Estimating Latent Variable Graphical Models with Moments and Likelihoods

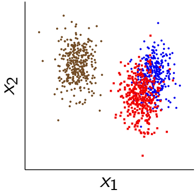
Arun Tejasvi Chaganty
Percy Liang

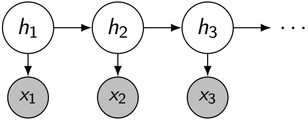
Stanford University

June 18, 2014

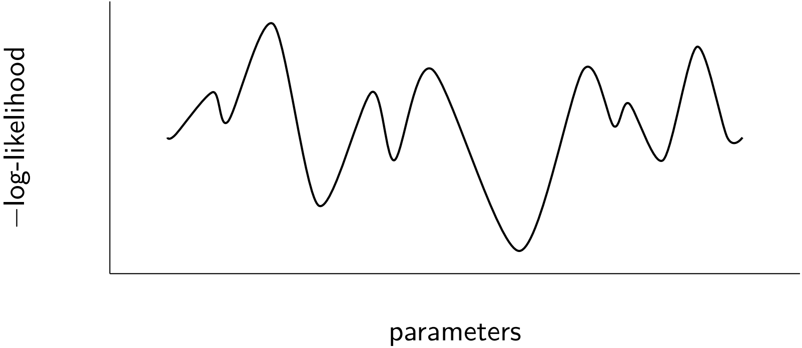
Latent Variable Graphical Models

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



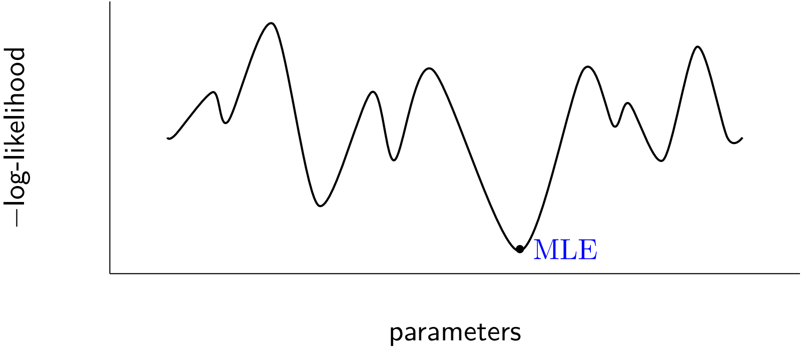


Parameter Estimation is Hard



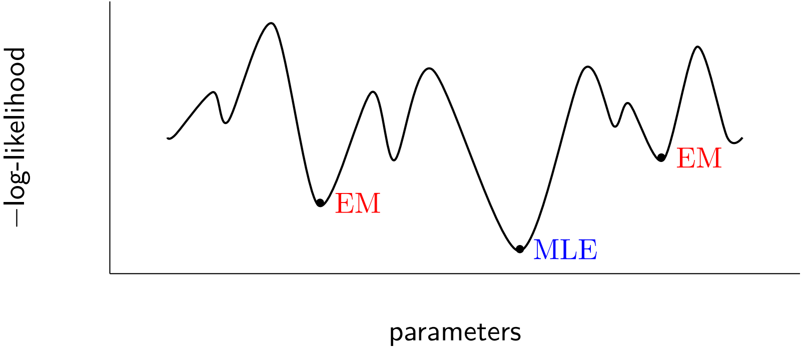
- Log-likelihood function is non-convex.

