Constraints Aware Learning and Inference

Daniel Khashabi

Based on

- Part 1:
  Contraint Driven Learning (Chang et al, 2008)
- Part 2:
  Measurements in Exponential Family (Liang, 2009)

Part 1

Constraints Driven Learning

Constrained Conditional Models (aka ILP Inference)

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)})$$

Penalty for violating the constraint.

(Soft) constraints component

Weight Vector for "local" models

Features, classifiers; log-linear models (HMM, CRF) or a combination

How far y is from a "legal" assignment

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an exact solution.

Cutting Planes, Dual Decomposition & other search techniques are possible

How to train?

Training is learning the objective function

Decouple? Decompose?

How to exploit the structure to minimize supervision?

Information extraction without Prior Knowledge

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis. DIKU , University of Copenhagen, May 1994 .

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y)$$

Prediction result of a trained HMM

[AUTHOR]
[TITLE]
[EDITOR]
[BOOKTITLE]
[TECH-REPORT]
[INSTITUTION]
[DATE]

Lars Ole Andersen . Program analysis and specialization for the C Programming language . PhD thesis . DIKU , University of Copenhagen , May 1994 .

Violates lots of natural constraints!

Examples of Constraints

- Each field must be a consecutive list of words and can appear at most once in a citation.
- State transitions must occur on punctuation marks.
- The citation can only start with *AUTHOR* or *EDITOR*.
- The words *pp., pages* correspond to *PAGE*.
- Four digits starting with 20xx and 19xx are *DATE*.
- Quotations can appear only in *TITLE*
- …….

Easy to express pieces of "knowledge"

Non Propositional; May use Quantifiers

Constraints Driven Learning (CoDL)

Several Training Paradigms

[Chang, Ratinov, Roth, ACL'07;ICML'08,ML, to appear]
Generalized by Ganchev et. al [PR work]

(w_0,ρ_0)=learn(L)

Supervised learning algorithm parameterized by (w,ρ). Learning can be justified as an optimization procedure for an objective function

For N iterations do

T=ϕ

For each x in unlabeled dataset

$h \leftarrow \text{argmax}_y\ w^T\ \phi(x,y) - \sum \rho_k\ d_C(x,y)$

$T=T \cup \{(x, h)\}$

Inference with constraints: augment the training set

$(w,\rho) = \gamma\ (w_0,\rho_0) + (1- \gamma)$ learn(T)

Learn from new training data Weigh supervised & unsupervised models.

Excellent Experimental Results showing the advantages of using constraints, especially with small amounts on labeled data [Chang et. al, Others]

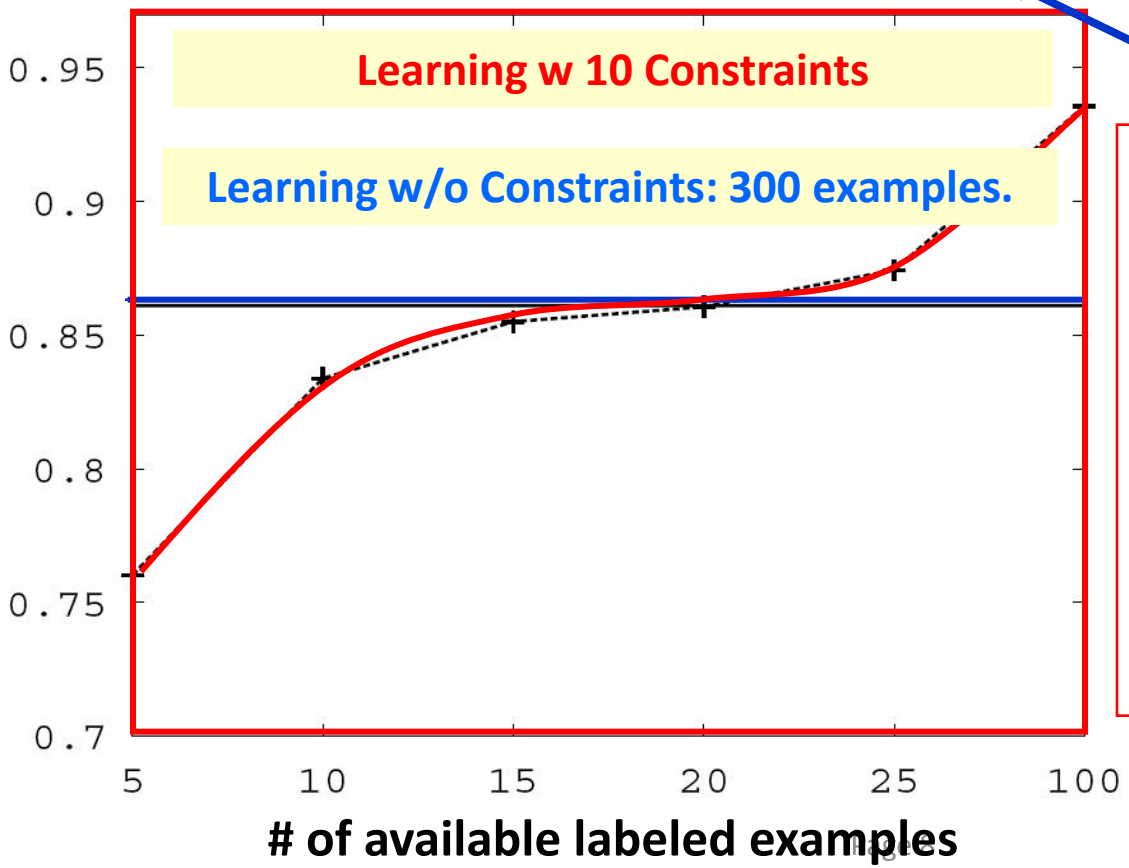
Page 7

Constraints Driven Learning (CODL)

[Chang, Ratinov, Roth, ACL'07;ICML'08,MLJ, to appear]
Generalized by Ganchev et. al [PR work]

- Semi-Supervised Learning Paradigm that makes use of constraints to bootstrap from a small number of examples

Objective function: $f_{\Phi,C}(\mathbf{x},\mathbf{y}) = \sum w_i \phi_i(\mathbf{x},\mathbf{y}) - \sum \rho_i d_{C_i}(\mathbf{x},\mathbf{y}).$



Poor model + constraints

Constraints are used to:

- Bootstrap a semi-supervised learner
- **Correct** weak models predictions on unlabeled data, which in turn are used to keep training the model.

