Alternative Specifications in Machine Learning

Percy Liang

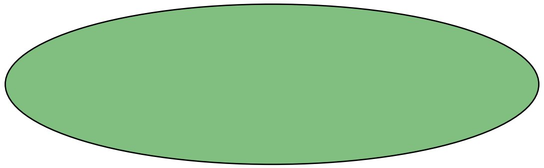
Jacob Steinhardt, Fereshte Khani

Stanford Artificial Intelligence

FLI Workshop

January 5, 2016

Foundations of machine learning



Train $p_0(x)$

Foundations of machine learning

Train $p_0(x)$ Test $p_1(x)$

Classic statistical learning theory:

training distribution $=$ test distribution

Foundations of machine learning

Train $p_0(x)$ Test $p_1(x)$

Classic statistical learning theory:

$$\text{training distribution} = \text{test distribution}$$

Relaxation: domain adaptation, mild adversaries

$$\text{training distribution} \approx \text{test distribution}$$

Foundations of machine learning

Train $p_0(x)$ Test $p_1(x)$

Classic statistical learning theory:

$$\text{training distribution} \;=\; \text{test distribution}$$

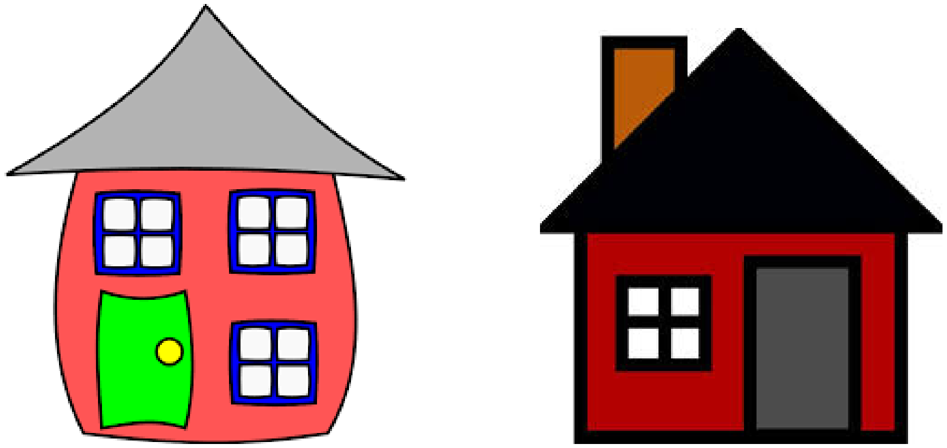
Relaxation: domain adaptation, mild adversaries

$$\text{training distribution} \;\approx\; \text{test distribution}$$

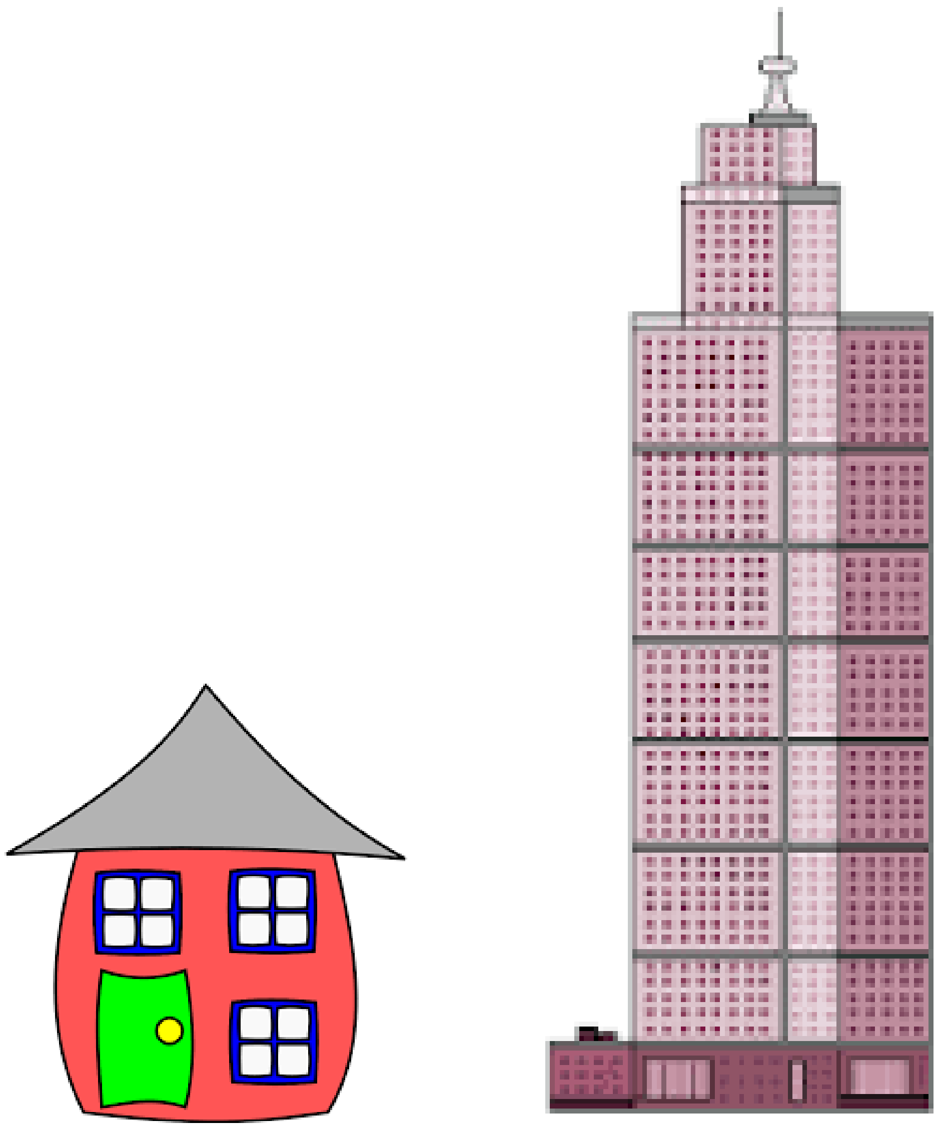
Issue:

doesn't address large changes (disasters, adversaries)

Changes and changes



Changes and changes



Long-term risks of AI: unknown unknowns

What's the right specification?

Specification: standard machine learning

Input: training data

Output: model that does obtains low **expected** test error

Is **expected** test error enough?

What's the right specification?

Specification: standard machine learning

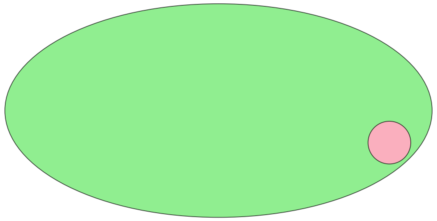
Input: training data

Output: model that does obtains low **expected** test error

Is **expected** test error enough?

Scenario:

- Err on 1% on instances
- Agents maximize, adversaries minimize, could drive us there!



New specification 1/2

Fereshte Khani

[ACL 2016]

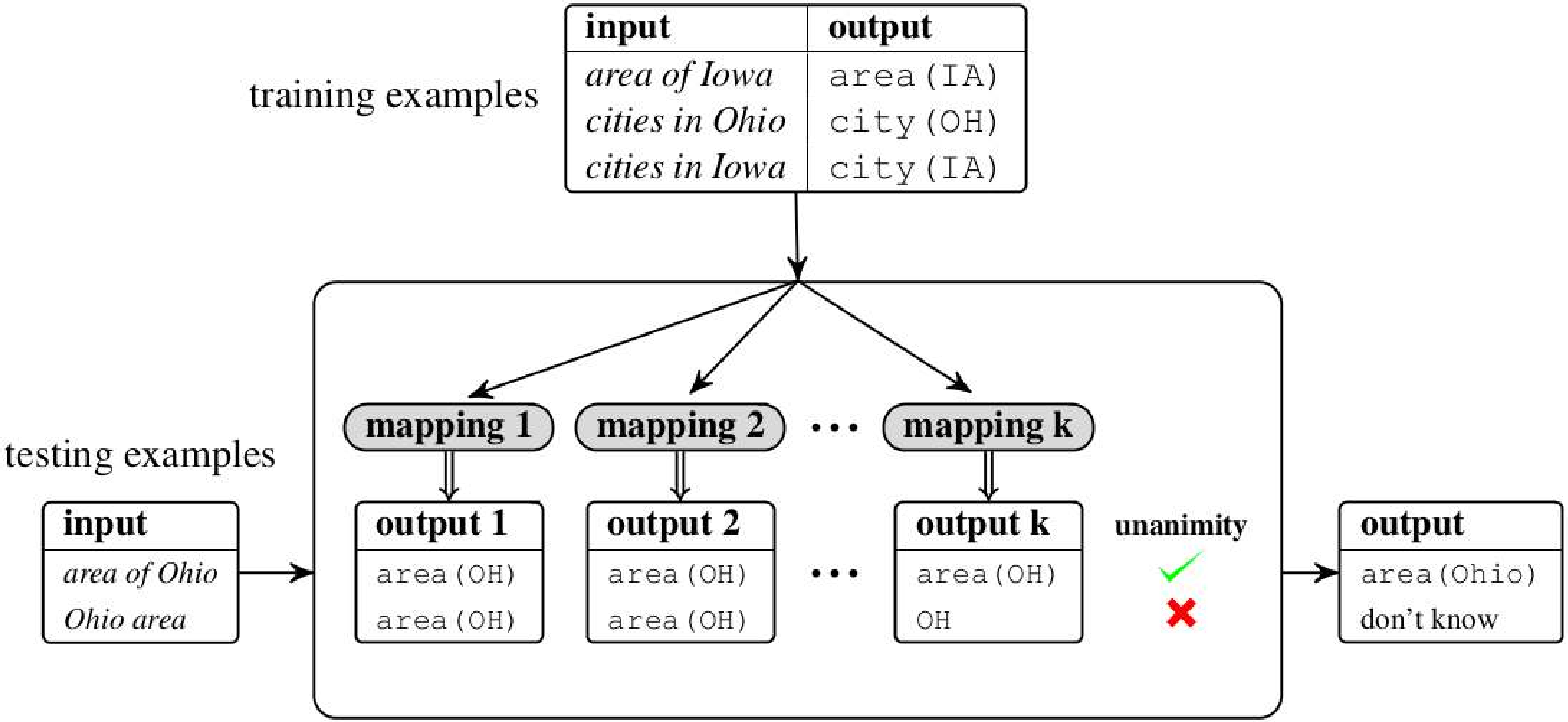
Specification: selective prediction

Input: training data

Output: model that outputs correct answer or "don't know"

