Scaling Laws for Learning High-Dimensional Markov Forest Distributions

Vincent Tan†, Animashree Anandkumar‡, Alan S. Willsky†

† Stochastic Systems Group, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology
‡ Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science, University of California, Irvine.

Allerton Conference (Sep 29, 2010)
Learning tree-structured graphical models given i.i.d. samples is well-known. The Chow-Liu algorithm (1968) provides an efficient implementation of maximum-likelihood estimation.

\[ P(x) = P_1(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3) \]

\[ x_1, \ldots, x_n \overset{i.i.d.}{\sim} P \]
Learning **tree-structured** graphical models given i.i.d. samples is well-known.

\[ P(\mathbf{x}) = P_1(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3) \]

\[ \mathbf{x}_1, \ldots, \mathbf{x}_n \sim P^{i.i.d.} \]

The Chow-Liu algorithm (1968) provides an efficient implementation of **maximum-likelihood** estimation.
Learning tree-structured graphical models given i.i.d. samples is well-known.

\[ P(x) = P_1(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3) \]

\[ x_1, \ldots, x_n \overset{i.i.d.}{\sim} P \]

The Chow-Liu algorithm (1968) provides an efficient implementation of maximum-likelihood estimation.

What if we want a larger class of acyclic models?
Motivation: Prevent Overfitting

- High-dimensional setting.

- If the number of samples $n$ is significantly fewer than the number of dimensions $d$, i.e.,

$$n \ll d$$

learning forest-structured distributions may reduce overfitting [Liu, Lafferty and Wasserman, 2010].
Motivation: Prevent Overfitting

- High-dimensional setting.
- If the number of samples $n$ is significantly fewer than the number of dimensions $d$, i.e., $n \ll d$
- Learning forest-structured distributions may reduce overfitting [Liu, Lafferty and Wasserman, 2010].
- Strategy: Remove “weak” edges to prevent overfitting.
Motivation: Prevent Overfitting

- High-dimensional setting.

- If the number of samples $n$ is significantly fewer than the number of dimensions $d$, i.e.,

$$n \ll d$$

-learning forest-structured distributions may reduce overfitting [Liu, Lafferty and Wasserman, 2010].

- Strategy: Remove “weak” edges to prevent overfitting.

$\Rightarrow$ Reduce Num Params $\Rightarrow$
Motivation: Prevent Overfitting

- High-dimensional setting.
- If the number of samples $n$ is significantly fewer than the number of dimensions $d$, i.e.,

$$n \ll d$$

learning forest-structured distributions may reduce overfitting [Liu, Lafferty and Wasserman, 2010].

- Strategy: Remove “weak” edges to prevent overfitting.

$$\Rightarrow \text{Reduce Num Params} \Rightarrow$$
Natural Questions

For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests consistently?
For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests \textbf{consistently}?

What are the \textbf{rates of convergence} for a particular $P$?
Natural Questions

- For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests **consistently**?

- What are the **rates of convergence** for a particular $P$?
For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests consistently?

What are the rates of convergence for a particular $P$?
Natural Questions

- For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests consistently?

- What are the rates of convergence for a particular $P$?

- How can following parameters scale with one another in the high-dimensional setting?
  1. Number of samples $n$
Natural Questions

- For a fixed model $P \in \mathcal{P}(\mathcal{X}^d)$, are there any simple modifications to Chow-Liu to learn forests **consistently**?

- What are the **rates of convergence** for a particular $P$?

- How can following parameters **scale** with one another in the high-dimensional setting?

  1. Number of samples $n$
  2. Number of variables $d$
For a fixed model $P \in P(\mathcal{X}^{d})$, are there any simple modifications to Chow-Liu to learn forests **consistently**?

What are the **rates of convergence** for a particular $P$?

How can following parameters **scale** with one another in the high-dimensional setting?

1. Number of samples $n$
2. Number of variables $d$
3. Number of edges $k \leq d - 1$
Main Contributions

- Propose CLThres, a thresholding algorithm, for consistently learning forest-structured models.

Prove convergence rates ("moderate deviations") for a fixed discrete graphical model \( P \in \mathcal{P}(X_d) \).

