AAAI 2018 Tutorial
Integrating Learning into Reasoning

Brendan Juba
Washington Univ. St. Louis
bjuba@wustl.edu
www.cse.wustl.edu/~bjuba

Loizos Michael
Open University of Cyprus
loizos@ouc.ac.cy
cognition.ouc.ac.cy/loizos

B. Juba supported by an AFOSR Young Investigator Award and NSF Award CCF-1718380

AAAI 2018 (New Orleans, Louisiana, U.S.A.) February 03, 2018

It Has Been Said That…

"For every belief comes either through syllogism or from induction"

Aristotle, Organon (Prior Analytics II, §23)

Artificial Intelligence research today?

Syllogism ***xor*** Induction

High-Level Tutorial Roadmap

A. **Why KRR Should Embrace Learning**
B. **Introduction to PAC-Semantics**
C. **Integrating Deduction and Induction**
D. **Reasoning Non-Monotonically**
E. **Overall Summary and Conclusions**

A

High-Level Tutorial Roadmap

A. **Why KRR Should Embrace Learning**
B. **Introduction to PAC-Semantics**
C. **Integrating Deduction and Induction**
D. **Reasoning Non-Monotonically**
E. **Overall Summary and Conclusions**

A. Why KRR Should Embrace Learning

Review of Standard KRR Problems

Representation and Reasoning

- Propositions (p, q, …). Connectives ($\wedge$, $\neg$, …).
  - Implications: $\varphi \Rightarrow x$. Equivalences: $\varphi \Leftrightarrow x$.
- Reasoning ***semantics*** through entailment $\vDash$.
- Proof procedures $\vdash$ to ***compute*** entailment.

- Given formulas in *KB* and an input *O*, ***deduce*** whether a result *R* is entailed ($KB \cup O \vDash R$).
- Given formulas in *KB* and an input *O*, ***abduce*** an explanation *E* that entails *O* ($KB \cup E \vDash O$).

Relational Representation and IQEs

- Predicates (p, q, …). Variables (x, y, …). ***Tokens* (t_i).** Connectives ($\wedge$, $\neg$, …), Quantifiers ($\forall$, $\exists$).

- From the class of implications / equivalences, consider only those whose body comprises **independently quantified expressions (IQEs)**.
  - $\exists y \exists z$ [*num(y)* $\wedge$ *num(z)* $\wedge$ *larger(z,y)* $\wedge$ $\neg$*div(y,z)*]
  - No tokens (they carry no meaning), small arity.
  - $\forall xs$ [*formula over IQEs* $\Leftrightarrow$ *head_predicate(xs)*]

(see later: these restrictions support learnability)

Non-Monotonic Reasoning

- Non-monotonicity typically viewed as property of ***extending input O for fixed KB***, and having result R become "smaller".
- Useful also when ***extending KB*** as a prerequisite to elaboration tolerance.

- Will use logic-based argumentation for NMR.
  - Most (all?) major NMR logics have been reformulated in terms of argumentation.
  - Compatible with human cognition **[Kakas+ '16]**.

Formal Argumentation in AI

- Abstract Argumentation framework <*Arg*,*Att*>
  - *Arg* is a set of arguments (no internal structure)
  - *Att* is a binary relation on *Arg* (lifted on sets)
- Goal: Find $S \subseteq Arg$ that defends all its attacks.
  - Several ways to make this precise **[Dung '95]**.

- Structured argumentation (e.g., ABA, ASPIC+): argument is a classical proof from inputs and KB rules. The overall reasoning is not classical.

Activity: Logic-Based Arguments

treated **implies** will_survive.
fatal **implies** ¬will_survive.
viral_meningitis **implies** meningitis.
bacterial_meningitis **implies** meningitis.
bacterial_meningitis **implies** fatal.
fatal **implies** ¬treatable.
meningitis **implies** ¬fatal.
meningitis **implies** treatable.
¬fatal **implies** will_survive.
true **implies** ¬meningitis.

total ordering (for simplicity)

↑ more preferred

What inference should we draw on each input?

{}, {viral_meningitis}, {bacterial_meningitis}, {bacterial_meningitis,treated}

Reasoning in Temporal Settings

- ***Frame Problem*** (commonsense law of inertia): Properties persist unless caused to stop.
  - If you see a bird flying, it is flying a bit later.
  - Persistence rules are weaker than causal rules.

- ***Ramification Problem***: Production of indirect effects as a way to satisfy state constraints.
  - If you shoot a bird, it stops being able to fly.
  - Encode ramifications as causal rules (since constraints could also qualify causal change).

Reasoning in Temporal Settings

- ***Qualification Problem***: Effects of actions are blocked if they would violate state constraints.
  - If you scare a dead bird, it does not fly away.
  - This constraint does not produce ramifications.
  - Encode constraints as preclusion / block rules.
- ***State Default Problem*** **[Kakas+ '08]**: If a state constraint is violated, the exception persists.
  - If you see a flying penguin, it remains one.
  - Persistence rules are stronger than constraints.

Activity: Story Understanding

What inferences follow from this story?

Papa Joe woke up early at dawn, and went off to the forest. He walked for hours, until the sight of a turkey in the distance made him stop.

A bird on a tree nearby was cheerfully chirping away, building its nest. He carefully aimed at the turkey, and pulled the trigger of his shotgun.

Undisturbed, the bird nearby continued chirping.

Q1: What is the condition of the turkey?

(a) Alive and unharmed. (b) Dead. (c) Injured.

… Through Argumentation

scene 1: forest(t_1) at 1. … 2

scene 2: bird(t_3) at 4. … nts.

gun($t_…$) at … t 7.

s… 9.

r_1…

r_2:

r_3: … (G,X)

r_4: f… (G,X), gun(G), turkey(X)…

r_5: fired_at(G… **causes** noise.

r_6: … nearby(B) **causes** -chirping(B).

r_7: -loaded(G), gun(G) **precludes** fired_at(G,X).

Uniform treatment thro…

Addition of **time** for causality and change.

RAC: inertia, … defaults.

Endogenous a… **tion**.

tokens have no meaning

rule bodies are IQEs

argum…

preferences

Try it online! ***http://cognition.ouc.ac.cy/star***

[Diakidoy+ '14,15]

A. Why KRR Should Embrace Learning

The Task of Knowledge Acquisition

The Declaration of Independence (of KRR from Machine Learning)

We hold these truths to be self-evident:

an appropriate knowledge base is given

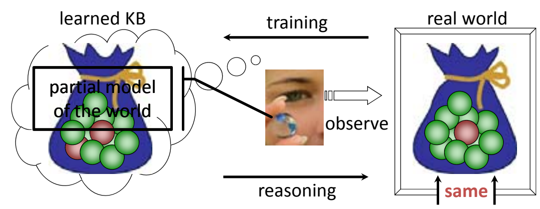
reasoning is evaluated against the KB

chaining of rules is trivially beneficial

KB rules can be applied in any order

acquisition of KB can be done a priori

What if KB not Given but Learned?



- Then, KB is ***P****robably* ***A****pproximately* ***C****orrect*:
  - improbable the samples were unrepresentative
  - future predictions will be approximately correct

Program, Memorize, or Induce?

- **Examples**: **Target Concept**: **Labels**:

Examples		Target Concept		Labels
1101001	→	$f: \{0,1\}^n \rightarrow \{0,1\}$	→	0
0101011				0
0001110				1
…				…

- **Concept Class *C***: bias, possible target concepts.
- **Hypothesis Class *H***: all acceptable hypotheses.
- **Complexity**: examples / time / memory to learn.

Typical Boolean Target Concepts

- **Conjunction function**: $x_3 \wedge x_5 \wedge x_7$
  - *1011101* the conjunction of x_3, x_5, x_7 is *1*
- **Parity function**: $x_2 \oplus x_3 \oplus x_5 \oplus x_6$
  - *0010101* the parity of x_2, x_3, x_5, x_6 is *0*
- **Decision list**: $\langle x_4,1\rangle, \langle x_2,0\rangle, \langle x_7,1\rangle, \langle \text{default},0\rangle$
  - *1010011* false: x_4, x_2 first true: x_7 value: *1*
- **Linear threshold** (hyberplane, perceptron): *weights*=⟨0.3, 0.1, 0.5, 0.2, 0.7, 0.9⟩, *threshold*=1
  - *0110100* *110100* · *weights* = 0.6 < *threshold*

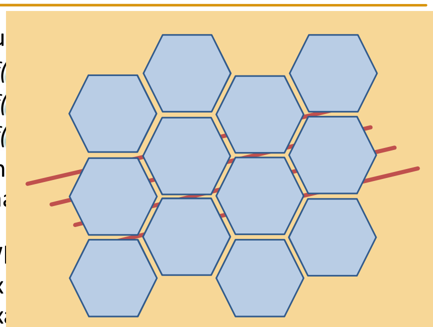
Activity: Is Learning Even Possible?

- Su
  f(
  f(
  f(
  an
  th

…
?
?
?
?

- Wh
  ex
  ex

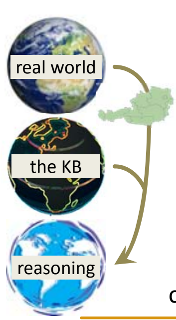
ary
th



Evaluation of Reasoning Process

- Evaluate reasoning process against given KB:
  - improve completeness, insisting on full soundness
  - okay since KB is considered the golden standard
  - not okay when KB is only *approximately* correct
- Evaluate reasoning process when KB *learned*:
  - improve completeness, without compromising soundness much more than what is *necessary*
  - soundness and completeness wrt an "ideal KB" (access only to its partial models during training)

Real World, the KB, and Reasoning



	traditional KRR	***with learned KB***
real world	no access(!)	partial models
the KB	globally correct (rules qualified)	individual rules approx. correct
reasoning	against the KB unsoundness = 0 completeness ++	against real world unsoundness << completeness ++

Is Chaining *Trivially* Beneficial?

- Yes, if rules in the KB are given / programmed.
  - Given: *"if you have fever then you are sick"*
  - Given: *"if you are sick then visit a doctor"*
  - Infer "visit a doctor" given "you have fever"?
- *Prima facie* no, if rules in the KB are learned.
  - Learning can render rule chaining superfluous (cf. shortcuts, heuristics, fast thinking, hunch).
  - Learned: *"if you have fever then visit a doctor"*

B

High-Level Tutorial Roadmap

B. Introduction to PAC-Semantics

Semantics for Propositional Logic

Motivating example: birds.com Analytics

birds.com™

¬[FOOD=FISH] ?

