Spectral Experts for Estimating Mixtures of Linear Regressions

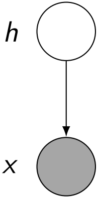
Arun Tejasvi Chaganty
Percy Liang

Stanford University

January 28, 2016

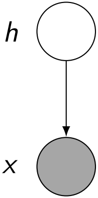
Latent Variable Models

- **Generative Models**



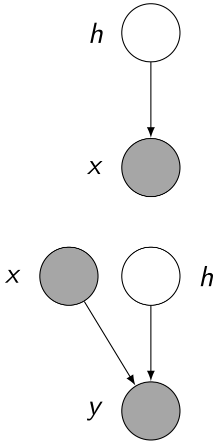
Latent Variable Models

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



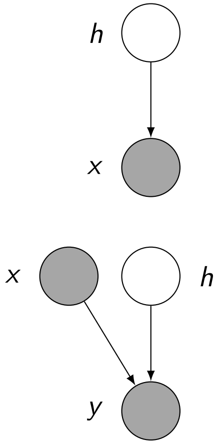
Latent Variable Models

- **Generative Models**
  - Gaussian Mixture Models
  - Hidden Markov Models
  - Latent Dirichlet Allocation
  - PCFGs
  - ...
- **Discriminative Models**



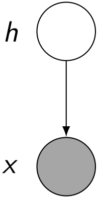
Latent Variable Models

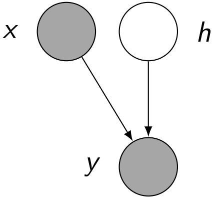
- **Generative Models**
  - Gaussian Mixture Models
  - Hidden Markov Models
  - Latent Dirichlet Allocation
  - PCFGs
  - ...
- **Discriminative Models**
  - Mixture of Experts
  - Latent CRFs
  - Discriminative LDA
  - ...



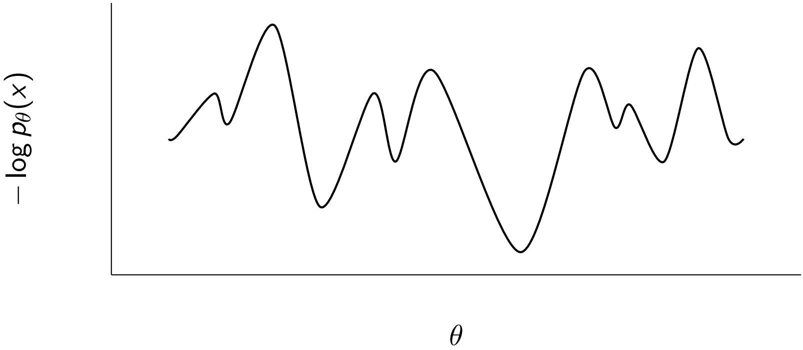
Latent Variable Models

- **Generative Models**
  - Gaussian Mixture Models
  - Hidden Markov Models
  - Latent Dirichlet Allocation
  - PCFGs
  - ...
- **Discriminative Models**
  - Mixture of Experts
  - Latent CRFs
  - Discriminative LDA
  - ...
- *Easy to include features and tend to be more accurate.*



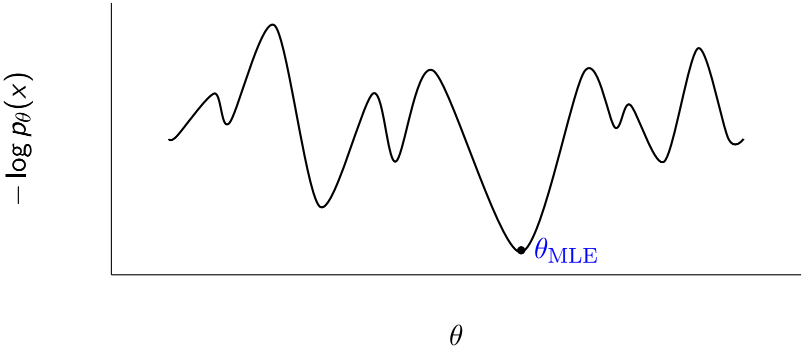


Parameter Estimation is Hard



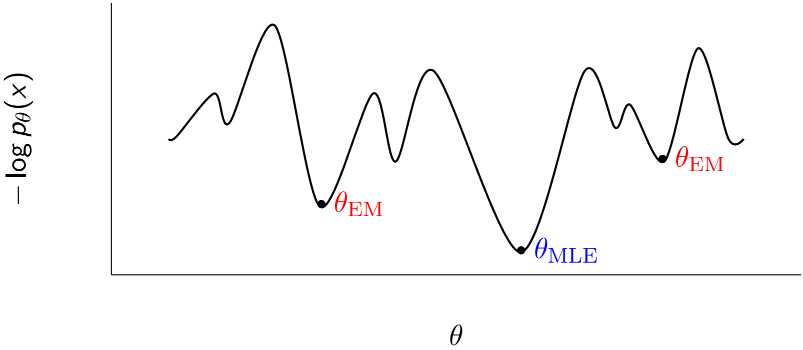
- Log-likelihood function is non-convex.

Parameter Estimation is Hard



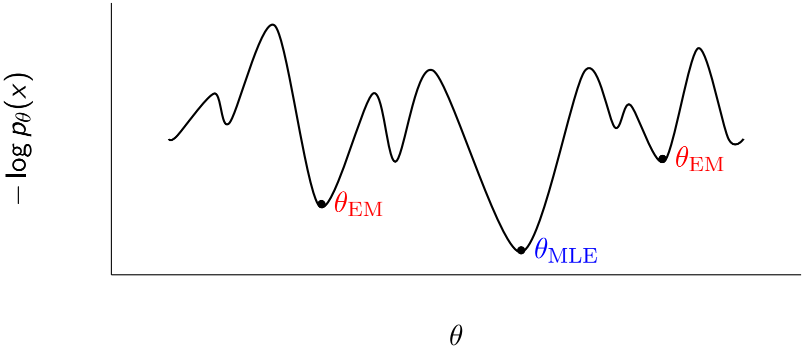
- Log-likelihood function is non-convex.
- MLE is consistent but intractable.

