Estimating Latent Variable Graphical Models with Moments and Likelihoods

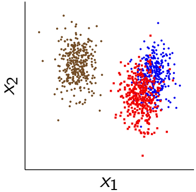
Arun Tejasvi Chaganty
Percy Liang

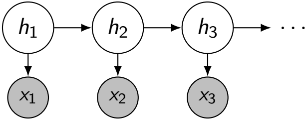
Stanford University

June 22, 2014

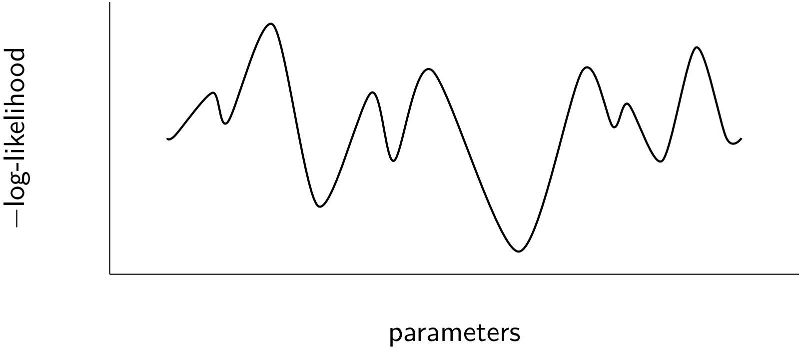
Latent Variable Graphical Models

- Gaussian Mixture Models
- Latent Dirichlet Allocation
- Hidden Markov Models
- PCFGs
- ...



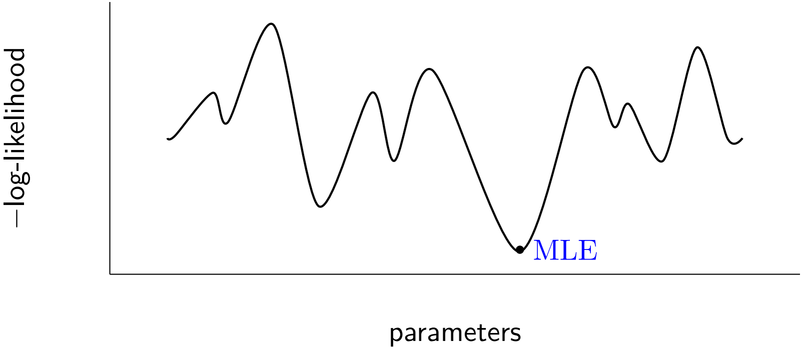


Parameter Estimation is Hard



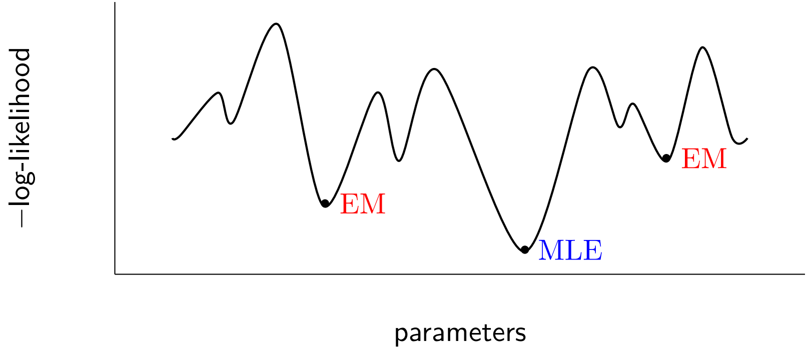
- Log-likelihood function is non-convex.

Parameter Estimation is Hard



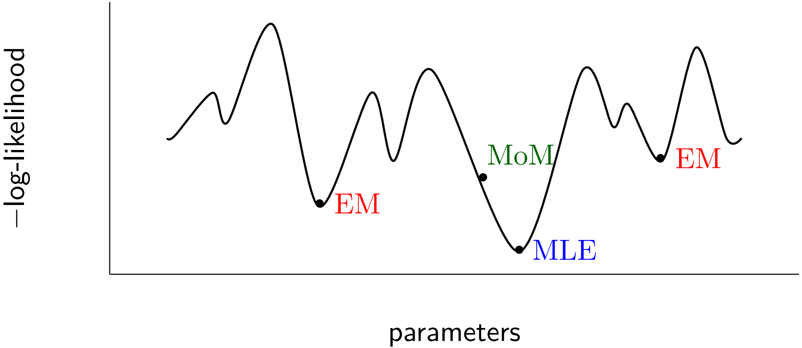
- Log-likelihood function is non-convex.
- MLE is consistent but intractable.

Parameter Estimation is Hard



- Log-likelihood function is non-convex.
- MLE is consistent but intractable.
- Local methods (EM, gradient descent, . . .) are tractable but inconsistent.

Parameter Estimation is Hard



- Log-likelihood function is non-convex.
- MLE is consistent but intractable.
- Local methods (EM, gradient descent, . . .) are tractable but inconsistent.
- *Method of moments* estimators can be consistent and computationally-efficient, but require more data.

Consistent estimation for general models

- Several estimators based on the method of moments.
  - **Phylogenetic trees:** Mossel and Roch 2005.
  - **Hidden Markov models:** Hsu, Kakade, and Zhang 2009
  - **Latent trees:** Anandkumar et al. 2011
  - **Latent Dirichlet Allocation:** Anandkumar et al. 2012
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors** Chaganty and Liang 2013
  - ...

Consistent estimation for general models

- Several estimators based on the method of moments.
  - **Phylogenetic trees:** Mossel and Roch 2005.
  - **Hidden Markov models:** Hsu, Kakade, and Zhang 2009
  - **Latent trees:** Anandkumar et al. 2011
  - **Latent Dirichlet Allocation:** Anandkumar et al. 2012
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors** Chaganty and Liang 2013
  - ...
- These estimators are applicable only to a specific type of model.

Consistent estimation for general models

- Several estimators based on the method of moments.
  - **Phylogenetic trees:** Mossel and Roch 2005.
  - **Hidden Markov models:** Hsu, Kakade, and Zhang 2009
  - **Latent trees:** Anandkumar et al. 2011
  - **Latent Dirichlet Allocation:** Anandkumar et al. 2012
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors** Chaganty and Liang 2013
  - ...
- These estimators are applicable only to a specific type of model.
- In contrast, EM and gradient descent apply for almost any model.