Previous work: Chow (1970); Tortorella (2000); El-Yaniv & Wiener (2010); Balsubramani (2016)

Unanimous prediction



Assumption: exists mapping with zero error

Unanimous prediction

$$S = \begin{array}{c|c} & \textit{area} \quad \textit{of} \quad \textit{Ohio} \quad \textit{cities} \quad \textit{in} \quad \textit{Iowa} \\ \begin{array}{r} \textit{area of Iowa} \\ \textit{cities in Ohio} \\ \textit{cities in Iowa} \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

$$M = \begin{array}{c|c} & \texttt{area} \quad \texttt{city} \quad \texttt{OH} \quad \texttt{IA} \\ \begin{array}{r} \textit{area} \\ \textit{of} \\ \textit{Ohio} \\ \textit{cities} \\ \textit{in} \\ \textit{Iowa} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$T = \begin{array}{c|c} & \texttt{area} \quad \texttt{city} \quad \texttt{OH} \quad \texttt{IA} \\ \begin{array}{r} \texttt{area(IA)} \\ \texttt{city(OH)} \\ \texttt{city(IA)} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

Unanimous prediction

$$S = \begin{array}{r l} & \begin{array}{cccccc} \textit{area} & \textit{of} & \textit{Ohio} & \textit{cities} & \textit{in} & \textit{Iowa} \end{array} \\ \begin{array}{r} \textit{area of Iowa} \\ \textit{cities in Ohio} \\ \textit{cities in Iowa} \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad M = \begin{array}{r l} & \begin{array}{cccc} \texttt{area} & \texttt{city} & \texttt{OH} & \texttt{IA} \end{array} \\ \begin{array}{r} \textit{area} \\ \textit{of} \\ \textit{Ohio} \\ \textit{cities} \\ \textit{in} \\ \textit{Iowa} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \quad T = \begin{array}{r l} & \begin{array}{cccc} \texttt{area} & \texttt{city} & \texttt{OH} & \texttt{IA} \end{array} \\ \begin{array}{r} \texttt{area(IA)} \\ \texttt{city(OH)} \\ \texttt{city(IA)} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

Models consistent with training data:

$$\mathcal{C} = \{M \geq 0 : SM = T\}$$

Unanimous prediction

$$S = \begin{array}{rc} & \begin{array}{cccccc} \textit{area} & \textit{of} & \textit{Ohio} & \textit{cities} & \textit{in} & \textit{Iowa} \end{array} \\ \begin{array}{r} \textit{area of Iowa} \\ \textit{cities in Ohio} \\ \textit{cities in Iowa} \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad M = \begin{array}{rc} & \begin{array}{cccc} \texttt{area} & \texttt{city} & \texttt{OH} & \texttt{IA} \end{array} \\ \begin{array}{r} \textit{area} \\ \textit{of} \\ \textit{Ohio} \\ \textit{cities} \\ \textit{in} \\ \textit{Iowa} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \quad T = \begin{array}{rc} & \begin{array}{cccc} \texttt{area} & \texttt{city} & \texttt{OH} & \texttt{IA} \end{array} \\ \begin{array}{r} \texttt{area(IA)} \\ \texttt{city(OH)} \\ \texttt{city(IA)} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

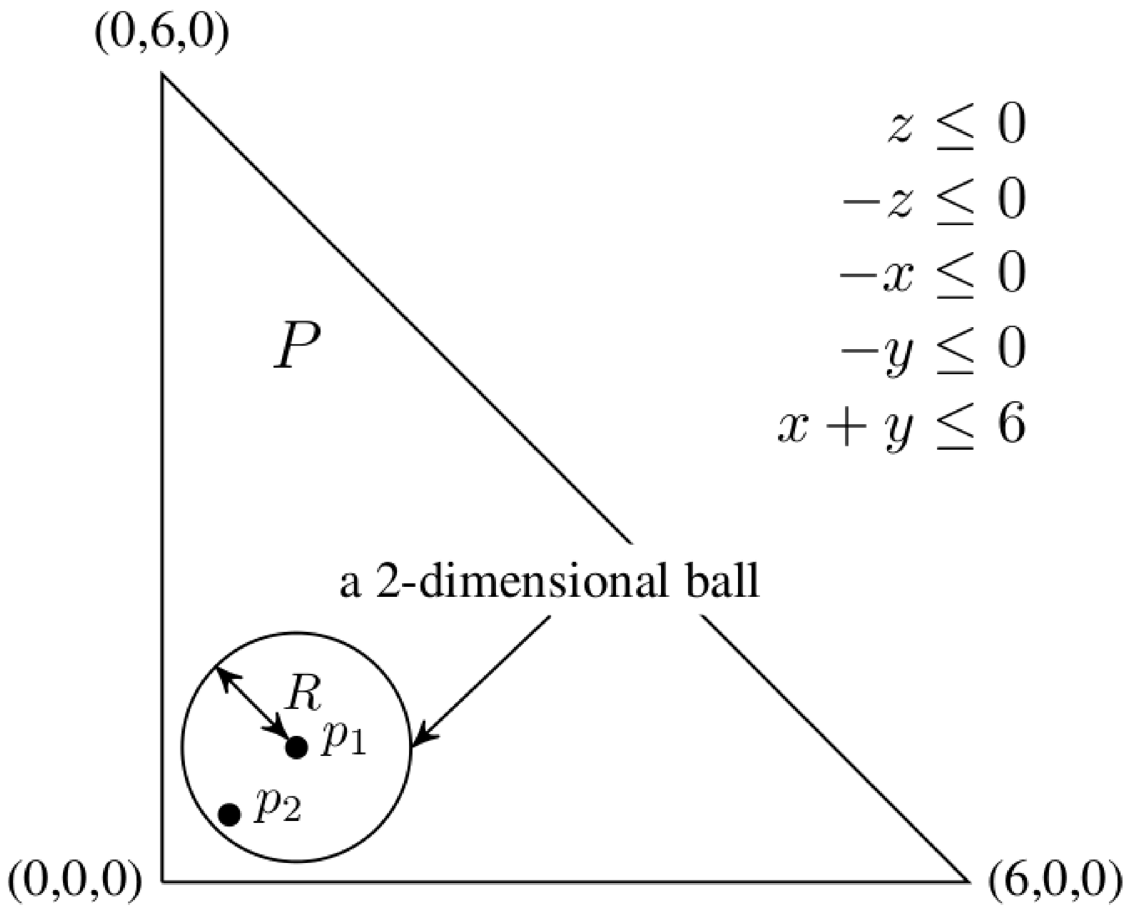
Models consistent with training data:

$$\mathcal{C} = \{M \geq 0 : SM = T\}$$

Challenge:

Checking all consistent $M \in \mathcal{C}$ is slow...

Fast two point scheme

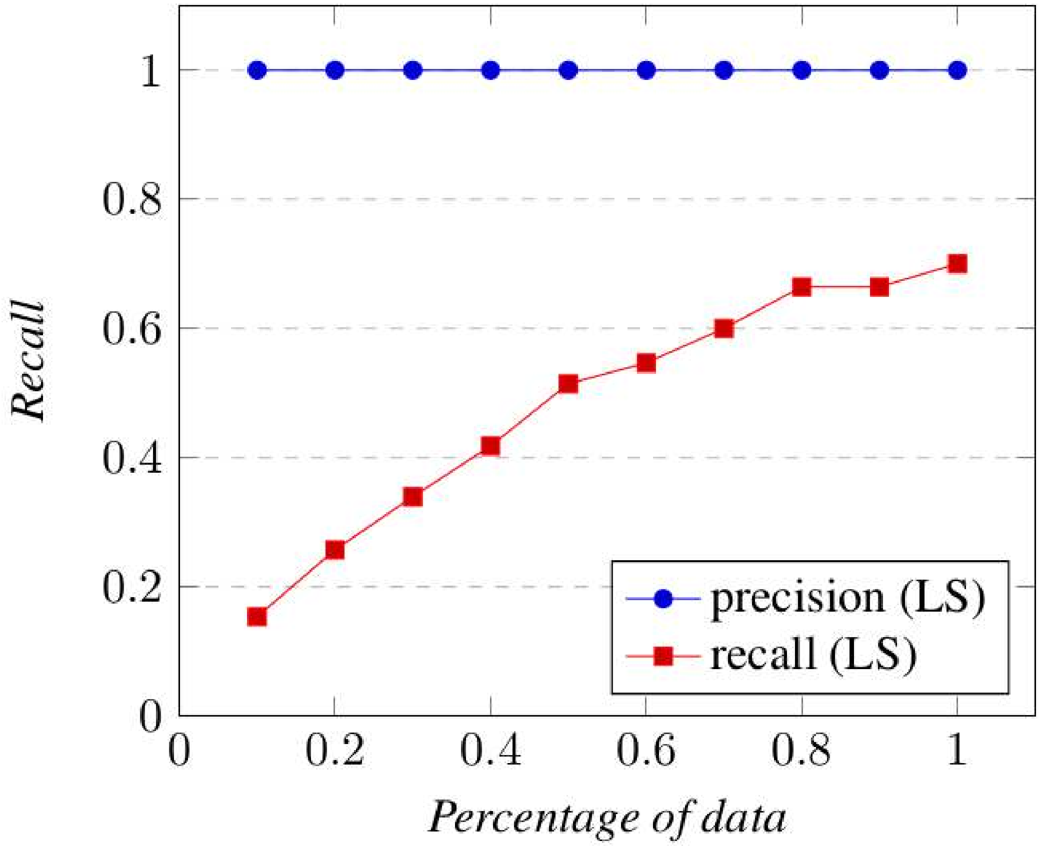


- Choose $M_1, M_2 \in \mathcal{C}$ randomly enough
- Return "don't know" iff M_1 and M_2 disagree

Experimental results

- GeoQuery semantic parsing dataset (800 train, 280 test)

What is the population of Texas?