Parameter Estimation is Hard



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

Parameter Estimation is Hard



- Log-likelihood function is non-convex.
- MLE is consistent but intractable.
- Local methods (EM, gradient descent, etc.) are tractable but inconsistent.
- Can we build an **efficient and consistent estimator**?

Related Work

- Method of Moments [Pearson, 1894]

Related Work

- Method of Moments [Pearson, 1894]
- Observable operators
  - Control Theory [Ljung, 1987]
  - Observable operator models [Jaeger, 2000; Littman/Sutton/Singh, 2004]
  - Hidden Markov models [Hsu/Kakade/Zhang, 2009]
  - Low-treewidth graphs [Parikh et al., 2012]
  - Weighted finite state automata [Balle & Mohri, 2012]

Related Work

- Method of Moments [Pearson, 1894]
- Observable operators
  - Control Theory [Ljung, 1987]
  - Observable operator models [Jaeger, 2000; Littman/Sutton/Singh, 2004]
  - Hidden Markov models [Hsu/Kakade/Zhang, 2009]
  - Low-treewidth graphs [Parikh et al., 2012]
  - Weighted finite state automata [Balle & Mohri, 2012]
- Parameter Estimation
  - Mixture of Gaussians [Kalai/Moitra/Valiant, 2010]
  - **Mixture models, HMMs [Anandkumar/Hsu/Kakade, 2012]**
  - Latent Dirichlet Allocation [Anandkumar/Hsu/Kakade, 2012]
  - Stochastic block models [Anandkumar/Ge/Hsu/Kakade, 2012]
  - Linear Bayesian networks [Anandkumar/Hsu/Javanmard/Kakade, 2012]

Outline

Introduction

Tensor Factorization for a Generative Model

Tensor Factorization for a Discriminative Model

Experimental Insights

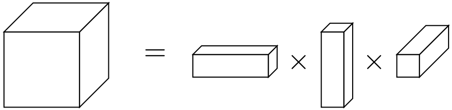
Conclusions

Aside: Tensor Operations

- Tensor Product

$$x^{\otimes 3} = x \otimes x \otimes x$$

$$x^{\otimes 3}_{ijk} = x_i x_j x_k$$

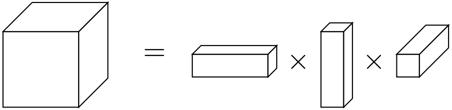


Aside: Tensor Operations

- Tensor Product

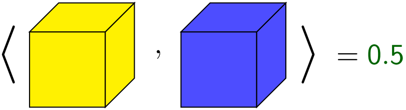
$$x^{\otimes 3} = x \otimes x \otimes x$$

$$x^{\otimes 3}_{ijk} = x_i x_j x_k$$



- Inner product

$$\langle A, B \rangle = \sum_{ijk} A_{ijk} B_{ijk}$$

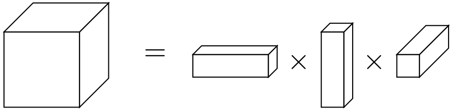


Aside: Tensor Operations

- Tensor Product

$$x^{\otimes 3} = x \otimes x \otimes x$$

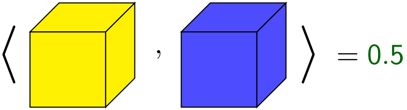
$$x^{\otimes 3}_{ijk} = x_i x_j x_k$$

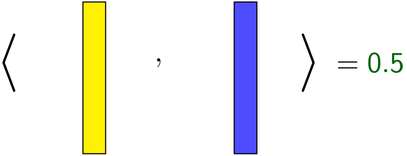


- Inner product

$$\langle A, B \rangle = \sum_{ijk} A_{ijk} B_{ijk}$$

$$= \langle \mathrm{vec}\, A, \mathrm{vec}\, B \rangle$$



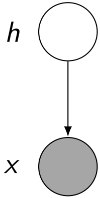


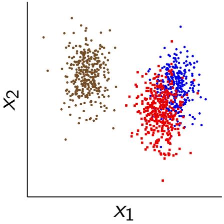
Example: Gaussian Mixture Model

anandkumar12moments

- Generative process:

$$h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$$
$$x \sim \mathcal{N}(\beta_h, \sigma^2).$$



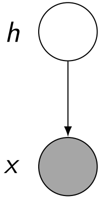


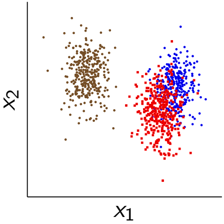
Example: Gaussian Mixture Model

anandkumar12moments

- Generative process:

$$h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$$
$$x \sim \mathcal{N}(\beta_h, \sigma^2).$$





- Moments:

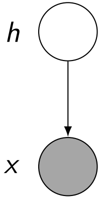
$$\mathbb{E}[x|h] = \beta_h$$

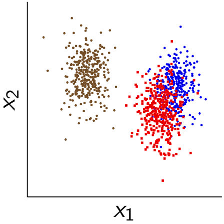
Example: Gaussian Mixture Model

anandkumar12moments

- Generative process:

$$h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$$
$$x \sim \mathcal{N}(\beta_h, \sigma^2).$$





- Moments:

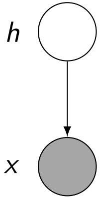
$$\mathbb{E}[x|h] = \beta_h$$
$$\mathbb{E}[x] = \sum_h \pi_h \beta_h$$

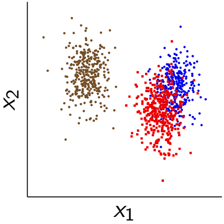
Example: Gaussian Mixture Model

anandkumar12moments

- Generative process:

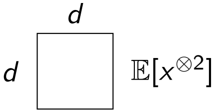
$$h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$$
$$x \sim \mathcal{N}(\beta_h, \sigma^2).$$