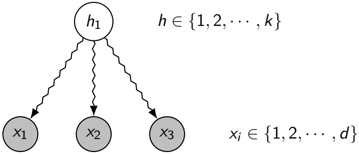
Consistent estimation for general models

- Several estimators based on the method of moments.
  - **Phylogenetic trees:** Mossel and Roch 2005.
  - **Hidden Markov models:** Hsu, Kakade, and Zhang 2009
  - **Latent trees:** Anandkumar et al. 2011
  - **Latent Dirichlet Allocation:** Anandkumar et al. 2012
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors** Chaganty and Liang 2013
  - ...
- These estimators are applicable only to a specific type of model.
- In contrast, EM and gradient descent apply for almost any model.
- Note: some work in the observable operator framework does apply to a more general model class.
  - **Weighted automata:** Balle and Mohri 2012.
  - **Junction trees:** Song, Xing, and Parikh 2011
  - ...

Consistent estimation for general models

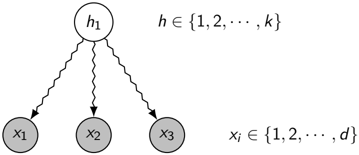
- Several estimators based on the method of moments.
  - **Phylogenetic trees:** Mossel and Roch 2005.
  - **Hidden Markov models:** Hsu, Kakade, and Zhang 2009
  - **Latent trees:** Anandkumar et al. 2011
  - **Latent Dirichlet Allocation:** Anandkumar et al. 2012
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors** Chaganty and Liang 2013
  - ...
- These estimators are applicable only to a specific type of model.
- In contrast, EM and gradient descent apply for almost any model.
- Note: some work in the observable operator framework does apply to a more general model class.
  - **Weighted automata:** Balle and Mohri 2012.
  - **Junction trees:** Song, Xing, and Parikh 2011
  - ...
- **How can we apply the method of moments to estimate *parameters efficiently* for a *general* model?**

Setup



- Discrete models with k hidden and $d \geq k$ observed values.

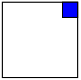
Setup



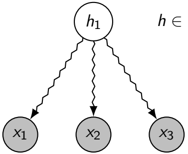
- Discrete models with k hidden and $d \geq k$ observed values.
- Parameters and marginals can be represented as matrices and tensors.

$$M_{12} \triangleq \mathbb{P}(x_1, x_2)$$

$$(M_{12})_{ij} \triangleq \mathbb{P}(x_1 = i, x_2 = j)$$



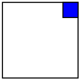
Setup



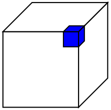
$h \in \{1, 2, \cdots, k\}$

$x_i \in \{1, 2, \cdots, d\}$

- Discrete models with k hidden and $d \geq k$ observed values.
- Parameters and marginals can be represented as matrices and tensors.



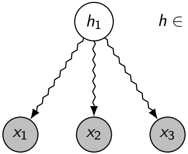
$$M_{12} \triangleq \mathbb{P}(x_1, x_2)$$
$$(M_{12})_{ij} \triangleq \mathbb{P}(x_1 = i, x_2 = j)$$



$$M_{123} \triangleq \mathbb{P}(x_1, x_2, x_3)$$
$$(M_{123})_{ijk} \triangleq \mathbb{P}(x_1 = i, x_2 = j, x_3 = k)$$

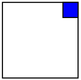
Setup

- Discrete models with k hidden and $d \geq k$ observed values.
- Parameters and marginals can be represented as matrices and tensors.

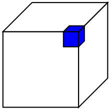


$h \in \{1, 2, \cdots, k\}$

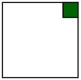
$x_i \in \{1, 2, \cdots, d\}$



$$M_{12} \triangleq \mathbb{P}(x_1, x_2)$$
$$(M_{12})_{ij} \triangleq \mathbb{P}(x_1 = i, x_2 = j)$$



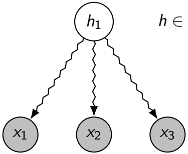
$$M_{123} \triangleq \mathbb{P}(x_1, x_2, x_3)$$
$$(M_{123})_{ijk} \triangleq \mathbb{P}(x_1 = i, x_2 = j, x_3 = k)$$



$$O^{(1|1)} \triangleq \mathbb{P}(x_1 \mid h_1)$$
$$(O^{(1|1)})_{ij} \triangleq \mathbb{P}(x_1 = i \mid h_1 = j)$$

Setup

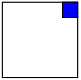
- Discrete models with k hidden and $d \geq k$ observed values.
- Parameters and marginals can be represented as matrices and tensors.
- Presented in terms of infinite data and exact moments.

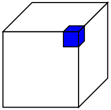


$h \in \{1, 2, \cdots, k\}$

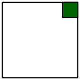
$x_i \in \{1, 2, \cdots, d\}$

$$M_{12} \triangleq \mathbb{P}(x_1, x_2)$$
$$(M_{12})_{ij} \triangleq \mathbb{P}(x_1 = i, x_2 = j)$$





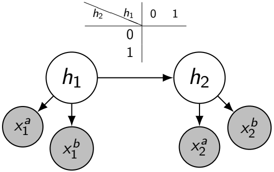
$$M_{123} \triangleq \mathbb{P}(x_1, x_2, x_3)$$
$$(M_{123})_{ijk} \triangleq \mathbb{P}(x_1 = i, x_2 = j, x_3 = k)$$



$$O^{(1|1)} \triangleq \mathbb{P}(x_1 \mid h_1)$$
$$(O^{(1|1)})_{ij} \triangleq \mathbb{P}(x_1 = i \mid h_1 = j)$$

Setup

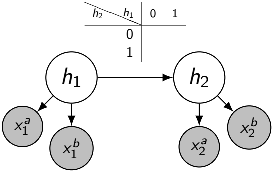
- Directed models parameterized by conditional probability tables.

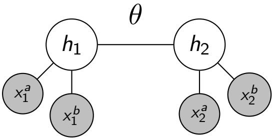


θ

Setup

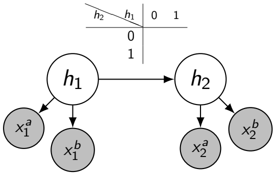
- Directed models parameterized by conditional probability tables.
- Undirected models parameterized as a log-linear model. Identify modulo $A(\theta)$.

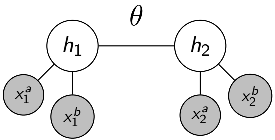




Setup

- Directed models parameterized by conditional probability tables.
- Undirected models parameterized as a log-linear model. Identify modulo $A(\theta)$.
- Focus on directed models, but return to undirected models later.

