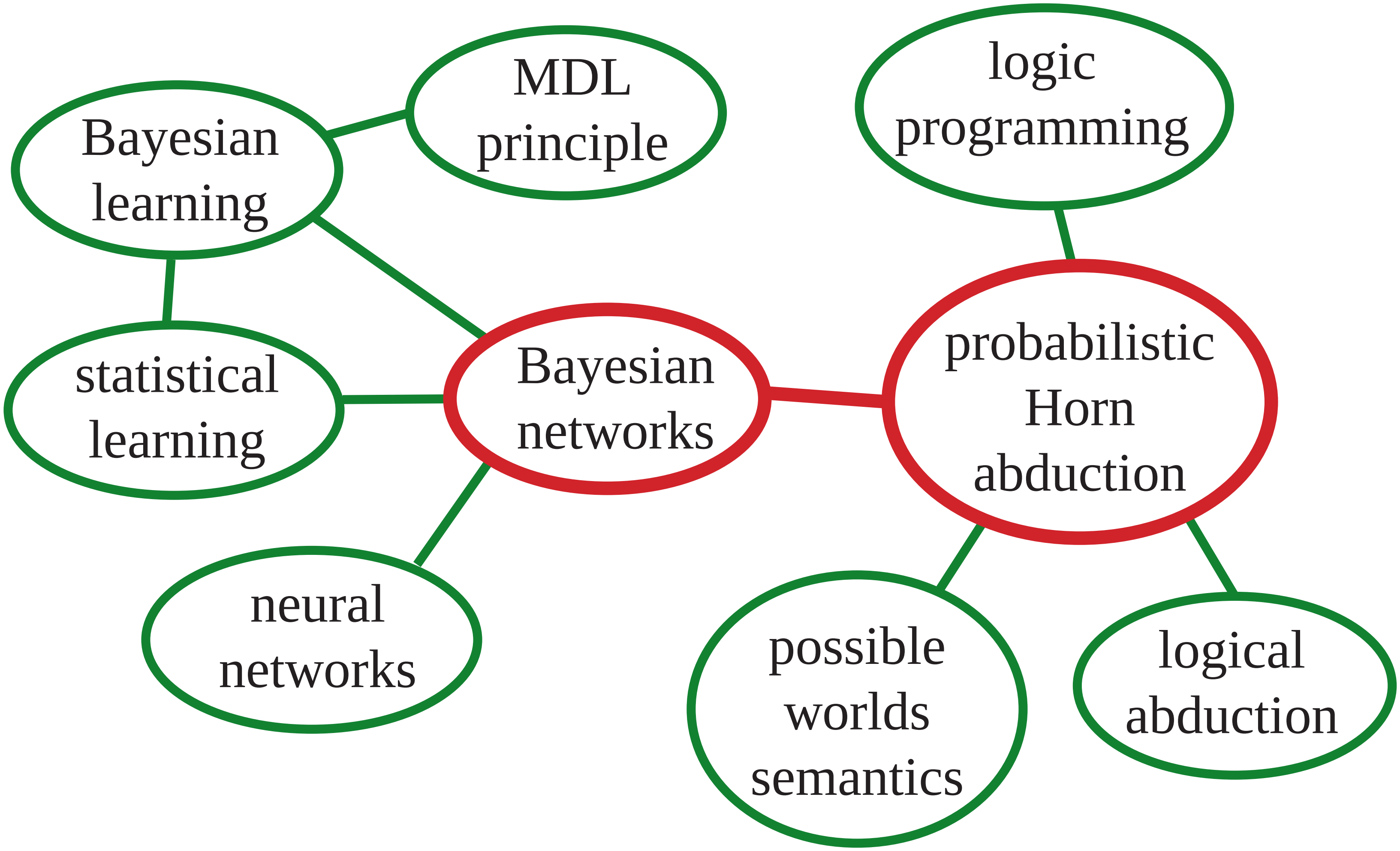
Learning, Bayesian Probability, Graphical Models, and Abduction

David Poole

University of British Columbia

Induction and abduction



Overview

- Causal and evidential modelling & reasoning
- Bayesian networks, Bayesian conditioning & abduction
- Noise, overfitting, and Bayesian learning