The **data**

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
107	54	Mouse
107	55	Grubs
⋮	⋮	⋮

To determine: *"true… with high probability for data source?"*

PAC-semantics for propositional logics

- Fixed, finite set of propositional variables: [FOOD=SEED], [FOOD=GRUBS], [FOOD=MOUSE], [FOOD=WORM], [FOOD=FISH], …, [FLIES], [SWIMS], [HAS_BILL], [HAS_BEAK], [COLOR=BLACK], [COLOR=RED], [COLOR=BROWN], [COLOR=WHITE], [COLOR=GREEN], [COLOR=BLUE], …
- Probability distribution D over Boolean valuations for the propositional variables
  - **NOTE:** generally not uniform, not independent

PAC-semantics for propositional logics

- Probability distribution D over Boolean valuations for the propositional variables
  - Usually propositional variables capture attributes of data entry, sensor values, etc.
  - D captures range of possible combinations of values and their relative frequency

☞**Definition.** [Valiant '00] A formula $\varphi(x_1,\ldots,x_n)$ is *(1-ε)-valid under D* if $\Pr_D[\varphi(x_1,\ldots,x_n)] \geq 1-\varepsilon$.

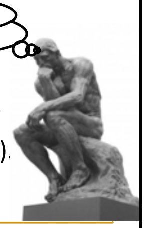
- *ε may or may not be* small (cf. [Adams '75])

PAC-semantics for propositional logics

☞**Definition.** [Valiant '00] A formula $\varphi(x_1,\dots,x_n)$ is *(1-ε)-valid under D* if $Pr_D[\varphi(x_1,\dots,x_n)] \geq 1-\varepsilon$.

- *ε may or may not be* small (cf. [Adams '75])

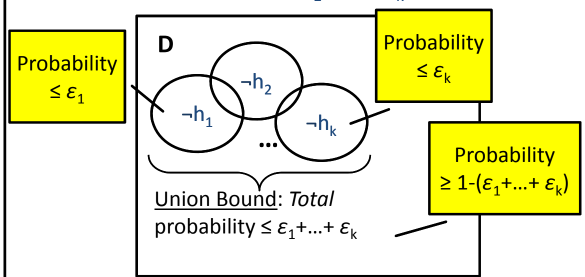
WHAT INFERENCES PRESERVE (1-ε)-VALIDITY?



- **Theorem.** [Juba '13] Suppose $h_1,\dots,h_k$ are, respectively, $1-\varepsilon_1, \dots, 1-\varepsilon_k$ –*valid*, and $\{h_1,\dots,h_k\} \models \varphi$ (*classical entailment*). Then φ is $1-(\varepsilon_1+\dots+\varepsilon_k)$-valid.

Classical entailment to PAC entailment: the **union bound**

- Consider the events $[\neg h_1],\dots,[\neg h_k]$



Classical entailment to PAC entailment: the **union bound**

- Consider the events $[\neg h_1],\dots,[\neg h_k]$
- Union bound: the event
  $[\neg h_1 \lor \cdots \lor \neg h_k] = [\neg(h_1 \land \cdots \land h_k)]$
  has total probability $\leq \varepsilon_1+\dots+\varepsilon_k$
- *Classical entailment*: $[h_1 \land \cdots \land h_k] \subseteq [\varphi]$
- Therefore: $Pr_D[\varphi] \geq Pr_D[h_1 \land \cdots \land h_k]$
  $\geq 1 - (\varepsilon_1+\dots+\varepsilon_k)$ ■
- **Summary**: *all* classical inferences preserved, but *each additional premise may incur a cost*

PAC-semantics does not mean Probability Logic (e.g. Nilsson'86)

PAC-semantics

- Classical Boolean object language
- Probability bounds in interpretation
  - Classical proof of a formula guarantees it holds with some probability under D

Logics of Probability

- Probability bounds in object language
- Classical (Tarskian) semantics
  - Classical proof of a probability bound on D that is true with certainty

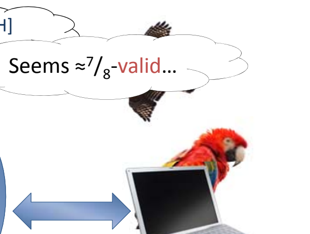
Motivating example: birds.com Analytics

birds.com™

¬[FOOD=FISH] ?

Seems ≈7/8-valid…

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
107	54	Mouse
107	55	Grubs
⋮	⋮	⋮



…JUSTIFIED ON WHAT GROUNDS??

B. Introduction to PAC-Semantics

Learning and Inductive Inference

Knowledge acquisition from data: birds.com Analytics

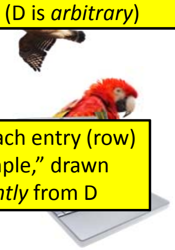
birds.com™

¬[FOOD=FISH] ?

To determine: ($1-\varepsilon$)-*valid for D?* (D is *arbitrary*)

The **data**

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
107	54	Mouse
107	55	Grubs
⋮	⋮	⋮



Assume: Each entry (row) is an "example," drawn *independently* from D

The "**i.i.d. data**" assumption

- **Assume**: data consists of *examples* – valuations of the propositional variables drawn independently from common D (*"i.i.d."*)
  - Will enable us to draw conclusions about D
- **Recall:** if A,B,C,... are (mutually) *independent*, then $\Pr[A \wedge B \wedge C \wedge \ldots] = \Pr[A]\Pr[B]\Pr[C]\ldots$
- Likewise, if *random variables* W,X,Y,... are *independent* then $\mathbf{E}[WXY\ldots] = \mathbf{E}[W]\,\mathbf{E}[X]\,\mathbf{E}[Y]\ldots$ (and $\mathbf{E}[f(W)g(X)h(Y)\ldots]=\mathbf{E}[f(W)]\mathbf{E}[g(X)]\mathbf{E}[h(Y)]\ldots$)

Inference from i.i.d. examples

- **Assume**: data consists of *examples* – valuations of the propositional variables drawn independently from common D (*"i.i.d."*)
  - Will enable us to draw conclusions about D
- Suppose φ is *not* ($1-\varepsilon$)-valid under D.
- Draw $X^{(1)} = (X_1^{(1)},X_2^{(1)},\ldots,X_n^{(1)})$, $X^{(2)} = (X_1^{(2)},X_2^{(2)},\ldots,X_n^{(2)})$, ..., $X^{(m)}= (X_1^{(m)},X_2^{(m)},\ldots,X_n^{(m)})$ } Data set of m "examples" independently from D.

Inference from i.i.d. examples

- Suppose φ is *not* ($1-\varepsilon$)-valid under D.
- Draw $X^{(1)}, X^{(2)}\ldots, X^{(m)}$ independently from D.
- $\Pr[\varphi(X^{(1)}), \varphi(X^{(2)}),\ldots, \text{and } \varphi(X^{(m)})]$
  $= \Pr[\varphi(X^{(1)})]\Pr[\varphi(X^{(2)})]\ldots\Pr[\varphi(X^{(m)})]$
- *By hypothesis*, each $\Pr[\varphi(X^{(i)})] \le 1-\varepsilon$
- *So,* the probability that we fail to observe that φ is false for some example $X^{(i)}$ is at most $(1-\varepsilon)^m \le e^{-\varepsilon m}$ (*...since for all* x, $1+x \le e^x$)
- Less than any given δ if $m \ge {}^1/_\varepsilon \ln {}^1/_\delta$

Inference from i.i.d. examples

- So we have shown
  Theorem. If φ is consistent with $m \ge {}^1/_\varepsilon \ln {}^1/_\delta$ examples drawn independently from D, then with probability $1-\delta$, φ is ($1-\varepsilon$)-valid under D. (*"Probably Approximately Correct"*)
- Only guaranteed to work if φ is actually always true. What if φ is only ($1-\varepsilon'$)-valid under D, for some $\varepsilon' < \varepsilon$? (*e.g., only* ${}^7/_8$*-valid, in the case of* birds.com?)

Chernoff/Hoeffding bounds: *sample averages are good estimates*

- Let $Y^{(1)}, Y^{(2)}\ldots, Y^{(m)}$ be independent random variables taking values in [0,1]. Let $\mu = \mathbf{E}[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)})]$.
- *Hoeffding bound*: for any $\gamma > 0$,
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) > \mu+\gamma] < e^{-2m\gamma^2}$
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) < \mu-\gamma] < e^{-2m\gamma^2}$
- *Chernoff bound*: for any $\gamma > 0$,
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) > (1+\gamma)\mu] < e^{-m\mu\gamma^2/3}$
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) < (1-\gamma)\mu] < e^{-m\mu\gamma^2/2}$

Inference using Chernoff/Hoeffding

- We draw $X^{(1)}, X^{(2)} \ldots, X^{(m)}$ independently from D
- For $p = {}^1/_m([\varphi(X^{(1)})]+[\varphi(X^{(2)})]+\ldots+[\varphi(X^{(m)})])$, *how large must m be to conclude that with probability 1-δ, φ is (p±γ)-valid under D?*
- *Use the Hoeffding bound*: for any γ > 0,
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) > \mu+\gamma] < e^{-2m\gamma^2}$
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) < \mu-\gamma] < e^{-2m\gamma^2}$

TRY IT!

Inference using Chernoff/Hoeffding

- We draw $X^{(1)}, X^{(2)} \ldots, X^{(m)}$ independently from D
- For $p = {}^1/_m([\varphi(X^{(1)})]+[\varphi(X^{(2)})]+\ldots+[\varphi(X^{(m)})])$, *how large must m be to conclude that with probability 1-δ, φ is (p±γ)-valid under D?*
- *Use the Hoeffding bound*: for any γ > 0,
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) > \mu+\gamma] < e^{-2m\gamma^2}$
  $\Pr[({}^1/_m)(Y^{(1)}+Y^{(2)}+\ldots+Y^{(m)}) < \mu-\gamma] < e^{-2m\gamma^2}$
- For $m \geq {}^1/_{2\gamma^2} \ln {}^2/_\delta$, can check that $e^{-2m\gamma^2} \leq {}^\delta/_2$; take a union bound of upper and lower bounds to conclude p is within ±γ of $\Pr_D[\varphi(X)]$.