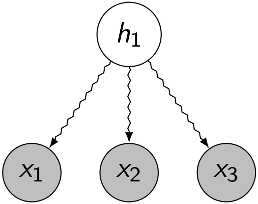




Background: Three-view mixture models aka bottlenecks

Definition (Bottleneck)

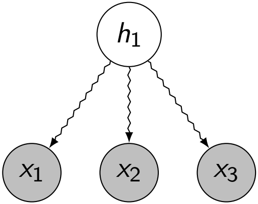
A hidden variable h is a **bottleneck** if there exist three observed variables (**views**) x_1, x_2, x_3 that are *conditionally independent* given h.



Background: Three-view mixture models aka bottlenecks

Definition (Bottleneck)

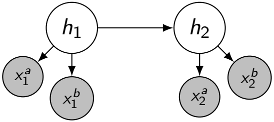
A hidden variable h is a **bottleneck** if there exist three observed variables (**views**) x_1, x_2, x_3 that are *conditionally independent* given h.



- Anandkumar et al. 2013 provide an algorithm to estimate conditional moments $O^{(i|1)} \triangleq \mathbb{P}(x_i \mid h_1)$ based on tensor eigendecomposition.
- In general, three views are necessary for identifiability (Kruskal 1977).

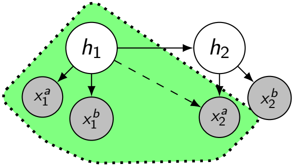
Example: a bridge, take I

- Each edge has a set of parameters.



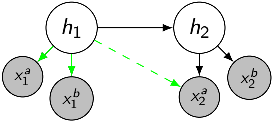
Example: a bridge, take I

- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.



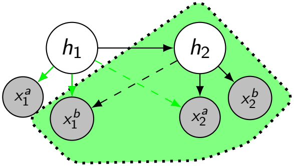
Example: a bridge, take I

- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.
- We can learn $\mathbb{P}(x_1^a | h_1), \mathbb{P}(x_1^b | h_1), \cdots$.



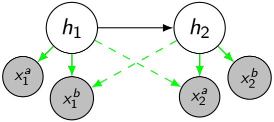
Example: a bridge, take I

- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.
- We can learn $\mathbb{P}(x_1^a | h_1), \mathbb{P}(x_1^b | h_1), \cdots$.



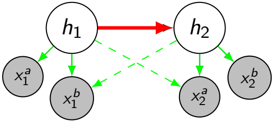
Example: a bridge, take I

- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.
- We can learn $\mathbb{P}(x_1^a | h_1), \mathbb{P}(x_1^b | h_1), \cdots$.

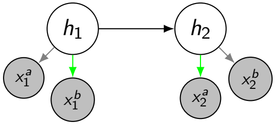


Example: a bridge, take I

- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.
- We can learn $\mathbb{P}(x_1^a|h_1), \mathbb{P}(x_1^b|h_1), \cdots$.
- However, we can't learn $\mathbb{P}(h_2|h_1)$ this way.



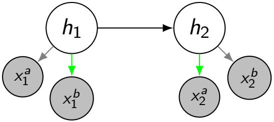
Example: a bridge, take II



- Observe the joint distribution of x_1, x_2,

$$\underbrace{\mathbb{P}(x_1^b, x_2^a)}_{M_{12}} = \sum_{h_1,h_2} \underbrace{\mathbb{P}(x_1^b \mid h_1)}_{O^{(1|1)}} \underbrace{\mathbb{P}(x_2^a \mid h_2)}_{O^{(2|2)}} \underbrace{\mathbb{P}(h_1, h_2)}_{Z_{12}}.$$

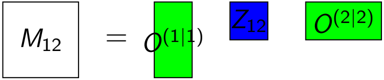
Example: a bridge, take II



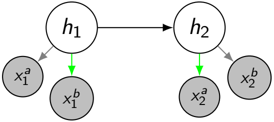
- Observe the joint distribution of x_1, x_2,

$$\underbrace{\mathbb{P}(x_1^b, x_2^a)}_{M_{12}} = \sum_{h_1, h_2} \underbrace{\mathbb{P}(x_1^b \mid h_1)}_{O^{(1|1)}} \underbrace{\mathbb{P}(x_2^a \mid h_2)}_{O^{(2|2)}} \underbrace{\mathbb{P}(h_1, h_2)}_{Z_{12}}.$$

- **Observed moments** $\mathbb{P}(x_1^b, x_2^a)$ are *linear* in the **hidden marginals** $\mathbb{P}(h_1, h_2)$.



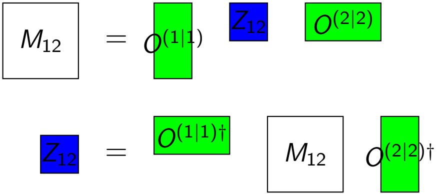
Example: a bridge, take II



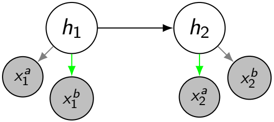
- Observe the joint distribution of x_1, x_2,

$$\underbrace{\mathbb{P}(x_1^b, x_2^a)}_{M_{12}} = \sum_{h_1, h_2} \underbrace{\mathbb{P}(x_1^b \mid h_1)}_{O^{(1|1)}} \underbrace{\mathbb{P}(x_2^a \mid h_2)}_{O^{(2|2)}} \underbrace{\mathbb{P}(h_1, h_2)}_{Z_{12}}.$$

- **Observed moments** $\mathbb{P}(x_1^b, x_2^a)$ are *linear* in the **hidden marginals** $\mathbb{P}(h_1, h_2)$.
- Solve for $\mathbb{P}(h_1, h_2)$ by pseudoinversion.