Parameter Estimation is Hard



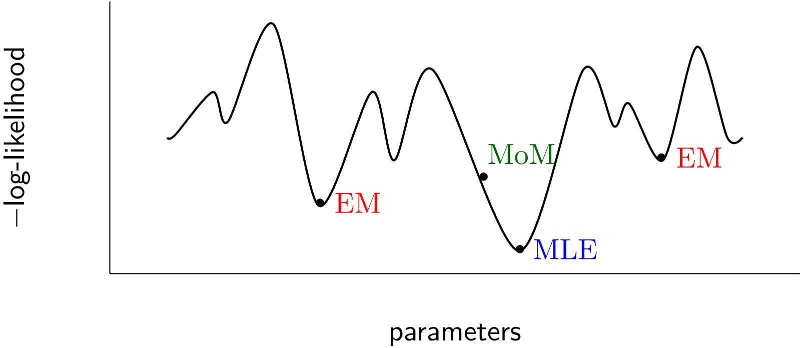
- Log-likelihood function is non-convex.
- MLE is consistent but intractable.

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 Dirichlet Allocation:** Anandkumar et al. 2012
  - **Latent trees:** Anandkumar et al. 2011
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors chaganty13regression**
  - **. . .**

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 Dirichlet Allocation:** Anandkumar et al. 2012
  - **Latent trees:** Anandkumar et al. 2011
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors chaganty13regression**
  - ...
- 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 Dirichlet Allocation:** Anandkumar et al. 2012
  - **Latent trees:** Anandkumar et al. 2011
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors chaganty13regression**
  - ...
- 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 Dirichlet Allocation:** Anandkumar et al. 2012
  - **Latent trees:** Anandkumar et al. 2011
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors chaganty13regression**
  - ...
- 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
  - ...
  - TODO: Check that this list is representative

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 Dirichlet Allocation:** Anandkumar et al. 2012
  - **Latent trees:** Anandkumar et al. 2011
  - **PCFGs:** Hsu, Kakade, and Liang 2012
  - **Mixtures of linear regressors chaganty13regression**
  - ...
- 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
  - ...
  - TODO: Check that this list is representative
- **How can we apply the method of moments to estimate *parameters efficiently* for a *general* model?**

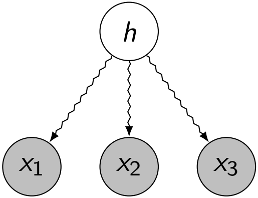
Setup

- Discrete models, d, k.
- Assume $d > k$.
- Parameters and marginals can be put into a matrix or tensor -¿ introduce notation.
- Assume infinite data.
- Highlight directed vs undirected - we focus on directed.

Background: Three-view Mixture Models

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.

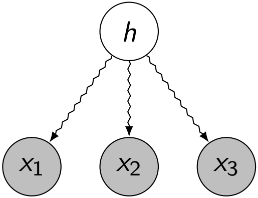


Background: Three-view Mixture Models

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, Hsu, and Kakade 2012 provide an algorithm to estimate conditional moments $\mathbb{P}(x_i \mid h)$ based on tensor eigendecomposition.
- In general, three views are necessary for identifiability (Kruskal 1977).

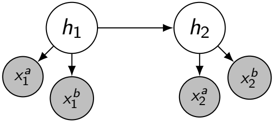


Outline

TODO: Make outline a diagram

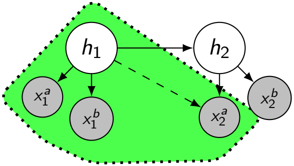
Example: a bridge, take I

- Each edge has a set of parameters.



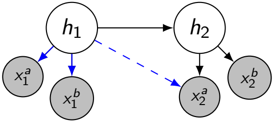
Example: a bridge, take I

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



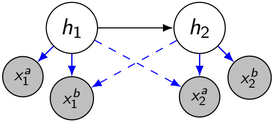
Example: a bridge, take I

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



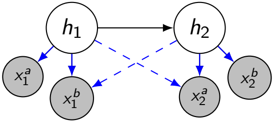
Example: a bridge, take I

- 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

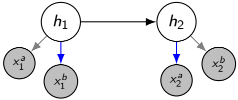
- Each edge has a set of parameters.
- h_1 and h_2 are bottlenecks.
- We can learn $\mathbb{P}(x_1^a|h_1), \mathbb{P}(x_1^b|h_1), \cdots$.
- However, we can't learn $\mathbb{P}(h_2|h_1)$ this way.