Part 2

Learning from Measurements in Exponential Families

The big picture

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

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

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

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

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

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

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

Measurements

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

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

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

$\sigma(\ X_1\ ,\ Y_1\ )$
$\sigma(\ X_2\ ,\ Y_2\ )$
$\sigma(\ X_3\ ,\ Y_3\ )$
$\quad\ldots\quad\ldots$
$\sigma(\ X_i\ ,\ Y_i\ )$
$\quad\ldots\quad\ldots$
$\sigma(\ X_n\ ,\ Y_n\ )$
$+$ noise

τ

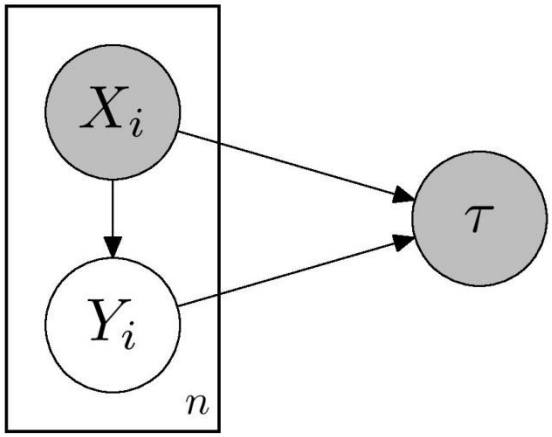
Measurements

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

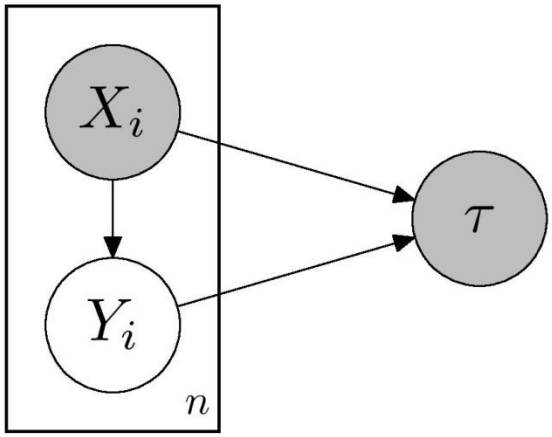
Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

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

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

$\sigma(\ X_1\ ,\ Y_1\ )$
$\sigma(\ X_2\ ,\ Y_2\ )$
$\sigma(\ X_3\ ,\ Y_3\ )$
$\quad\ldots\quad\ldots$
$\sigma(\ X_i\ ,\ Y_i\ )$
$\quad\ldots\quad\ldots$
$\sigma(\ X_n\ ,\ Y_n\ )$
$+$ noise

τ



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

Examples of measurements

Fully-labeled example:

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

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:

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

Partially-labeled example:

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

Labeled predicate:

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

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:

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

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:

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

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:

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

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:

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

Label preference:

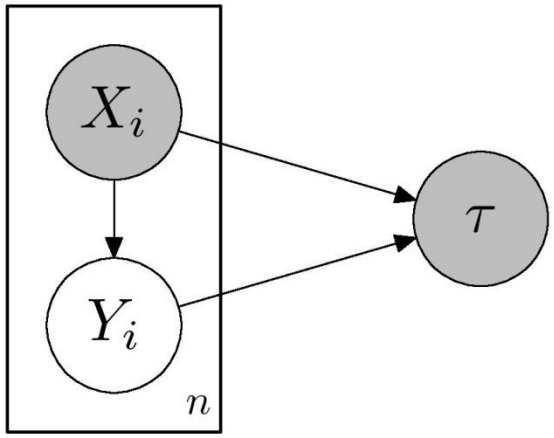
$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}]$$

Can get measurement values τ without looking at all examples

Next: How to combine these diverse measurements coherently?

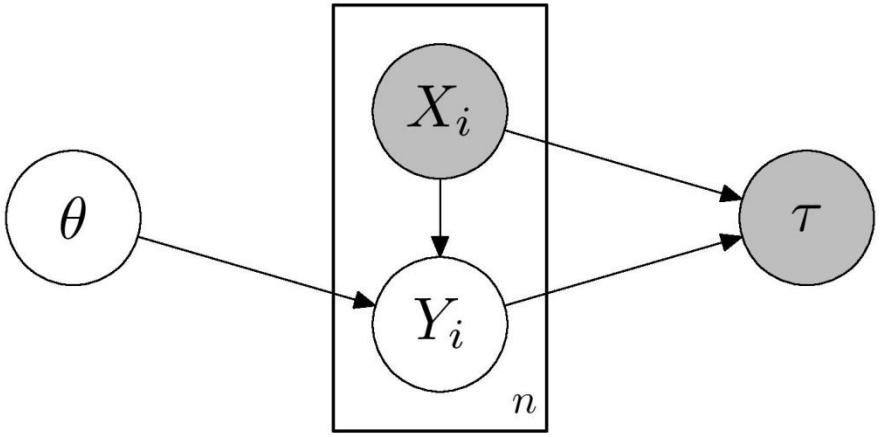
Prediction model

Bayesian framework:



Prediction model

Bayesian framework:

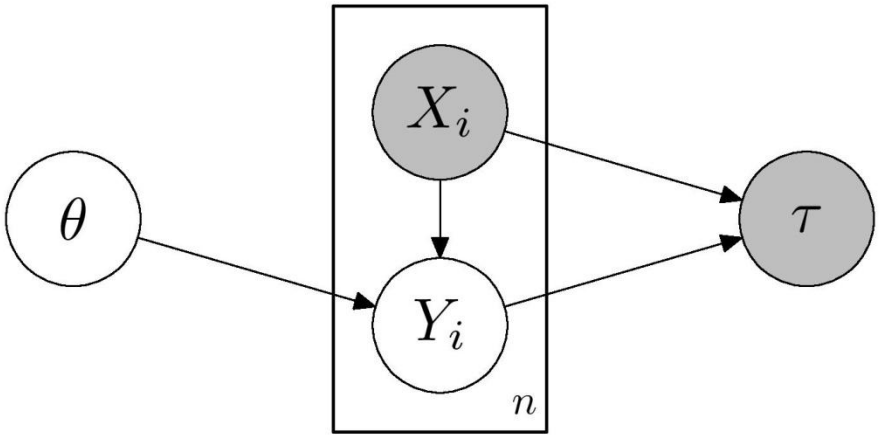


Exponential families:

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

Prediction model

Bayesian framework:



Exponential families:

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

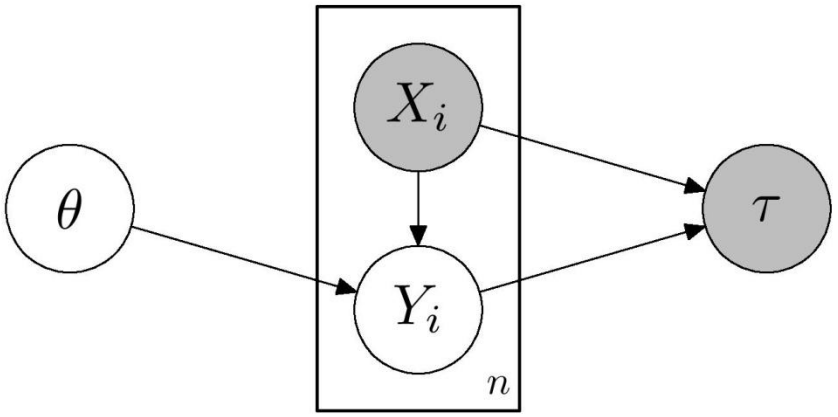
$\phi(x,y) \in \mathbb{R}^d$: model features

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

$A(\theta; x) = \log \int \exp\{\langle \phi(x,y), \theta \rangle\} dy$: log-partition function

Learning via Bayesian inference

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



Variational formulation:

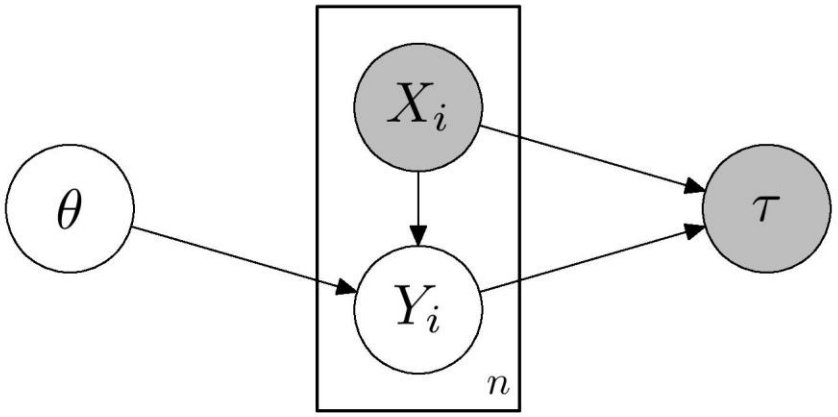
$$\min_{q \in \mathcal{Q}_{\theta,Y}} \mathsf{KL}\left(q(\theta, Y) \,||\, p(\theta, Y \mid \tau, X)\right)$$

Approximations:

- $\mathcal{Q}_{\theta,Y}$: 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:

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

Approximations:

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

Algorithm:

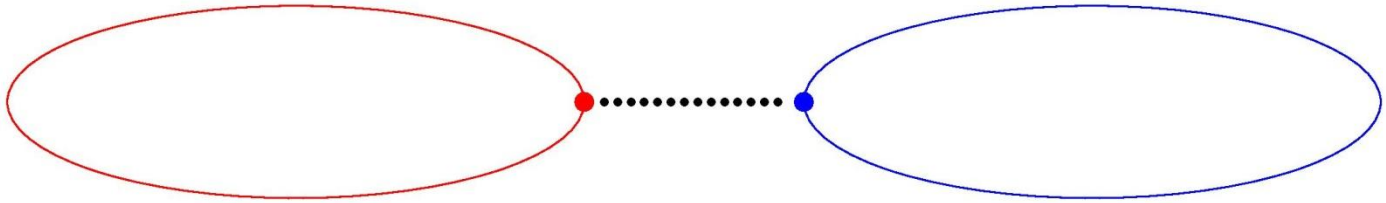
Apply Fenchel duality $\rightarrow$ saddlepoint problem

Take alternating stochastic gradient steps

Information geometry viewpoint

(assume zero measurement noise)