Prove achievable scaling laws on \((n,d,k)\) for consistent recovery in high-dimensions. Roughly speaking, \( n > C_1 \log \frac{1}{\delta} (d - k) \), for all \( \delta > 0 \), is achievable.
Main Contributions

- Propose **CLThres**, a thresholding algorithm, for consistently learning forest-structured models.

- Prove convergence rates ("moderate deviations") for a fixed discrete graphical model $P \in \mathcal{P}(\mathcal{X}^d)$.

Roughly speaking, $n > C_1 \log\frac{1}{\delta} (d - k)$, $\forall \delta > 0$ is achievable.
Propose **CLThres**, a thresholding algorithm, for consistently learning forest-structured models.

Prove **convergence rates** ("moderate deviations") for a fixed discrete graphical model $P \in \mathcal{P}(\mathcal{X}^d)$.

Prove **achievable scaling laws** on $(n, d, k)$ for consistent recovery in high-dimensions. Roughly speaking,

$$n > C_1 \log^{1+\delta}(d - k), \quad \forall \delta > 0$$

is achievable.
Let $\mathcal{X}$ be a finite set and let $\mathcal{P}(\mathcal{X}^d)$ be the probability simplex over $\mathcal{X}^d$.

We say that $P \in \mathcal{P}(\mathcal{X}^d)$ is a forest-structured model if it factorizes as

$$P(x) = \prod_{i \in V} P(x_i) \prod_{(i,j) \in E_P} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

where $V = [1 : d]$ and $E_P \subset \binom{V}{2}$ and note $|E_P| \leq d - 1$. 

Given $n$ i.i.d. samples $\{x_1, \ldots, x_n\}$ drawn from $P$, output an estimate of the structure $\hat{E}_P$. 

Vincent Tan

Scaling Laws for Learning High-Dimensional Markov Forests
Let $\mathcal{X}$ be a finite set and let $\mathcal{P}(\mathcal{X}^d)$ be the probability simplex over $\mathcal{X}^d$.

We say that $P \in \mathcal{P}(\mathcal{X}^d)$ is a forest-structured model if it factorizes as

$$P(x) = \prod_{i \in V} P(x_i) \prod_{(i,j) \in E_P} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

where $V = [1 : d]$ and $E_P \subset \binom{V}{2}$ and note $|E_P| \leq d - 1$.

Given $n$ i.i.d. samples $\{x_1, \ldots, x_n\}$ drawn from $P$, a forest-structured model with edge set $E_P$. 
Let $\mathcal{X}$ be a finite set and let $\mathcal{P}(\mathcal{X}^d)$ be the probability simplex over $\mathcal{X}^d$.

We say that $P \in \mathcal{P}(\mathcal{X}^d)$ is a forest-structured model if it factorizes as

$$P(x) = \prod_{i \in V} P(x_i) \prod_{(i,j) \in E_P} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

where $V = [1 : d]$ and $E_P \subset \binom{V}{2}$ and note $|E_P| \leq d - 1$.

Given $n$ i.i.d. samples $\{x_1, \ldots, x_n\}$ drawn from $P$, a forest-structured model with edge set $E_P$.

Output an estimate of the structure $\hat{E}$. 
Main Difficulty

- Unknown minimum mutual information $I_{\text{min}}$ in the forest model.
- Unknown minimum mutual information $I_{\text{min}}$ in the forest model.
- Markov order estimation.
- Unknown minimum mutual information $I_{\text{min}}$ in the forest model.

- Markov order estimation.

- If known, can easily use a threshold, i.e,

  \[
  \text{if } \hat{I}(X_i; X_j) < I_{\text{min}}, \text{ remove } (i,j)
  \]
- Unknown minimum mutual information $I_{\min}$ in the forest model.
- Markov order estimation.
- If known, can easily use a threshold, i.e,
  \[
  \text{if } \hat{I}(X_i; X_j) < I_{\min}, \text{ remove } (i, j)
  \]
- How to deal with classic tradeoff between over- and underestimation errors?
Compute the set of empirical mutual information $\hat{I}(X_i; X_j)$ for all $(i,j) \in V \times V$. 