Comparison: PAC-semantics versus Inductive Logic Programming

PAC-semantics	**Inductive Logic Programming**

• Examples drawn from larger distribution D
  • D mostly unseen
	• Examples define domain
	  • Closed-world assumption

• Rules partially capture D	• Rules fully capture domain
• Rules can be (a little) inconsistent with examples	• Rules must be faithful to defining examples

Occam's Razor [Blumer et al. '87]

- **Theorem**. Let $\mathcal{H}$ be a set of formulas that can be specified using at most B bits. Suppose we draw $X^{(1)}, X^{(2)} \ldots, X^{(m)}$ independently from D for $m \geq {}^1/_{2\gamma^2}((B+1)\ln 2 + \ln {}^2/_\delta)$. With probability 1-δ, ${}^1/_m([h(X^{(1)})]+[h(X^{(2)})]+\ldots+[h(X^{(m)})])$ is within ±γ of $\Pr_D[h(X)]$ for every h in $\mathcal{H}$.
- *We know:* ${}^1/_m([h(X^{(1)})]+[h(X^{(2)})]+\ldots+[h(X^{(m)})])$ is within ±γ of $\Pr_D[h(X)]$ for any *single* h in $\mathcal{H}$ with probability $1-{}^\delta/_{2^{B+1}}$.
- There are < 2^{B+1} strings of B bits, so there are fewer than 2^{B+1} h in $\mathcal{H}$. We take a union bound.■

(Simplified) **Learning to reason** using Occam's Razor [Khardon-Roth '97]

- *Recall*: Deciding if a 3-DNF (ORs of ANDs of at most three literals) is valid is NP-complete.
- There are $2^{O(n^3)}$ 3-DNFs.
- **Theorem**. Suppose we draw $X^{(1)}, X^{(2)} \ldots, X^{(m)}$ independently from D for $m = O({}^1/_{\gamma^2}(n^3+\ln {}^1/_\delta))$. Then with probability 1-δ, ${}^1/_m([h(X^{(1)})]+[h(X^{(2)})]+\ldots+[h(X^{(m)})])$ is within ±γ of $\Pr_D[h(X)]$ for every 3-DNF h.
- In particular, if all $h(X^{(i)})=1$, we guarantee h is at least (1-γ)-valid with probability 1-δ.

Learning a chaining rule using **Elimination** [Valiant '84]

- Suppose x_t is defined by a *conjunction*, i.e., $x_t \Leftrightarrow \ell_1 \wedge \ell_2 \wedge \cdots \wedge \ell_k$ where $\ell_1, \ell_2, \ldots, \ell_k$ are literals.
- There are only 2^{2n} conjunctions of literals on n propositional variables.
- <u>*Occam's Razor*</u>: *any* rule $x_t \Leftrightarrow \ell_1 \wedge \ell_2 \wedge \cdots \wedge \ell_k$ that is consistent with $m = O({}^1/_{\gamma^2}(n+\ln {}^1/_\delta))$ examples is (1-γ)-valid with probability 1-δ.
- We only need to <u>find</u> such a rule.

Learning a chaining rule using **Elimination** [Valiant '84]

- We only need to find $x_t \Leftrightarrow \ell_1 \wedge \ell_2 \wedge \cdots \wedge \ell_k$ consistent with $m = O(\frac{1}{\gamma^2}(n+\ln\frac{1}{\delta}))$ examples
- *Elimination*: for each ith example, if $X_t^{(i)} = 1$, then any literal ℓ with $\ell(X^{(i)}) = 0$ cannot be included in the rule. Delete it.
- The rule given by the conjunction of the remaining literals must contain the actual defining conjunction $\ell_1 \wedge \ell_2 \wedge \cdots \wedge \ell_k$.
- Therefore, we find a conjunction that is consistent with all m examples, as needed.■

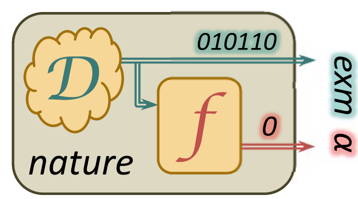
Learning a chaining rule using **Elimination** [Valiant '84]

- We find a $(1-\gamma)$-valid rule using $O(\frac{1}{\gamma^2}(n+\ln\frac{1}{\delta}))$ examples by taking the conjunction of all literals for which whenever $\ell(X^{(i)}) = 0$, $X_t^{(i)} = 0$.
- As stated, the algorithm runs in time $O(\frac{n}{\gamma^2}(n+\ln\frac{1}{\delta}))$ so this is efficient.
  - *Note*: actually, $O(\frac{1}{\gamma}(n+\ln\frac{1}{\delta}))$ examples will do

B. Introduction to PAC-Semantics

Coping with Partial Information

PAC Learning (with Complete Info)



Given access to *(exm,α)'s* drawn during training, w.h.p. and efficiently produce *h* s.t. on *(exm',α')* drawn during testing, $h(exm') \neq \alpha'$ w.p. at most ε.

Learning and Prediction Scenarios

- Construct hypothesis to predict target x_1.
  - 1101100011011 learning
  - ?011101001010 prediction
- What if we want to predict multiple x_j's?
  - 1101100011011 learning
  - ?01?1??0010?0 prediction
- What if we want to learn **autonomously**?
  - 1?011?001?0?? learning
  - ?01?10?001010 prediction

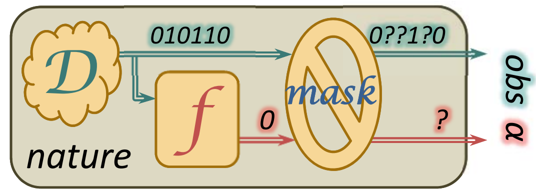
Complete the Missing Value

Marital Status	Age	Smoking Habits	Drinking Habits	Blood Pressure	Cholesterol Level	Exercising Habits
single			3 weekly	130/90		
married parent	49	cigarettes 10 daily			high	
	32		no drinking		normal	jogging daily
divorced			2 daily	145/100		
	68	cigars 3 daily		?		no exercise
single parent		no smoking		125/80	slightly elevated	
	25		rarely		normal	gym 3 weekly

Structure in Missingness?

- Q: What is the number of your credit card?
  - The responder may not wish to share it.
  - What if the responder does not have one?
- Q: When was the last time you ate apples?
  - The responder may have a poor memory.
- Q: Where you ever convicted for murder?
  - The responder may not wish to answer…
  - … especially if the answer would be "yes"!

Autodidactic Learning (PAC + Missing)



Given access to *(obs,α)'s* drawn during training, w.h.p. and efficiently produce *h* s.t. on *(obs',α')* drawn during testing, $h(obs') \neq \alpha'$ w.p. at most ε.

Reading Between the Lines

[Michael '09]

Learning the Hidden Reality

- Need learning algorithms able to:
  - identify structure in the hidden example,
  - given access only to partial observations,
  - without knowing how masking works.
- Autodidactic learning [Michael '08,10]:
  - Suffices to learn rules ***consistent*** with obs.
    - e.g., predict that $x_1 = 0$ in *?010??11*
  - W.h.p., these rules make ***accurate*** predictions.

Evaluation of an Individual Rule

example	observation	evaluate rule $(x_5 \vee \neg x_7) \Leftrightarrow x_3$		
		inference	consistent	accurate
0010110	0?10??0	1	Yes	Yes
0100100	?10??00	1	No	No
0100100	?10??0?	?	Yes	Abstain
0010110	0??01??	1	Yes	Yes
0100100	0??01??	1	Yes	No

Theorem: For each mask there is η s.t. each $(1-\eta\cdot\varepsilon)$-consistent rule is $(1-\varepsilon)$-accurate. The "discount" is tight.

When to Abstain from Predictions

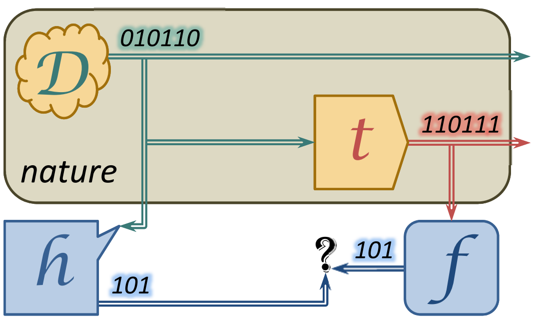
- ***Choosing*** to abstain would trivialize learning.
- Hypotheses are still Boolean functions…
- Abstain only if hypothesis not fully determined.
  - Assume hypothesis is $h = x_1 \wedge \neg x_3 \wedge x_7$
  - If observation is x = *α10?1101* then $h(x) = ?$
  - If observation is x = *α10?1100* then $h(x) = 0$
- Does missing info "kill" our ability to predict?

Activity: Predict the Hidden Animal

Autodidactic Learnability [Michael '08,10]

- **Theorem:** All classes of monotone and read-once formulas that are PAC-learnable, are also autodidactically learnable (***arbitrary*** masking).
- Learning algorithm:
  - Ignore observations where the label is masked.
  - During training map *α0??1?0* to *α0αα1α0*.
  - PAC-learn from resulting (complete) instances.
- **Theorem:** Parities and monotone term 1-decision lists are ***not properly learnable***.

Causal Learnability [Michael '11]



B. Introduction to PAC-Semantics

Dealing with First-Order Expressions

An Example of Learned Knowledge

- NL query : "members share something"
- logic form : *member(t_1), something(t_3), share(t_1,t_3)*
- knowledge : *(trained on "spyware" web pages)*

  file(x) ⇔ **threshold(1.0)** % pos:1852 neg:1831
  ∃v : scan(v,x) ∧ rogue(v) **weight(0.962710)** % pos:16 neg:1
  ∃v : share(v,x) **weight(1.627098)** % pos:11 neg:1
  ∃v : have(x,v) ∧ program(v) **weight(0.645691)** % pos:19 neg:0
  ∃v : open(v,x) **weight(1.593269)** % pos:27 neg:2
- inference : *file(t_3)*
- NL answer : "something is a file" **[Michael '13]**

How To Learn Relational Rules?

- Looking to learn a rule in the correct form:
  - No tokens (they carry no meaning), small arity.
  - ∀xs [*formula over IQEs* ⇔ *head_predicate(xs)*]
- The formula belongs in a concept class that is known to be PAC learnable (cf. **[Valiant '00]**).
  - Linear thresholds with propositional features.
  - Recall: also learnable from partial observations.