Causal & Evidential Modelling

Causal modelling:

$$\begin{array}{ccc} \textit{causes} & \longrightarrow & \textit{effects} \\ \uparrow & & \uparrow \\ \text{of interest} & & \text{observed} \end{array}$$

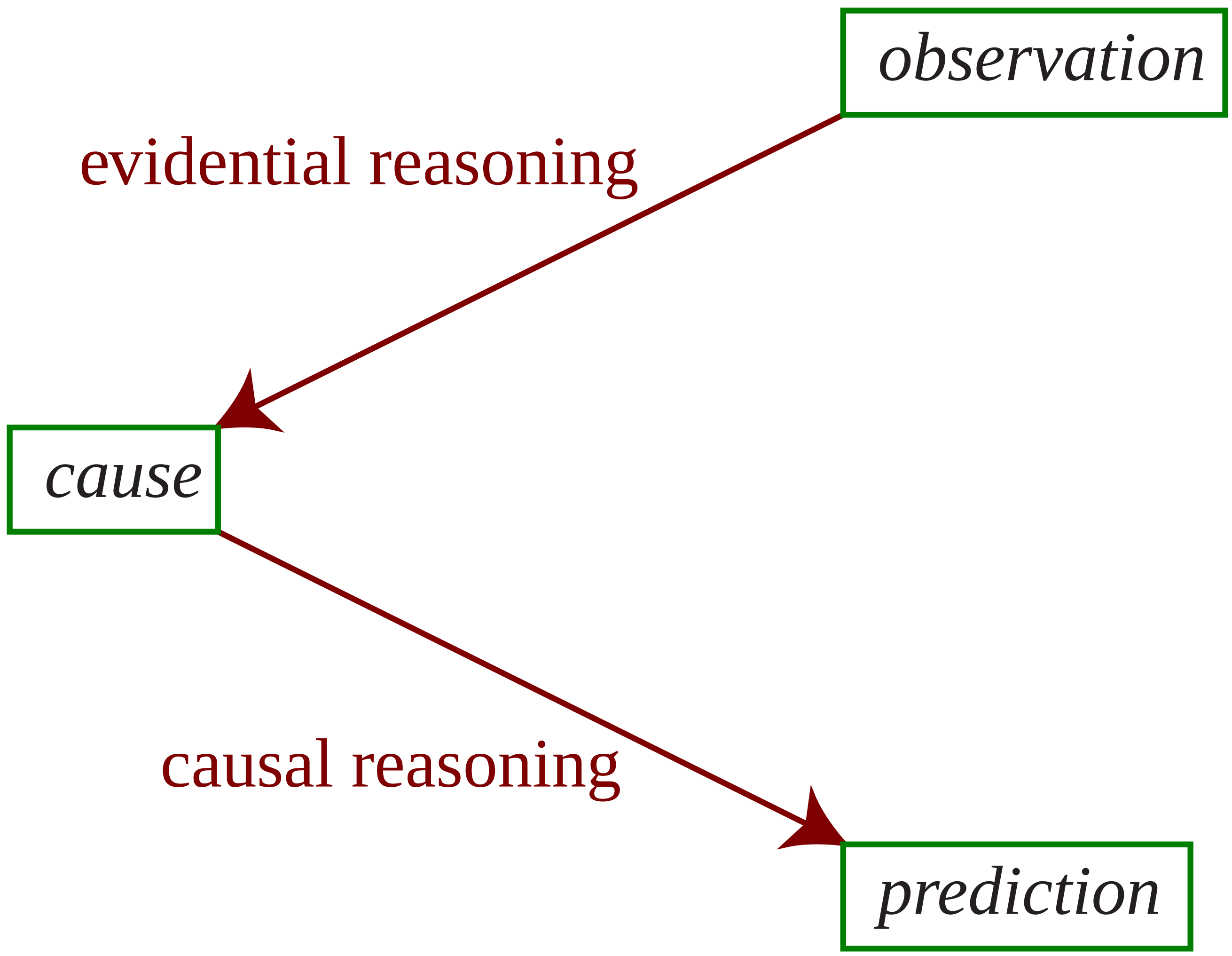
vision:	*scene* $\longrightarrow$ *image*
diagnosis:	*disease* $\longrightarrow$ *symptoms*
learning:	*model* $\longrightarrow$ *data*

Evidential modelling:

$$\textit{effects} \longrightarrow \textit{causes}$$

vision:	$\textit{image} \longrightarrow \textit{scene}$
diagnosis:	$\textit{symptoms} \longrightarrow \textit{diseases}$
learning:	$\textit{data} \longrightarrow \textit{model}$

Causal & Evidential Reasoning



Reasoning & Modelling Strategies

How do we do causal and evidential reasoning, given modelling strategies?