- Moments:

$$\mathbb{E}[x|h] = \beta_h$$
$$\mathbb{E}[x] = \sum_h \pi_h \beta_h$$
$$\mathbb{E}[x^{\otimes 2}] = \sum_h \pi_h (\beta_h \beta_h^T) + \sigma^2$$
$$= \sum_h \pi_h \beta_h^{\otimes 2} + \sigma^2$$

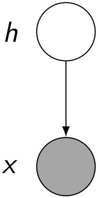


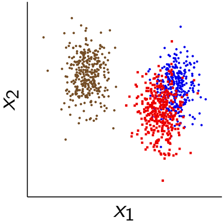
Example: Gaussian Mixture Model

anandkumar12moments

- Generative process:

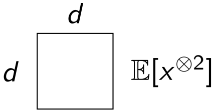
$$h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$$
$$x \sim \mathcal{N}(\beta_h, \sigma^2).$$

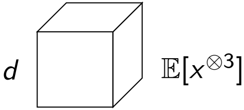




- Moments:

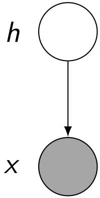
$$\mathbb{E}[x|h] = \beta_h$$
$$\mathbb{E}[x] = \sum_h \pi_h \beta_h$$
$$\mathbb{E}[x^{\otimes 2}] = \sum_h \pi_h (\beta_h \beta_h^T) + \sigma^2$$
$$= \sum_h \pi_h {\beta_h}^{\otimes 2} + \sigma^2$$
$$\mathbb{E}[x^{\otimes 3}] = \sum_h \pi_h \beta_h^{\otimes 3} + \mathrm{bias}.$$

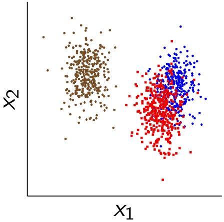




Solution: Tensor Factorization

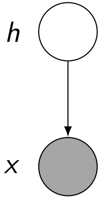
- $\mathbb{E}[x^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$.

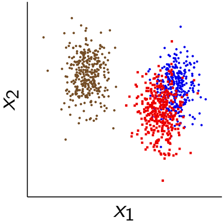


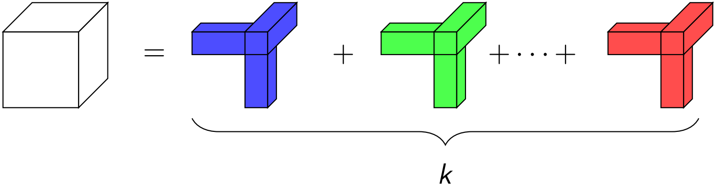


Solution: Tensor Factorization

- $\mathbb{E}[x^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$.





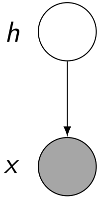


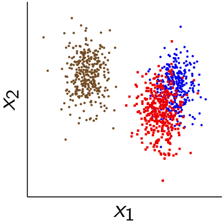
Solution: Tensor Factorization

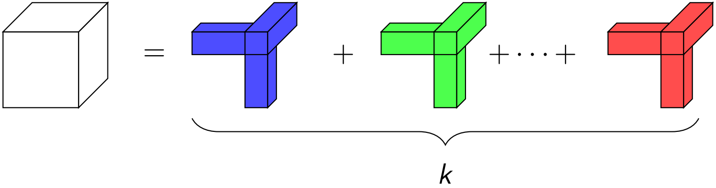
AnandkumarGeHsu2012

- $\mathbb{E}[x^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$.
- If β_h are orthogonal, they are eigenvectors!

$$\mathbb{E}[x^{\otimes 3}](\beta_h, \beta_h) = \pi_h \beta_h.$$







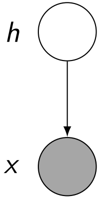
Solution: Tensor Factorization

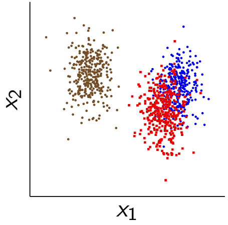
AnandkumarGeHsu2012

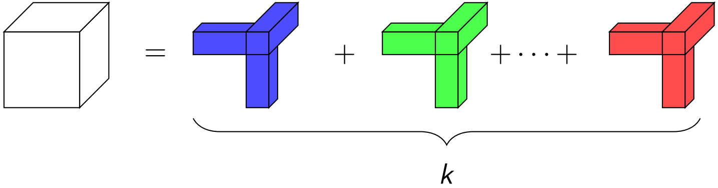
- $\mathbb{E}[x^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$.
- If β_h are orthogonal, they are eigenvectors!

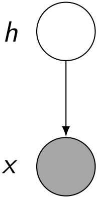
$$\mathbb{E}[x^{\otimes 3}](\beta_h, \beta_h) = \pi_h \beta_h.$$

- In general, whiten $\mathbb{E}[x^{\otimes 3}]$ first.