Example: a bridge, take II

- Observe the joint distribution, TODO: Use cartoon matrices

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



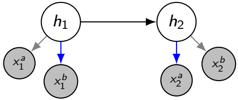
Example: a bridge, take II

- Observe the joint distribution, TODO: Use cartoon matrices

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

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

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



Example: a bridge, take II

- Observe the joint distribution, TODO: Use cartoon matrices

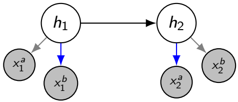
$$\underbrace{\mathbb{P}(x_1^b, x_2^a)}_{M_{12}} = \sum_{h_1, h_2} \underbrace{\mathbb{P}(x_1^b \mid h_1)}_{O^{(1|1)}} \underbrace{\mathbb{P}(x_2^a \mid h_2)}_{O^{(2|2)}} \underbrace{\mathbb{P}(h_1, h_2)}_{Z_{12}}.$$

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

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

- Solve for $\mathbb{P}(h_1, h_2)$ using pseudoinversion.

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



Example: a bridge, take II

- Observe the joint distribution, TODO: Use cartoon matrices

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

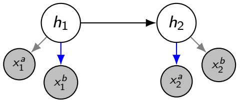
- **Observed moments** $\mathbb{P}(x_1^b, x_2^a)$ are *linear* in the **hidden marginals** $\mathbb{P}(h_1, h_2)$.

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

- Solve for $\mathbb{P}(h_1, h_2)$ using pseudoinversion.

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

- $\mathbb{P}(h_2 \mid h_1)$ can be recovered by normalization.



Outline

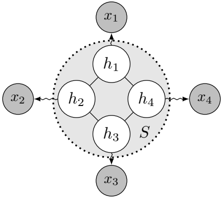
TODO: Make outline a diagram

Exclusive Views

Definition (Exclusive views)

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

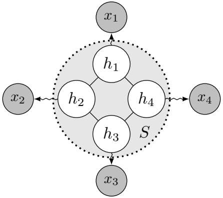
1. There exists some observed variable x_v which is conditionally independent of the others $S \backslash \{h_i\}$ given h_i.
2. The conditional moment matrix $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$ has full column rank k and can be recovered.
3. TODO: Exclusive views for a clique



Exclusive views give parameters

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

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

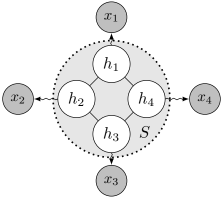


Exclusive views give parameters

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

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

$$M = Z(O^{(1|1)}, \cdots, O^{(m|m)})$$

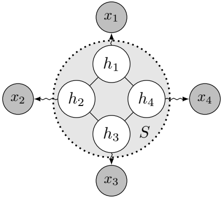


Exclusive views give parameters

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

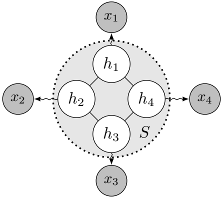
$$\underbrace{\mathbb{P}(x_1, \ldots, x_m)}_{M} = \sum_{h_1,\ldots,h_m} \underbrace{P(x_1|h_1)}_{O^{(1|1)}} \cdots \underbrace{{\color{blue}P(h_1, \cdots, h_m)}}_{Z}$$

$$M = Z(O^{(1|1)}, \cdots, O^{(m|m)})$$
$$Z = M(O^{(1|1)\dagger}, \cdots, O^{(m|m)\dagger}).$$



Bottlenecked graphs

- When are we assured exclusive views?

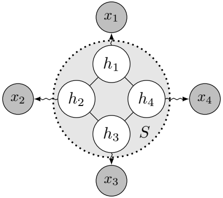


Bottlenecked graphs

- When are we assured exclusive views?