- Evidential modelling & only evidential reasoning (Mycin, Neural Networks).
- Model evidentially + causally (problem: consistency, redundancy, knowledge acquisition)
- Model causally; use different reasoning strategies for causal & evidential reasoning. (deduction + abduction or Bayes' theorem)

Bayes' Rule

[de Moivre 1718, Bayes 1763, Laplace 1774]

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

Proof:

$$\begin{aligned} P(h \wedge e) &= P(e|h)P(h) \\ &= P(h|e)P(e) \end{aligned}$$

Lesson #1

You should know the difference between

- evidential & causal modelling
- evidential & causal reasoning

There seems to be a relationship between Bayes' theorem and abduction — used for the same task.

Overview

- Causal and evidential modelling & reasoning
- Bayesian networks, Bayesian conditioning & abduction
- Noise, overfitting, and Bayesian learning

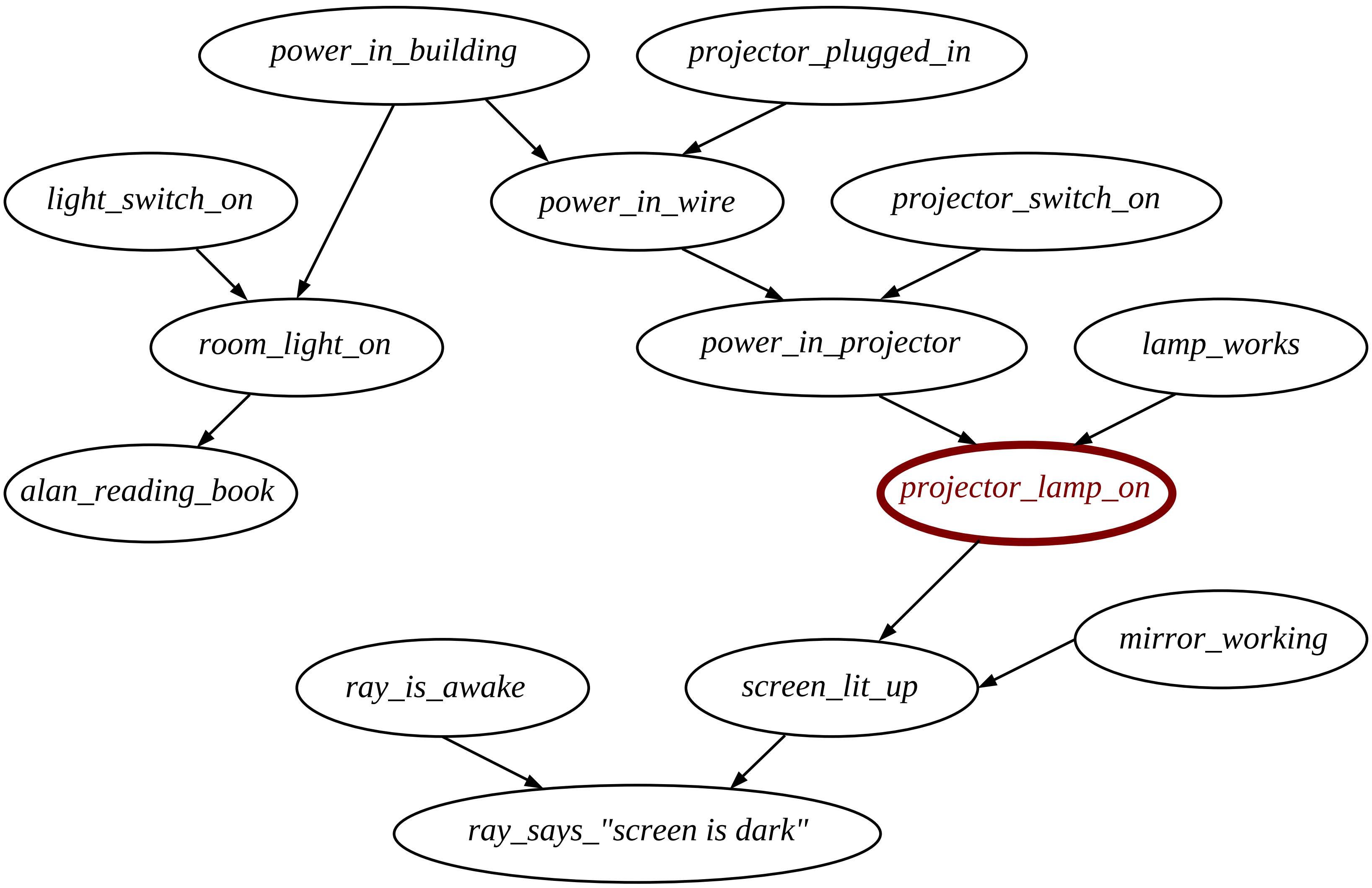
Bayesian Networks

- Graphical representation of independence.
- DAGs with nodes representing random variables.
- Embed independence assumption:
  If $b_1, \cdots, b_n$ are the parents of a then

$$P(a|b_1, \cdots, b_n, V) = P(a|b_1, \cdots, b_n)$$

  if V is not a descendant of a.

Bayesian Network for Overhead Projector



Bayesian networks as logic programs

$projector\_lamp\_on \leftarrow$

$\quad power\_in\_projector \wedge$

$\quad lamp\_works \wedge$

$\quad projector\_working\_ok. \longleftarrow$ possible hypothesis with associated probability

$projector\_lamp\_on \leftarrow$

$\quad power\_in\_projector \wedge$