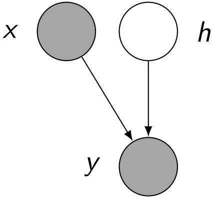




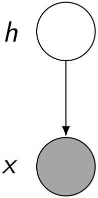




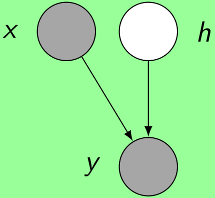
Generative Models



Discriminative Models

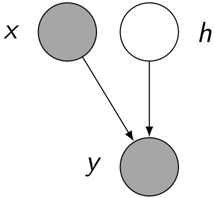


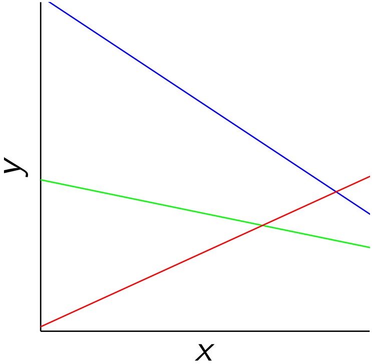
Generative Models



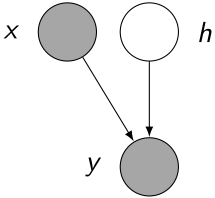
Discriminative Models

Mixture of Linear Regressions

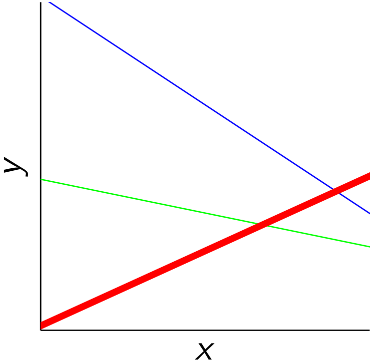




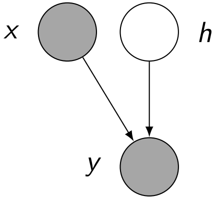
Mixture of Linear Regressions



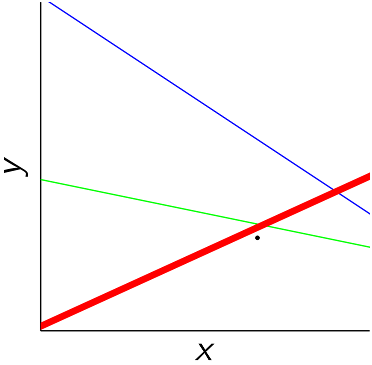
- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.



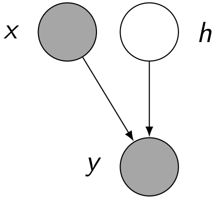
Mixture of Linear Regressions

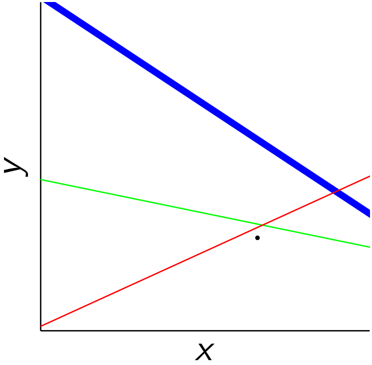


- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.



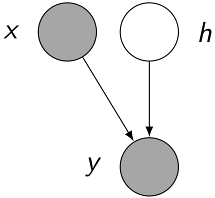
Mixture of Linear Regressions



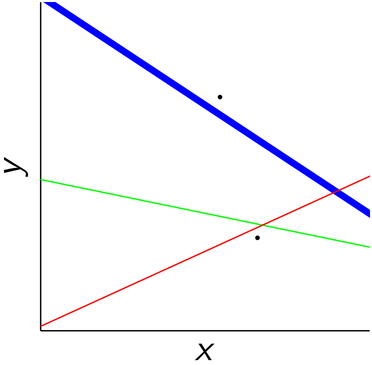


- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.

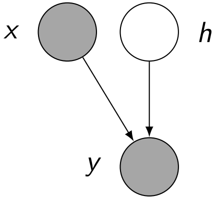
Mixture of Linear Regressions



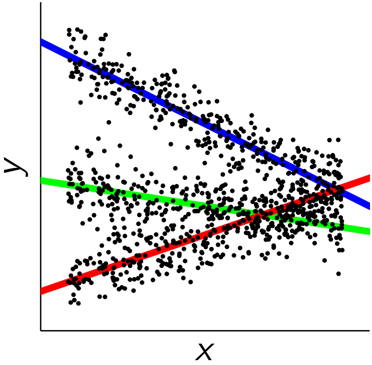
- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.



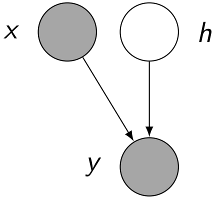
Mixture of Linear Regressions



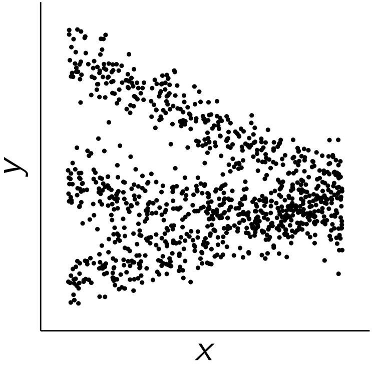
- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.



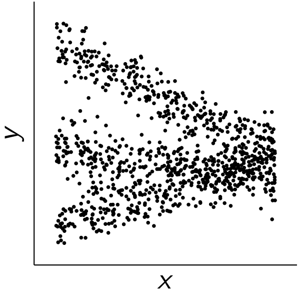
Mixture of Linear Regressions



- Given x
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.



Mixture of Linear Regressions



$$\xrightarrow{\quad ? \quad} \begin{bmatrix}\pi_1\\ \pi_2\\ \vdots\\ \pi_k\end{bmatrix} \underbrace{\begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_k \end{bmatrix}}_{B}$$

Finding Tensor Structure

$$y = \langle \beta_h , x \rangle + \epsilon$$

Finding Tensor Structure

$$y = \langle \underbrace{\beta_h}_{\text{random}}, x \rangle + \epsilon$$

Finding Tensor Structure

$$
\begin{aligned}
y &= \langle \underbrace{\beta_h}_{\text{random}}, x \rangle + \epsilon \\
&= \langle \mathbb{E}[\beta_h], x \rangle + \langle (\beta_h - \mathbb{E}[\beta_h]), x \rangle + \epsilon
\end{aligned}
$$

$\boxed{\mathbb{E}[\beta_h] = \sum_h \pi_h \beta_h.}$

Finding Tensor Structure

$$y = \langle \underbrace{\beta_h}_{\text{random}}, x \rangle + \epsilon$$

$$= \underbrace{\langle \mathbb{E}[\beta_h], x \rangle}_{\text{linear measurement}} + \langle (\beta_h - \mathbb{E}[\beta_h]), x \rangle + \epsilon$$

$$\boxed{\mathbb{E}[\beta_h] = \sum_h \pi_h \beta_h.}$$

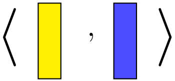
Finding Tensor Structure

$$
\begin{aligned}
y &= \langle \underbrace{\beta_h}_{\text{random}}, x \rangle + \epsilon \\
&= \underbrace{\langle \mathbb{E}[\beta_h], x \rangle}_{\text{linear measurement}} + \underbrace{\langle (\beta_h - \mathbb{E}[\beta_h]), x \rangle + \epsilon}_{\text{noise}}
\end{aligned}
$$

$$\boxed{\mathbb{E}[\beta_h] = \sum_h \pi_h \beta_h.}$$

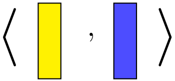
Finding Tensor Structure

$$y = \overbrace{\langle \mathbb{E}[\beta_h], x \rangle}^{\text{linear measurement}} + \overbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}^{\text{noise}}$$