New specification 2/2

Jacob Steinhardt

[NIPS 2016]

Specification: unsupervised risk estimation

Input: **unlabeled** examples and model

Output: estimate of **labeled** accuracy

Previous work: Donmez et al. (2010); Dawid/Skene (1979); Zhang et al. (2014); Jaffe et al. (2015); Balasubramanian et al. (2011)

Is this possible?

model θ

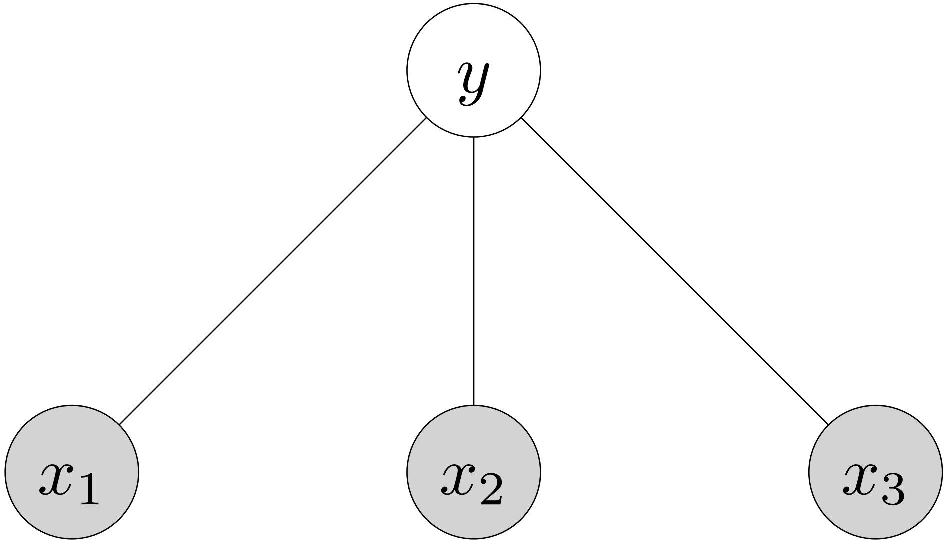
? ? ? ? ?

$x^{(1)}$ $x^{(2)}$ $x^{(3)}$ $x^{(4)}$ $x^{(5)}$

Compute $\mathbb{E}[\text{loss}(x, y; \theta)]$

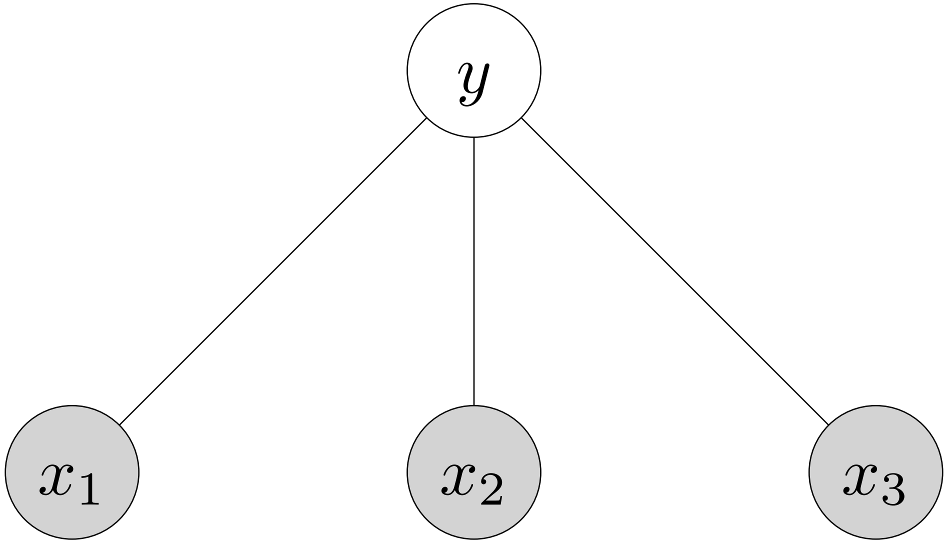
Assumptions

Conditional independence:



Assumptions

Conditional independence:

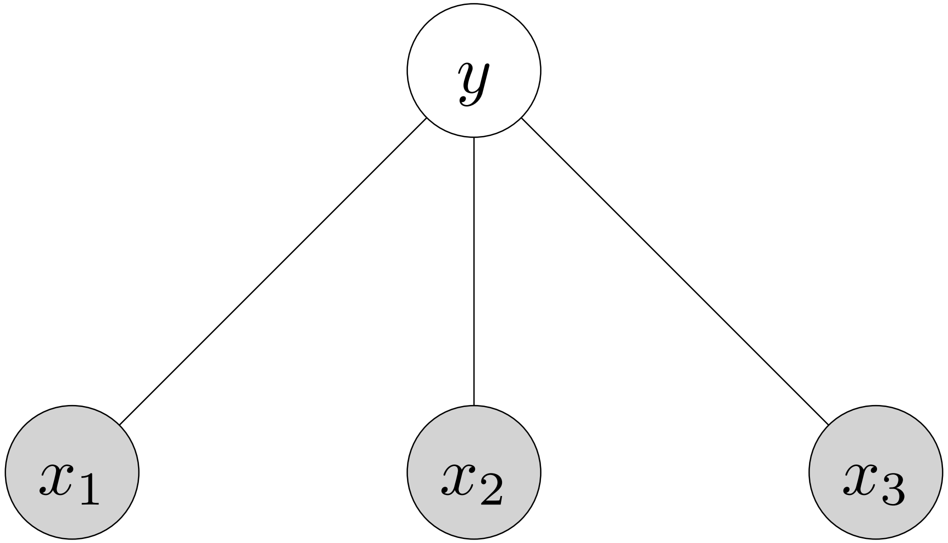


Loss function decomposes:

$$A(x;\theta) - f_1(x_1, y; \theta) - f_2(x_2, y; \theta) - f_3(x_3, y; \theta)$$

