Learning from Measurements in Exponential Families

ICML – Montreal

June 16, 2009

Percy Liang   Michael Jordan   Dan Klein

University of California Berkeley

The big picture

target predictor p^*

human

The big picture

target predictor p^* human

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

The big picture

target predictor p^* human $\longrightarrow$ information $\longrightarrow$ learning algorithm $\longrightarrow$ learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

The big picture

target predictor p^* human $\longrightarrow$ information $\longrightarrow$ learning algorithm $\longrightarrow$ learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

Types of information:

Labeled examples (specific) [standard supervised learning]

The big picture

target predictor p^* → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	…			
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	…			
	AVAIL	AVAIL	AVAIL	…	SIZE	SIZE	SIZE	SIZE	…
	Available	*July*	*1*	…	*2*	*bedroom*	*1*	*bath*	…

Types of information:

Labeled examples (specific) [standard supervised learning]

Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

The big picture

target predictor p^* → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

Types of information:

Labeled examples (specific) [standard supervised learning]

Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

Measurements: our unifying framework

The big picture

target predictor p^* → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

Types of information:

Labeled examples (specific) [standard supervised learning]

Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

Measurements: our unifying framework

Outline:

1. Coherently learn from diverse measurements

The big picture

target predictor p^* → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

y:	FEAT	FEAT	FEAT	FEAT	FEAT	...
x:	*View*	*of*	*Los*	*Gatos*	*Foothills*	...

AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
Available	*July*	*1*	...	*2*	*bedroom*	*1*	*bath*	...

Types of information:

Labeled examples (specific) [standard supervised learning]

Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

Measurements: our unifying framework

Outline:

1. Coherently learn from diverse measurements
2. Actively select the best measurements

Measurements

X_1 , Y_1
X_2 , Y_2
X_3 , Y_3
... ...
X_i , Y_i
... ...
X_n , Y_n

Measurements

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

$$\sigma(\; X_1 \;,\; Y_1 \;)$$
$$\sigma(\; X_2 \;,\; Y_2 \;)$$
$$\sigma(\; X_3 \;,\; Y_3 \;)$$
$$\ldots \qquad \ldots$$
$$\sigma(\; X_i \;,\; Y_i \;)$$
$$\ldots \qquad \ldots$$
$$\sigma(\; X_n \;,\; Y_n \;)$$

Measurements

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

Measurement values: $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}$$

$$\begin{array}{l}
\sigma(\ X_1\ ,\ Y_1\ ) \\
\sigma(\ X_2\ ,\ Y_2\ ) \\
\sigma(\ X_3\ ,\ Y_3\ ) \\
\quad \ldots \quad \ldots \\
\sigma(\ X_i\ ,\ Y_i\ ) \\
\quad \ldots \quad \ldots \\
\sigma(\ X_n\ ,\ Y_n\ ) \\
+ \text{ noise} \\
\hline
\tau
\end{array}$$

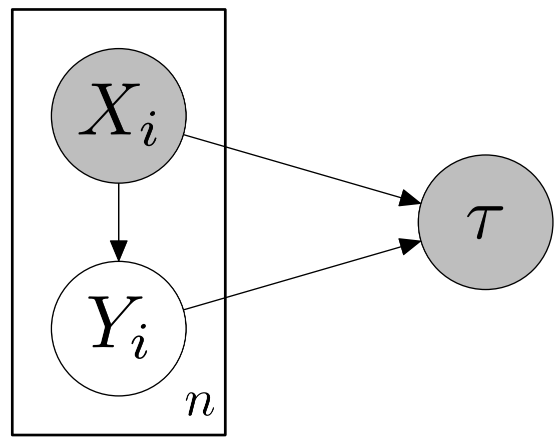
Measurements

$$
\begin{array}{l}
\quad \sigma(\ X_1\ ,\ Y_1\ ) \\
\quad \sigma(\ X_2\ ,\ Y_2\ ) \\
\quad \sigma(\ X_3\ ,\ Y_3\ ) \\
\qquad\quad \dots \qquad \dots \\
\quad \sigma(\ X_i\ ,\ Y_i\ ) \\
\qquad\quad \dots \qquad \dots \\
\quad \sigma(\ X_n\ ,\ Y_n\ ) \\
+\ \text{noise} \\
\hline
\quad \tau
\end{array}
$$

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

Measurement values: $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}$$



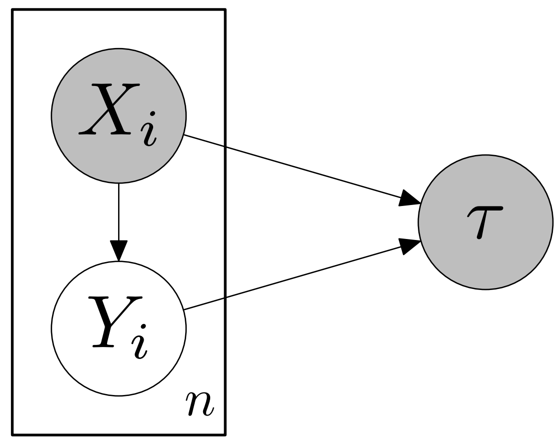
Measurements

$$
\begin{array}{l}
\sigma(\ X_1\ ,\ Y_1\ ) \\
\sigma(\ X_2\ ,\ Y_2\ ) \\
\sigma(\ X_3\ ,\ Y_3\ ) \\
\quad \dots \quad \dots \\
\sigma(\ X_i\ ,\ Y_i\ ) \\
\quad \dots \quad \dots \\
\sigma(\ X_n\ ,\ Y_n\ ) \\
+ \text{ noise} \\
\hline
\tau
\end{array}
$$

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

Measurement values: $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}$$



Set σ to reveal various types of information about Y through τ

Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ \ldots, y = * \ * \ * \ \ldots]$$

Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los ...}, y = * * * \; ...]$$

Partially-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los ...}, y_1 = *]$$

Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y = * * * \ldots]$$

Partially-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y_1 = *]$$

Labeled predicate:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \textit{View}, y_i = *]$$

Examples of measurements

Fully-labeled example:

$$\sigma_j(x,y) = \mathbb{I}[x = \textit{View of Los ...}, y = *\ *\ *\ ...]$$

Partially-labeled example:

$$\sigma_j(x,y) = \mathbb{I}[x = \textit{View of Los ...}, y_1 = *]$$

Labeled predicate:

$$\sigma_j(x,y) = \sum_i \mathbb{I}[x_i = \textit{View}, y_i = *]$$

Label proportions:

$$\sigma_j(x,y) = \sum_i \mathbb{I}[y_i = *]$$

Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y = * * * \ldots]$$

Partially-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y_1 = *]$$

Labeled predicate:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \textit{View}, y_i = *]$$

Label proportions:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = *]$$

Label preference:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}]$$

Examples of measurements

Fully-labeled example:

$$\sigma_j(x,y) = \mathbb{I}[x = \textit{View of Los} \ldots, y = * * * \ldots]$$

Partially-labeled example:

$$\sigma_j(x,y) = \mathbb{I}[x = \textit{View of Los} \ldots, y_1 = *]$$

Labeled predicate:

$$\sigma_j(x,y) = \sum_i \mathbb{I}[x_i = \textit{View}, y_i = *]$$

Label proportions:

$$\sigma_j(x,y) = \sum_i \mathbb{I}[y_i = *]$$

Label preference:

$$\sigma_j(x,y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}]$$

Can get measurement values τ without looking at all examples

Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y = * * * \ldots]$$

Partially-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of Los} \ldots, y_1 = *]$$

Labeled predicate:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \textit{View}, y_i = *]$$

Label proportions:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = *]$$

Label preference:

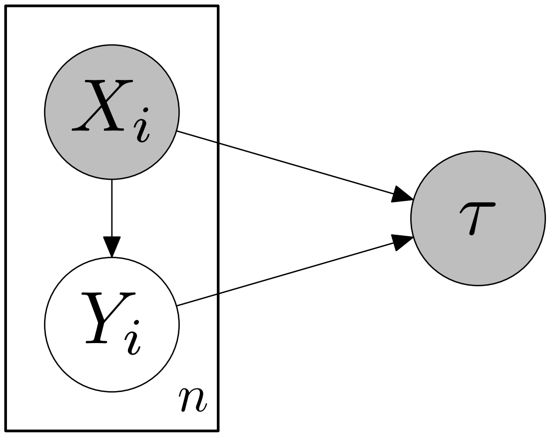
$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}]$$

Can get measurement values τ without looking at all examples

Next: How to combine these diverse measurements coherently?