Example: a bridge, take II



- Observe the joint distribution of x_1, x_2,

$$\underbrace{\mathbb{P}(x_1^b, x_2^a)}_{M_{12}} = \sum_{h_1,h_2} \underbrace{\mathbb{P}(x_1^b \mid h_1)}_{O^{(1|1)}} \underbrace{\mathbb{P}(x_2^a \mid h_2)}_{O^{(2|2)}} \underbrace{\mathbb{P}(h_1, h_2)}_{Z_{12}}.$$

- **Observed moments** $\mathbb{P}(x_1^b, x_2^a)$ are *linear* in the **hidden marginals** $\mathbb{P}(h_1, h_2)$.
- Solve for $\mathbb{P}(h_1, h_2)$ by pseudoinversion.
- Normalize for $\mathbb{P}(h_2 \mid h_1)$.

$$M_{12} = O^{(1|1)} \; Z_{12} \; O^{(2|2)}$$

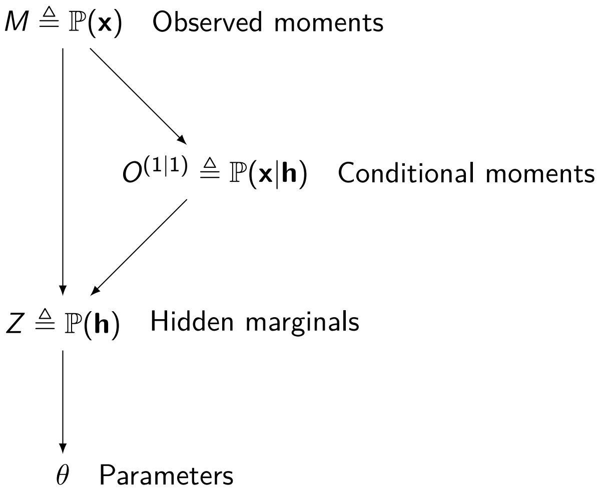
$$Z_{12} = O^{(1|1)\dagger} \; M_{12} \; O^{(2|2)\dagger}$$

Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

θ Parameters

Outline


$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments
$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments
$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals
θ Parameters

Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

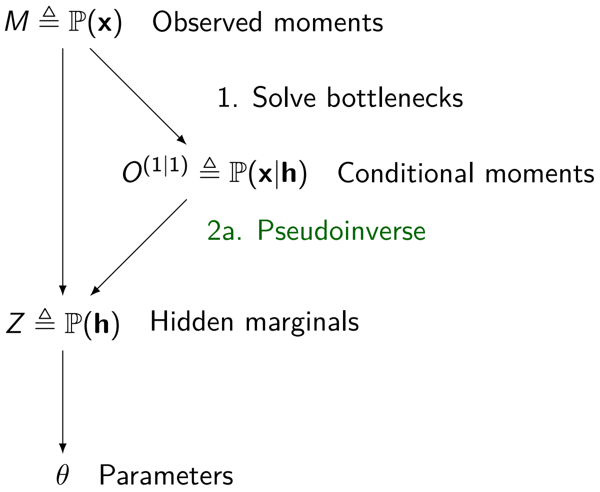
1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

θ Parameters

Outline

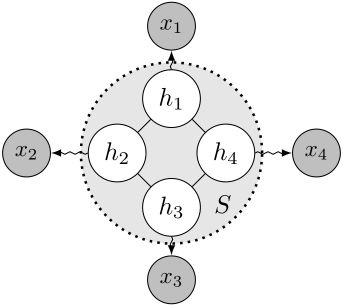

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments
1. Solve bottlenecks
$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments
2a. Pseudoinverse
$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals
θ Parameters

Exclusive Views

Definition (Exclusive views)

We say $h_i \in S \subseteq \mathbf{h}$ has an **exclusive view** x_v if

1. There exists *some observed variable* x_v which is *conditionally independent of the others* $(S \backslash \{h_i\})$ given h_i.

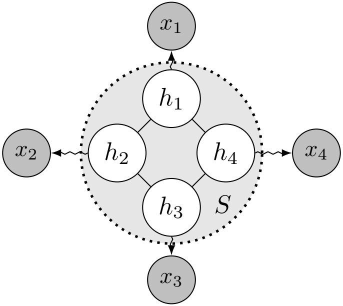


Exclusive Views

Definition (Exclusive views)

We say $h_i \in S \subseteq \mathbf{h}$ has an **exclusive view** x_v if

1. There exists *some observed variable* x_v which is *conditionally independent of the others* $(S \backslash \{h_i\})$ given h_i.
2. The conditional moment matrix $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$ has full column rank k and can be recovered.

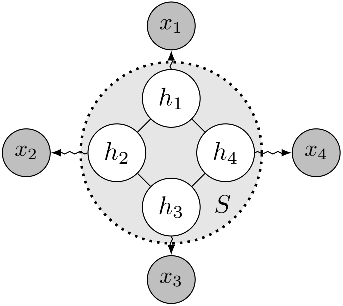


Exclusive Views

Definition (Exclusive views)

We say $h_i \in S \subseteq \mathbf{h}$ has an **exclusive view** x_v if

1. There exists *some observed variable* x_v which is *conditionally independent of the others* $(S \backslash \{h_i\})$ given h_i.
2. The conditional moment matrix $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$ has full column rank k and can be recovered.
3. A set has exclusive views if each $h_i \in S$ has an exclusive view.



Exclusive views give parameters

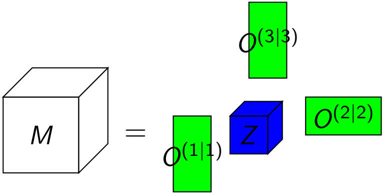
- Given *exclusive views*, $\mathbb{P}(x \mid h)$, learning cliques is solving a linear equation!

$$\underbrace{\mathbb{P}(x_1, \dots, x_m)}_{M} = \sum_{h_1,\dots,h_m} \underbrace{P(x_1|h_1)}_{O^{(1|1)}} \cdots \underbrace{P(h_1, \cdots, h_m)}_{Z}$$

Exclusive views give parameters

- Given *exclusive views*, $\mathbb{P}(x \mid h)$, learning cliques is solving a linear equation!

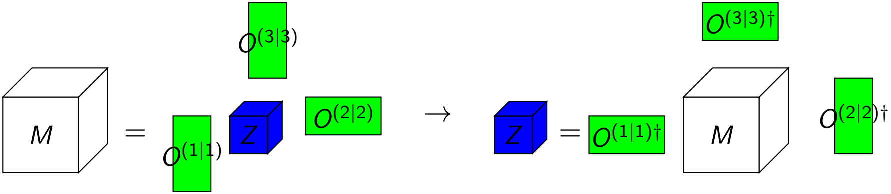
$$\underbrace{\mathbb{P}(x_1, \dots, x_m)}_{M} = \sum_{h_1,\dots,h_m} \underbrace{P(x_1|h_1)}_{O^{(1|1)}} \cdots \underbrace{P(h_1, \cdots, h_m)}_{Z}$$