Relational Scenes → Propositional Examples → Propositional Learning → Relational Hypothesis

Activity: Learn a Linear Threshold

- Initially assign weight 0.8 to every proposition.
- When a negative example is predicted true:
  divide by 2 weights of true propositions.
- When a positive example is predicted false:
  multiply by 2 weights of true propositions.
- Hidden target concept:

  000110, *011101*, *101011*, *001110*, *110010*,
  011001, *111111*, *101011*, *001101*, *011101*.
- Computed hypothesis: ⟨*0.2, 0.1, 0.05, 0.8, 0.2*⟩

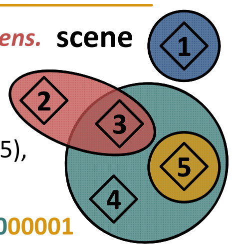
Examples from Relational Scenes

- Instead of Boolean learning examples, we get ***relational scenes*** giving instances of relations:
  file(t_2), *scan*(t_7,t_2), *rogue*(t_7), ¬*file*(t_5), ¬*open*(t_5,t_2).
- For a head predicate (e.g., *file(x)*), ***consider all IQEs*** that follow a schema (e.g., *scan* ∧ *rogue*):

scan(x,x) ∧ *rogue(x)*, ∃v : *scan(x,x)* ∧ *rogue(v)*,
∃v : *scan(x,v)* ∧ *rogue(x)*, ∃v : *scan(x,v)* ∧ *rogue(v)*,
∃v : *scan(v,x)* ∧ *rogue(x)*, ∃v : *scan(v,x)* ∧ *rogue(v)*,
∃u,v : *scan(u,v)* ∧ *rogue(v)*, ∃u,v : *scan(u,v)* ∧ *rogue(u)*.

Propositionalizaton of Learning

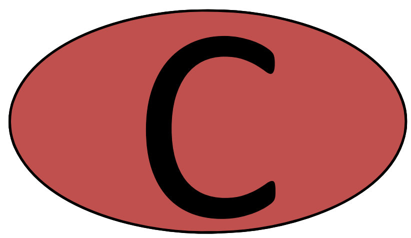
Full info. Small arity. Few tokens. **scene**



P(1), P(2), …, P(5),
P(1,1), …, P(2,3), …, P(5,5),
P(1,1,1), …, P(3,5,4), …, P(5,5,5),
P(1), P(2), …, P(5).

10000000000010…00…010…000001

x=3 –P(**x**) with ∃P(**x**,a,b), ∃P(a,**x**), …
x=5 +P(**x**) with ∃P(a,**x**,b), ∃P(a,**x**,b)∧P(c,a), …



High-Level Tutorial Roadmap

A. **Why KRR Should Embrace Learning**

B. **Introduction to PAC-Semantics**

C. **Integrating Deduction and Induction**

D. **Reasoning Non-Monotonically**

E. **Overall Summary and Conclusions**

C. **Integrating Deduction and Induction**

Interleaving Learning and Reasoning

The Declaration of Independence (of KRR from Machine Learning)

We hold these truths to be self-evident:
an appropriate knowledge base is given
reasoning is evaluated against the KB
chaining of rules is trivially beneficial
KB rules can be applied in any order
acquisition of KB can be done a priori

Reasoning ≈ Fill-In Missing Values

Marital Status	Age	Smoking Habits	Drinking Habits	Blood Pressure	Cholesterol Level	Exercising Habits
single			3 weekly	130/90		
married parent	49	cigarettes 10 daily			high	
	32		no drinking		normal	jogging daily
divorced			2 daily	145/100		
	68	cigars 3 daily		?		no exercise
single parent		no smoking		125/80	slightly elevated	
	25		rarely		normal	gym 3 weekly

KB Performance Evaluation

- Example knowledge base:
  - (x_3 and not x_5) determines the value of x_2
  - (not x_4 or not x_1) determines the value of x_6

example	observation	KB inference	KB evaluation	
101011	1?101?	101011	no disagreement no "don't know"	50% sound 75% complete
001001	0?100?	011001	disagreement no "don't know"	
010101	0??101	0??101	no disagreement "don't know"	
010001	110000	110001	disagreement no "don't know"	

Multiple Targets and Autonomy

- Predict multiple x_j's and learn autonomously.
  - 1?011?001?0?? learning
  - ?01?10?001010 prediction
- Options on how to tackle this task:
  - Learn hypotheses first, then apply in parallel.
  - Learn hypotheses first, then apply by chaining.
  - Simultaneously learn hypotheses and predict.
- Which of these approaches is appropriate?

Basic Hypotheses / Predictors

- Given a specified attribute x_i of the world, and sufficient resources during training,
- produce a PAC ***predictor*** (=rule) P_j for x_i
  - e.g., $x_7 \equiv x_1 \wedge \neg x_3$ is a predictor for x_7
- Predictor P_j is ***"classical"*** and could potentially
  - abstain, when its body is not fully determined
  - predict incorrectly (cf. approximately correct)
  - but generally improve completeness of the DB

Reasoning Process Semantics

- $exm = \langle -1, 0, 5, 0, 4, -7, 9\rangle \in S$
- $obs = \langle -1, 0, *, 0, 4, *, 9\rangle \in S^*$

Policy $\mathbb{P}$ determines how predic[tors are applied]

$(P_1)\ x_3 \equiv \sqrt{x_5} + \sqrt{x_7}$ $(P_2)\ x_4 \equiv x_2 \cdot x_6$ ... $x_5/2$

endogenous and exogenous qualification (***hints of NMR?***)

e.g., $\mathbb{P} = \langle\{P_1, P_2, P_3\}\rangle$ ***"flat/parallel"*** policy

$\mathbb{P}(obs) = \langle -1, 0, 5, 0, 4, *, 9\rangle$

sound against *exm* but ***incomplete***

e.g., $\mathbb{P} = \langle\{P_1, P_2\}, \{P_3\}\rangle$ ***"chaining"*** of predictors

$\mathbb{P}(obs) = \langle -1, 0, 5, 0, 4, 3, 9\rangle$

unsound against *exm* but complete

Activity: Chaining & Completeness

- Example knowledge base:
  - (noon time) ⇔ lunch time
  - (lunch time and Bob hungry) ⇔ Bob eats
- Statement:
  - "It was not noon time yet, but Bob was hungry."

- Multi-layered application:
- Single-layered application:
- Chaining is better than ***any*** single-layered KB.

Chaining is *Provably* Beneficial

Definition: ***Chaining collapses*** on S^* against S if

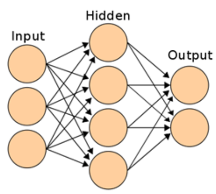
- for every policy $\mathbf{P}$ with some ***performance*** (soundness + completeness) on S^* against S,
- there exists a flat policy $\mathbf{P}'$ (not a reordering of $\mathbf{P}$, necessarily) with equally high performance.

Theorem: **[Michael '14]** There exist S^*, S such that chaining does not collapse on S^* against S.

Proof: Chaining can ***simulate non-monotonic reasoning***, which is beyond individual predictors.

Analogy to Neural Networks

There are multilayered NNs that compute more complex functions than those by ***any*** NN without hidden layers.



Trivially, because of larger hypothesis space.

What if each neuron can compute any function?
What if neurons can abstain from predictions?

Learn First, Then Predict (& Chain)

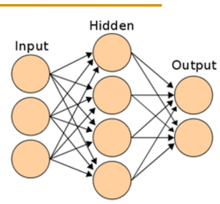
- Consider rules R_1 and R_2 obtained to make highly accurate predictions on distribution D.
- On future examples from D, apply R_1 then R_2.

- ***No! R_2 is applied on distribution $R_1 \circ D$ ($\neq D$). There are no guarantees on that distribution!***

- Are there situations that justify this approach?
  - If information during learning is complete.
  - Could happen, but less realistic assumption.

Analogy to Neural Networks

Cannot train each neuron independently and then assign neurons in layers.



Trivially, because the target to be learned by each specific neuron is not directly observable.

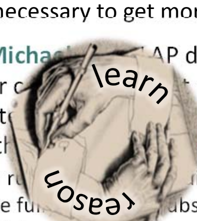
What if one could train each neuron in isolation?

Simultaneously Learn and Predict

- Learn rules for all attributes during training.
- Apply those rules to enhance the training set.
- Repeat as necessary to get more KB "layers".

- **Theorem:** **[Micha** … AP does not reduce soundness or c … may produce KBs with bett … completeness (combined) t … ***lassical***) KB.

Proof: Train each r … istribution, until its predictions are fu … ubsequent training. Doubly-exponential depe … on reasoning depth.



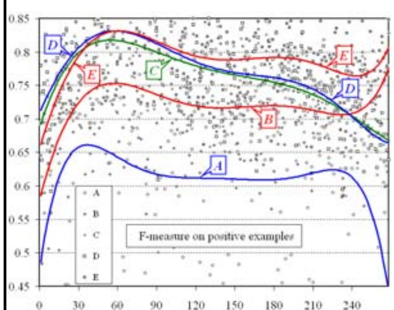
An Experimental Impasse?

- Scene is an observation. e.g., 1?10010??10
- Goal: Predict what information holds in scene.
- Predictions useful exactly when information is missing from scene. ***How to evaluate them?!***
- Evaluate another task with known answers!
  - Use standard approach to solve the task. (1)
  - Make predictions, and give them as input.
  - See whether task performance improves. (2)

Guess the Missing Word [Michael+ '08]

- Training T_0 & testing E_0 from news text corpus.
  - Learn ***underlying text structure*** **[Michael '09]**.
- Using T_0 learn ***relational*** thresholds for 268 frequently-used words: *price*, *market*, *stock*, …
  - Train on T_0 and test on E_0 for each of 268 tasks.
- Using T_0 learn ***relational*** thresholds for 599 verbs: *buy*$_{sbj,obj}$, *charge*$_{obj}$, *coerce*$_{obj,prd}$, …
  - ***Use verb rules on T_0 to get T_1 / on E_0 to get E_1.***
  - Train on T_1 and test on E_1 for each of 268 tasks.

Experimental SLAP Results [Michael+ '08]



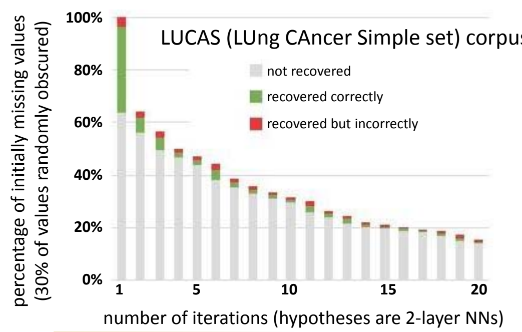
"missing word" task
train on T_0 & test on E_0

A: word/pos + verb
C: word/pos + prox.
D: use all instances

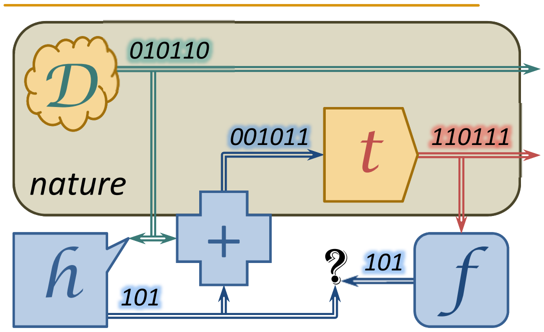
"missing word" task
train on T_1 & test on E_1

B: word/pos + verb
E: use all instances

Multiple Iteration SLAP [Skouteli+ '16]



Introspective Forecasting [Michael '15]



C. Integrating Deduction and Induction

Implicit Learning of Testable KBs

A problem requiring learning and deduction: birds.com Analytics

birds.com™

The **data**

FLY?

