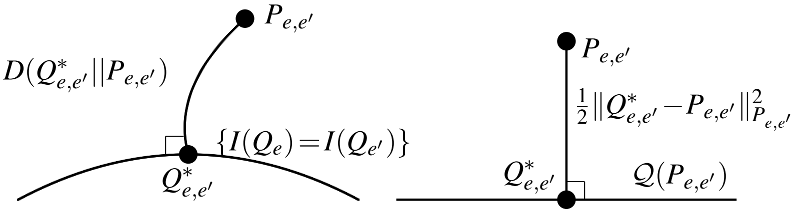
The Crossover Rate V

Convexifying the optimization problem by linearizing the constraints.



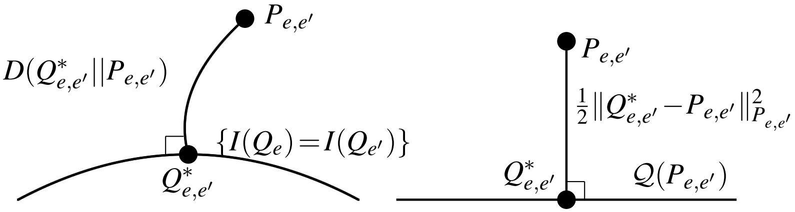
The Crossover Rate V

Convexifying the optimization problem by linearizing the constraints.



The Crossover Rate V

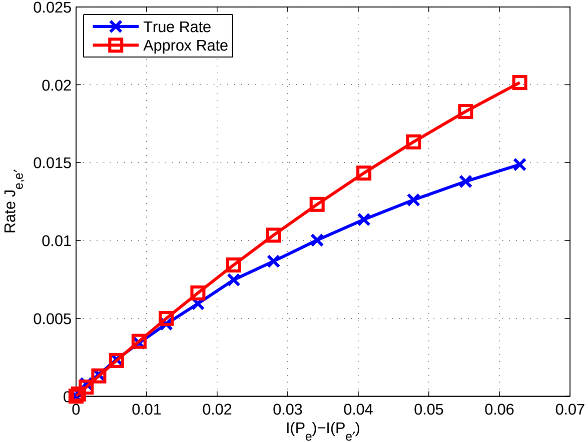
Convexifying the optimization problem by linearizing the constraints.



Non-Convex problem becomes a Least-Squares problem.

The Crossover Rate V

How good is the approximation? We consider a binary model.

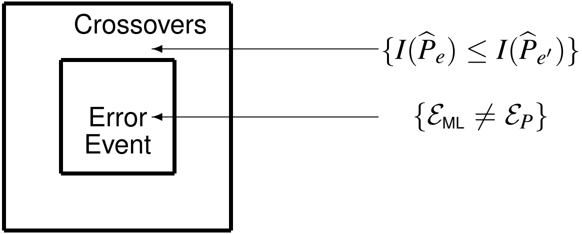


Error Exponent for Structure Learning I

- We have characterized the rate for the crossover event $\left\{ I(\widehat{P}_e) \leq I(\widehat{P}_{e'}) \right\}$. We called the rate $J_{e,e'}$.

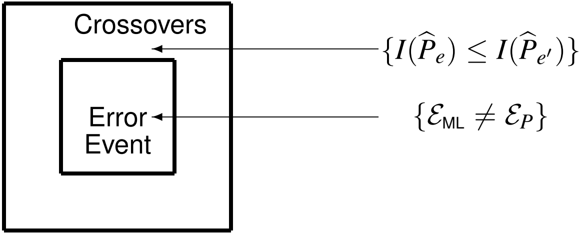
Error Exponent for Structure Learning I

- We have characterized the rate for the crossover event $\left\{ I(\widehat{P}_e) \leq I(\widehat{P}_{e'}) \right\}$. We called the rate $J_{e,e'}$.



Error Exponent for Structure Learning I

- We have characterized the rate for the crossover event $\left\{ I(\widehat{P}_e) \leq I(\widehat{P}_{e'}) \right\}$. We called the rate $J_{e,e'}$.



- How to relate this to K_P, the overall error exponent?

$$\mathbb{P}\left(\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}\right) \doteq \exp(-nK_P)$$

Error Exponent for Structure Learning II

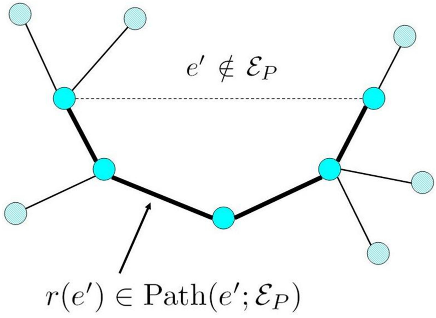
Theorem

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

Error Exponent for Structure Learning II

Theorem

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



Positivity of the Error Exponent

Theorem

The following statements are equivalent:

(a) *The error probability decays exponentially i.e.,*

$$K_P > 0.$$

(b) T_P *is a spanning tree, i.e., not a proper forest.*

Positivity of the Error Exponent

Theorem

The following statements are equivalent:

(a) *The error probability decays exponentially i.e.,*

$$K_P > 0.$$

(b) T_P *is a spanning tree, i.e., not a proper forest.*

Conclusions

- Goal: Find the rate of decay of the probability of error:

$$K_P := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left(\{\mathcal{E}_{\mathsf{ML}} \neq \mathcal{E}_P\}\right).$$

- Employed tools from Large-Deviation theory and basic properties of trees.

Conclusions

- Goal: Find the rate of decay of the probability of error:

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}\right).$$

- Employed tools from Large-Deviation theory and basic properties of trees.
- Found the dominant error tree that relates crossover events to overall error event.

Conclusions

- Goal: Find the rate of decay of the probability of error:

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{\mathcal{E}_{\mathsf{ML}} \neq \mathcal{E}_P\}\right).$$

- Employed tools from Large-Deviation theory and basic properties of trees.
- Found the dominant error tree that relates crossover events to overall error event.
- Used Euclidean Information Theory to obtain an intuitive signal-to-noise ratio expression for the crossover rate.

Conclusions

- Goal: Find the rate of decay of the probability of error:

$$K_P := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left(\{\mathcal{E}_{\mathsf{ML}} \neq \mathcal{E}_P\}\right).$$

- Employed tools from Large-Deviation theory and basic properties of trees.
- Found the dominant error tree that relates crossover events to overall error event.
- Used Euclidean Information Theory to obtain an intuitive signal-to-noise ratio expression for the crossover rate.
- More details available on arXiv (http://arxiv.org/abs/0905.0940).