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

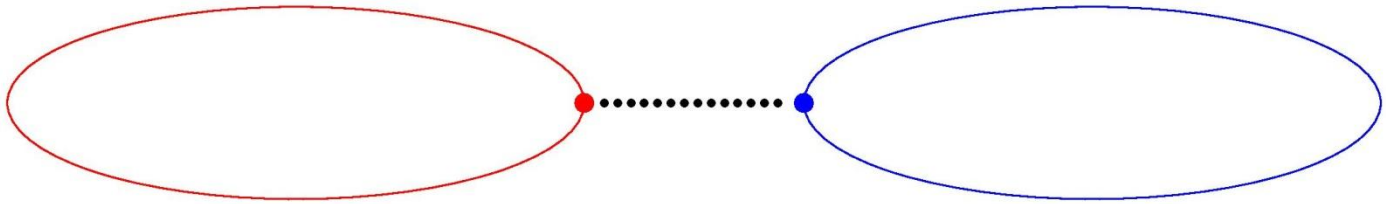


$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathsf{KL}\left(q \,||\, p\right)$$

Information geometry viewpoint

(assume zero measurement noise)

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

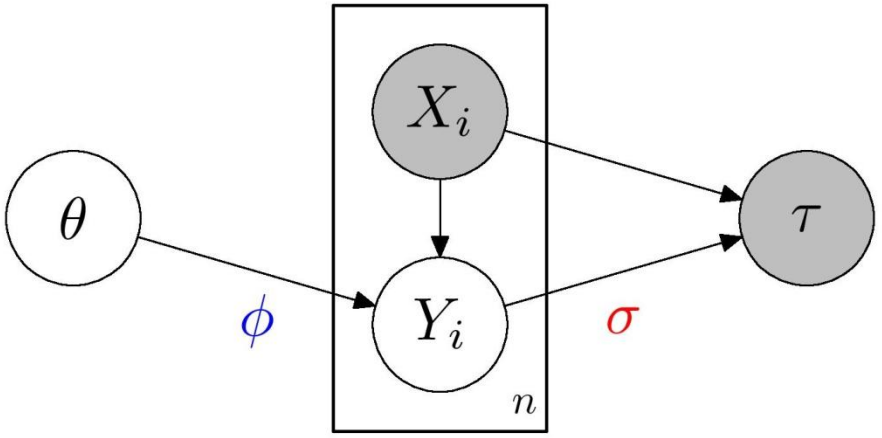


$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathsf{KL}\left(q \,||\, p\right)$$

Interpretation:

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

Model features ϕ versus measurement features σ

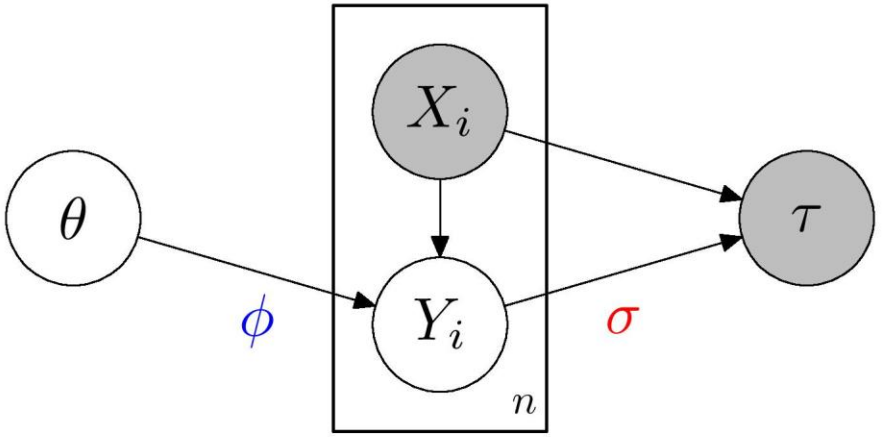


Guidelines:

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

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

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., $\sum_i \mathbb{I}[x_i = View, y_i = \text{FEAT}]$)

Per-position test accuracy (on 100 examples):

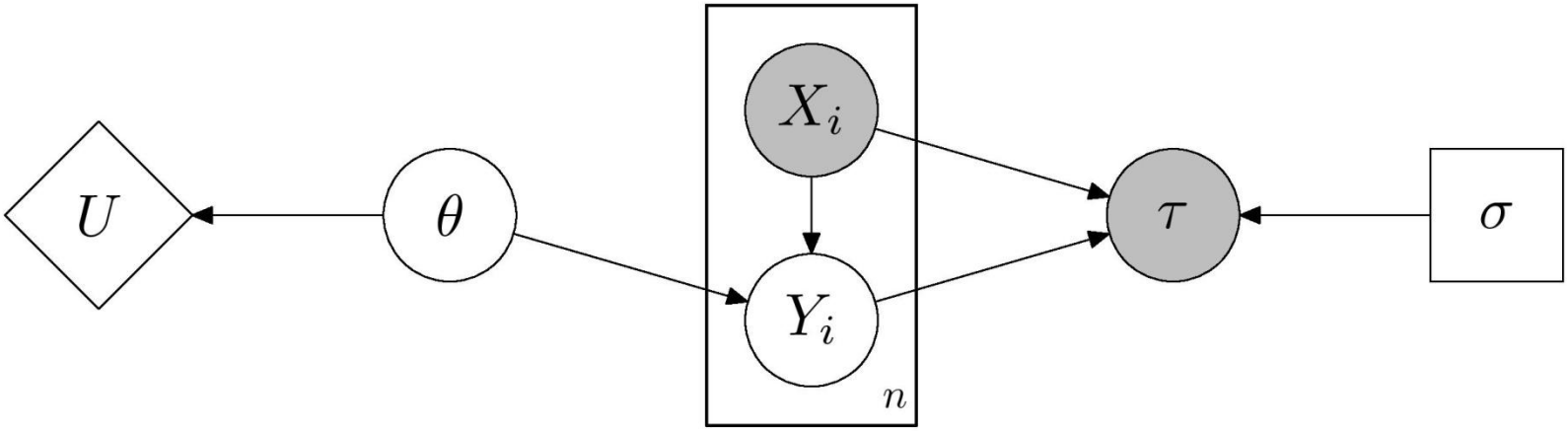
# labeled examples	10	25	100
General Expectation Criteria	74.6	77.2	80.5
Constraint-Driven Learning	**74.7**	**78.5**	81.7
Measurements	71.4	76.5	**82.5**

Able to integrate labeled examples and predicates gracefully

So far: given measurements, how to learn

Next: how to choose measurements?