Definition (Bottlenecked set)

A set of hidden variables S is said to be *bottlenecked* if each $h \in S$ is a bottleneck.



Bottlenecked graphs

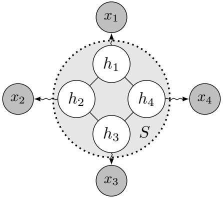
- When are we assured exclusive views?

Definition (Bottlenecked set)

A set of hidden variables S is said to be *bottlenecked* if each $h \in S$ is a bottleneck.

- **Theorem:** A bottlenecked clique has exclusive views.

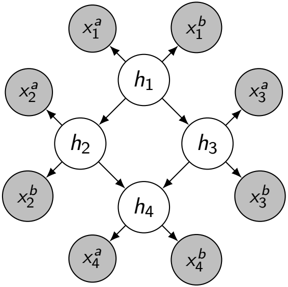
  TODO: Say show in paper.



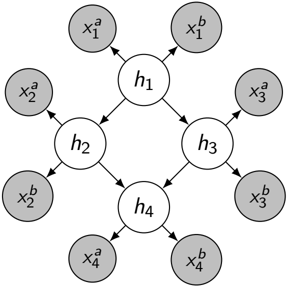
Outline

TODO: Make outline a diagram

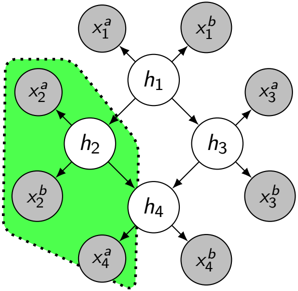
Example



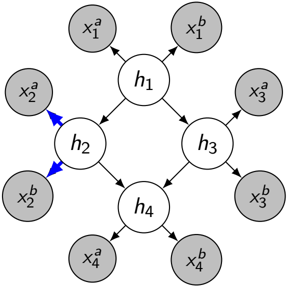
Example



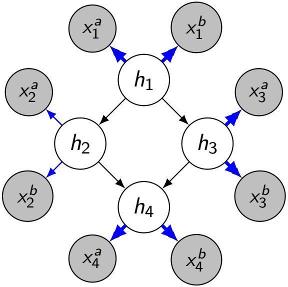
Example



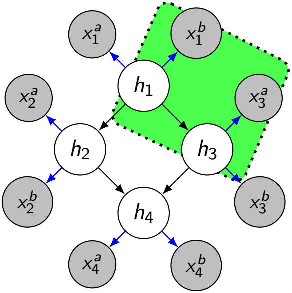
Example



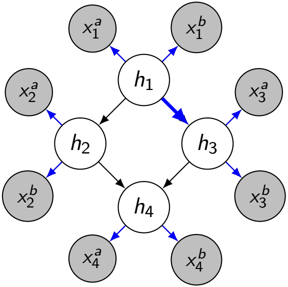
Example



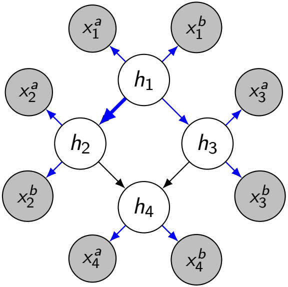
Example



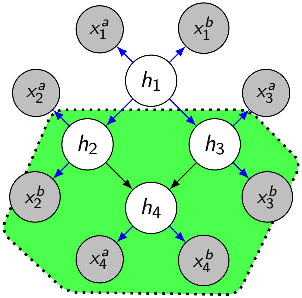
Example



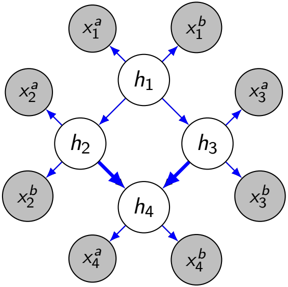
Example



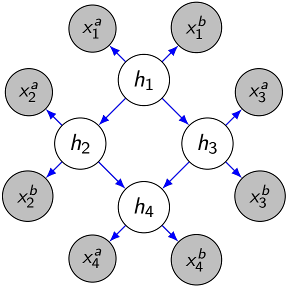
Example



Example

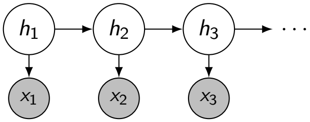


Example

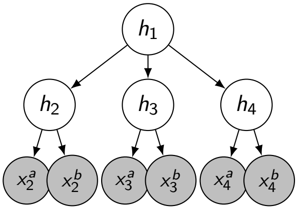


More Bottlenecked Examples

Hidden Markov models

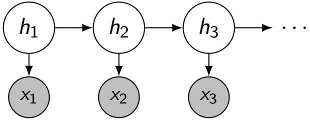


Latent Tree models

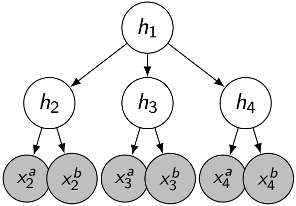


More Bottlenecked Examples