$\quad \sim lamp\_works \wedge$

$\quad working\_with\_faulty\_lamp.$

Probabilities of hypotheses

$$
\begin{aligned}
&P(projector\_working\_ok) \\
&\quad = \; P(projector\_lamp\_on \mid \\
&\qquad\qquad power\_in\_projector \land lamp\_works)
\end{aligned}
$$

— provided as part of Bayesian network

$$
\begin{aligned}
&P(\sim projector\_working\_ok) \\
&\quad = \; 1 - P(projector\_working\_ok)
\end{aligned}
$$

What do these logic programs mean?

- Possible world for each assignment of truth value to a possible hypothesis:
  {*projector_working_ok*, *working_with_faulty_lamp*}
  {*projector_working_ok*, ∼*working_with_faulty_lamp*}
  {∼*projector_working_ok*, *working_with_faulty_lamp*}
  {∼*projector_working_ok*, ∼*working_with_faulty_lamp*}
- Probability of a possible world is the product of the probabilities of the associated hypotheses.
- Logic program specifies what else is true in each possible world.

Probabilistic logic programs & abduction

Semantics is abductive in nature

— set of explanations of a proposition characterizes the possible worlds in which it is true.

(assume possible hypotheses and their negations).

$$P(g) = \sum_{E \text{ is an explanation of } g} P(E)$$

$$P(E) = \prod_{h \in E} P(h)$$

↑ given with logic program

Conditional Probabilities

$$P(g|e) = \frac{P(g \wedge e)}{P(e)} \quad \begin{array}{l} \longleftarrow \text{explain } g \wedge e \\ \longleftarrow \text{explain } e \end{array}$$

Given evidence e, explain e, then try to explain g from these explanations.