Finding Tensor Structure

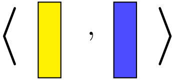
$$y = \overbrace{\langle \mathbb{E}[\beta_h], x \rangle}^{\text{linear measurement}} + \overbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}^{\text{noise}}$$



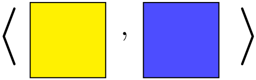
$$y^2 = (\langle \beta_h, x \rangle + \epsilon)^2$$

Finding Tensor Structure

$$y = \overbrace{\langle \mathbb{E}[\beta_h], x\rangle}^{\text{linear measurement}} + \overbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}^{\text{noise}}$$

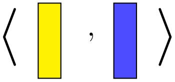


$$\begin{aligned} y^2 &= (\langle \beta_h, x\rangle + \epsilon)^2 \\ &= \langle \mathbb{E}[\beta_h^{\otimes 2}], x^{\otimes 2}\rangle + \text{bias}_2 + \text{noise}_2 \end{aligned}$$

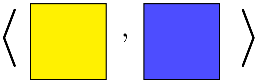


Finding Tensor Structure

$$y = \overbrace{\langle \mathbb{E}[\beta_h], x \rangle}^{\text{linear measurement}} + \overbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}^{\text{noise}}$$



$$\begin{aligned} y^2 &= (\langle \beta_h, x \rangle + \epsilon)^2 \\ &= \langle \underbrace{\mathbb{E}[\beta_h^{\otimes 2}]}_{M_2}, x^{\otimes 2} \rangle + \text{bias}_2 + \text{noise}_2 \end{aligned}$$

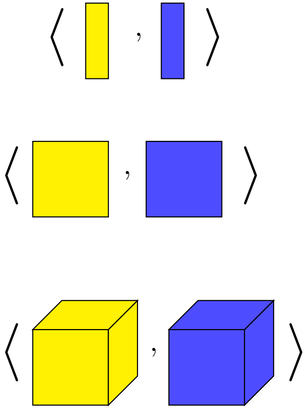


Finding Tensor Structure

$$y = \overbrace{\langle \mathbb{E}[\beta_h], x \rangle}^{\text{linear measurement}} + \overbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}^{\text{noise}}$$

$$y^2 = (\langle \beta_h, x \rangle + \epsilon)^2$$
$$= \langle \underbrace{\mathbb{E}[\beta_h^{\otimes 2}]}_{M_2}, x^{\otimes 2} \rangle + \text{bias}_2 + \text{noise}_2$$

$$y^3 = \langle \underbrace{\mathbb{E}[\beta_h^{\otimes 3}]}_{M_3}, x^{\otimes 3} \rangle + \text{bias}_3 + \text{noise}_3$$

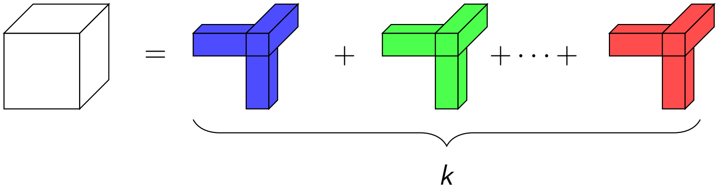


Recovering Parameters

- $M_3 \stackrel{\text{def}}{=} \mathbb{E}[\beta_h^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$

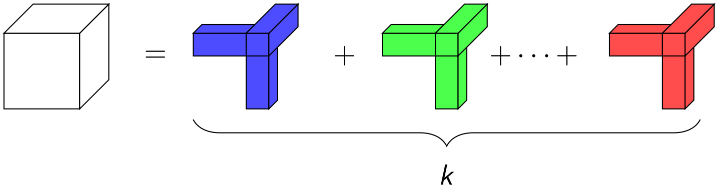
Recovering Parameters

- $M_3 \stackrel{\text{def}}{=} \mathbb{E}[\beta_h^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$

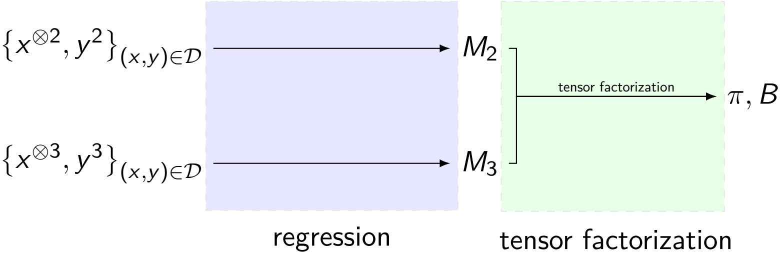


Recovering Parameters

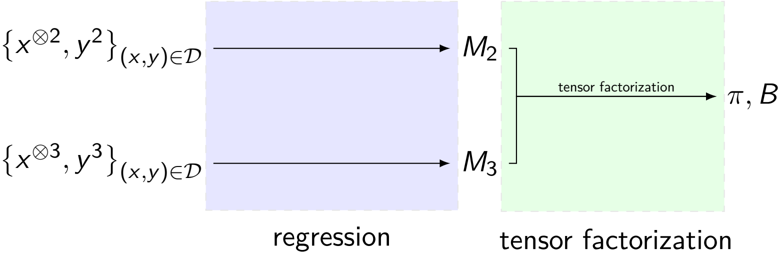
- $M_3 \stackrel{\text{def}}{=} \mathbb{E}[\beta_h^{\otimes 3}] = \sum_{h=1}^{k} \pi_h \beta_h^{\otimes 3}$
- Apply tensor factorization!