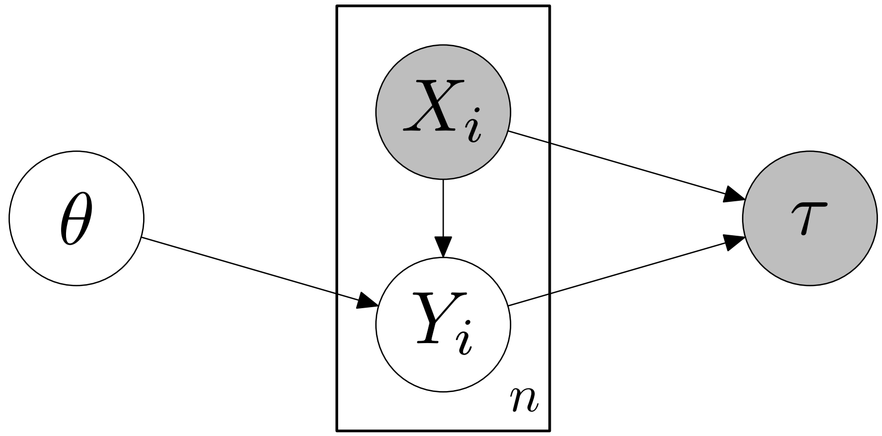
Prediction model

Bayesian framework:



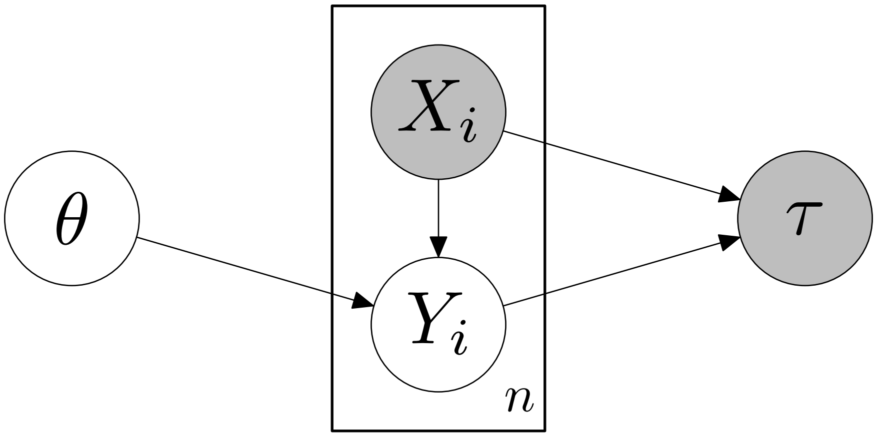
Prediction model

Bayesian framework:



Prediction model

Bayesian framework:

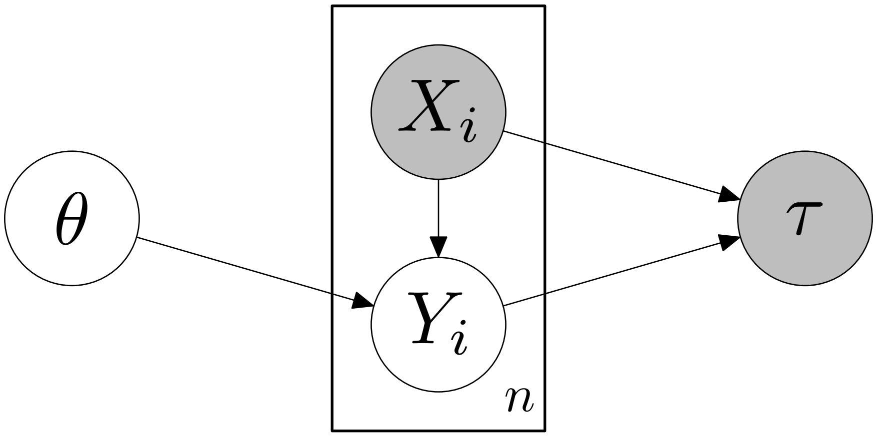


Exponential families:

$$p_\theta(y \mid x) = \exp\{\langle \phi(x, y), \theta \rangle - A(\theta; x)\}$$

Prediction model

Bayesian framework:



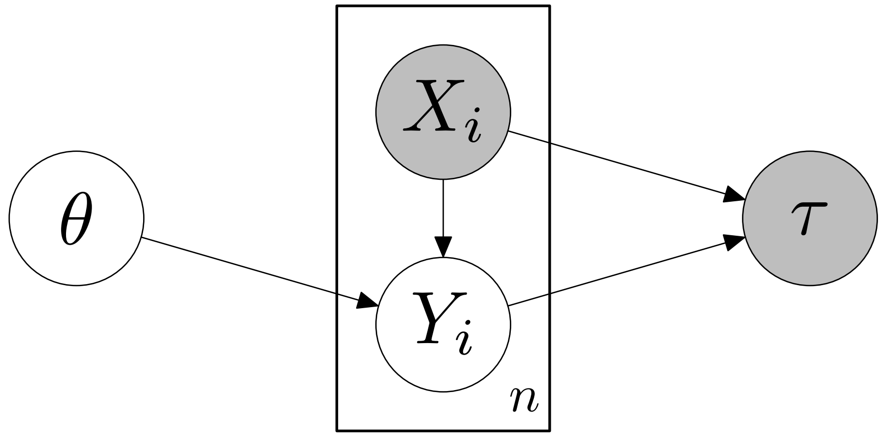
Exponential families:

$$p_\theta(y \mid x) = \exp\{\langle \phi(x, y), \theta \rangle - A(\theta; x)\}$$

$\phi(x, y) \in \mathbb{R}^d$: model features

Prediction model

Bayesian framework:



Exponential families:

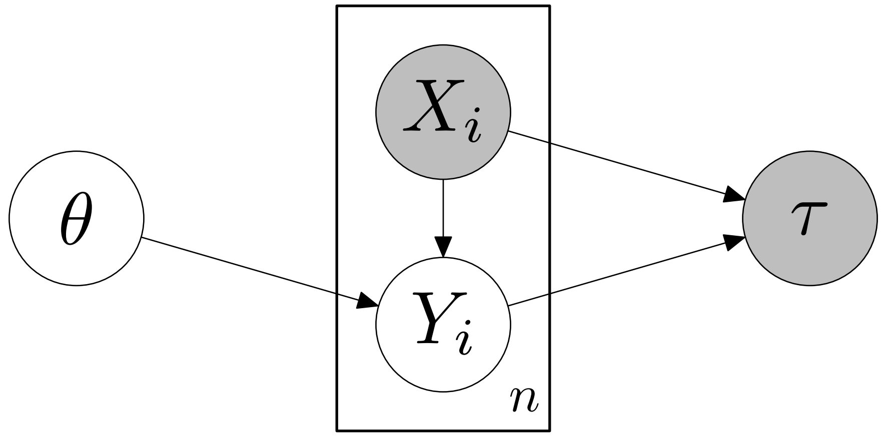
$$p_\theta(y \mid x) = \exp\{\langle \phi(x,y), \theta \rangle - A(\theta; x)\}$$