The (statistical) **query**: *"true with high probability for data source?"*

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
107	54	Mouse
107	55	Grubs
⋮	⋮	⋮

THE DATA ALONE DOESN'T ANSWER THE QUERY

birds.com: Inferring *Flying* from *Food?*

- **Data:**

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
⋮	⋮	⋮

- **Query**: FLY?
- **Background knowledge:**
  PENGUIN ⋁ FLY
  ¬PENGUIN ⋁ EAT(FISH)
  ¬EAT(FISH) ⋁ ¬EAT(GRUBS)
  ¬EAT(FISH) ⋁ ¬EAT(MOUSE) } *foods are mutually exclusive…*
  ¬EAT(FISH) ⋁ ¬EAT
  …etc…

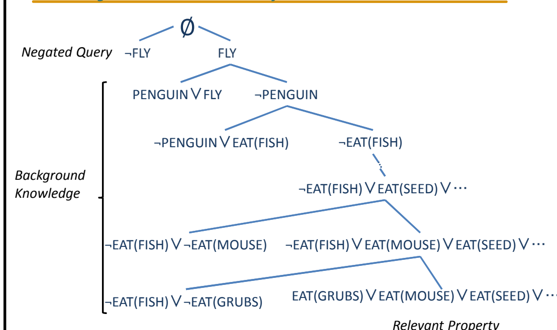
THE KNOWLEDGE ALONE ALSO DOESN'T ANSWER THE QUERY

Relevant knowledge we aim to *discover* from the data

- **Relevant property:** the birds eat anything but fish—
  EAT(GRUBS) ⋁ EAT(MOUSE) ⋁ EAT(SEED) ⋁ ⋯
  - *Usually* satisfied by *Data* ✔
  - Completing a *Proof* of *Query*…

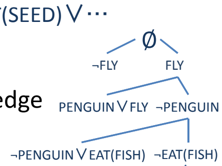
Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
⋮	⋮	⋮

The *Relevant Property* completes a *Proof* of the *Query*



Relevant knowledge we aim to *discover* from the data

- **Relevant property:** the birds eat anything but fish—
  EAT(GRUBS) ⋁ EAT(MOUSE) ⋁ EAT(SEED) ⋁ ⋯
  - *Usually* satisfied by *Data* ✔
  - Completing a *Proof* of *Query* ✔
- The **existence** of such knowledge establishes the query (FLY)
  ☞*It isn't important what the actual property says!*



☞We are not told what property to search for…

Approximate Query Answering with Implicit Learning

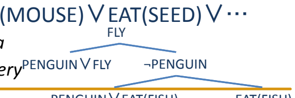
- **Data**:

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
⋮	⋮	⋮

- **Background knowledge**: PENGUIN ⋁ FLY ¬PENGUIN ⋁ EAT(FISH)
  ¬EAT(FISH) ⋁ ¬EAT(GRUBS) …etc…
- **Query**: FLY?

DON'T NEED TO SAY WHAT PROPERTY

- **Verdict:** *Does there exist a* **relevant property**
  e.g., EAT(GRUBS) ⋁ EAT(MOUSE) ⋁ EAT(SEED) ⋁ ⋯
  - *Usually* satisfied by *Data*
  - Completing *Proof* of *Query*



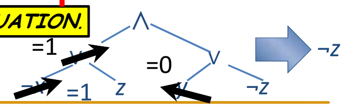
Proposed Algorithm
for Approximate Query Answering

Given a query φ, background knowledge KB, target ε, and i.i.d. partial examples $\rho^{(1)},\ldots,\rho^{(m)}$

- For each partial example $\rho^{(i)}$,
  - If $\varphi|_{\rho^{(i)}}$ provable from $\mathrm{KB}|_{\rho^{(i)}}$, increment *SUCCESS* (initially, 0)
- If $^{SUCCESS}/_m$ > 1-ε, ***ACCEPT***, otherwise ***REJECT***

PARTIAL EVALUATION

ρ: x=0, y=0



Proposed Algorithm
for Approximate Query Answering

Given a query φ, background knowledge KB, target ε, and i.i.d. partial examples $\rho^{(1)},\ldots,\rho^{(m)}$

- For each partial example $\rho^{(i)}$,
  - If $\varphi|_{\rho^{(i)}}$ provable from $\mathrm{KB}|_{\rho^{(i)}}$, increment *SUCCESS* (initially, 0)
- If $^{SUCCESS}/_m$ > 1-ε, ***ACCEPT***, otherwise ***REJECT***

PARTIAL EVALUATION: *PLUG IN $\rho^{(i)}$ AND RECURSIVELY **REPLACE EACH LOCALLY DETERMINED** CONNECTIVE BY ITS **VALUE**.*

Algorithm on birds.com Problem

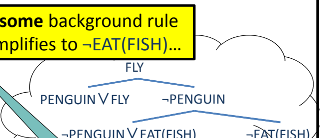
birds.com™ FLY?

On **most** partial examples…

107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
107	54	Mouse
107	55	Grubs
⋮	⋮	⋮

PENGUIN ⋁ FLY
¬PENGUIN ⋁ EAT(FISH)
¬EAT(FISH) ⋁ ¬EAT(GRUBS)
¬EAT(FISH) ⋁ ¬EAT(MOUSE)
¬EAT(FISH) ⋁ ¬EAT(SEED)
…etc…

… **some** background rule simplifies to ¬EAT(FISH)…



ACCEPT

… completing a *proof* of the *query*

Approximate Queries: The theorem
[Juba '13]

Theorem. *Our algorithm distinguishes:*

I. The query is only satisfied by *D* w.p. < 1-ε-γ

II. **There exists** a *"(1-ε+γ)-testable"* formula ψ for which **there exists** a proof of the query from ψ and any background knowledge

w.p. 1-δ, *given* $^1/_{\gamma^2}$ ln $^1/_\delta$ *partial examples from M(D)*

Key Definition: Testable formulas
Relevant Properties we can detect [Juba '13]

- **Definition.** A formula ψ is (1-ε)-*testable* under a distribution over partial examples M(D) if

$$\Pr_{M(D)}[\psi|_\rho=1] \geq 1\text{-}\varepsilon$$

PARTIAL EVALUATION: *PLUG IN ρ AND RECURSIVELY **REPLACE EACH LOCALLY DETERMINED** CONNECTIVE BY ITS **VALUE**.*

Key Definition: Testable formulas
Relevant Properties we can detect [Juba '13]

- **Definition.** A formula ψ is (1-ε)-*testable* under a distribution over partial examples M(D) if

$$\Pr_{M(D)}[\psi|_\rho=1] \geq 1\text{-}\varepsilon$$

birds.com™

Day	Bird no.	Food
107	48	Seed
107	49	Grubs
107	50	Mouse
107	51	Mouse
107	52	Worm
107	53	Fish
⋮	⋮	⋮

EAT(GRUBS) ⋁ EAT(MOUSE) ⋁ EAT(SEED) ⋁ ⋯

Key Definition: Testable formulas

Relevant Properties we can detect [Juba '13]

- **Definition.** A formula ψ is (1-ε)-*testable* under a distribution over partial examples M(D) if
  $\Pr_{M(D)}[\psi|_\rho=1] \geq 1-\varepsilon$
- Require *more* than truth of Relevant Property
  - *Standard cases (clause/linear inequality premises):* actually **no** more demanding

Approximate Queries: The theorem

[Juba '13]

Theorem. *Our algorithm distinguishes:*

I. The query is only satisfied by *D* w.p. < 1-ε-γ

II. **There exists** a (1-ε+γ)-testable formula ψ for which **there exists** a proof of the query from ψ and any background knowledge

w.p. 1-δ, *given* $^1/_{\gamma^2} \ln {^1/_\delta}$ *partial examples from M(D)*

Proposed Algorithm for Approximate Query Answering

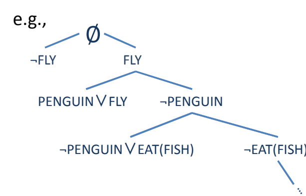
Given a query φ, background knowledge KB, target ε, and i.i.d. partial examples $\rho^{(1)},\ldots,\rho^{(m)}$

- For each partial example $\rho^{(i)}$,
  - If $\varphi|_{\rho^{(i)}}$ provable from $\mathrm{KB}|_{\rho^{(i)}}$, increment SUCCESS (initially, 0)
- If $^{SUCCESS}/_m$ > 1-ε, ***ACCEPT***, otherwise ***REJECT***

Detecting a Relevant Property…

Consider: Tractable Proof Systems

- Bounded-width resolution
- Treelike, bounded clause space resolution

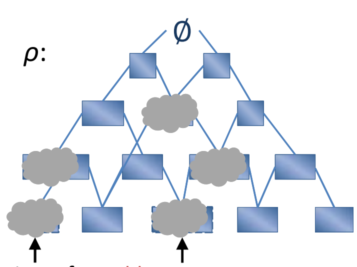


A Key Property of the Example Tractable Proof Systems

- Bounded-width resolution
- Treelike, bounded clause space resolution

☞*Partial evaluation of proofs of these forms yields proofs of the same form*
(from a proof of a query φ, we obtain a proof of $\varphi|_\rho$ of the same syntactic form)

Insight: *The Testable Premises Drop Out of the Proof!*



Premises of testable property

Recall: Proposed Algorithm Detects These Residual Proofs

Given a query φ, background knowledge KB, target ε, and i.i.d. partial examples $\rho^{(1)},\ldots,\rho^{(m)}$

- For each partial example $\rho^{(i)}$,
  - If $\varphi|_{\rho^{(i)}}$ provable from $KB|_{\rho^{(i)}}$, increment *SUCCESS* (initially 0)
- If $^{SUCCESS}/_{m} > 1\text{-}\varepsilon$, ***ACCEPT***, otherwise ***REJECT***