Assumptions

Conditional independence:

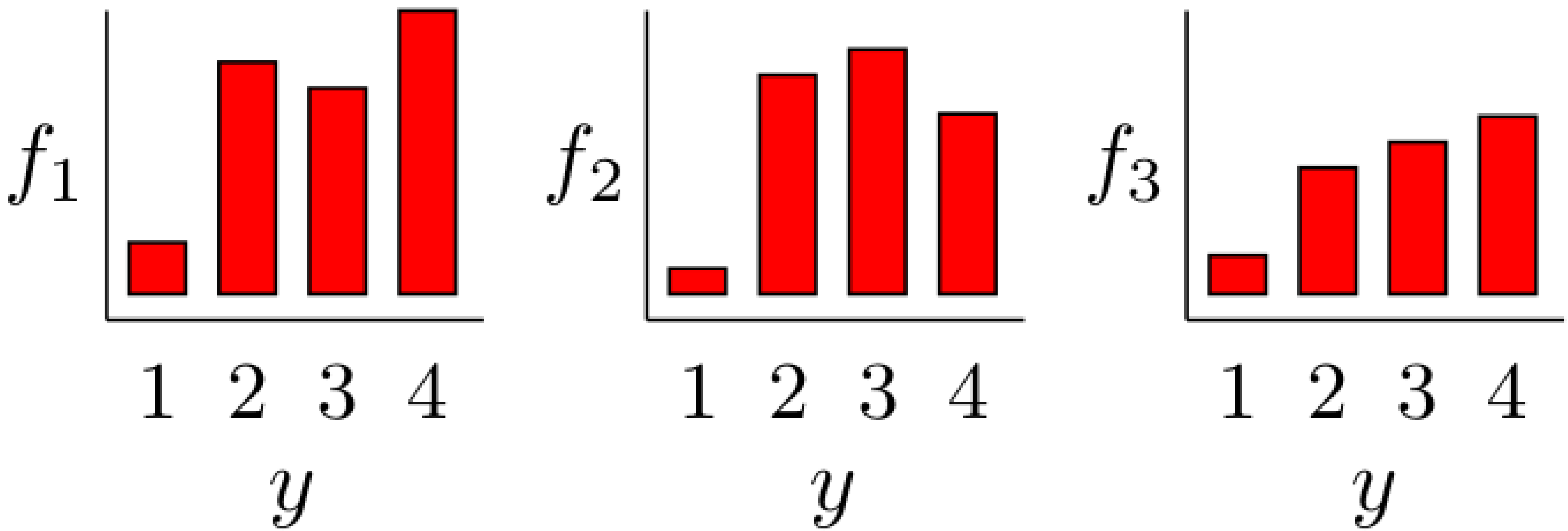


Loss function decomposes:

$$A(x;\theta) - f_1(x_1, y;\theta) - f_2(x_2, y;\theta) - f_3(x_3, y;\theta)$$

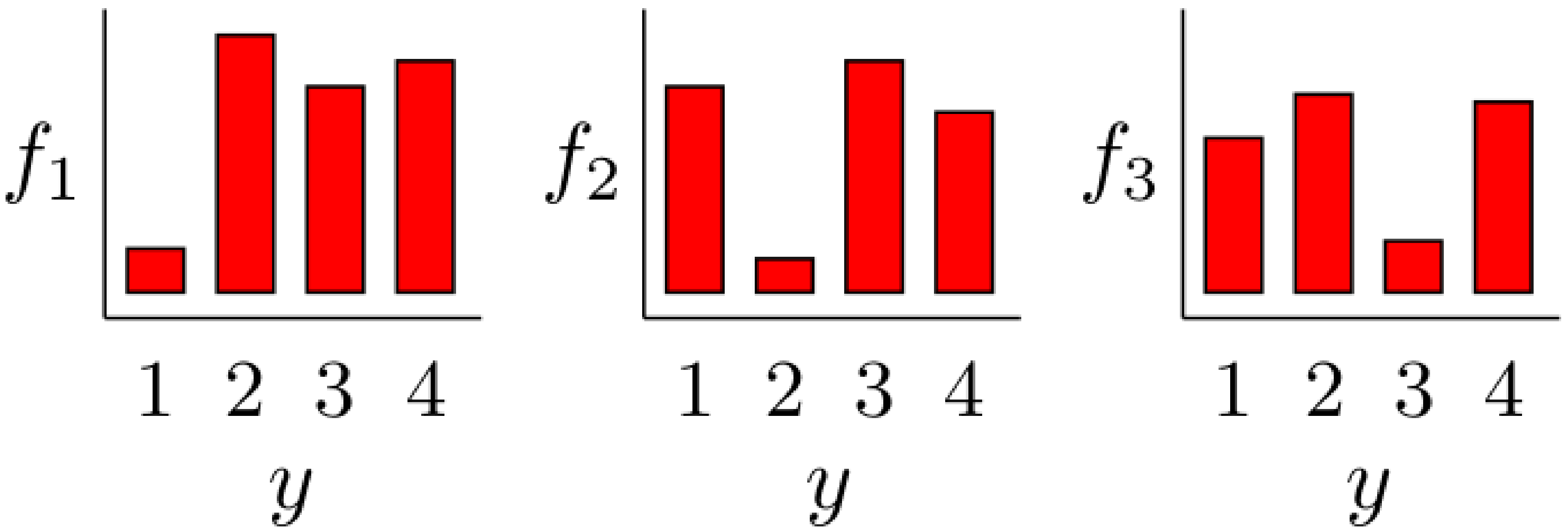
only conditional independence structure

Intuition



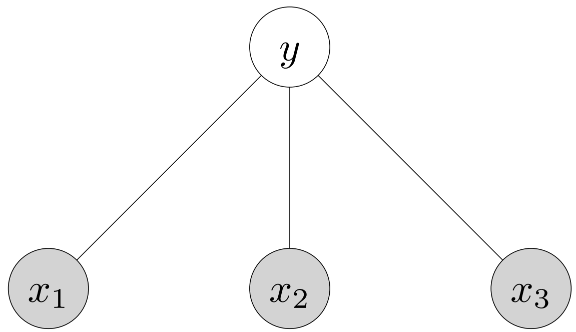
Three views agree $\rightarrow$ (probably) low error

Intuition



Three views disagree $\rightarrow$ high error

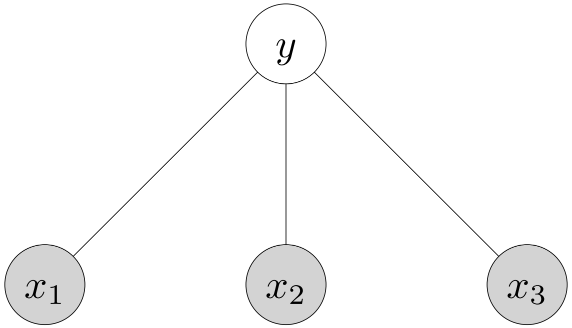
Tensor factorization



(k labels, views $v = 1, 2, 3$)

$$\begin{bmatrix} f_v(x, 1) \\ \dots \\ f_v(x, k) \end{bmatrix}$$

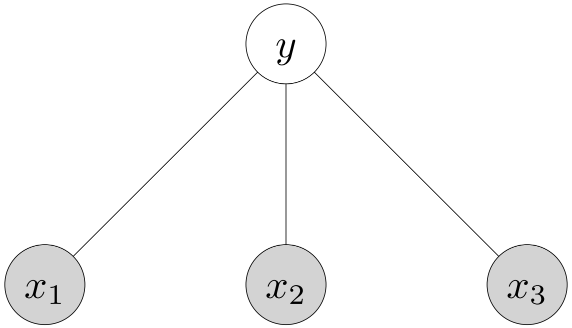
Tensor factorization



(k labels, views $v = 1, 2, 3$)

$$M_v = \begin{bmatrix} \mathbb{E}[f_v(x,1) \mid y=1] & \dots & \mathbb{E}[f_v(x,1) \mid y=k] \\ \dots & & \dots \\ \mathbb{E}[f_v(x,k) \mid y=1] & \dots & \mathbb{E}[f_v(x,k) \mid y=k] \end{bmatrix}$$

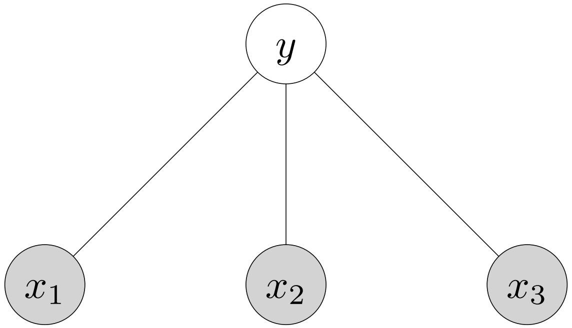
Tensor factorization



(k labels, views $v = 1, 2, 3$)

- Observe $\mathbb{E}[f_1(x, a) f_2(x, b)]$

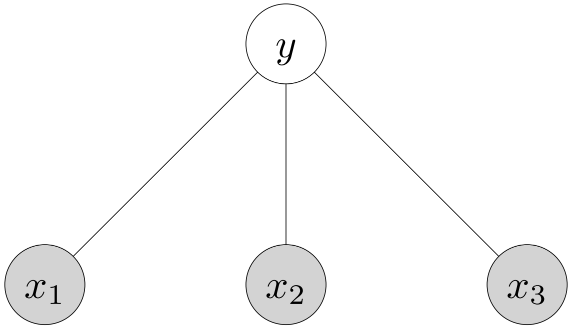
Tensor factorization



(k labels, views $v = 1, 2, 3$)

- Observe $\mathbb{E}[f_1(x, a) f_2(x, b) f_3(x, c)]$

Tensor factorization



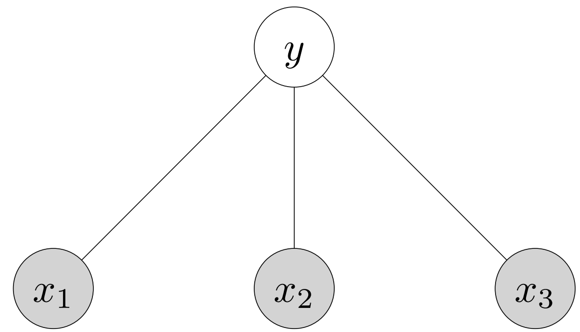
(k labels, views $v = 1, 2, 3$)