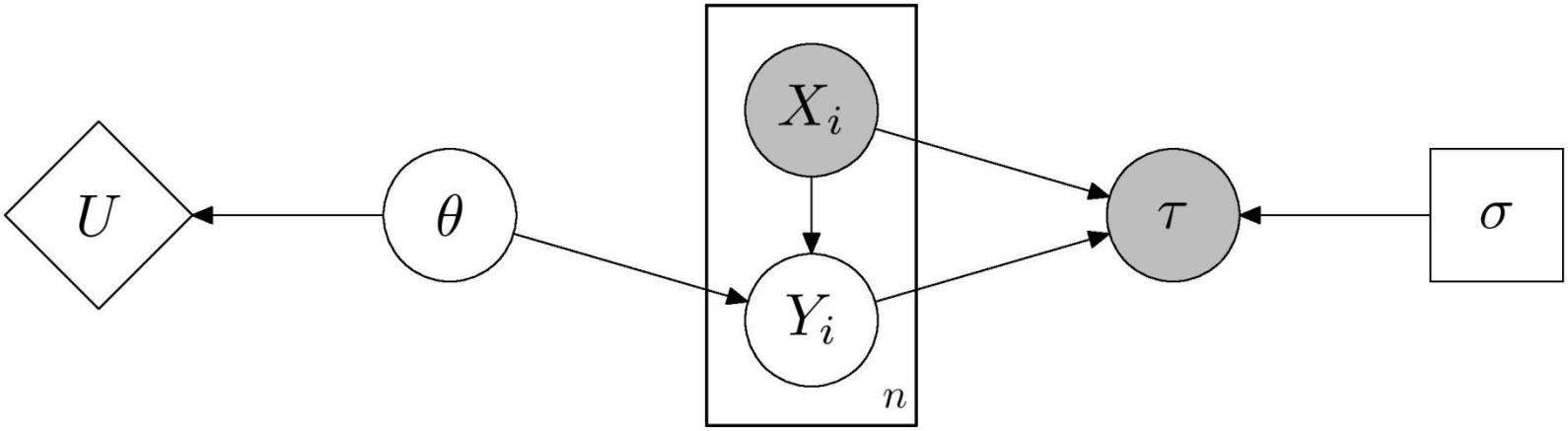
Experimental design (active learning)



Utility of measurement (σ, τ):

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

Experimental design (active learning)



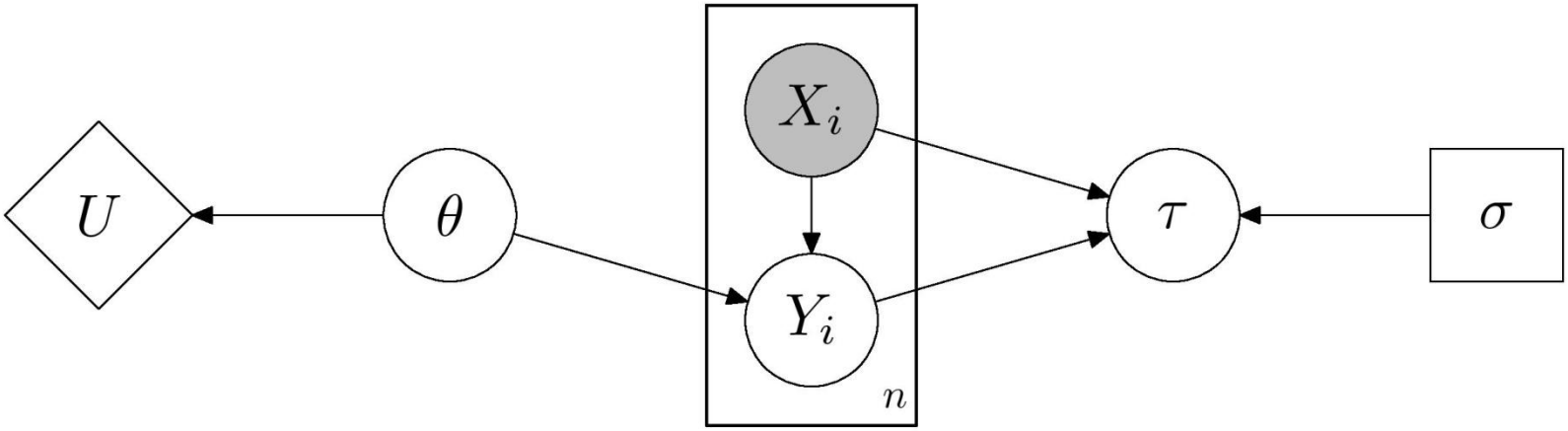
Utility of measurement (σ, τ):

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

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

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

Experimental design (active learning)