THE THEOREM THEREFORE FOLLOWS IMMEDIATELY FROM HOEFFDING'S INEQUALITY

Pros and cons of Implicit Learning

- **Pro**: utilizes rules with imperfect validity
  - Usually intractable to learn explicitly
  - Captures kinds of commonsense reasoning *(next part)*
- **Pro**: reasoning time complexity independent of size of implicit KB
  - Actually, may reduce reasoning complexity in some circumstances [Juba '15]
- **Con**: cannot report rules used to support conclusion

D

High-Level Tutorial Roadmap

A. Why KRR Should Embrace Learning

B. Introduction to PAC-Semantics

C. Integrating Deduction and Induction

D. Reasoning Non-Monotonically

E. Overall Summary and Conclusions

D. Reasoning Non-Monotonically

Conditional Probability and NMR

Nonmonotonicity in learning to reason [Roth '95]

- *We have seen*: deciding queries by counting the frequency with which they are provable
  - Allows us to answer "hard" queries
  - Draws on a potentially large KB of implicitly learned knowledge as needed
- *New twist*: suppose we incorporate a hypothesis by **filtering out examples that do not satisfy the hypothesis**
  - Produces desirable non-monotonic inferences, appropriate for *"commonsense reasoning"*

Nonmonotonicity in learning to reason [Roth '95]

- **Example**: suppose we are reasoning about "birds" – we filter the set of examples to *only* include bird = 1:

bird	fly	has_beak	red	purple	penguin
1	1	1	0	0	0
1	1	1	1	0	0
1	0	1	0	0	1
1	0	1	0	0	1
1	1	1	0	0	0
1	1	1	1	0	0
1	1	1	0	0	0

- We find has_beak is $(1-\gamma)$-valid, fly is $({}^{5}/_{7}-\gamma)$-valid, but red and penguin are at most $({}^{2}/_{7}+\gamma)$-valid…

Nonmonotonicity in learning to reason [Roth '95]

- **Example**: Now suppose we consider specifically red birds, filtering to *only* include bird $\wedge$ red = 1:

bird	fly	has_beak	red	purple	penguin
[redacted]					
1	1	1	1	0	0
[redacted]					
1	1	1	1	0	0
[redacted]					

- Now has_beak and fly are (still) $(1-\gamma')$-valid, and penguin is at most γ'-valid (for some $\gamma' > \gamma$)…

Nonmonotonicity in learning to reason [Roth '95]

- **Example**: now suppose we are considering penguins, filtering to only include penguin= 1:

bird	fly	has_beak	red	purple	penguin
[redacted]					
1	0	1	0	0	1
1	0	1	0	0	1
[redacted]					

- We find has_beak is still $(1-\gamma')$-valid, but fly and red are now at most γ'-valid

Nonmonotonicity in learning to reason [Roth '95]

- **Example**: now considering specifically "penguins with beaks," we only include penguin $\wedge$ has_beak = 1:

bird	fly	has_beak	red	purple	penguin
[redacted]					
1	0	1	0	0	1
1	0	1	0	0	1
[redacted]					

- Again, we find fly and red are at most γ'-valid – not affected by has_beak

Nonmonotonicity in learning to reason [Roth '95]

- **Example**: if we instead consider "purple penguins," we only include examples with penguin $\wedge$ purple = 1:

bird	fly	has_beak	red	purple	penguin
[redacted]					

- With no examples remaining, we cannot draw *any* conclusions (except perhaps from a given KB)

Nonmonotonicity and conditional probability distributions

- A set of examples filtered to satisfy a formula h has the conditional probability distribution D|[h(X)=1] (we *"condition on h"*)
- *So*: as long as we consider $(1-\varepsilon)$-validity for some $\varepsilon>0$ (e.g., $\varepsilon = {}^{1}/_{3}$ could have sufficed), conditioning may have a *non-monotonic* effect
- *Note*: this requires the use of non-negligibly large ε
  - Since: we must have enough examples in the conditional distribution to support inferences

Classical issues in commonsense reasoning, in PAC-semantics

- **Qualification problem**: taking $\varepsilon>0$ permits a rule to fail for any number of unspecified and unmodeled reasons. The commonsense rule we implicitly use is never written out.
- **Elaboration tolerance**: we can simply add a new example to our set of examples, and the next time we answer a query the count will be slightly different. But the implicit KB does not need to be "edited" in any way.

Classical issues in commonsense reasoning, in PAC-semantics

- **Ramification problem**: any further consequences r to a hypothesis h are included: for any example x satisfying h, we are given that x also satisfies r, so r will be highly valid in D|[h(X)=1]. So, r is included in the implicit KB of D|[h(X)=1] without further consideration.
  - E.g.: ¬fly is highly valid in D|[penguin(X)=1].

Context in reasoning using "**preconditions**" [Valiant '94,95,00]

- The condition h in D|[h(X)=1] is sometimes called a *"precondition"*
- Valiant proposed: when answering a query, a precondition capturing the current context should be used to filter examples.
  - E.g., given by the units currently firing in the "neuroidal" cognitive model [Valiant '94]
- **Problem:** may be too specific.
- *How should we choose a precondition?*

D. Reasoning Non-Monotonically

Preconditions and Abduction

Reasoning with **preconditions** [Juba '16]

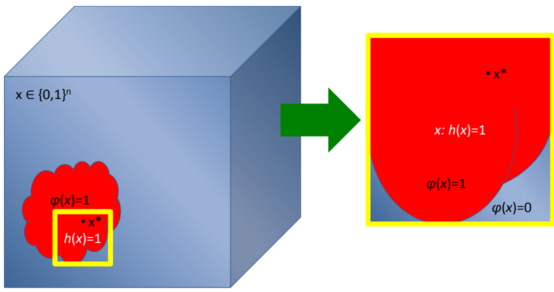
- One possible formulation of reasoning with preconditions:

  "Does there exist a precondition h from a class of representations H such that…
  1. h supports the query φ
  2. h is common
  3. h is consistent with the current context x?"
- If *any* precondition supports the query, we will find one.

Reasoning with **preconditions**: formalization [Juba '16]

- Fix a class of Boolean representations $\mathcal{H}$
- **Given** query formula φ; context assignment x*, ε, δ, μ∈(0,1); access to examples from D,
- Suppose that there exists a $h^*\in\mathcal{H}$ such that
  1. *h^* supports φ:* $\Pr_D[\varphi(X)|h^*(X)] = 1$
  2. *h^* is common:* $\Pr_D[h^*(X)] \geq \mu$
  3. *h^* is consistent with context x^*:* $h^*(x^*) = 1$
- **Find** a h *(ideally in $\mathcal{H}$)* such that with prob. 1-δ,
  1. $\Pr_D[\varphi(X)|h(X)] \geq 1-\varepsilon$
  2. $\Pr_D[h(X)] \geq \mu'$ for some μ' *(ideally close to μ)*
  3. $h(x^*) = 1$

The precondition task, in pictures…



Reasoning with k-DNF preconditions is **possible** *(complete information)*

Theorem. If there is a k-DNF h^* such that

1. *h^* supports φ:* $\Pr_D[\varphi(X)\,|\,h^*(X)] = 1$
2. *h^* is common:* $\Pr_D[h^*(X)] \geq \mu$
3. *h^* is consistent with context x^*:* $h^*(x^*) = 1$

then using $m = O(^1/_{\mu\varepsilon}\,(n^k+\log{}^1/_\delta))$ examples, in time $O(mn^k)$ we can find a k-DNF h such that with probability $1-\delta$,

1. *h supports φ:* $\Pr_D[\varphi(X)\,|\,h(X)] \geq 1-\varepsilon$
2. *h is common:* $\Pr_D[h(X)] \geq \mu$
3. *h is consistent with context x^*:* $h(x^*) = 1$

Finding a supporting k-DNF precondition using Elimination

- Start with h as an OR over <u>all</u> terms of size k
- For each example $x^{(1)},\ldots,x^{(m)}$
  - If $\varphi(x^{(i)}) = 0$,
    delete all terms T from h such that $T(x^{(i)}) = 1$
- If $h(x^*)=1$, return h
- Else return *FAIL (no supporting precondition)*

Running time is still clearly $O(mn^k)$

Analysis pt 1: $\Pr_D[h(X)] \geq \mu$, $h(x^*) = 1$

- We are given that some k-DNF h^* has
  1. *h^* supports φ:* $\Pr_D[\varphi(X)\,|\,h^*(X)] = 1$
  2. *h^* is common:* $\Pr_D[h^*(X)] \geq \mu$
  3. *h^* is consistent with context x^*:* $h^*(x^*) = 1$
- Initially, <u>every</u> term of h^* is in h
- Terms of h^* are <u>never</u> true when $\varphi(x)=0$ by *1*.

☞<u>every</u> term of h^* remains in h

☞*h^* implies h,* so $\Pr_D[h(X)] \geq \Pr_D[h^*(X)] \geq \mu$ and $h(x^*)=1$ since $h^*(x^*)=1$ by *3*.

Analysis pt 2: $\Pr_D[\varphi(X)\,|\,h(X)] \geq 1-\varepsilon$

- Rewrite conditional probability:
  $\Pr_D[[\neg\varphi(X)]\wedge h(X)] \leq \varepsilon \Pr_D[h(X)]$
- We'll show: $\Pr_D[[\neg\varphi(X)]\wedge h(X)] \leq \varepsilon\mu$
  ($\leq \varepsilon \Pr_D[h(X)]$ by part 1)
- Consider any h' s.t. $\Pr_D[[\neg\varphi(X)]\wedge h'(X)] > \varepsilon\mu$
  - Since each $X^{(i)}$ is drawn independently from D
    $\Pr_D[\text{no } i \text{ has } [\neg\varphi(X^{(i)})]\wedge h'(X^{(i)})] < (1-\varepsilon\mu)^m$
  - A term of h' is deleted when $\varphi=0$ and $h'=1$
  - So, h' is only possibly output w.p. $< (1-\varepsilon\mu)^m$

Analysis pt 2, cont'd: $\Pr_D[\varphi(X)|h(X)] \geq 1-\varepsilon$

- We'll show: $\Pr_D[[\neg\varphi(X)]\wedge h(X)] \leq \varepsilon\mu$
- Consider any h' s.t. $\Pr_D[[\neg\varphi(X)]\wedge h'(X)] > \varepsilon\mu$
  - h' is only possibly output w.p. $< (1-\varepsilon\mu)^m$
- There are only $2^{O(n^k)}$ possible k-DNF h'
- *Since* $1-z \leq e^{-z}$, $m = O(^1/_{\mu\varepsilon}\,(n^k+\log{}^1/_\delta))$ ex's suffice to guarantee that each such h' is only possible to output w.p. $< \delta/2^{O(n^k)}$