$\phi(x,y) \in \mathbb{R}^d$: model features

$\theta \in \mathbb{R}^d$: model parameters

Prediction model

Bayesian framework:



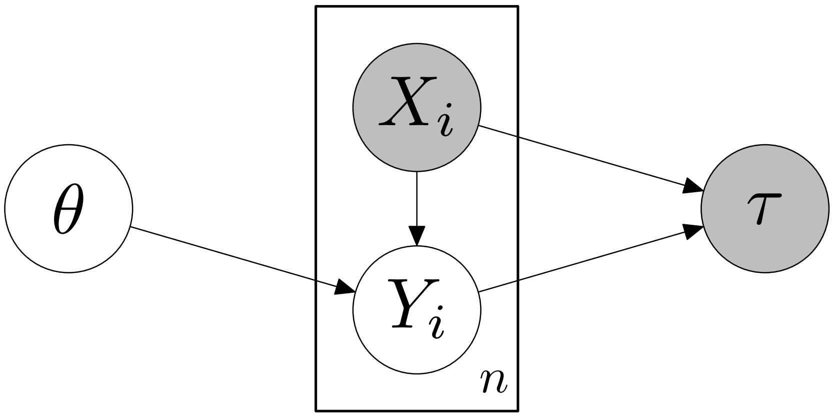
Exponential families:

$$p_\theta(y \mid x) = \exp\{\langle \phi(x,y), \theta \rangle - A(\theta; x)\}$$

$\phi(x,y) \in \mathbb{R}^d$: model features
$\theta \in \mathbb{R}^d$: model parameters
$A(\theta; x) = \int \exp\{\langle \phi(x,y), \theta \rangle\} dy$: log-partition function

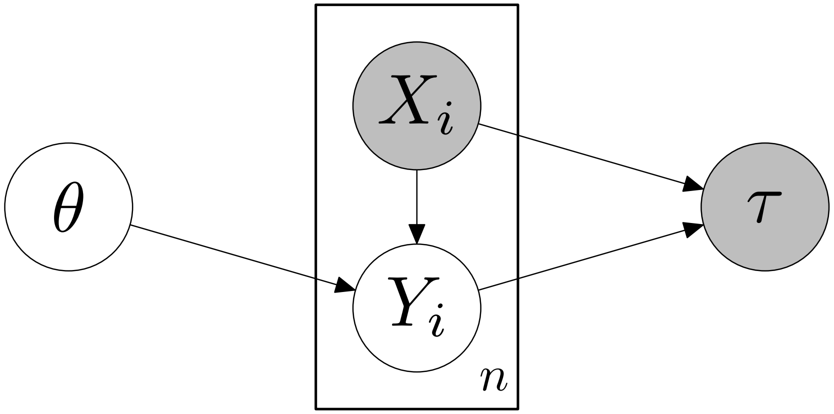
Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$



Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$

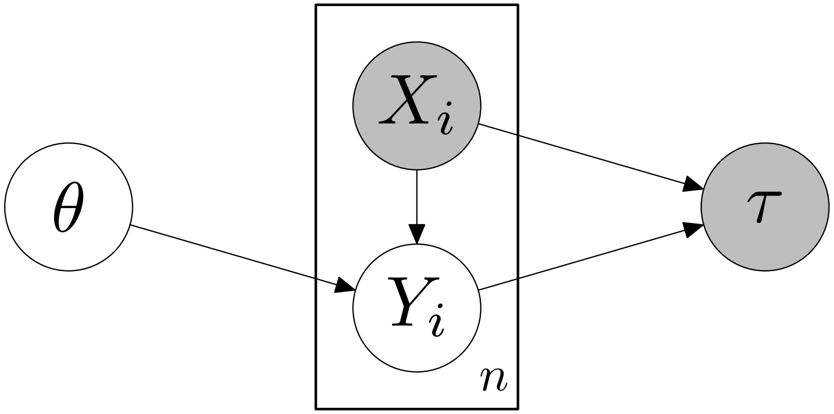


Variational formulation:

$$\min_{q \in \mathcal{Q}_{Y\theta}} \mathsf{KL}\left(q(Y,\theta) \,||\, p(\theta, Y \mid \tau, X)\right)$$

Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$



Variational formulation:

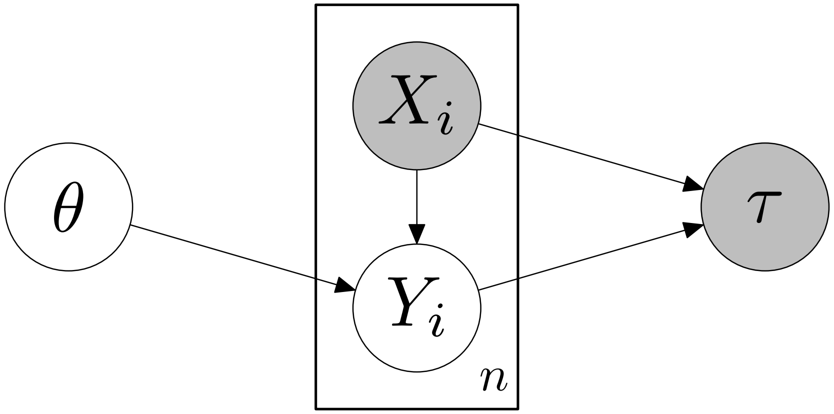
$$\min_{q \in \mathcal{Q}_{Y\theta}} \mathsf{KL}\left(q(Y,\theta) \,||\, p(\theta, Y \mid \tau, X)\right)$$

Approximations:

- $\mathcal{Q}_{Y\theta}$: mean-field factorization of $q(Y)$ and degenerate $\tilde{\theta}$

Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$



Variational formulation:

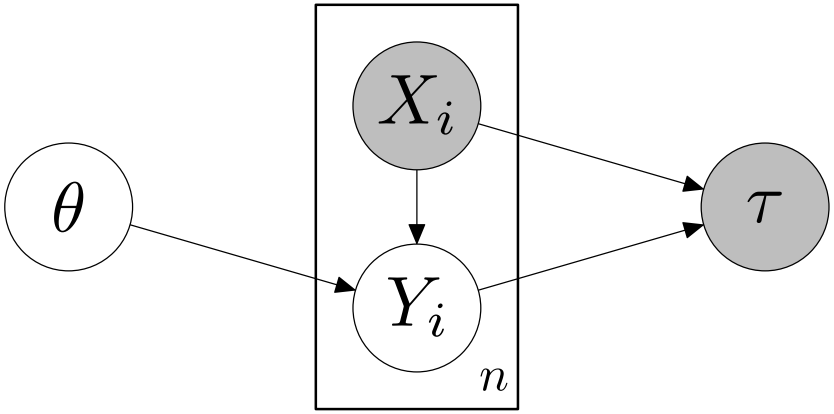
$$\min_{q \in \mathcal{Q}_{Y\theta}} \mathsf{KL}\left(q(Y,\theta) \,||\, p(\theta, Y \mid \tau, X)\right)$$

Approximations:

- $\mathcal{Q}_{Y\theta}$: mean-field factorization of $q(Y)$ and degenerate $\tilde{\theta}$
- KL: measurements only hold in expectation (w.r.t. $q(Y)$)

Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$



Variational formulation:

$$\min_{q \in \mathcal{Q}_{Y\theta}} \mathsf{KL}\left(q(Y,\theta) \,||\, p(\theta, Y \mid \tau, X)\right)$$

Approximations:

- $\mathcal{Q}_{Y\theta}$: mean-field factorization of $q(Y)$ and degenerate $\tilde{\theta}$
- KL: measurements only hold in expectation (w.r.t. $q(Y)$)

Algorithm:

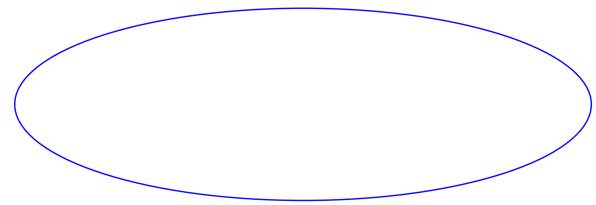
Apply Fenchel duality $\rightarrow$ saddlepoint problem

Take alternating stochastic gradient steps

Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



Information geometry viewpoint

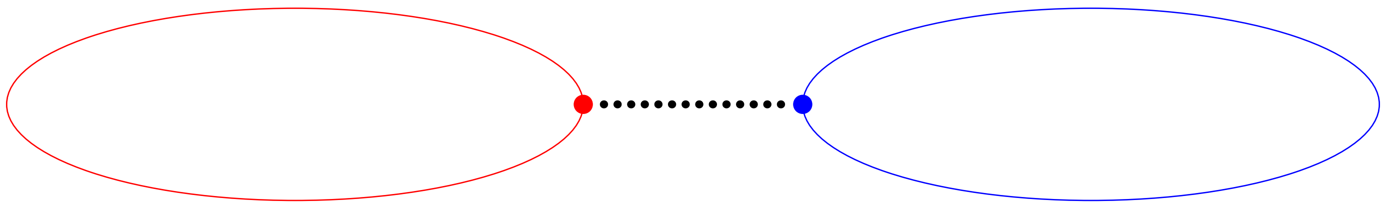
(assume zero measurement noise)

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\} \qquad \mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$

Information geometry viewpoint

(assume zero measurement noise)

$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\}$ $\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$

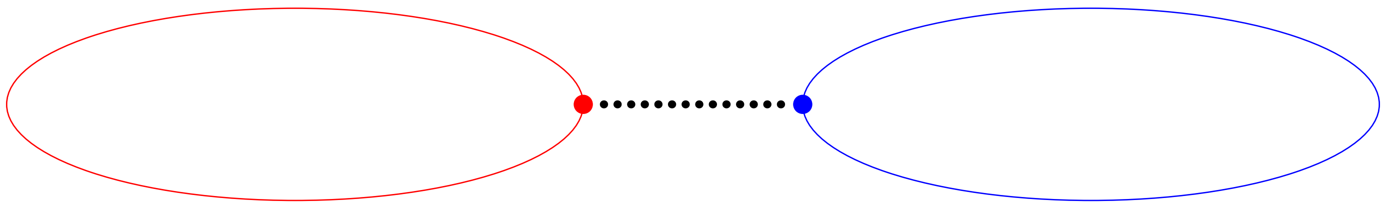


$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathsf{KL}\left(q \,||\, p\right)$$

Information geometry viewpoint

(assume zero measurement noise)

$$\color{red}{\mathcal{Q}} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\} \qquad \color{blue}{\mathcal{P}} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



$$\min_{\color{red}{q \in \mathcal{Q}}, \color{blue}{p \in \mathcal{P}}} \mathsf{KL}\left(\color{red}{q} \,||\, \color{blue}{p}\right)$$

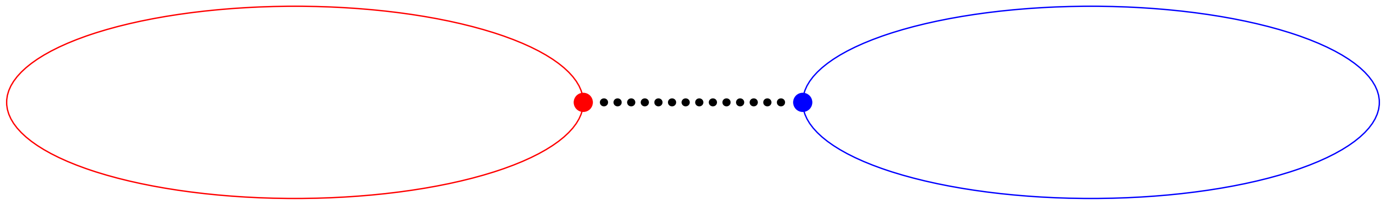
Interpretation:

Measurements shape $\color{red}{\mathcal{Q}}$ Find model in $\color{blue}{\mathcal{P}}$ with best fit

Information geometry viewpoint

(assume zero measurement noise)

$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\}$ $\qquad$ $\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$



$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathsf{KL}\left(q \,||\, p\right)$$

Interpretation:

Measurements shape $\mathcal{Q}$ $\qquad$ Find model in $\mathcal{P}$ with best fit

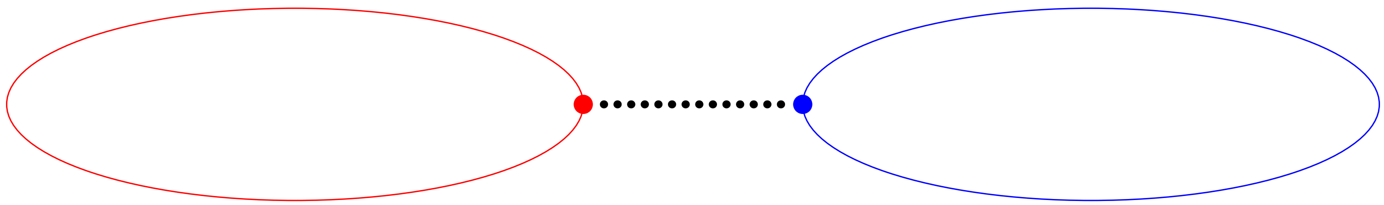
Two ways to recover supervised learning:

1. Measure $\sigma = \phi$: $\mathcal{P} \cap \mathcal{Q}$ is the unique solution

Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\} \qquad \mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathsf{KL}\left(q \,||\, p\right)$$

Interpretation:

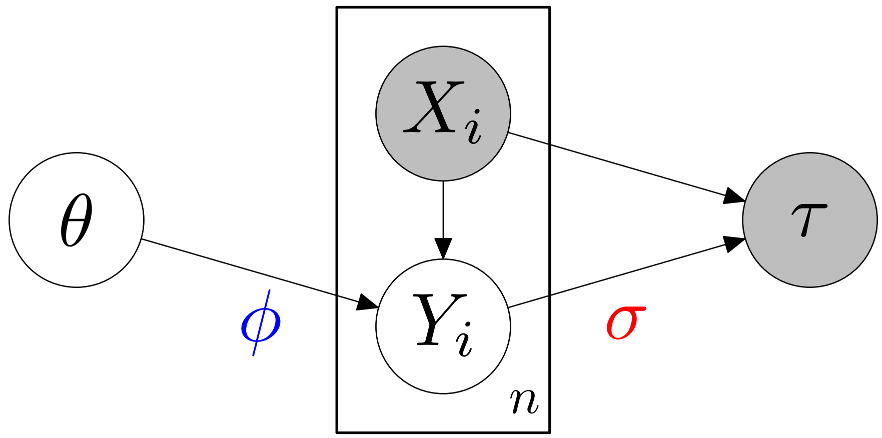
Measurements shape $\mathcal{Q}$ Find model in $\mathcal{P}$ with best fit

Two ways to recover supervised learning:

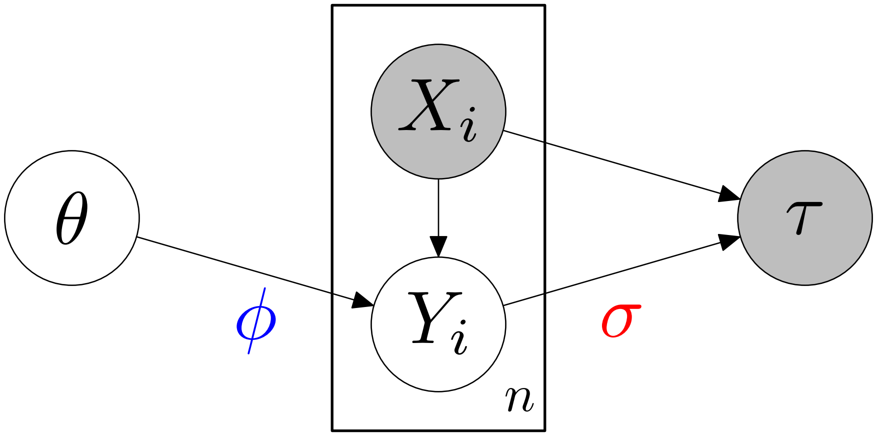
1. Measure $\sigma = \phi$: $\mathcal{P} \cap \mathcal{Q}$ is the unique solution
2. Measure $\sigma = \{\mathbb{I}[x = a, y = b]\}$:

   $\mathcal{Q} = \{$empirical distribution$\}$, project onto $\mathcal{P}$

Model features ϕ versus measurement features σ



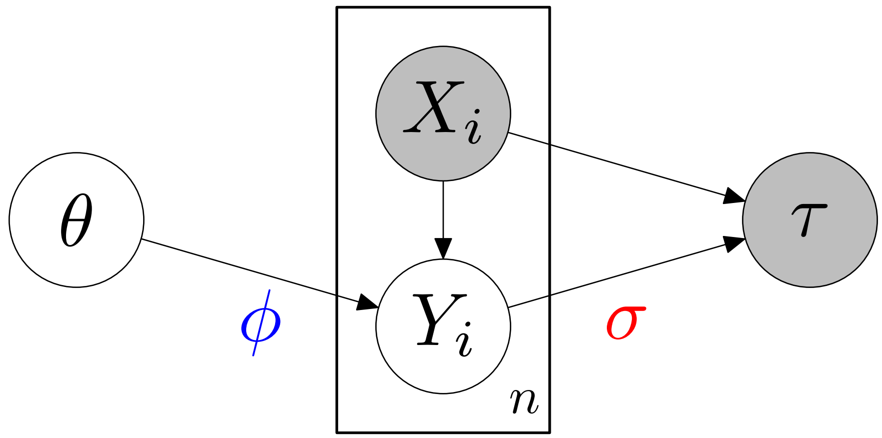
Model features ϕ versus measurement features σ



Guidelines:

To set σ, consider human (e.g., full labels)

Model features ϕ versus measurement features σ

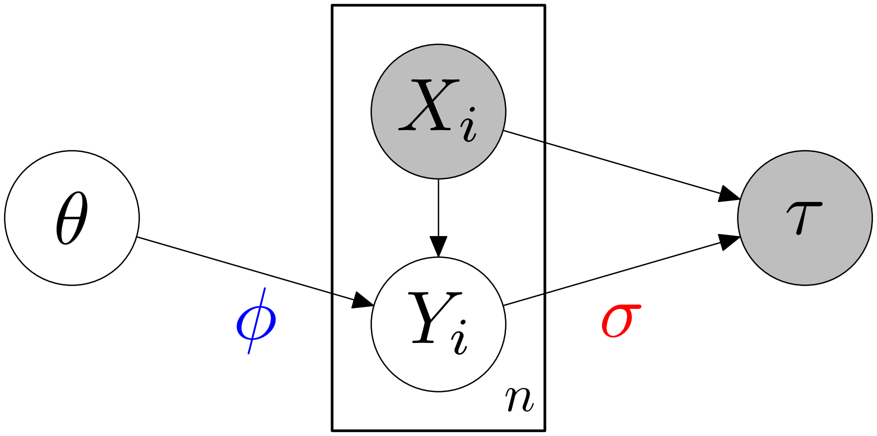


Guidelines:

To set σ, consider human (e.g., full labels)

To set ϕ, consider statistical generalization (e.g., word suffixes)

Model features ϕ versus measurement features σ



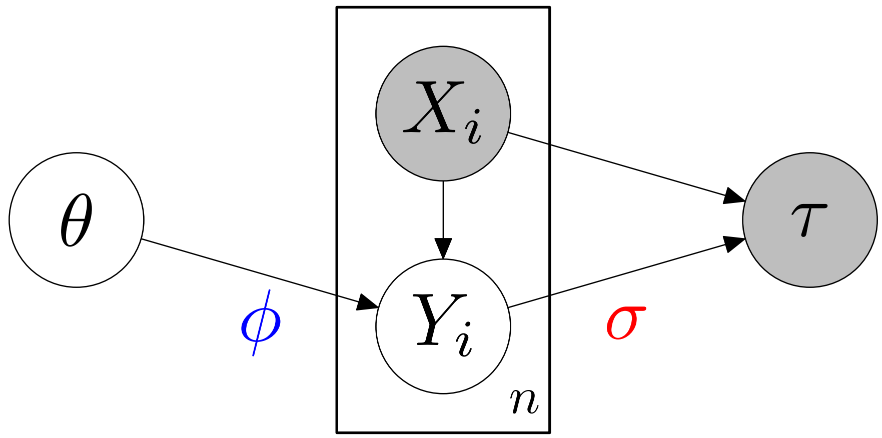
Guidelines:

To set σ, consider human (e.g., full labels)

To set ϕ, consider statistical generalization (e.g., word suffixes)

Intuition: consider feature $f(x, y) = \mathbb{I}[x \in A, y = 1]$

Model features ϕ versus measurement features σ



Guidelines:

To set σ, consider human (e.g., full labels)

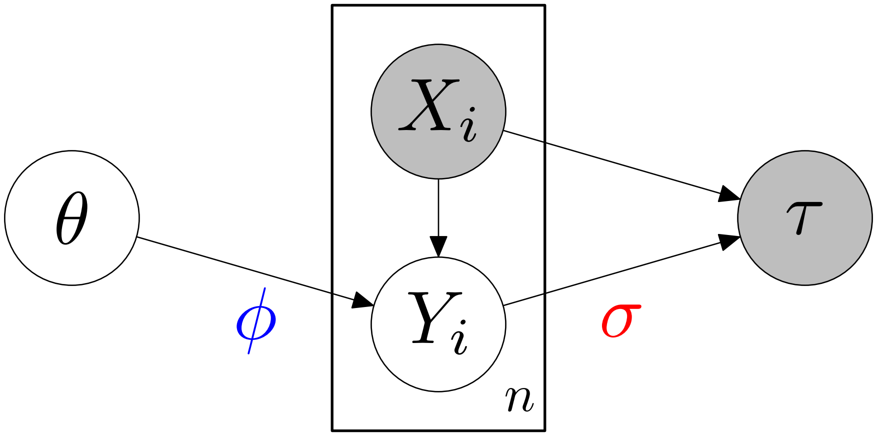
To set ϕ, consider statistical generalization (e.g., word suffixes)

Intuition: consider feature $f(x, y) = \mathbb{I}[x \in A, y = 1]$

If f is a measurement feature (direct):

"inputs in A should be labeled according to τ"

Model features ϕ versus measurement features σ



Guidelines:

To set σ, consider human (e.g., full labels)

To set ϕ, consider statistical generalization (e.g., word suffixes)

Intuition: consider feature $f(x, y) = \mathbb{I}[x \in A, y = 1]$

If f is a measurement feature (direct):

“inputs in A should be labeled according to τ”

If f is a model feature (indirect):

“inputs in A should be labeled similarly”

Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

Measurements:

- fully-labeled examples
- 33 labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \mathit{View}, y_i = \textsc{feat}]$)

Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

Measurements:

- fully-labeled examples
- 33 labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \mathit{View}, y_i = \textsc{feat}]$)

Per-position test accuracy (on 100 examples):