- Observe $\mathbb{E}[f_1(x,a) f_2(x,b) f_3(x,c)]$
- Perform tensor factorization to obtain

$$M_{vba} = \mathbb{E}[f_v(x,b) \mid y = a]$$

[Anandkumar et al., 2013]

Tensor factorization



(k labels, views $v = 1, 2, 3$)

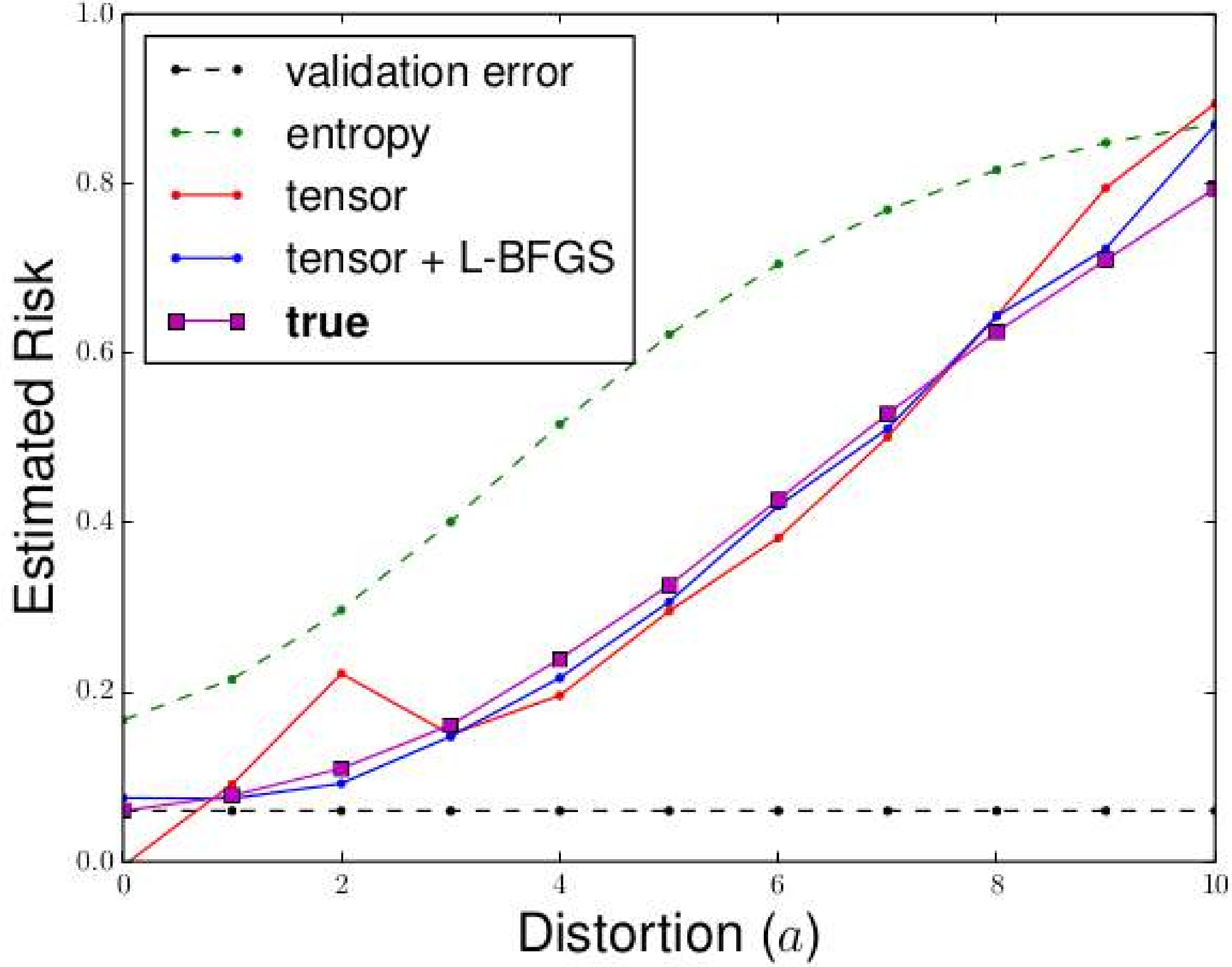
- Observe $\mathbb{E}[f_1(x, a) f_2(x, b) f_3(x, c)]$
- Perform tensor factorization to obtain

$$M_{vba} = \mathbb{E}[f_v(x, b) \mid y = a]$$

- Use to compute risk (up to label permutation)

$$\mathbb{E}[A(x; \theta) - f_1(x_1, y; \theta) - f_2(x_2, y; \theta) - f_3(x_3, y; \theta)]$$

Results


validation error
entropy
tensor
tensor + L-BFGS
true
Estimated Risk
Distortion (a)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

Discussion

- Maximize expected accuracy $\Rightarrow$ selective prediction, unsupervised risk estimation

- Key question: Can we weaken the assumptions?

Code and data

worksheets.codalab.org

Collaborators

Fereshte Khani

Jacob Steinhardt

Thank you!