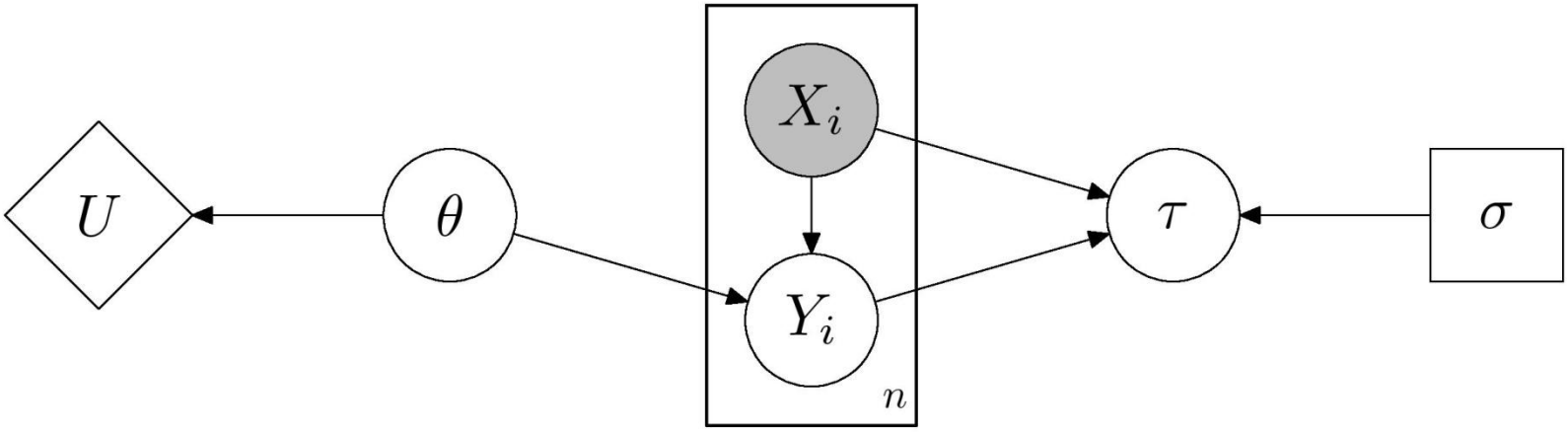
Utility of measurement (σ, τ):

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

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

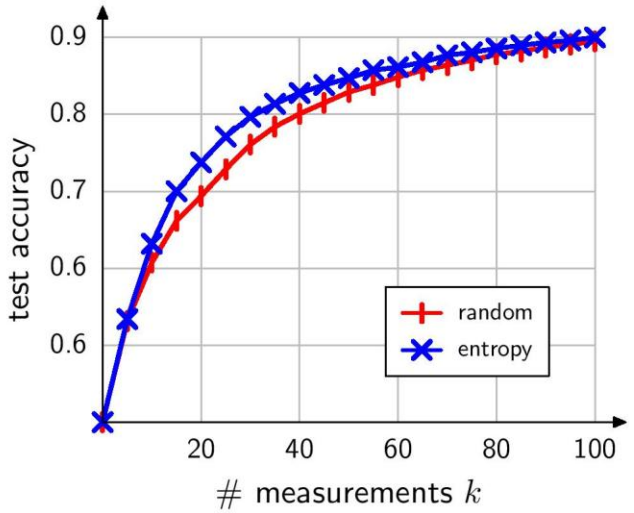
Model: Indep. logistic regression with standard NLP features

Measurements:

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

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

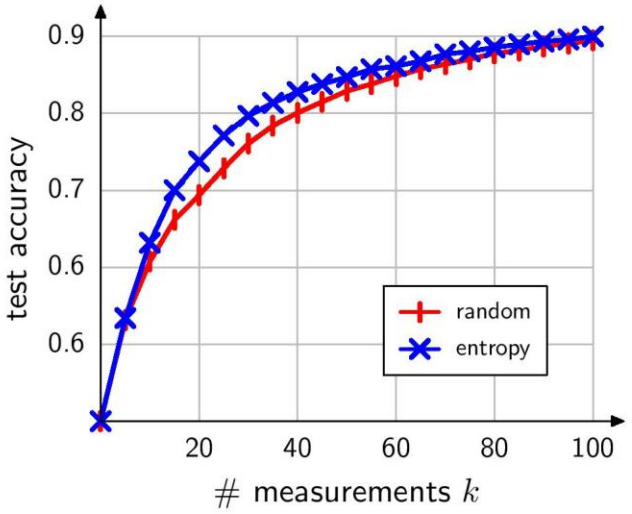
Model: Indep. logistic regression with standard NLP features

Measurements:

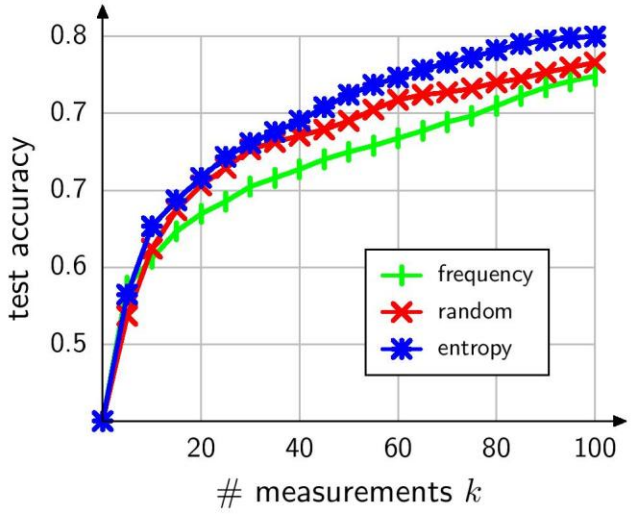
- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = the, y_i = \text{DT}]$)

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):