# labeled examples	10	25	100
General Expectation Criteria	74.6	77.2	80.5
Constraint-Driven Learning	**74.7**	**78.5**	81.7
Measurements	71.4	76.5	**82.5**

Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

Measurements:

- fully-labeled examples
- 33 labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \mathit{View}, y_i = \textsc{feat}]$)

Per-position test accuracy (on 100 examples):

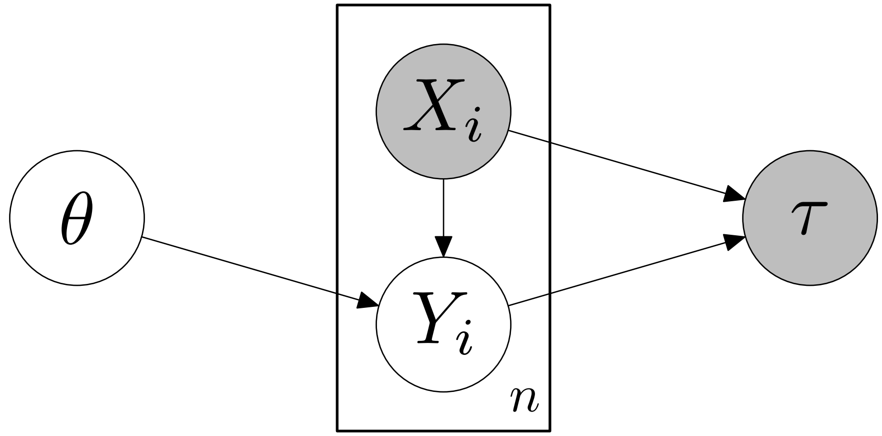
# labeled examples	10	25	100
General Expectation Criteria	74.6	77.2	80.5
Constraint-Driven Learning	**74.7**	**78.5**	81.7
Measurements	71.4	76.5	**82.5**

Able to integrate labeled examples and predicates gracefully

So far: given measurements, how to learn

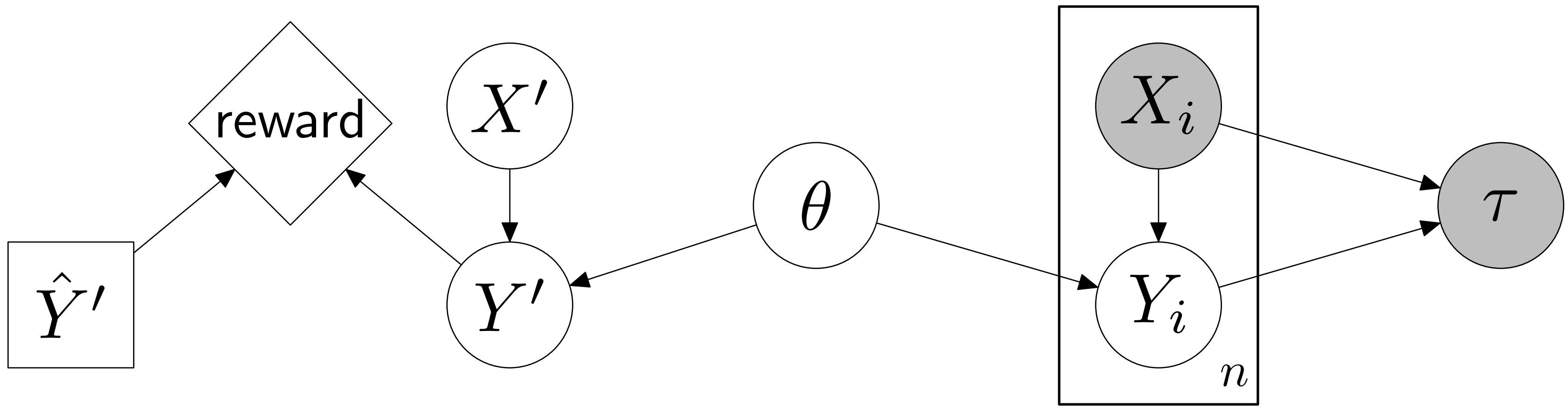
Next: how to choose measurements?

Bayesian decision theory



What do we do with an (approximate) posterior $p(\theta, Y \mid X, \tau)$?

Bayesian decision theory

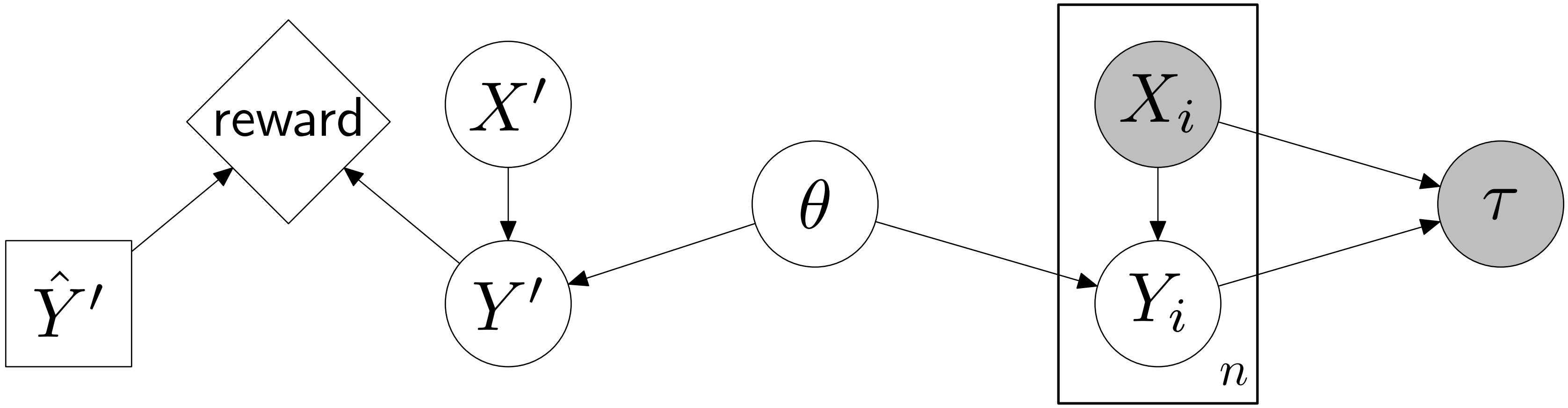


What do we do with an (approximate) posterior $p(\theta, Y \mid X, \tau)$?

Bayes-optimal predictor:

average over X', max over $\hat{Y}'$, average over Y' of reward

Bayesian decision theory



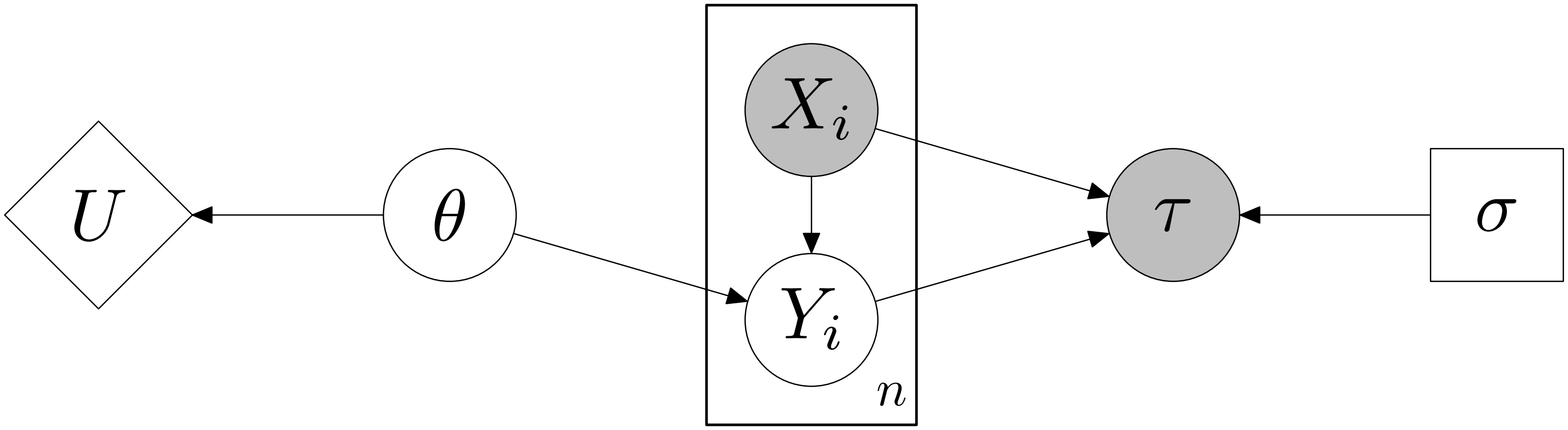
What do we do with an (approximate) posterior $p(\theta, Y \mid X, \tau)$?