Max-weight spanning tree $\hat{E}_{d-1} := \arg\max_{E: \text{Tree}} \sum_{(i,j) \in E} \hat{I}(X_i; X_j)$

Estimate number of edges given threshold $\epsilon$

\[
\hat{k}_n := |\{(i,j) \in \hat{E}_{d-1} : \hat{I}(X_i; X_j) \geq \epsilon\}|
\]

Output the forest with the top $\hat{k}_n$ edges.

Computational Complexity $= O((n + \log d) d^2)$. 

Scaling Laws for Learning High-Dimensional Markov Forests
The CLThres Algorithm

- Compute the set of empirical mutual information $\hat{I}(X_i; X_j)$ for all $(i, j) \in V \times V$.
- Max-weight spanning tree

$$\hat{E}_{d-1} := \arg\max_{E: \text{Tree}} \sum_{(i, j) \in E} \hat{I}(X_i; X_j)$$
The CLThres Algorithm

- Compute the set of empirical mutual information $\hat{I}(X_i; X_j)$ for all $(i,j) \in V \times V$.

- Max-weight spanning tree

  $$\hat{E}_{d-1} := \arg\max_{E: \text{Tree}} \sum_{(i,j) \in E} \hat{I}(X_i; X_j)$$

- Estimate number of edges given threshold $\epsilon_n$

  $$\hat{k}_n := \left| \left\{ (i,j) \in \hat{E}_{d-1} : \hat{I}(X_i; X_j) \geq \epsilon_n \right\} \right|$$

Computational Complexity = $O((n + \log d) d^2)$.
The CLThres Algorithm

- Compute the set of empirical mutual information \( \hat{I}(X_i; X_j) \) for all \((i, j) \in V \times V\).

- Max-weight spanning tree

\[
\hat{E}_{d-1} := \arg\max_{E:Tree} \sum_{(i, j) \in E} \hat{I}(X_i; X_j)
\]

- Estimate number of edges given threshold \( \epsilon_n \)

\[
\hat{k}_n := \left| \left\{ (i, j) \in \hat{E}_{d-1} : \hat{I}(X_i; X_j) \geq \epsilon_n \right\} \right|
\]

- Output the forest with the top \( \hat{k}_n \) edges.

- Computational Complexity = \( O((n + \log d)d^2) \).
The CLThres Algorithm with $\epsilon_n = 1$
The CLThres Algorithm with $\epsilon_n = 1$

True MI $I(X_i; X_j)$

Empirical MI $\hat{I}(X_i; X_j)$
The CLThres Algorithm with $\epsilon_n = 1$

True MI $I(X_i; X_j)$

Empirical MI $\hat{I}(X_i; X_j)$

Max-weight spanning tree $\hat{E}_{d-1}$

Thresholded Forest $\hat{E}_{kn}$
We first assume that $P \in \mathcal{P}(\mathcal{X}^d)$ is a fixed distribution, i.e., $d$ does not grow with $n$. 

**Theorem ("Moderate Deviations")**

Assume that the sequence $\{\epsilon_n\}_{n=1}^{\infty}$ satisfies

$$\lim_{n \to \infty} \epsilon_n = 0, \quad \lim_{n \to \infty} n \epsilon_n \log n = \infty$$

Then

$$\limsup_{n \to \infty} \frac{1}{n} \epsilon_n \log P(\hat{E} \neq E_P) \leq -1$$

Roughly speaking, $P(\hat{E} \neq E_P) \approx \exp(-n \epsilon_n)$

Also have a "liminf" lower bound.
We first assume that $P \in \mathcal{P}(\mathcal{X}^d)$ is a fixed distribution, i.e., $d$ does not grow with $n$.

**Theorem ("Moderate Deviations")**

Assume that the sequence $\{\epsilon_n\}_{n=1}^{\infty}$ satisfies

\[
\lim_{n \to \infty} \epsilon_n = 0, \quad \lim_{n \to \infty} \frac{n\epsilon_n}{\log n} = \infty
\]

Roughly speaking, $P(\hat{E}_{\hat{k}n} \neq E_P) \approx \exp(-n\epsilon_n)$

Also have a "liminf" lower bound.
We first assume that $P \in \mathcal{P}(\mathcal{X}^d)$ is a fixed distribution, i.e., $d$ does not grow with $n$.