Exclusive views give parameters

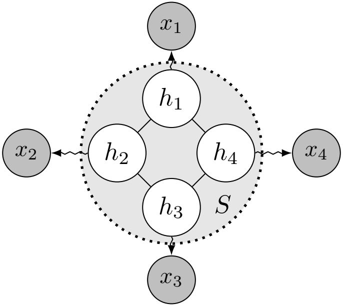
- Given *exclusive views*, $\mathbb{P}(x \mid h)$, learning cliques is solving a linear equation!

$$\underbrace{\mathbb{P}(x_1, \dots, x_m)}_{M} = \sum_{h_1,\dots,h_m} \underbrace{P(x_1|h_1)}_{O^{(1|1)}} \cdots \underbrace{P(h_1, \cdots, h_m)}_{Z}$$



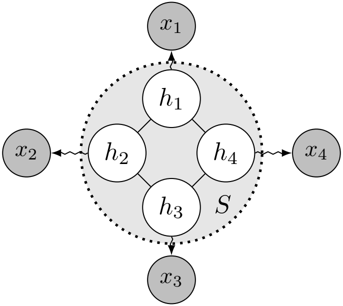
Bottlenecked graphs

- When are we assured exclusive views?



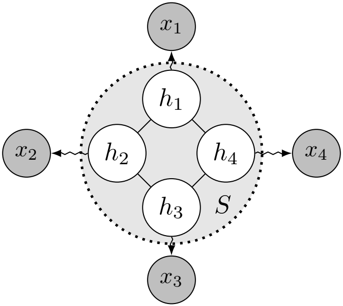
Bottlenecked graphs

- When are we assured exclusive views?
- **Theorem:** A clique in which **each hidden variable is a bottleneck** has exclusive views.



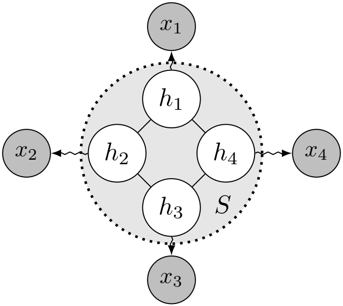
Bottlenecked graphs

- When are we assured exclusive views?
- **Theorem:** A clique in which **each hidden variable is a bottleneck** has exclusive views.
    - Follows by graph independence conditions.

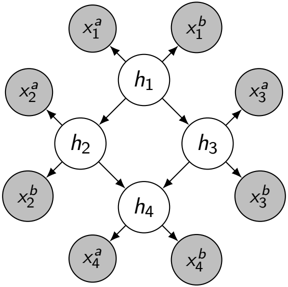


Bottlenecked graphs

- When are we assured exclusive views?
- **Theorem:** A clique in which **each hidden variable is a bottleneck** has exclusive views.
  - Follows by graph independence conditions.
  - We say that the clique is "bottlenecked".