Bayes-optimal predictor:

average over X', max over $\hat{Y}'$, average over Y' of reward

$R(\sigma, \tau) =$ expected reward of Bayes-optimal predictor
(i.e., how happy we are with the given situation)

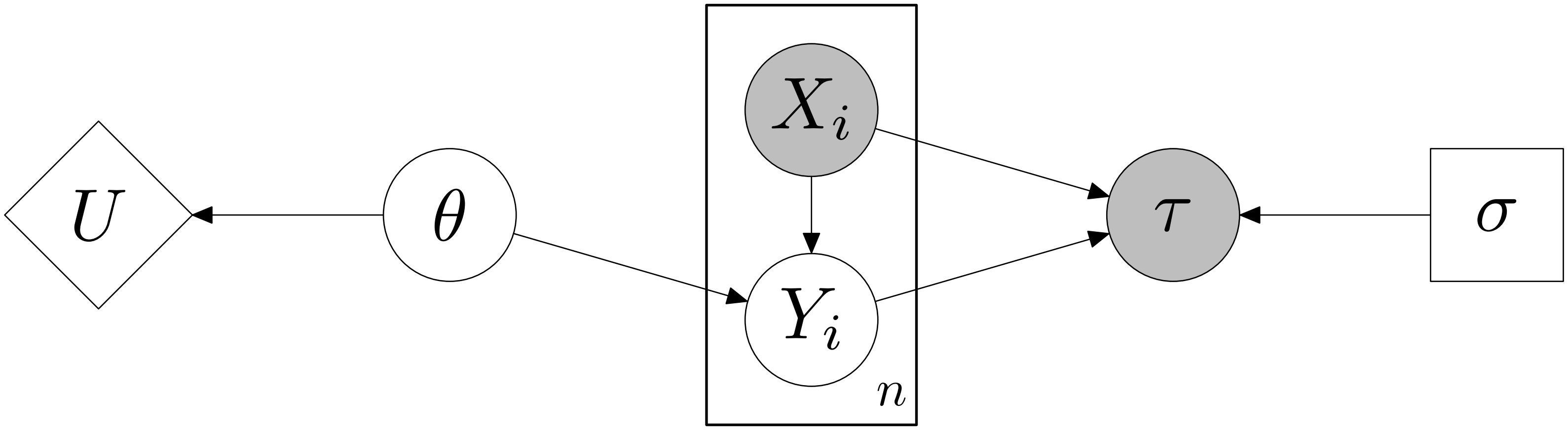
Experimental design (active learning)



Utility of measurement (σ, τ):

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

Experimental design (active learning)



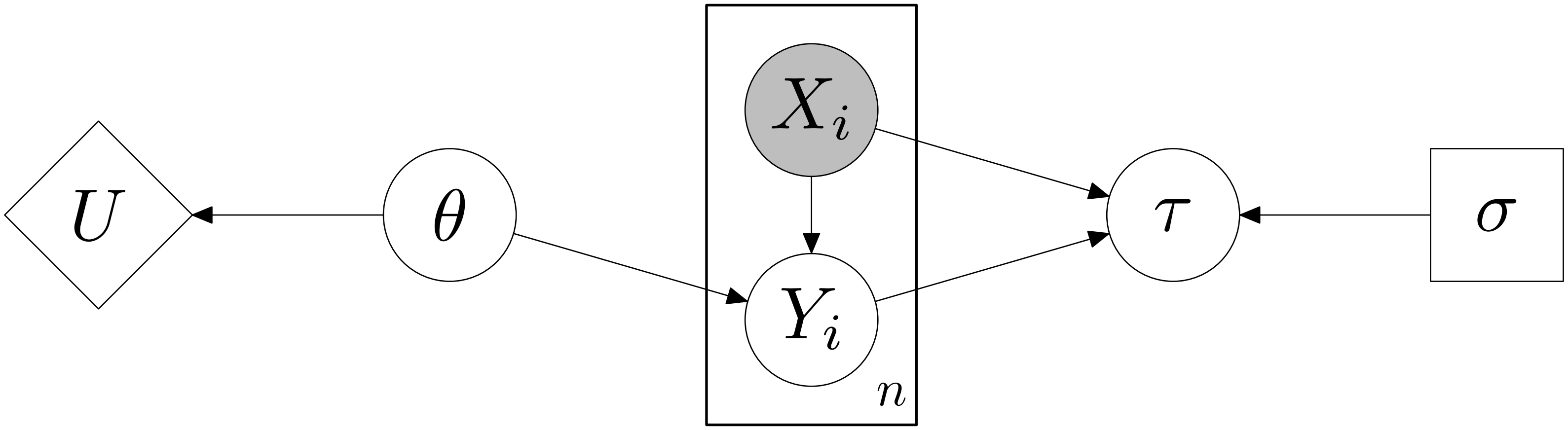
Utility of measurement (σ, τ):

$$U(\sigma,\tau) = \underbrace{R(\sigma,\tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

When considering σ, don't know τ, so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma,\tau)]$$

Experimental design (active learning)



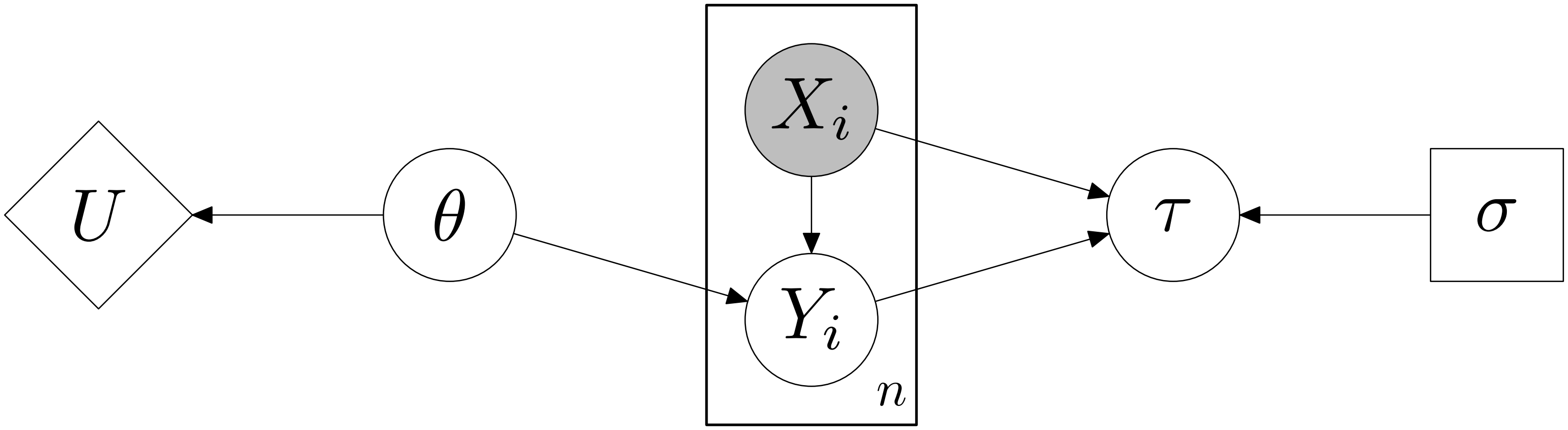
Utility of measurement (σ, τ):