☞w.p. $>1-\delta$, h has $\Pr_D[[\neg\varphi(X)]\wedge h(X)] \leq \varepsilon\mu$.■

Reasoning with k-DNF preconditions is **possible** *(complete information)*

Theorem. If there is a k-DNF h^* such that

1. *h^* supports φ:* $\Pr_D[\varphi(X)\,|\,h^*(X)] = 1$
2. *h^* is common:* $\Pr_D[h^*(X)] \geq \mu$
3. *h^* is consistent with context x^*:* $h^*(x^*) = 1$

then using $m = O(^1/_{\mu\varepsilon}\,(n^k + \log{}^1/_\delta))$ examples, in time $O(mn^k)$ we can find a k-DNF h such that with probability $1-\delta$,

1. *h supports φ:* $\Pr_D[\varphi(X)\,|\,h(X)] \geq 1-\varepsilon$
2. *h is common:* $\Pr_D[h(X)] \geq \mu$
3. *h is consistent with context x^*:* $h(x^*) = 1$

Extension: Finding preconditions **tolerating ε>0**

- **Given only** that some h^* achieves
  1. *h^* supports φ:* $\Pr_D[\varphi(X)\,|\,h^*(X)] \geq 1-\varepsilon$
  2. *h^* is common:* $\Pr_D[h^*(X)] \geq \mu$
  3. *h^* is consistent with context x^*:* $h^*(x^*) = 1$

Find an h such that for some other μ' & ε',

1. $\Pr_D[\varphi(X)\,|\,h^*(X)] \geq 1-\varepsilon'$
2. $\Pr_D[h(x)] \geq \mu'$ for some μ' *(ideally close to μ)*
3. $h(x^*) = 1$

- *Extension of algorithm for k-DNF achieves* $\mu'=\mu$, $\varepsilon'=O(n^k\varepsilon)$ (only delete T making $\varepsilon\mu m$ errors)

Abductive reasoning: *making plausible guesses*

- This formulation of precondition search is a form of ***abductive reasoning***
  - **Given** a *conclusion c*, **find** a *"plausible" h* that implies/leads to/… *c*
  - Proposing a precondition supporting the query
- Two varieties of *"plausibility"* in common use
  - *Syntactic*: a small *h* from which *c* follows
  - *Bayesian*: a *h* which has large posterior probability when given *c* *ie., "*Pr [*h actual rule used*|*c true*] > *…"*

Why might we **want** a new model?

- Existing models only **tractable** in *simple cases*
  - *E.g.* Horn rules ($a \wedge b \wedge c \Rightarrow d$ …no negations), "nice" (conjugate) priors
- The *choice* of formulation, prior distribution, etc. **really matters**
  - And, they are difficult to specify by hand

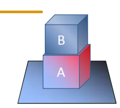
New model: *abductive reasoning from random examples [Juba '16]*

- Task: for a *conclusion c*, find a *h* such that
  1. *Plausibility:* $\Pr_D[h(X)] \geq \mu$ (for some given μ)
  2. *h almost entails c:* $\Pr_D[c(X)\,|\,h(X)] \geq 1-\varepsilon$

- *Note:* D now captures both the entailment relation and the measure of *"plausibility"*

- *Distinction from earlier precondition search*: no "context" assignment x*

Example: identifying a subgoal



- *Consider*: blocks world. For *t*=1,2,…,T
  - Propositional state vars. ("fluents") $ON_t(A,B)$, $ON_t(A,TABLE)$, $ON_t(C,A)$, etc.
  - Actions also encoded by propositional vars. $PUT_t(B,A)$, $PUT_t(C,TABLE)$, etc.
- *Given many examples of interaction…*
- *Our goal c*: $ON_T(A,TABLE) \wedge ON_T(B,A) \wedge ON_T(C,B)$
- A perhaps plausibly good "subgoal" *h*: $[ON_{T-1}(B,A) \wedge PUT_T(C,B)] \vee [PUT_{T-1}(B,A) \wedge PUT_T(C,B)]$

Formally: *abductive reasoning from random examples* for a class $\mathcal{H}$

- Fix a class of Boolean representations $\mathcal{H}$
- **Given** Boolean formula c; ε, δ, $\mu \in (0,1)$; *independent* examples $x^{(1)},\ldots,x^{(m)} \in D$,
- Suppose that there exists a $h^* \in \mathcal{H}$ such that
  1. *Plausibility:* $\Pr_D[h^*(X)] \geq \mu$
  2. *h^* entails c:* $\Pr_D[c(X) | h^*(X)] = 1$
- **Find** h *(ideally in $\mathcal{H}$)* such that with prob. $1-\delta$,
  1. *Plausibility:* $\Pr_D[h(X)] \geq \mu'$ for some $\mu'(\mu,n,\varepsilon,\delta)$
  2. *h almost entails c:* $\Pr_D[c(X) | h(X)] \geq 1-\varepsilon$

Abducing k-DNFs is also **possible** *(complete information)*

Theorem. If there is a k-DNF h^* such that

1. *Plausibility:* $\Pr_D[h^*(X)] \geq \mu$
2. *h^* entails c:* $\Pr_D[c(X) | h^*(X)] = 1$

then using $m = O(\frac{1}{\mu\varepsilon}(n^k + \log\frac{1}{\delta}))$ examples, in time $O(mn^k)$ we can find a k-DNF h such that with probability $1-\delta$,

1. *Plausibility:* $\Pr_D[h(X)] \geq \mu$
2. *h almost entails c:* $\Pr_D[c(X) | h(X)] \geq 1-\varepsilon$

Elimination Algorithm also solves k-DNF abduction

- Start with h as an OR over <u>all</u> terms of size k
- For each example $x^{(1)},\ldots,x^{(m)}$
  - If $c(x^{(i)}) = 0$, delete all terms T from h such that $T(x^{(i)}) = 1$
- Return h

*JUST **OMIT** THE FINAL TEST FOR **CONSISTENCY WITH** x^*. THE ANALYSIS IS ALMOST **IDENTICAL**.*

Extension: *Abductive Reasoning* **tolerating ε>0**

- **Given only** that some h^* achieves
  1. *Plausibility:* $\Pr_D[h^*(X)] \geq \mu$
  2. *h^* almost entails c:* $\Pr_D[c(X) | h^*(X)] \geq 1-\varepsilon$

Find an h such that for some other μ' & ε',

1. *Plausibility:* $\Pr_D[h(x)] \geq \mu'$
2. *h almost entails c:* $\Pr_D[c(X) | h(X)] \geq 1-\varepsilon'$

- *Improved algorithm for k-DNF achieves* $\mu'=(1-\gamma)\mu$, $\varepsilon'=\tilde{O}(n^{k/2}\varepsilon)$ [Zhang-Mathew-Juba'17]
  - *Cf. only obtained* $\varepsilon'=O(n^k\varepsilon)$ *for preconditions…*

*But, what about abducing **conjunctions**?? [Juba '16]*

Theorem. Suppose that a polynomial-time algorithm exists for learning abduction from random examples for conjunctions with $\mu'=C((1-\gamma)\mu/n)^d$ for some C, d. Then there is a polynomial-time algorithm for PAC-learning DNF.

So what?
1. **Central open problem in computational learning theory raised in original paper by Valiant (1984)**
2. **Recent work by Daniely and Shalev-Shwartz (2016) shows that algorithms for PAC-learning DNF would have other surprising consequences.**

***In summary* – an algorithm for our problem would constitute an unlikely breakthrough.**

Extension to abduction with **partial examples**

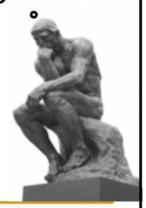
- Subtle: *do we condition on…*
  - h(X) true?
  - h(X) provable (under ρ)?
  - $h|_\rho=1$ (h "witnessed" on ρ)?

  (↑ Generally more desirable…)
- Come see our poster *"Learning Abduction Under Partial Observability"* (Juba, Li, Miller)
  - Short version: condition on *some term T of h provable under ρ* (i.e. $T|_\rho$ provable from $KB|_\rho$)
  - Can use Elimination; incorportates implicit KB

Extent of commonsense reasoning with PAC-semantics

*COULD THIS CAPTURE **ALL** COMMONSENSE REASONING?*

- *Unlikely*!
  - So far, generally doesn't capture "law of inertia," naïve theories, …
- We still seem to require the use of a non-monotonic logic as a foundation
  - At minimum, for "law of inertia"
- **Question**: can simulations provide examples for naïve theories?



Naïve use of preconditions is *problematic*

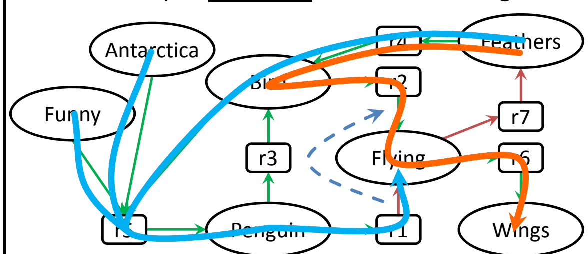
- Our reasoning with preconditions is too credulous
- **Example**: consider the query fly for x* with penguin(x*)=1… then h^* = bird
  - Is reasonably common (Pr_D[bird] moderate)
  - Is consistent with x* (bird(x*)=1)
  - Supports the query (Pr[fly|bird] high)
- *So,* we will return a precondition such as h = bird supporting fly

D. Reasoning Non-Monotonically

Learning with Non-Monotonic Logics

Activity: Find the Arguments

This animal has Feathers, lives in Antarctica, and looks Funny. Question: Does it have Wings?



Why Abandon Equivalences?

- r_1: *penguin* ⇒ ¬*flying*, r_2: *bird* ⇒ *flying*, $r_1 > r_2$
  Formula: (***u*** ∨ *bird*) ∧ ¬*penguin* ⇔ *flying*
  - Good on "full" scenes. Still abstains on {¬*p*, ¬*b*}.
  - Infers too little… Does not infer *f* on {*b*}. ***Bad!***
- r_1: *b* ⇒ *a*, r_2: ¬*b* ⇒ *a*   Formula: *true* ⇔ *a*
  - Infers too much… *a* by case analysis on { }. ***Bad!***
- NP-hard reasoning. Still not 1 rule/atom. ***Bad!***
- ***Thus:*** logic-based arguments with preferences.

How to Learn Arguments?