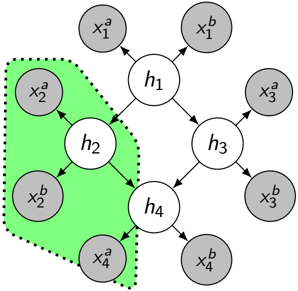


Example

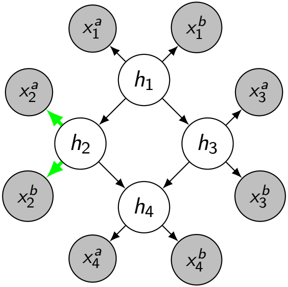


Example

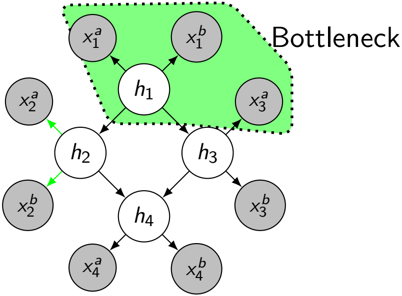
Bottleneck



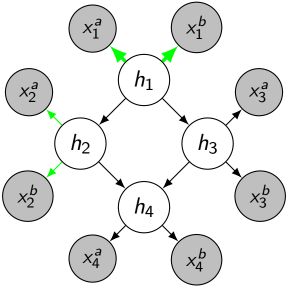
Example



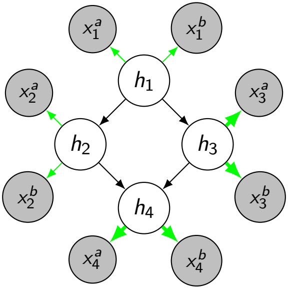
Example



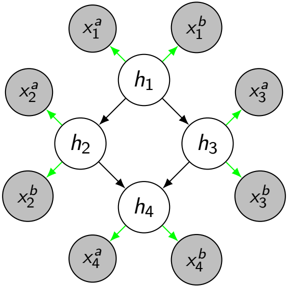
Example



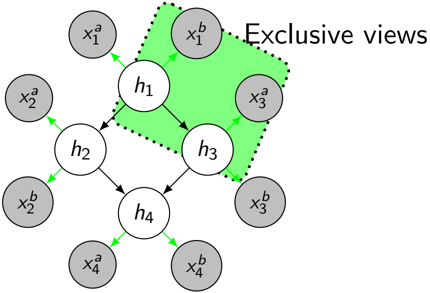
Example



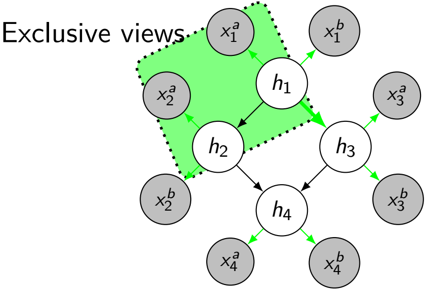
Example



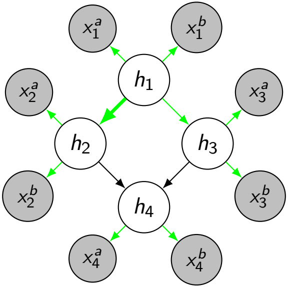
Example



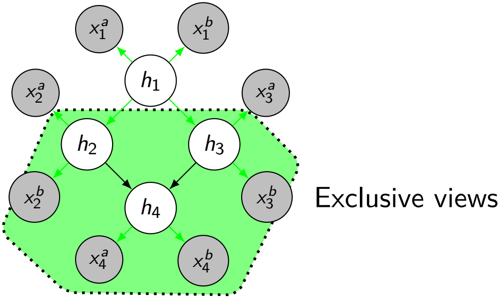
Example



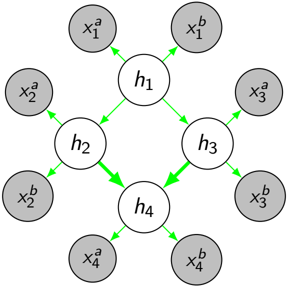
Example



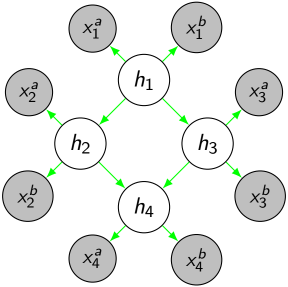
Example



Example

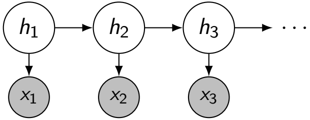


Example

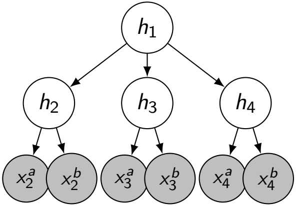


More Bottlenecked Examples

Hidden Markov models

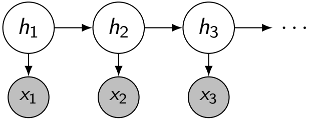


Latent Tree models

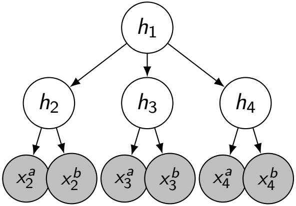


More Bottlenecked Examples

Hidden Markov models

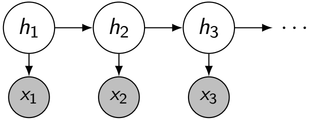


Latent Tree models

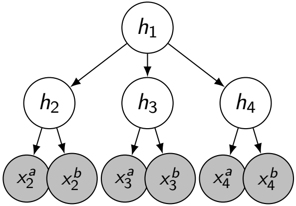


More Bottlenecked Examples

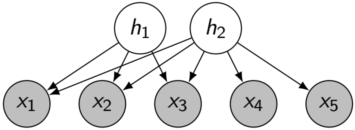
Hidden Markov models



Latent Tree models



Noisy Or (non-example) (Halpern and Sontag 2013)



Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

θ Parameters

Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse

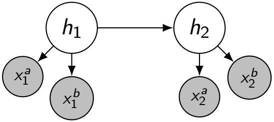
2b. Composite likelihood

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

θ Parameters

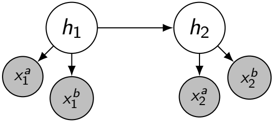
Convex marginal likelihoods

- The MLE is statistically most efficient, but usually non-convex.



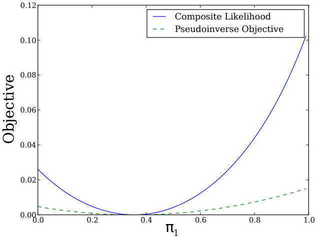
$$\log \mathbb{P}(\mathbf{x}) = \log \sum_{h_1, h_2} \mathbb{P}(\mathbf{x}_1 | h_1)\, \mathbb{P}(\mathbf{x}_2 | h_2) \mathbb{P}(h_1, h_2)$$

Convex marginal likelihoods



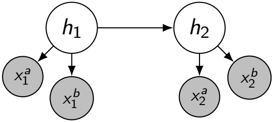
- The MLE is statistically most efficient, but usually non-convex.
- If we fix the conditional moments, $-\log \mathbb{P}(x)$ is convex in θ.

$$\log \mathbb{P}(\mathbf{x}) = \log \sum_{h_1,h_2} \underbrace{\mathbb{P}(\mathbf{x}_1|h_1)\,\mathbb{P}(\mathbf{x}_2|h_2)}_{\text{known}} \mathbb{P}(h_1, h_2)$$

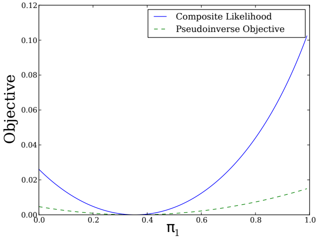


Convex marginal likelihoods

- The MLE is statistically most efficient, but usually non-convex.
- If we fix the conditional moments, $-\log \mathbb{P}(x)$ is convex in θ.
- No closed form solution, but a local method like EM is guaranteed to converge to the global optimum.

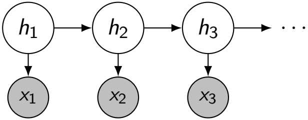


$$\log \mathbb{P}(\mathbf{x}) = \log \sum_{h_1,h_2} \underbrace{\mathbb{P}(\mathbf{x}_1|h_1)\,\mathbb{P}(\mathbf{x}_2|h_2)}_{\text{known}}\mathbb{P}(h_1,h_2)$$



Composite likelihoods

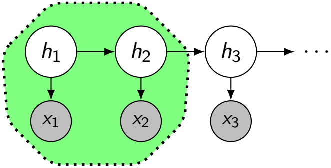
- In general, the full likelihood is still non-convex.



$$\log \mathbb{P}(\mathbf{x}_{123}) = \log \sum_{h_1, h_2, h_3} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\, \mathbb{P}(\mathbf{x}_2 \mid h_2) \mathbb{P}(\mathbf{x}_3 \mid h_3)}_{\text{known}}$$

$$\mathbb{P}(h_3 \mid h_2)\mathbb{P}(h_1, h_2)$$

Composite likelihoods



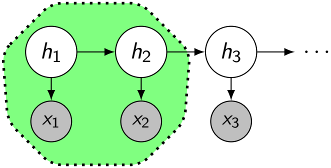
- In general, the full likelihood is still non-convex.
- Consider *composite likelihood* on a subset of observed variables.

$$\log \mathbb{P}(\mathbf{x}_{123}) = \log \sum_{h_1, h_2, h_3} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\, \mathbb{P}(\mathbf{x}_2 \mid h_2) \mathbb{P}(\mathbf{x}_3 \mid h_3)}_{\text{known}}$$

$$\mathbb{P}(h_3 \mid h_2)\mathbb{P}(h_1, h_2)$$

Composite likelihoods

- In general, the full likelihood is still non-convex.
- Consider *composite likelihood* on a subset of observed variables.