(a) Labeling examples



(b) Labeling word types

Summary

Measurements

Summary

Measurements

|

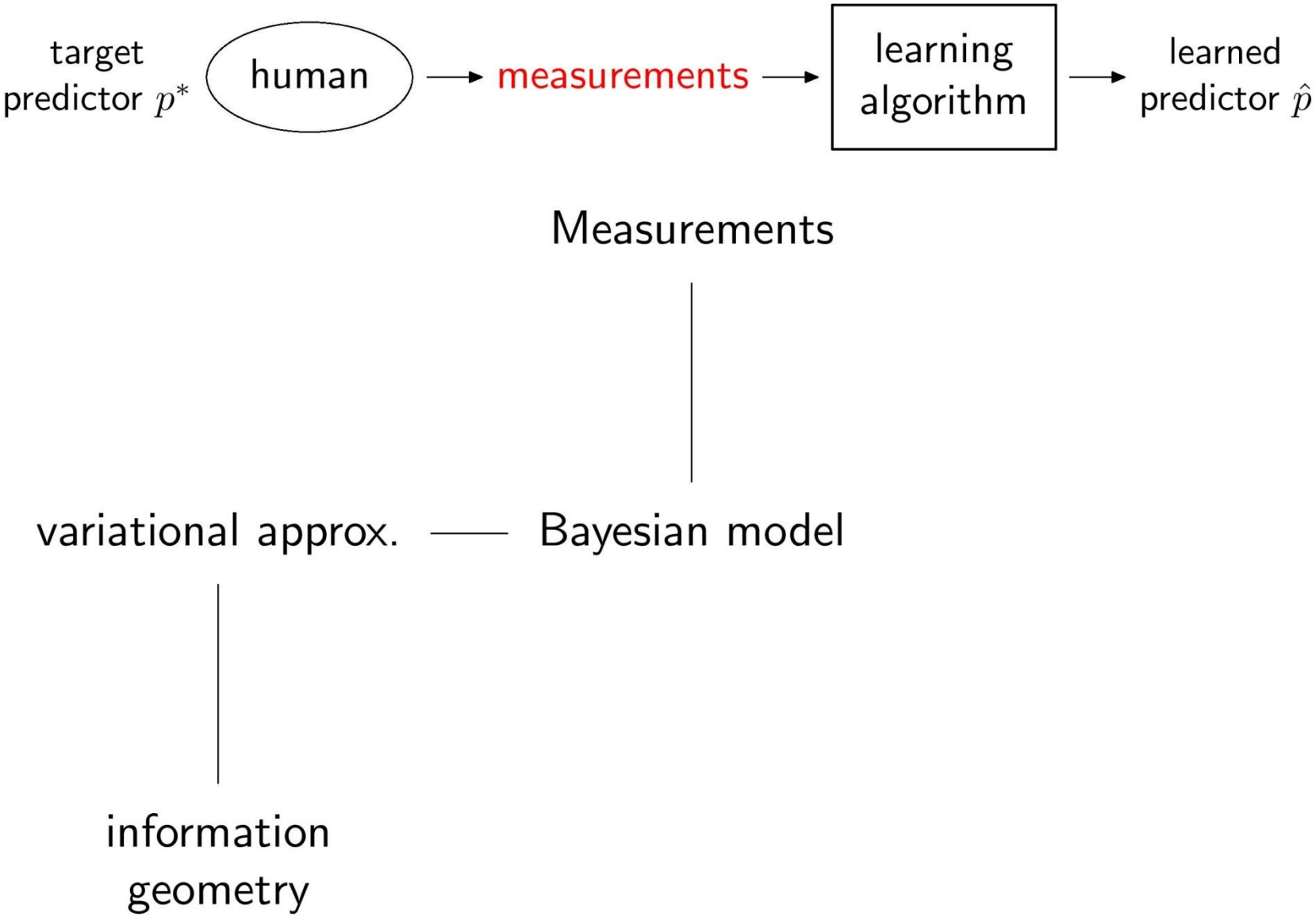
Bayesian model

Summary

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

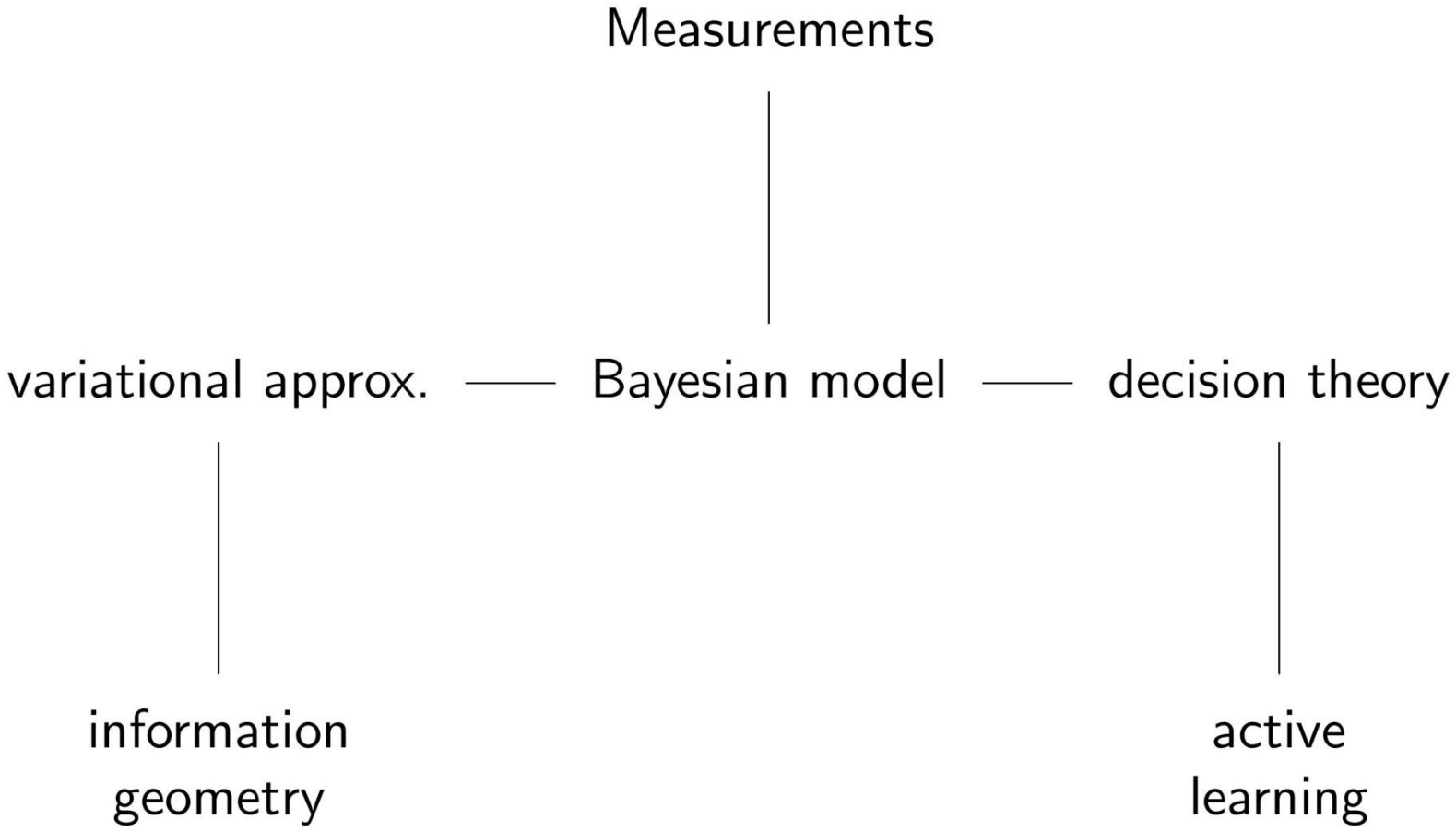
Measurements

variational approx. — Bayesian model — decision theory

information
geometry

Summary


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


Measurements
variational approx.
Bayesian model
decision theory
information
geometry
active
learning

Approximate Inference

- $$\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(Y, \theta) \,||\, p(Y, \theta \mid \tau, X, \sigma)\right).$$

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(Y, \theta) : q(Y, \theta) = q(Y)\delta_{\tilde{\theta}}(\theta)\}.$$

The probabilistic model

$$p(\theta, Y, \tau \mid X, \sigma) \stackrel{\text{def}}{=} p(\theta) \prod_{i=1}^{n} p_\theta(Y_i \mid X_i) p(\tau \mid X, Y, \sigma).$$

Log-concave prior:

$$\log p(\theta) = -h_\phi(\theta) + \text{constant}$$

$$\log p(\tau \mid X, Y, \sigma) = -h_\sigma(\tau - \sigma^X(Y)) + \text{constant}$$

Approximate Inference

- The probabilistic model

$$p(\theta, Y, \tau \mid X, \sigma) \stackrel{\text{def}}{=} p(\theta) \prod_{i=1}^{n} p_\theta(Y_i \mid X_i) p(\tau \mid X, Y, \sigma).$$

Log-concave prior:

$$\log p(\theta) = -h_\phi(\theta) + \text{constant}$$

$$\log p(\tau \mid X, Y, \sigma) = -h_\sigma(\tau - \sigma^X(Y)) + \text{constant}$$

For example:

- Gaussian: $h_\phi(\theta) = \frac{\lambda}{2}||\theta||^2$
- Box: $h_\sigma(u) = \mathbf{W}[\forall j, |u_j| \leq \epsilon_j]$

Approximate Inference

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

$$\min_{q(Y),\theta} -H(q(Y)) + E_{q(Y)}[h_\sigma(\tau - \sigma^X(Y))]$$

$$- \sum_{i=1}^{n} E_{q(Y)} \log p_\theta(Y_i \mid X_i) + h_\phi(\theta).$$

$$q(Y) = \prod_{i=1}^{n} q_{\beta,\theta}(Y_i \mid X_i)$$

$$q_{\beta,\theta}(y \mid x) = \exp\{\langle \sigma(x, y), \beta \rangle + \langle \phi(x, y), \theta \rangle - B(\beta, \theta; x)\},$$

Fenchel's Duality Theorem

- Let *f* be convex and *g* be concave function

$$\min_x (f(x) - g(x)) = \max_p (g_\star(p) - f^\star(p)).$$

Then:

$$f^\star(x^*) := \sup\{\langle x^*, x\rangle - f(x) \mid x \in \mathbb{R}^n\}$$

$$g_\star(x^*) := \inf\{\langle x^*, x\rangle - g(x) \mid x \in \mathbb{R}^n\}$$

Finding a convex lower bound on posterior

$$\min_{\theta \in \mathbb{R}^d} \max_{\beta \in \mathbb{R}^k} L(\beta, \theta),$$

$$L(\beta, \theta) = \langle \tau, \beta \rangle - \sum_{i=1}^{n} B(\beta, \theta; X_i) + \sum_{i=1}^{n} A(\theta; X_i) - h_\sigma^*(\beta) + h_\phi(\theta),$$

$$h_\sigma^*(\beta) = \sup_{u \in \mathbb{R}^k} \{ \langle u, \beta \rangle - h_\sigma(u) \}$$

$$\frac{\partial L(\beta, \theta)}{\partial \beta} \qquad \frac{\partial L(\beta, \theta)}{\partial \theta}$$

References

- Slides: Dan Roth: http://l2r.cs.uiuc.edu/~danr/Talks/IndirectSup-MSR-06011.ppt
- Slides: Percy Liang: http://cs.stanford.edu/~pliang/papers/measurements-icml2009-talk.pdf