Hidden Markov models



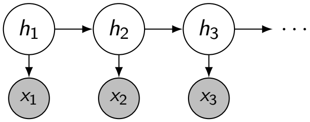
Latent Tree models



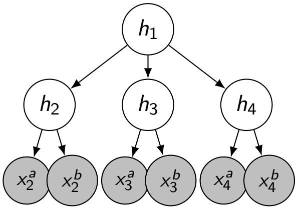
More Bottlenecked Examples

Halpern and Sontag 2013

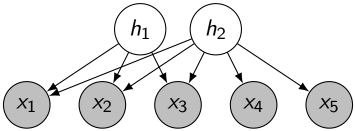
Hidden Markov models



Latent Tree models



Noisy Or (non-example)

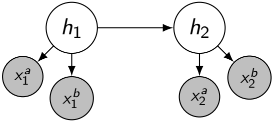


Outline

TODO: Make outline a diagram

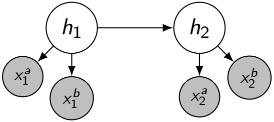
Convex marginal likelihoods

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



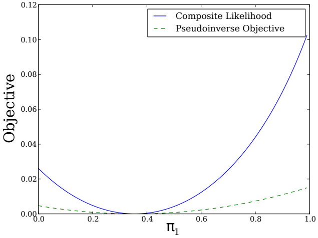
$$\log \mathbb{P}(\mathbf{x}) = \log \sum_{h_1,h_2} \mathbb{P}(\mathbf{x}_1|h_1)\,\mathbb{P}(\mathbf{x}_2|h_2)\mathbb{P}(h_1, h_2)$$

Convex marginal likelihoods

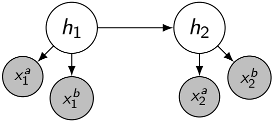


- The MLE is statistically most efficient, but usually non-convex.
- If we fix the conditional moments, $-\log \mathbb{P}(x)$ is convex in θ.

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

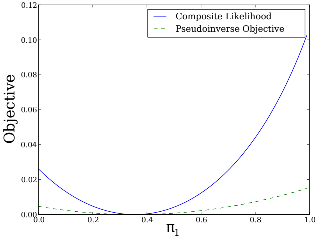


Convex marginal likelihoods



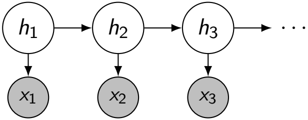
- The MLE is statistically most efficient, but usually non-convex.
- If we fix the conditional moments, $-\log \mathbb{P}(x)$ is convex in θ.
- No closed form solution, but a local method like EM is guaranteed to converge to the global optimum.