- Rules with different heads in each argument, therefore, one has to deal with ***partial observability***, which then requires ***SLAP***.
  - Sufficient to get consistency in predictions.
- Some ***relational expressivity*** comes for "free".
  - Each rule has an IQE body that is efficiently testable for satisfaction on observations.
- Following ***linear thresholds*** and ***decision lists***.
  - Online, efficient, dealing with priorities, etc.

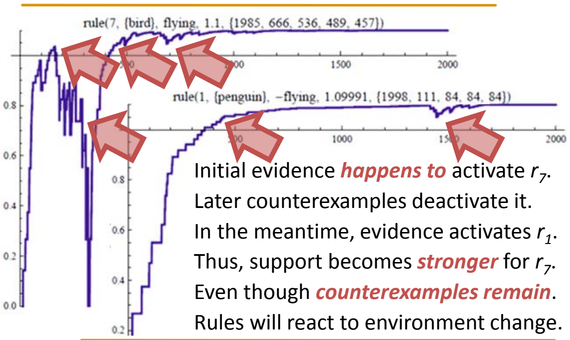
Never-Ending Rule Discovery [Michael '16]

1. Get observation *obs*, and ***reason with active rules*** to get an enhanced observation *obs**.
2. Find a literal *x* that is observed in *obs* but ***not inferred*** by active rules triggered in *obs**. Add *body* $\Rightarrow$ *x*, for random *body* satisfied by *obs**.
3. Increase / decrease weight of rules triggered in *obs** that ***concur with*** / ***oppose*** *obs*.
4. Newly active rules are ***weaker than*** existing.
5. Newly inactive rules, have ***no preferences***.

What Does it End Up Learning?

- Hide some attributes from states drawn from:
  {¬*bird*, ¬*penguin*, ¬*plane*, ¬*flying*} w.p. *5/16*
  {¬*bird*, ¬*penguin*, ¬*plane*, *flying*} w.p. *5/16*
  {¬*bird*, ¬*penguin*, *plane*, *flying*} w.p. *2/16*
  { *bird*, *penguin*, ¬*plane*, ¬*flying*} w.p. *1/16*
  { *bird*, ¬*penguin*, ¬*plane*, *flying*} w.p. *3/16*
- "Intended" learned rules with head (¬)*flying*:
  penguin $\Rightarrow$ ¬*flying*, *bird* $\Rightarrow$ *flying*, *plane* $\Rightarrow$ *flying*
- But also "picks up": *plane* $\Rightarrow$ ¬*penguin* (mutually exclusive), ¬*bird* $\Rightarrow$ ¬*penguin* (contrapositive), *penguin* $\Rightarrow$ *bird*, ¬*flying* $\Rightarrow$ ¬*bird* (explaining away).

Demo of NERD's Online Behavior



Initial evidence ***happens to*** activate r_7.
Later counterexamples deactivate it.
In the meantime, evidence activates r_1.
Thus, support becomes ***stronger*** for r_7.
Even though ***counterexamples remain***.
Rules will react to environment change.

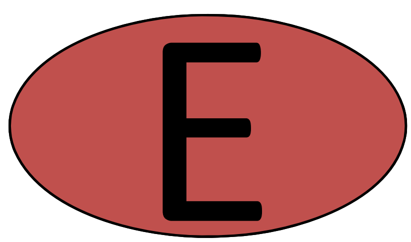
Probably Approximately Correct?

- Equivalences are PAC learnable, but suffer from Goldilocks effect: infer too much / little.
- Are logic-based arguments PAC learnable?

\+ Learn with unknown atoms. Learn priorities.
− Nested if-then-else's unlearnable from scenes.
\+ Non-adversarial environments. Equiv ≠ Args.
− Learning requires reasoning, restricts depth.
\+ Psychological evidence on bounded reasoning.
\+ **On edge of learnable. Ev****re.**

Machine Coaching [Michael '17]

- Integrate user-provided rules ***while*** learning.
  - Like online learning, but instead of providing the correct label, "question" part of the argument that leads to the wrong prediction.
  - Possible to give PAC-semantics to the process!
- <u>**Theorem:**</u> Arguments ***(ASPIC+ type: grounded semantics, axiomatic premises, defeasible rules, rebutting attacks, last link preferences)*** are PAC learnable via ***machine coaching*** alone.



High-Level Tutorial Roadmap

A. Why KRR Should Embrace Learning

B. Introduction to PAC-Semantics

C. Integrating Deduction and Induction

D. Reasoning Non-Monotonically

E. **Overall Summary and Conclusions**

E. Overall Summary and Conclusions

Recap and Open Problems

Semantics for Learned Knowledge

- PAC-semantics is a suitable choice:
  - models the world as a probability distribution
  - treats knowledge as high-probability properties
- Offers simple solutions to classic KR problems:
  - non-monotonicity from conditioning of rules
  - exogenous qualification from validity defect
  - elaboration tolerance through implicit learning
  - ramifications incorporated in "implicit KB"
  - natural formulation of abductive inference
  - explicit solutions through argument learning

KRR and Learning Integration

- Traditional view of reasoning and learning as independent processes must be abandoned.
- When combined under PAC-semantics they:
  - soundly achieve greater completeness
  - circumvent computational barriers in learning
  - enable fast and compact access to "implicit KB"
  - may sometimes reduce reasoning complexity
- Also, seamlessly supports user intervention during learning through machine coaching.

Open Problems with KRR Flavor

- What kind of first-order expressions can we learn from partial examples in reasoning?
  - IQEs via propositionalization. Classical barriers (e.g., Haussler) apply to integrated problem?
- What kind of first-order integrated learning and reasoning is possible?
  - May want first-order expressions to refer to a limited domain. Seems closely related to selection of "preconditions" (see point in next slide) but for limiting the domain of quantifiers.

Open Problems with KRR Flavor

- When, and how broadly, is the complexity of reasoning reduced in the integrated problem?
  - More natural condition than in [Juba '15]?
  - More broadly: which fragments are tractable?
- Incorporating naive theories into the model?
  - E.g., treat naive simulations as populating data set for implicit KB — how well does this work?
- Preconditions for commonsense reasoning?
  - Perhaps select an "unchallenged" precondition; suggests argument semantics for precondition.

E. Overall Summary and Conclusions

Bibliography of Related Work

References

- Adams. (1975), 'The Logic of Conditionals: An Application of Probability to Deductive Logic', Springer.
- Blumer, Ehrenfeucht, Haussler, Warmuth. (1987), 'Occam's Razor', Information Processing Letters, 24:377–380.
- Daniely, Shalev-Shwartz. (2016), 'Complexity Theoretic Limtations on Learning DNF's', Proc. COLT, volume 49 of JMLR W&CP.
- Diakidoy, Kakas, Michael, Miller. (2014), 'Story Comprehension through Argumentation', Proc. COMMA, p. 31–42.
- Diakidoy, Kakas, Michael, Miller. (2015), 'STAR: A System of Argumentation for Story Comprehension and Beyond', Proc. Commonsense, p. 64–70.
- Dimopoulos, Michael, Athienitou. (2009), 'Ceteris Paribus Preference Elicitation with Predictive Guarantees', Proc. IJCAI, p. 1890–1895.
- Dung. (1995), 'On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games', Artificial Intelligence 77:321–357.

References

- Haussler. (1989), 'Learning Conjunctive Concepts in Structural Domains', Machine Learning 4(1):7–40.
- Juba. (2013), 'Implicit Learning of Common Sense for Reasoning', Proc. IJCAI, p. 939–946.
- Juba. (2015), 'Restricted Distribution Automatizability in PAC-Semantics', Proc. ITCS, p. 93–102.
- Juba. (2016), 'Learning Abductive Reasoning Using Random Examples', Proc. AAAI, p. 999–1007.
- Kakas, Michael. (2016), 'Cognitive Systems: Argument and Cognition', IEEE Intelligent Informatics Bulletin 17(1):14–20.
- Kakas, Michael, Miller. (2008), 'Fred meets Tweety', Proc. ECAI, p. 747–748.
- Khardon, Roth. (1997), 'Learning to Reason', J. ACM, 44(5):697–725.
- Michael. (2008), 'Autodidactic Learning and Reasoning', PhD thesis, Harvard.
- Michael. (2009), 'Reading Between the Lines', Proc. IJCAI, p. 1525–1530.
- Michael. (2010), 'Partial Observability and Learnability', Artificial Intelligence 174(11):639–669.

References

- Michael. (2011), 'Causal Learnability', Proc. IJCAI, p. 1014–1020.
- Michael. (2013), 'Machines with WebSense', Proc. Commonsense.
- Michael. (2014), 'Simultaneous Learning and Prediction', Proc. KR, p. 348–357.
- Michael. (2015a), 'The Disembodied Predictor Stance', Pattern Recognition Letters 64(C):21–29.
- Michael. (2015b), 'Introspective Forecasting', Proc. IJCAI, p. 3714–3720.
- Michael. (2016), 'Cognitive Reasoning and Learning Mechanisms', Proc. AIC, p. 2–23.
- Michael. (2017), 'The Advice Taker 2.0', Proc. Commonsense.
- Michael, Papageorgiou. (2013), 'An Empirical Investigation of Ceteris Paribus Learnability', Proc. IJCAI, p. 1537–1543.
- Michael, Valiant. (2008), 'A First Experimental Demonstration of Massive Knowledge Infusion', Proc. KR, p. 378–388.
- Modgil, Prakken. (2014), 'The ASPIC+ Framework for Structured Argumentation: A Tutorial', Argument & Computation 5(1):31–62.

References

- Nilsson. (1986), 'Probabilistic Logic', Artificial intelligence, 28(1):71–87.
- Roth. (1995), 'Learning to Reason: The Non-monotonic Case', Proc. IJCAI, vol. 2, p. 1178–1184.
- Skouteli, Michael. (2016), 'Empirical Investigation of Learning-Based Imputation Policies', Proc. GCAI, p. 161–173.
- Toni. (2014), 'A Tutorial on Assumption-based Argumentation', Argument & Computation 5(1):89–117.
- Valiant. (1984), 'A Theory of the Learnable', Communications of the ACM, 18(11):1134– 1142.
- Valiant. (1994), 'Circuits of the Mind', Oxford University Press.
- Valiant. (1995), 'Rationality', Proc. COLT, p. 3–14.
- Valiant. (2000), 'Robust Logics', Artificial Intelligence, 117:231–253.
- Valiant. (2000b), 'A Neuroidal Architecture for Cognitive Computation', J. ACM, 47(5):854–882.
- Zhang, Mathew, Juba. (2017), 'An Improved Algorithm for Learning to Perform Exception-Tolerant Abduction', Proc. AAAI, p. 1257–1265.

end