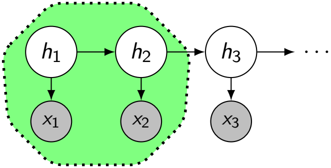


$$\log \mathbb{P}(\mathbf{x}_{12}) = \log \sum_{h_1, h_2} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\, \mathbb{P}(\mathbf{x}_2 \mid h_2)}_{\text{known}}$$

$\mathbb{P}(h_1, h_2)$

Composite likelihoods

- In general, the full likelihood is still non-convex.
- Consider *composite likelihood* on a subset of observed variables.
- Can be shown that estimation with composite likelihoods is consistent (Lindsay 1988).

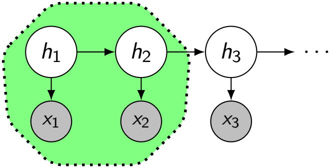


$$\log \mathbb{P}(\mathbf{x}_{12}) = \log \sum_{h_1, h_2} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\,\mathbb{P}(\mathbf{x}_2 \mid h_2)}_{\text{known}}$$

$\mathbb{P}(h_1, h_2)$

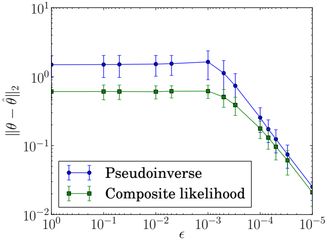
Composite likelihoods

- In general, the full likelihood is still non-convex.
- Consider *composite likelihood* on a subset of observed variables.
- Can be shown that estimation with composite likelihoods is consistent (Lindsay 1988).
- Asymptotically, the composite likelihood estimator is more efficient.



$$\log \mathbb{P}(\mathbf{x}_{12}) = \log \sum_{h_1,h_2} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\,\mathbb{P}(\mathbf{x}_2 \mid h_2)}_{\text{known}}$$

$\mathbb{P}(h_1, h_2)$



Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse

2b. Composite likelihood

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

θ Parameters

Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse
2b. Composite likelihood

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

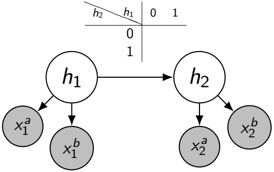
3a. Renormalization
3b. Convex optimization

θ Parameters

Recovering parameters in directed models

- Conditional probability tables are the default for a directed model.
- Can be recovered by normalization:

$$\mathbb{P}(h_2 \mid h_1) = \frac{\mathbb{P}(h_1, h_2)}{\sum_{h_2} \mathbb{P}(h_1, h_2)}.$$

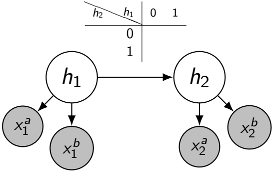


Recovering parameters in directed models

- Conditional probability tables are the default for a directed model.
- Can be recovered by normalization:

$$\mathbb{P}(h_2 \mid h_1) = \frac{\mathbb{P}(h_1, h_2)}{\sum_{h_2} \mathbb{P}(h_1, h_2)}.$$

- No dependence on tree-width. Memory, computation and samples depend linearly on the size of each clique.



Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\sum_{\mathcal{C} \in \mathcal{G}} \theta^\top \phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}) - A(\theta)\right).$$

Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\sum_{\mathcal{C} \in \mathcal{G}} \theta^\top \phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}) - A(\theta) \right).$$

- The *unsupervised* negative log-likelihood is non-convex,

$$\mathcal{L}_{\text{unsup}}(\theta) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[- \log \sum_{\mathbf{h} \in \mathcal{H}} p_\theta(\mathbf{x}, \mathbf{h})].$$

Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\sum_{\mathcal{C}\in\mathcal{G}} \theta^\top \phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}) - A(\theta)\right).$$

- The *unsupervised* negative log-likelihood is non-convex,

$$\mathcal{L}_{\text{unsup}}(\theta) \triangleq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[-\log \sum_{\mathbf{h}\in\mathcal{H}} p_\theta(\mathbf{x}, \mathbf{h})].$$

- However, the *supervised* negative log-likelihood is convex,

$$\begin{aligned}\mathcal{L}_{\text{sup}}(\theta) &\triangleq \mathbb{E}_{(\mathbf{x},\mathbf{h})\sim\mathcal{D}_{\text{sup}}}\left[-\log p_\theta(\mathbf{x}, \mathbf{h})\right] \\ &= -\theta^\top \left(\sum_{\mathcal{C}\in\mathcal{G}} \mathbb{E}_{(\mathbf{x},\mathbf{h})\sim\mathcal{D}_{\text{sup}}}[\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})]\right) + A(\theta).\end{aligned}$$

Recovering parameters in undirected log-linear models

- Recall, the marginals can typically estimated from supervised data.

$$\mathcal{L}_{\text{sup}}(\theta) = -\theta^\top \underbrace{\left(\sum_{\mathcal{C} \in \mathcal{G}} \mathbb{E}_{(\mathbf{x},\mathbf{h}) \sim \mathcal{D}_{\text{sup}}} [\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})] \right)}_{\mu_\mathcal{C}} + A(\theta).$$

Recovering parameters in undirected log-linear models

- Recall, the marginals can typically estimated from supervised data.

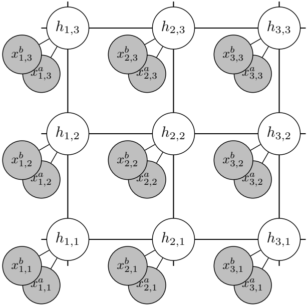
$$\mathcal{L}_{\text{sup}}(\theta) = -\theta^\top \underbrace{\left(\sum_{\mathcal{C} \in \mathcal{G}} \mathbb{E}_{(\mathbf{x},\mathbf{h}) \sim \mathcal{D}_{\text{sup}}} [\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})] \right)}_{\mu_\mathcal{C}} + A(\theta).$$

- However, the marginals can also be *consistently* estimated by moments!

$$\mu_\mathcal{C} = \sum_{\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}} \underbrace{\mathbb{P}(\mathbf{x}_\mathcal{C} \mid \mathbf{h}_\mathcal{C})}_{\text{cond. moments}} \underbrace{\mathbb{P}(\mathbf{h}_\mathcal{C})}_{\text{hidden marginals}} \phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}).$$

Optimizing pseudolikelihood

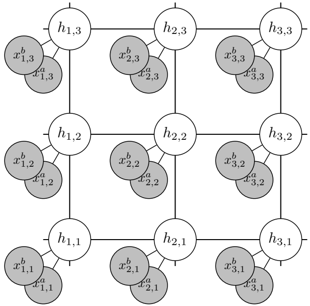
- Estimating $\mu_{\mathcal{C}}$: independent of treewidth.