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

When considering σ, don't know τ, so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

Experimental design (active learning)



Utility of measurement (σ, τ):

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

When considering σ, don't know τ, so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

Choose best measurement feature σ:

$$\sigma^* = \text{argmax}_\sigma U(\sigma)$$

Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features

Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features

Measurements:

- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = the, y_i = \text{DT}]$)

Use label entropy as surrogate for assessing measurements

Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

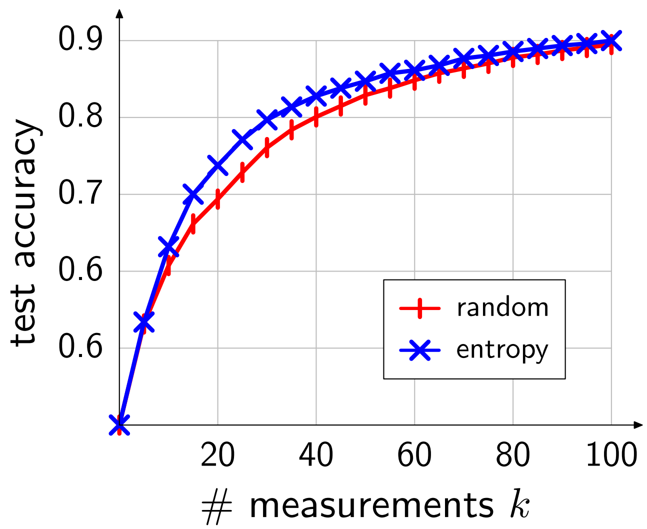
Model: Indep. logistic regression with standard NLP features

Measurements:

- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \mathit{the}, y_i = \mathrm{DT}]$)

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):



(a) Labeling examples

Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

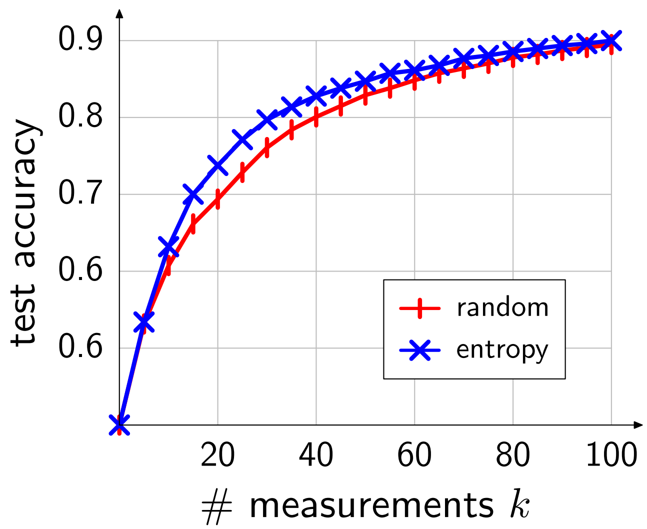
Model: Indep. logistic regression with standard NLP features

Measurements:

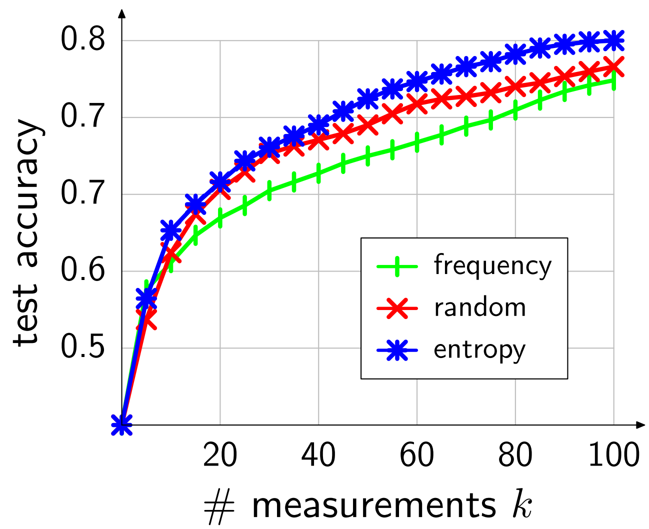
- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = the, y_i = \text{DT}]$)

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):



(a) Labeling examples



(b) Labeling word types

Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

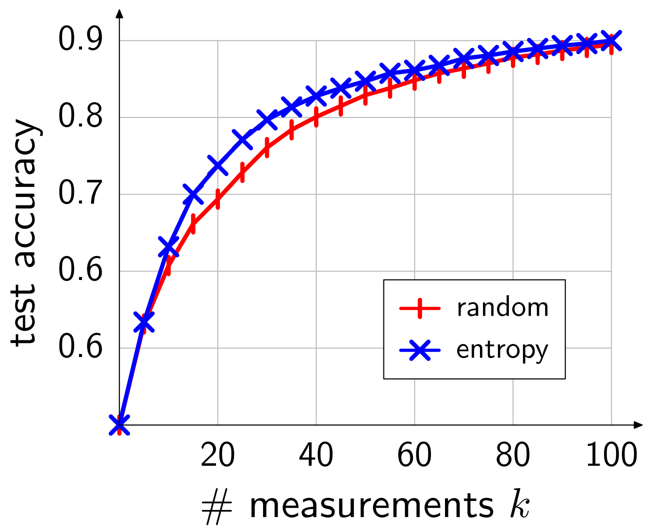
Model: Indep. logistic regression with standard NLP features

Measurements:

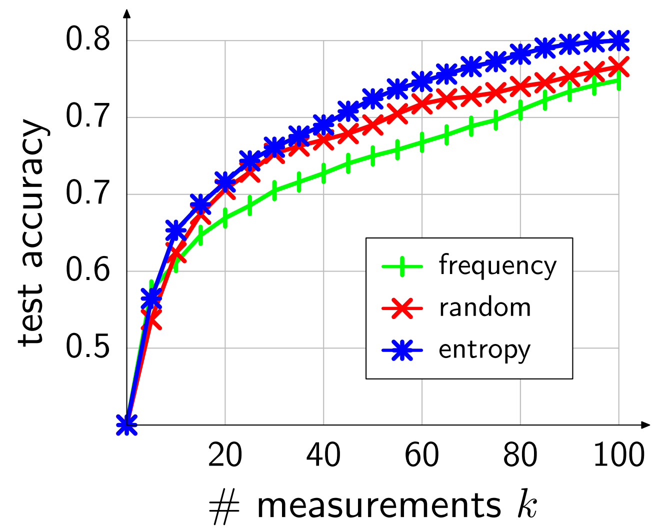
- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \mathit{the}, y_i = \mathrm{DT}]$)

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):



(a) Labeling examples



(b) Labeling word types

Summary

target predictor p^* human $\longrightarrow$ measurements $\longrightarrow$ learning algorithm $\longrightarrow$ learned predictor $\hat{p}$

Measurements

Summary

target predictor p^* → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

Bayesian model

Summary

target predictor p^* → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. — Bayesian model

Summary

target predictor p^* → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. — Bayesian model

information geometry

Summary

target predictor p^* → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. — Bayesian model — decision theory

information geometry

Summary

target predictor p^* — human $\longrightarrow$ measurements $\longrightarrow$ learning algorithm $\longrightarrow$ learned predictor $\hat{p}$

Measurements

variational approx. — Bayesian model — decision theory

information geometry

active learning