**Theorem ("Moderate Deviations")**

Assume that the sequence $\{\epsilon_n\}_{n=1}^{\infty}$ satisfies

$$
\lim_{n \to \infty} \epsilon_n = 0, \quad \lim_{n \to \infty} \frac{n\epsilon_n}{\log n} = \infty
$$

Then

$$
\limsup_{n \to \infty} \frac{1}{n\epsilon_n} \log \mathbb{P}(\hat{E}_{kn} \neq E_P) \leq -1
$$

Roughly speaking, $\mathbb{P}(\hat{E}_{kn} \neq E_P) \approx \exp(-n\epsilon_n)$

Also have a "liminf" lower bound.
A Convergence Result for CLThres

We first assume that $P \in \mathcal{P}(\mathcal{X}^d)$ is a fixed distribution, i.e., $d$ does not grow with $n$.

**Theorem ("Moderate Deviations")**

Assume that the sequence $\{\epsilon_n\}_{n=1}^\infty$ satisfies

$$\lim_{n \to \infty} \epsilon_n = 0, \quad \lim_{n \to \infty} \frac{n\epsilon_n}{\log n} = \infty$$

Then

$$\limsup_{n \to \infty} \frac{1}{n\epsilon_n} \log \mathbb{P}(\hat{E}_{kn} \neq E_P) \leq -1$$

Roughly speaking, $\mathbb{P}(\hat{E}_{kn} \neq E_P) \approx \exp(-n\epsilon_n)$

Also have a "liminf" lower bound.
Remarks: A Convergence Result for CLThres

- The Chow-Liu phase is consistent with exponential rate of convergence [Tan, Anandkumar, Tong and Willsky 2009].
The Chow-Liu phase is consistent with exponential rate of convergence [Tan, Anandkumar, Tong and Willsky 2009].

The sequence can be taken to be $\epsilon_n := n^{-\beta}$ for $\beta \in (0, 1)$. 
The Chow-Liu phase is consistent with \textit{exponential rate of convergence} [Tan, Anandkumar, Tong and Willsky 2009].

The sequence can be taken to be $\epsilon_n := n^{-\beta}$ for $\beta \in (0, 1)$.

For all $n$ sufficiently large,

\[ \epsilon_n < I_{\min} \]

implies no underestimation asymptotically.
The Chow-Liu phase is consistent with exponential rate of convergence [Tan, Anandkumar, Tong and Willsky 2009].

The sequence can be taken to be $\epsilon_n := n^{-\beta}$ for $\beta \in (0, 1)$.

For all $n$ sufficiently large,

$$\epsilon_n < I_{\min}$$

implies no underestimation asymptotically.

Note that for two independent random variables $X_i$ and $X_j$ with product pmf $Q_i \times Q_j$,

$$\text{std}(\hat{I}(X_i; X_j)) = \Theta(1/n)$$

Since the sequence $\epsilon_n = \omega(\log n/n)$ decays slower than $\text{std}(\hat{I}(X_i; X_j))$, no overestimation asymptotically.
Pruning Away Weak Empirical Mutual Informations

\[ I_{\min} (\text{unknown}) \]

Asymptotically, \( \epsilon_n \) will be smaller than \( I_{\min} \) and larger than \( \hat{I}(X_i; X_j) \) with high probability.
Pruning Away Weak Empirical Mutual Informations

Asymptotically, $\epsilon_n$ will be smaller than $I_{\min}$ and larger than $\hat{I}(X_i; X_j)$ with high probability.

\[ \hat{I}(X_i; X_j) \approx \frac{1}{n} \]
Asymptotically, $\epsilon_n$ will be smaller than $I_{\text{min}}$ and larger than $\hat{I}(X_i; X_j)$ with high probability.
Proof Idea

Based fully on the method of types [Csiszár and Körner].
Proof Idea

Based fully on the method of types [Csiszár and Körner].