Optimizing pseudolikelihood

- Estimating $\mu_{\mathcal{C}}$: independent of treewidth.
- Computing $A(\theta)$: dependent on treewidth.

$$A(\theta) \triangleq \log \sum_{\mathbf{x},\mathbf{h}} \exp\left(\theta^\top \phi(\mathbf{x},\mathbf{h})\right).$$

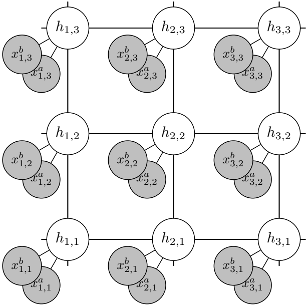


Optimizing pseudolikelihood

- Estimating $\mu_{\mathcal{C}}$: independent of treewidth.
- Computing $A(\theta)$: dependent on treewidth.

$$A(\theta) \triangleq \log \sum_{\mathbf{x},\mathbf{h}} \exp\left(\theta^\top \phi(\mathbf{x}, \mathbf{h})\right).$$

- Instead, use pseudolikelihood (Besag 1975) to consistently estimate distributions over local neighborhoods.



$$A_{\text{pseudo}}(\theta; \mathcal{N}(a)) \triangleq \log \sum_{a} \exp\left(\theta^\top \phi(\mathbf{x}_{\mathcal{N}}, \mathbf{h}_{\mathcal{N}})\right).$$

Outline

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse
2b. Composite likelihood

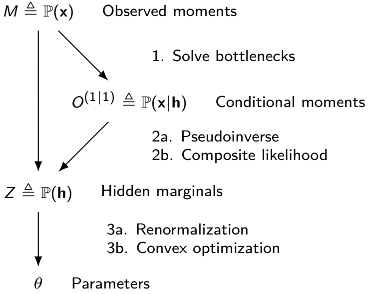
$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

3a. Renormalization
3b. Convex optimization

θ Parameters

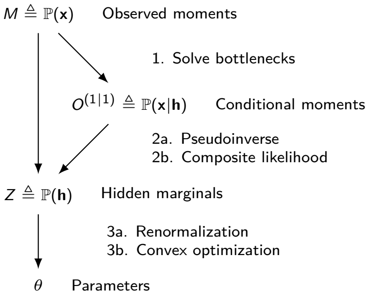
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.



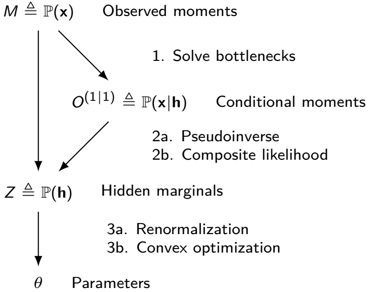
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.



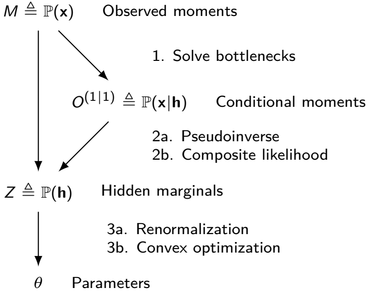
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.



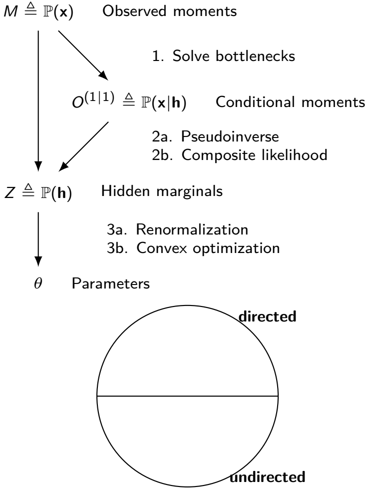
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.



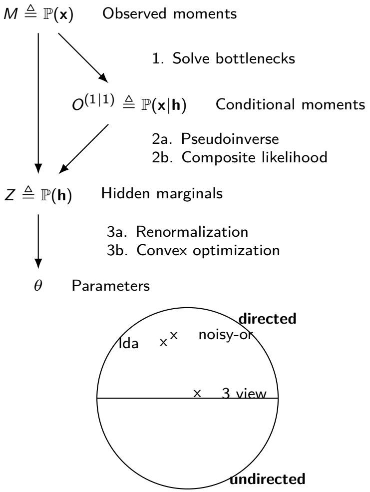
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.



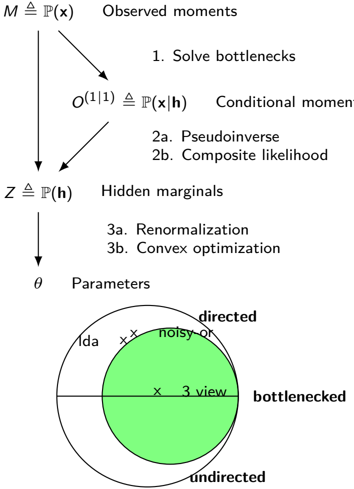
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.



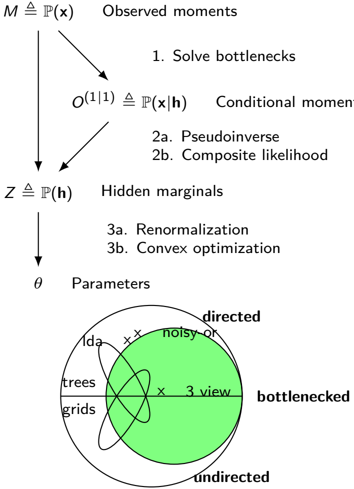
Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.



Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.



Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.

$M \triangleq \mathbb{P}(\mathbf{x})$ Observed moments

1. Solve bottlenecks

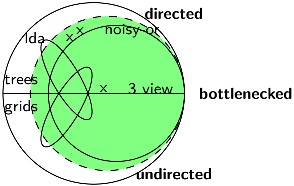
$O^{(1|1)} \triangleq \mathbb{P}(\mathbf{x}|\mathbf{h})$ Conditional moments

2a. Pseudoinverse
2b. Composite likelihood

$Z \triangleq \mathbb{P}(\mathbf{h})$ Hidden marginals

3a. Renormalization
3b. Convex optimization

θ Parameters



Conclusions

- An algorithm for any **bottlenecked discrete graphical model**.
- Combine moment estimators with likelihood estimators.
- Extends to **log-linear models**.
- Efficiently learns models with **high-treewidth**.
- **Thank you! Poster: M58**