Overview: Spectral Experts

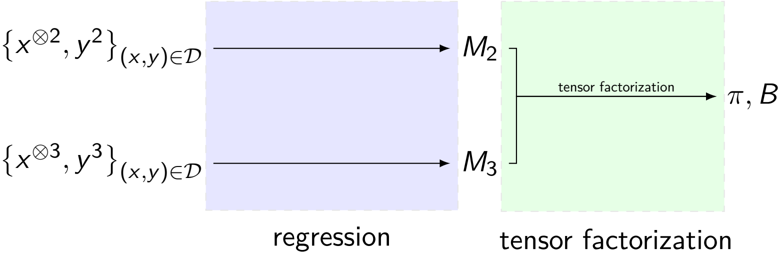


Overview: Spectral Experts



Assumptions:

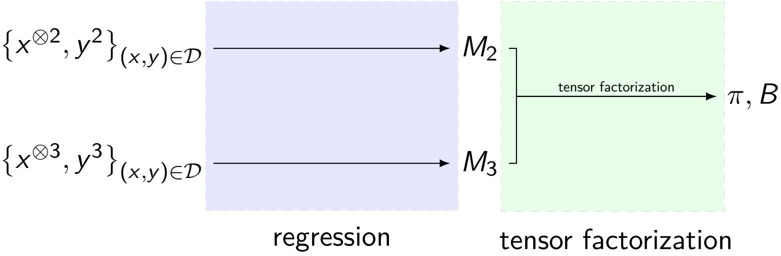
Overview: Spectral Experts



Assumptions:

$$\hat{\mathbb{E}}[\mathrm{vec}(x^{\otimes 2})^{\otimes 2}] \succ 0$$
$$\hat{\mathbb{E}}[\mathrm{vec}(x^{\otimes 3})^{\otimes 2}] \succ 0.$$

Overview: Spectral Experts



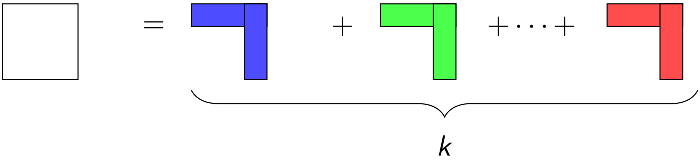
Assumptions:

$$\hat{\mathbb{E}}[\mathrm{vec}(x^{\otimes 2})^{\otimes 2}] \succ 0 \qquad \pi \succ 0$$

$$\hat{\mathbb{E}}[\mathrm{vec}(x^{\otimes 3})^{\otimes 2}] \succ 0. \qquad rank(B) = k \leq d$$

Exploiting Low-rank Structure.

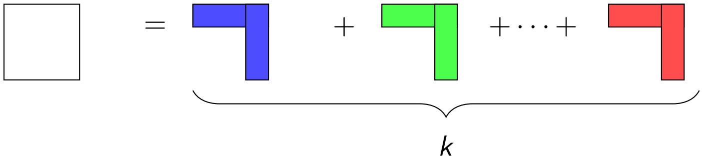
$$\hat{M}_2 = \arg\min_{M} \sum_{(x,y)\in\mathcal{D}} \left(y^2 - \left\langle M, x^{\otimes 2} \right\rangle - \mathrm{bias}_2\right)^2$$



Exploiting Low-rank Structure.

fazel2002matrix

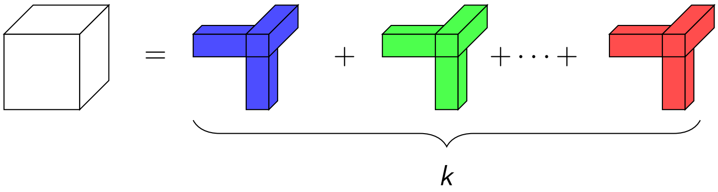
$$\hat{M}_2 = \arg\min_{M} \sum_{(x,y)\in\mathcal{D}} \left(y^2 - \left\langle M, x^{\otimes 2}\right\rangle - \mathrm{bias}_2\right)^2 + \underbrace{\|M\|_*}_{\sum_i \sigma_i(M)}$$



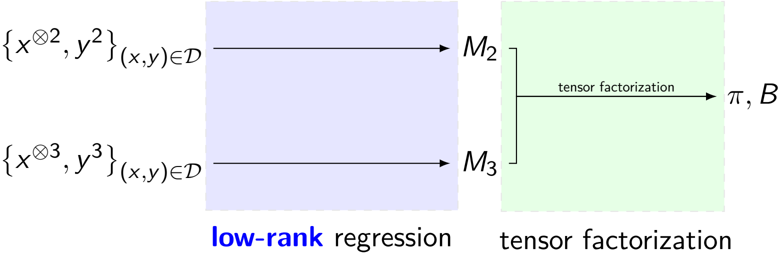
Exploiting Low-rank Structure.

fazel2002matrix
tomioka2010estimation

$$\hat{M}_3 = \arg\min_{M} \sum_{(x,y)\in\mathcal{D}} \left(y^3 - \left\langle M, x^{\otimes 3} \right\rangle - \mathrm{bias}_3\right)^2 + \|M\|_*$$

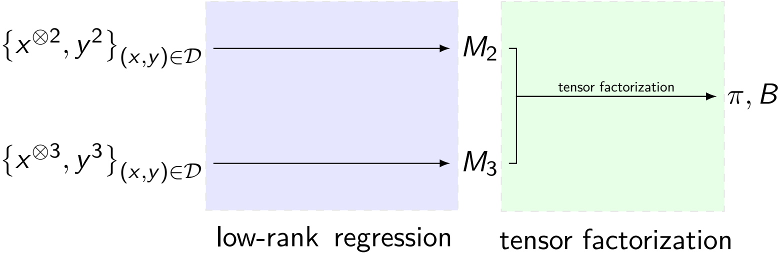


Sample Complexity



Sample Complexity

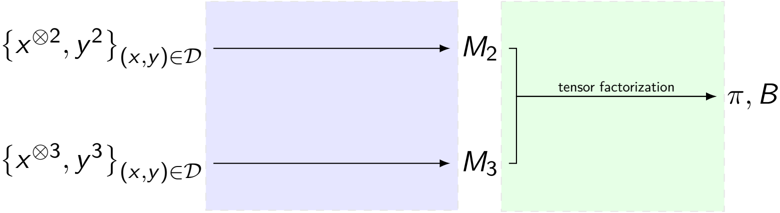
NegahbanWainwright2009; **Tomioka2011**



$$O\left(k \, \|x\|^{12} \, \|\beta\|^{6} \, \|\mathbb{E}[\epsilon^2]\|^{6}\right)$$