- Estimate Chow-Liu learning error.
Proof Idea

Based fully on the method of types [Csiszár and Körner].

- Estimate Chow-Liu learning error.
- Estimate underestimation error.

\[
P(\hat{k}_n < k) = \exp(-nL_P).
\]
Proof Idea

Based fully on the method of types [Csiszár and Körner].

- Estimate Chow-Liu learning error.
- Estimate underestimation error.

\[ \mathbb{P}(\hat{k}_n < k) = \exp(-nL_P). \]

- Estimate overestimation error:

  This can be shown to decay subexponentially but faster than any polynomial:

\[ \mathbb{P}(\hat{k}_n > k) \approx \exp(-n\epsilon_n). \]

Upper bound has no dependence on \( P \).
Proof Idea

Based fully on the method of types [Csiszár and Körner].

- Estimate **Chow-Liu** learning error.
- Estimate **underestimation** error.

\[ P(\hat{k}_n < k) = \exp(-nL_P). \]

- Estimate **overestimation** error:

  This can be shown to decay subexponentially but faster than any polynomial:

\[ P(\hat{k}_n > k) \approx \exp(-n\epsilon_n). \]

Upper bound has no dependence on \( P \).

Additional Technique: Ideas from **Euclidean Information Theory** [Borade and Zheng 2008].
Consider a sequence of structure learning problems indexed by number of samples $n$. For each particular problem, we have data $x_n = \{x_i\}_{i=1}^n$. Each sample $x_i \in X_d$ is drawn independently from a forest-structured model with $d$ nodes and $k$ edges.

Sequence of tuples $\{ (n, d_n, k_n) \}_{n=1}^\infty$.

Assumptions:

- **(A1)** $I_{\inf} := \inf_{d \in \mathbb{N}} \min_{(i,j) \in E} I(P_i, j) > 0$
- **(A2)** $\kappa := \inf_{d \in \mathbb{N}} \min_{(x_i, x_j) \in X^2} P_i, j(x_i, x_j) > 0$
Consider a sequence of structure learning problems indexed by number of samples $n$. 
Consider a sequence of structure learning problems indexed by number of samples $n$.

For each particular problem, we have data $x^n = \{x_i\}_{i=1}^n$. 
Consider a sequence of structure learning problems indexed by number of samples $n$.

For each particular problem, we have data $x^n = \{x_i\}_{i=1}^n$.

Each sample $x_i \in \mathcal{X}^d$ is drawn independently from a forest-structured model with $d$ nodes and $k$ edges.
Consider a sequence of structure learning problems indexed by number of samples $n$.

For each particular problem, we have data $x^n = \{x_i\}_{i=1}^n$.

Each sample $x_i \in \mathcal{X}^d$ is drawn independently from a forest-structured model with $d$ nodes and $k$ edges.

Sequence of tuples $\{(n, d_n, k_n)\}_{n=1}^{\infty}$.
Consider a sequence of structure learning problems indexed by number of samples $n$.

For each particular problem, we have data $\mathbf{x}^n = \{x_i\}_{i=1}^n$.

Each sample $x_i \in \mathcal{X}^d$ is drawn independently from a forest-structured model with $d$ nodes and $k$ edges.

Sequence of tuples $\{(n, d_n, k_n)\}_{n=1}^\infty$.

Assumptions

(A1) $I_{\inf} := \inf_{d \in \mathbb{N}} \min_{(i,j) \in E_P} I(P_{i,j}) > 0$

(A2) $\kappa := \inf_{d \in \mathbb{N}} \min_{(x_i,x_j) \in \mathcal{X}^2} P_{i,j}(x_i, x_j) > 0$
Theorem ("Achievability")

Assume (A1) and (A2). Fix $\delta > 0$. Then if

$$n > \max \left\{ C_1 \log d, C_2 \log k, \right\}$$

the error probability of structure learning $P(error) \to 0$ as $(n, d, k) \to \infty$. 