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



Composite likelihoods

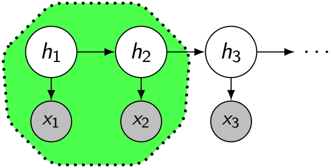
- In general, the full likelihood is still non-convex. TODO: Specify which x.



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

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

Composite likelihoods



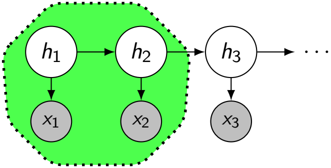
- In general, the full likelihood is still non-convex. TODO: Specify which x.
- Consider *composite likelihood* on a subset of observed variables.

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

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

Composite likelihoods

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

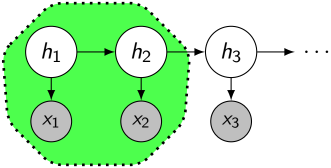


$$\log \mathbb{P}(\mathbf{x}) = \log \sum_{h_1, h_2} \underbrace{\mathbb{P}(\mathbf{x}_1 \mid h_1)\, \mathbb{P}(\mathbf{x}_2 \mid h_2)}_{\text{known}}$$

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

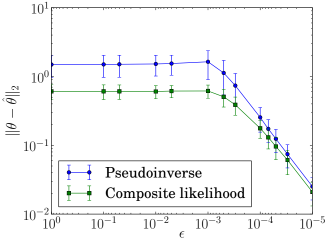
Composite likelihoods

- In general, the full likelihood is still non-convex. TODO: Specify which x.
- 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.



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

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



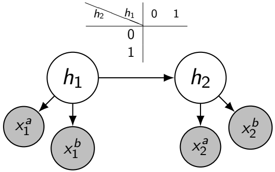
Outline

TODO: Make outline a diagram

Recovering parameters in directed models

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

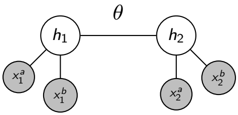
$$\mathbb{P}(h_2 \mid h_1) = \frac{\mathbb{P}(h_1, h_2)}{\sum_{h_2} \mathbb{P}(h_1, h_2)}.$$



Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,
  TODO: use sum over cliques - talk through.

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\theta^\top \phi(\mathbf{x}, \mathbf{h}) - A(\theta)\right).$$



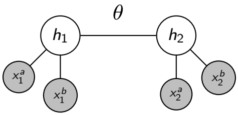
Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,
  TODO: use sum over cliques - talk through.

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\theta^\top \phi(\mathbf{x}, \mathbf{h}) - A(\theta)\right).$$

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

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



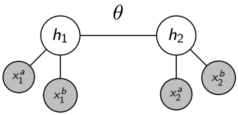
Recovering parameters in undirected log-linear models

- Assume a log-linear parameterization,
  TODO: use sum over cliques - talk through.

$$p_\theta(\mathbf{x}, \mathbf{h}) = \exp\left(\theta^\top \phi(\mathbf{x}, \mathbf{h}) - A(\theta)\right).$$

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

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



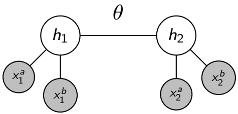
- However, the *supervised* negative log-likelihood is convex,

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

Recovering parameters in undirected log-linear models

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

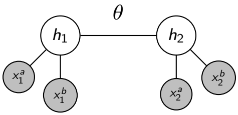
$$\mathcal{L}_{\text{sup}}(\theta) = -\theta^\top \underbrace{\left(\sum_{\mathcal{C} \in \mathcal{G}} \mathbb{E}_{(\mathbf{x},\mathbf{h}) \sim \mathcal{D}_{\text{sup}}}[\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})] \right)}_{\mu_\mathcal{C}} + A(\theta).$$



Recovering parameters in undirected log-linear models

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

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



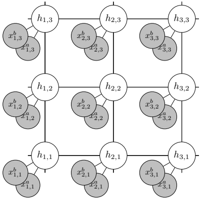
- However, the marginals can also be *consistently* estimated by moments!