Sample Complexity

NegahbanWainwright2009; **Tomioka2011** **AnandkumarGeHsu2012**



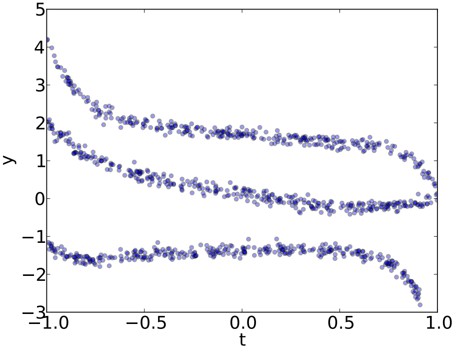
low-rank regression

$$O\left(k \|x\|^{12} \|\beta\|^{6} \|\mathbb{E}[\epsilon^2]\|^{6}\right)$$

tensor factorization

$$O\left(\frac{k\pi_{\max}^2}{\sigma_k(M_2)^5}\right)$$

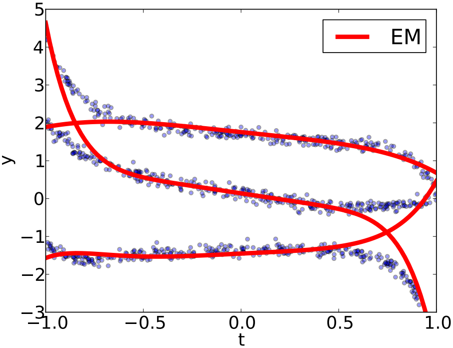
Experimental Insights



$$y = \beta^T \underbrace{\begin{bmatrix} 1 \\ t \\ t^4 \\ t^7 \end{bmatrix}}_{x} + \epsilon$$

$$k = 3, d = 4, n = 10^5$$

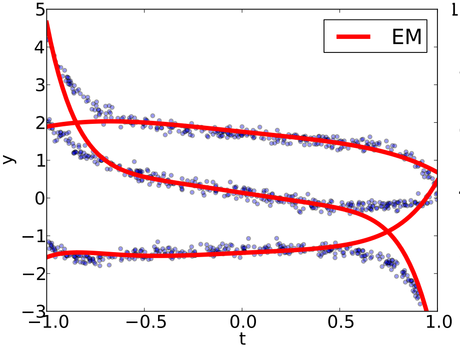
Experimental Insights

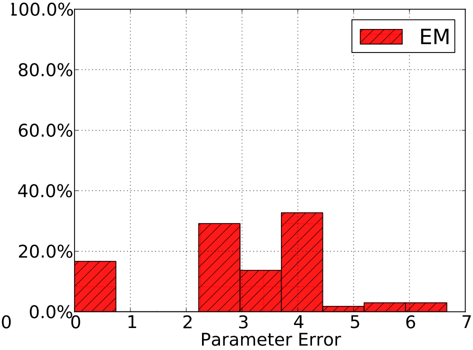


$$y = \beta^T \underbrace{\begin{bmatrix} 1 \\ t \\ t^4 \\ t^7 \end{bmatrix}}_{x} + \epsilon$$