An Achievable Scaling Law for CLThres
Theorem ("Achievability")

Assume (A1) and (A2). Fix $\delta > 0$. Then if

$$n > \max \left\{ C_1 \log d, C_2 \log k, (2 \log(d - k))^{1+\delta} \right\}$$

the error probability of structure learning $P(error) \to 0$ as $(n, d, k) \to \infty$. 
Theorem ("Achievability")

Assume (A1) and (A2). Fix $\delta > 0$. Then if

$$n > \max \left\{ C_1 \log d, C_2 \log k, \left(2 \log (d - k)\right)^{1+\delta}\right\}$$

the error probability of structure learning

$$\mathbb{P}(\text{error}) \rightarrow 0$$

as $(n, d_n, k_n) \rightarrow \infty$. 

Vincent Tan

Scaling Laws for Learning High-Dimensional Markov Forests
Remarks on the Achievable Scaling Law for CLThres

- If the model parameters \((n, d, k)\) grow with \(n\) but if

\[
\begin{align*}
  d & \text{ subexponential} \\
  k & \text{ subexponential} \\
  d - k & \text{ subexponential}
\end{align*}
\]

structure recovery is \textit{asymptotically possible}.
Remarks on the Achievable Scaling Law for CLThres

- If the model parameters \((n, d, k)\) grow with \(n\) but if
  
  \[
  \begin{align*}
  d & \text{ subexponential} \\
  k & \text{ subexponential} \\
  d - k & \text{ subexponential}
  \end{align*}
  \]

  structure recovery is asymptotically possible.

- \(d\) can grow much faster than \(n\).
Remarks on the Achievable Scaling Law for CLThres

- If the model parameters \((n, d, k)\) grow with \(n\) but if
  
  \[
  d \quad \text{subexponential} \\
  k \quad \text{subexponential} \\
  d - k \quad \text{subexponential}
  \]

  structure recovery is \textit{asymptotically possible}.

- \(d\) can grow \textbf{much faster} than \(n\).

- Close to the \textbf{strong converse} lower bound.
Remarks on the Achievable Scaling Law for CLThres

- If the model parameters \((n, d, k)\) grow with \(n\) but if
  \[
  d \quad \text{subexponential} \\
  k \quad \text{subexponential} \\
  d - k \quad \text{subexponential}
  \]
  structure recovery is asymptotically possible.

- \(d\) can grow much faster than \(n\).

- Close to the strong converse lower bound.

- Proof uses:
  1. Previous fixed \(d\) result.
  2. Exponents in the limsup upper bound do not vanish with increasing problem size as \((n, d_n, k_n) \to \infty\).
There exists a **tradeoff** between under- and overestimation in the finite-sample case:

But asymptotically, overestimation error **dominates**.

Design of $\epsilon_n := n^{-\beta}$ takes into account the tradeoff.
Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests **consistently**.
Concluding Remarks

- Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests **consistently**.

- Derived precise error rates in the form of a “**moderate deviations**” result.
Concluding Remarks

- Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests consistently.

- Derived precise error rates in the form of a “moderate deviations” result.

- Derived scaling laws on \((n, d, k)\) for structural consistency in high dimensions.

Extensions:
- Risk consistency has also been analyzed. See manuscript on arXiv.
- Need to find the right balance between over- and underestimation for the finite sample case.
Concluding Remarks

- Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests consistently.

- Derived precise error rates in the form of a “moderate deviations” result.

- Derived scaling laws on \((n, d, k)\) for structural consistency in high dimensions.

Extensions:

Concluding Remarks

- Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests consistently.
- Derived precise error rates in the form of a “moderate deviations” result.
- Derived scaling laws on \((n, d, k)\) for structural consistency in high dimensions.

Extensions:

- **Risk consistency** has also been analyzed. See manuscript on arXiv.
Concluding Remarks

- Proposed a simple extension of Chow-Liu’s max-weight spanning tree algorithm to learn forests consistently.

- Derived precise error rates in the form of a “moderate deviations” result.

- Derived scaling laws on \((n, d, k)\) for structural consistency in high dimensions.

Extensions:

- **Risk consistency** has also been analyzed. See manuscript on arXiv.

- Need to find the right balance between over- and underestimation for the finite sample case.