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

Optimizing pseudolikelihood

- Estimating marginals $\mu_{\mathcal{C}}$ is independent of treewidth, but computing the normalization constant is: TODO: convex but not easy

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



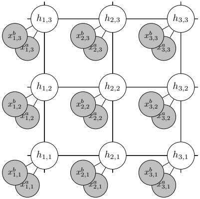
Optimizing pseudolikelihood

- Estimating marginals $\mu_{\mathcal{C}}$ is independent of treewidth, but computing the normalization constant is: TODO: convex but not easy

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

- We can use pseudolikelihood (**besag75pseudo**) to consistently estimate distributions over local neighborhoods.

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



Optimizing pseudolikelihood

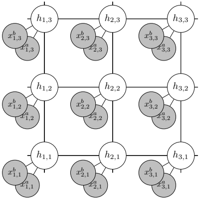
- Estimating marginals $\mu_{\mathcal{C}}$ is independent of treewidth, but computing the normalization constant is: TODO: convex but not easy

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

- We can use pseudolikelihood (**besag75pseudo**) to consistently estimate distributions over local neighborhoods.

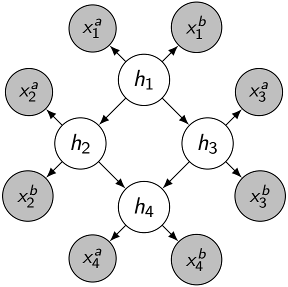
$$A_{\text{pseudo}}(\theta; \mathcal{N}(a)) \triangleq \log \sum_{a} \exp\left(\theta^\top \phi(\mathbf{x}_{\mathcal{N}}, \mathbf{h}_{\mathcal{N}})\right).$$

- Clique marginals not sufficient statistics, but we can still estimate them.



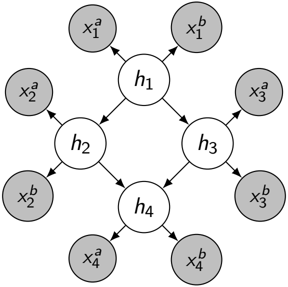
Conclusions

- TODO: Use outline slide.
- TODO: Show the venn diagram on progress on generality..
- An algorithm for any (non-degenerate) **bottlenecked discrete graphical models**.



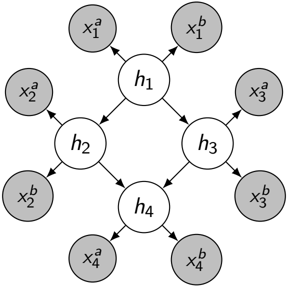
Conclusions

- TODO: Use outline slide.
- TODO: Show the venn diagram on progress on generality..
- An algorithm for any (non-degenerate) **bottlenecked discrete graphical models**.
- Efficiently learns models with **high-treewidth**.



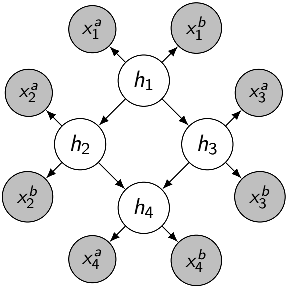
Conclusions

- TODO: Use outline slide.
- TODO: Show the venn diagram on progress on generality..
- An algorithm for any (non-degenerate) **bottlenecked discrete graphical models**.
- Efficiently learns models with **high-treewidth**.
- Combine moment estimators with composite likelihood estimators.



Conclusions

- TODO: Use outline slide..
- TODO: Show the venn diagram on progress on generality..
- An algorithm for any (non-degenerate) **bottlenecked discrete graphical models**.
- Efficiently learns models with **high-treewidth**.
- Combine moment estimators with composite likelihood estimators.
- Extends to **log-linear models**.
  - Allows for easy regularization, missing data,



Thank you!