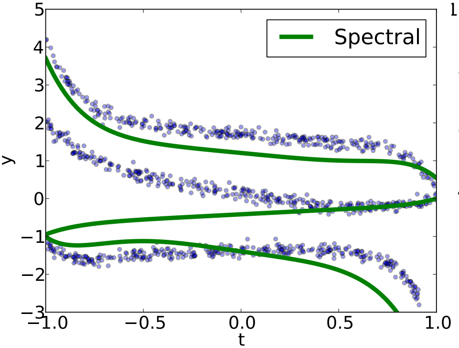
$$k = 3, d = 4, n = 10^5$$

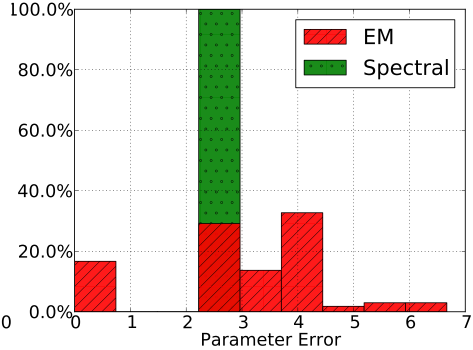
Experimental Insights



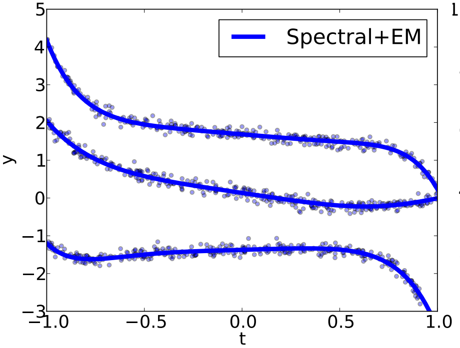


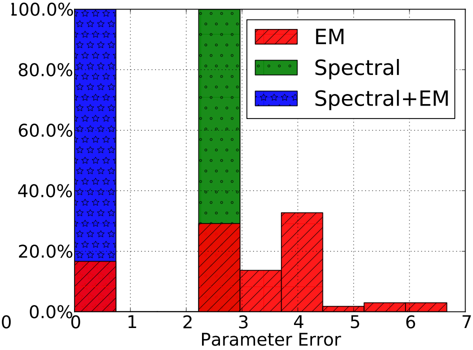
Experimental Insights



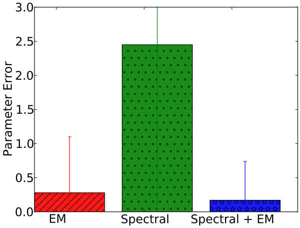


Experimental Insights

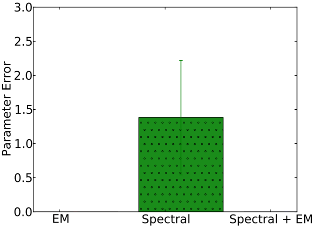




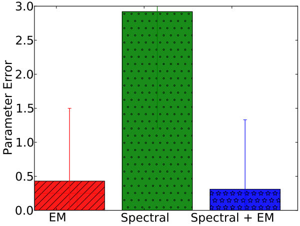
Experimental Insights



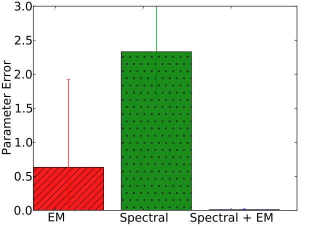
$d = 4, k = 2$



$d = 5, k = 2$

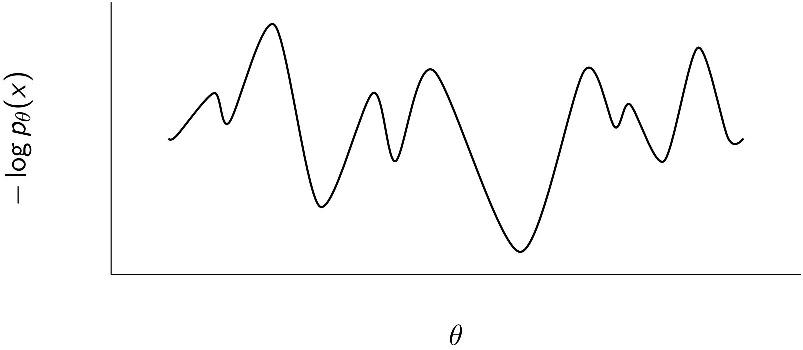


$d = 5, k = 3$

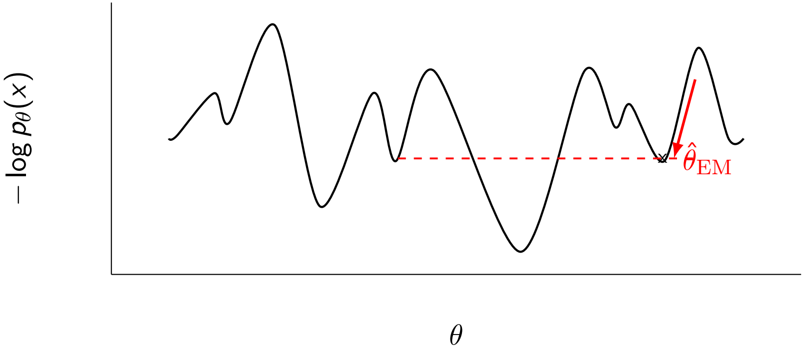


$d = 6, k = 2$

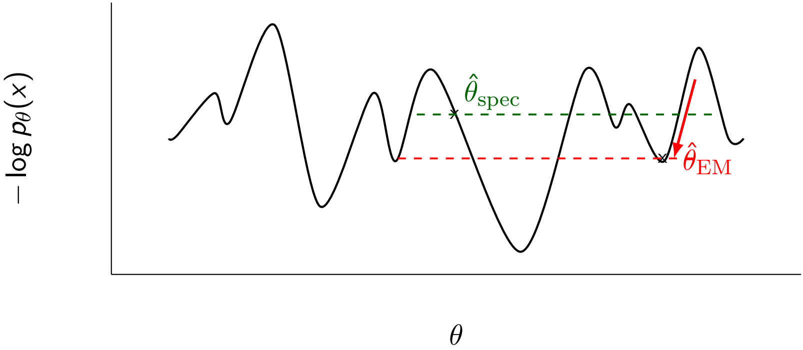
On Initialization (Cartoon)



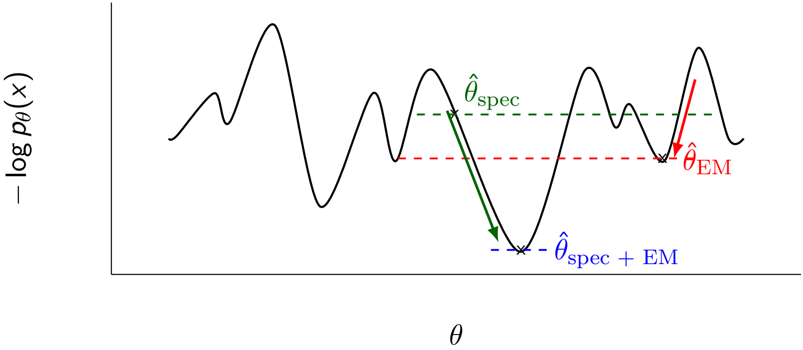
On Initialization (Cartoon)



On Initialization (Cartoon)



On Initialization (Cartoon)



Conclusions

- Consistent estimator for the mixture of linear regressions

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.
- **Theory:** Polynomial sample and computational complexity.

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.
- **Theory:** Polynomial sample and computational complexity.
- **Experiments:** Method of moment estimates can be a good initialization for EM.

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.
- **Theory:** Polynomial sample and computational complexity.
- **Experiments:** Method of moment estimates can be a good initialization for EM.
- **Future Work:** How can we handle other discriminative models?

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.
- **Theory:** Polynomial sample and computational complexity.
- **Experiments:** Method of moment estimates can be a good initialization for EM.
- **Future Work:** How can we handle other discriminative models?
  - Dependencies between h and x (mixture of experts).

Conclusions

- Consistent estimator for the mixture of linear regressions
- **Key Idea:** Expose tensor factorization structure through regression.
- **Theory:** Polynomial sample and computational complexity.
- **Experiments:** Method of moment estimates can be a good initialization for EM.
- **Future Work:** How can we handle other discriminative models?
  - Dependencies between h and x (mixture of experts).
  - Non-linear link functions (hidden variable logistic regression).

Thank you!