Lessons #2

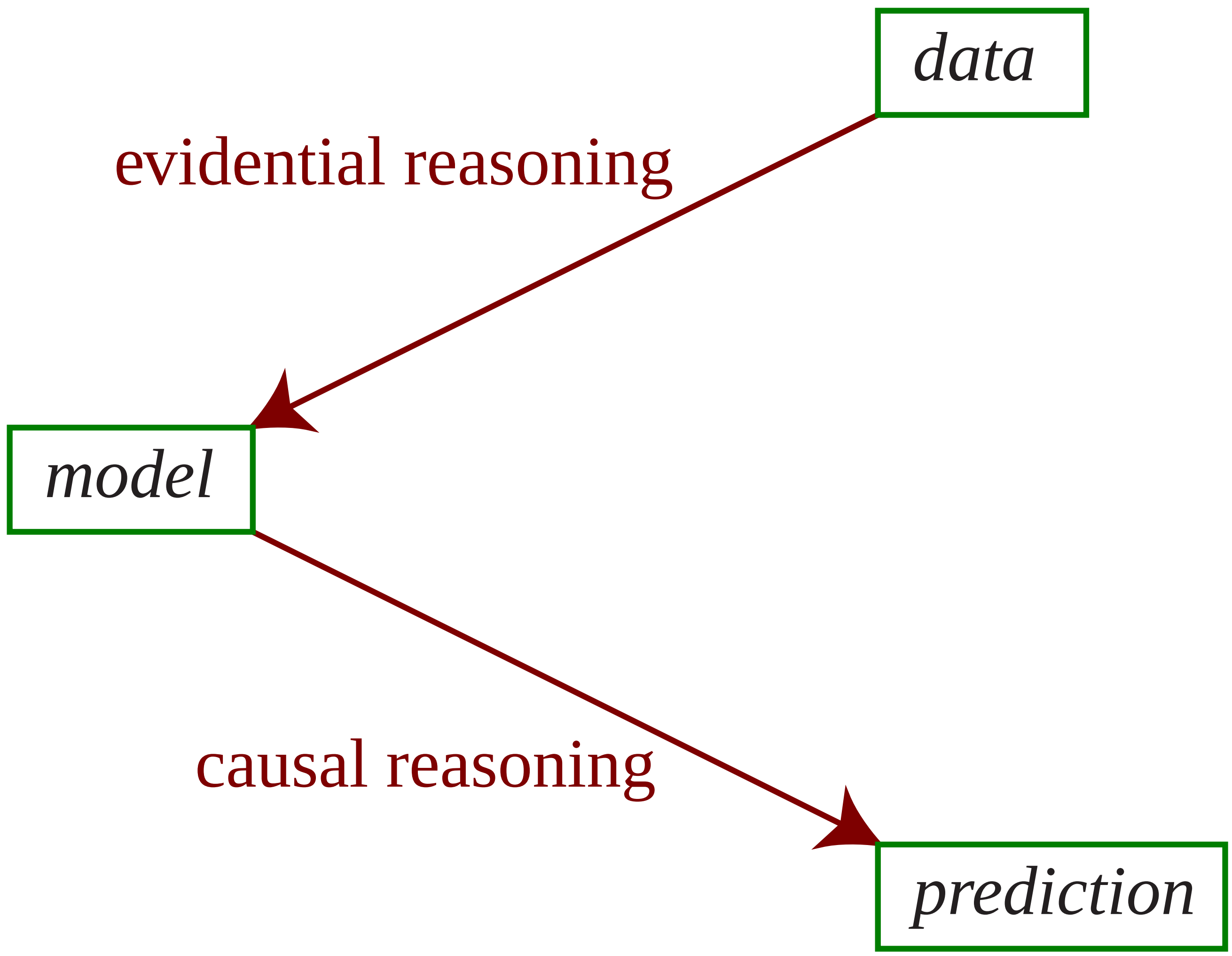
- Bayesian conditioning = abduction
- The evidence of Bayesian conditioning is what is to be explained.
  - Condition on all information obtained since the knowledge base was built.

$$P(h|e \wedge k) \longrightarrow P_k(h|e)$$

Overview

- Causal and evidential modelling & reasoning
- Bayesian networks, Bayesian conditioning & abduction
- Noise, overfitting, and Bayesian learning

Induction


data
evidential reasoning
model
causal reasoning
prediction

Potential Confusion

Often the data is about some evidential reasoning task. e.g., classification, diagnosis, recognition …

Example: We can do Bayesian learning where the hypotheses are decision trees, neural networks, parametrized distribution, or Bayesian networks.

Example: We can do hill climbing learning with cross validation where the hypotheses are decision trees, neural networks, parametrized distribution, or Bayesian networks.

Noise and Overfitting

Most data contains noise (errors, inadequate attributes, spurious correlations)

$\Longrightarrow$ overfitting — the model learned fits random correlations in the data

Example: A more detailed decision tree *always* fits the data better, but usually smaller decision trees provides better predictive value.

Need tradeoff between

model simplicity + fit to data.

Overfitting and Bayes' theorem

fit to data

bias

$$P(h|e) = \frac{P(e|h)\,P(h)}{P(e)}$$

normalizing constant

Minimum description length principle

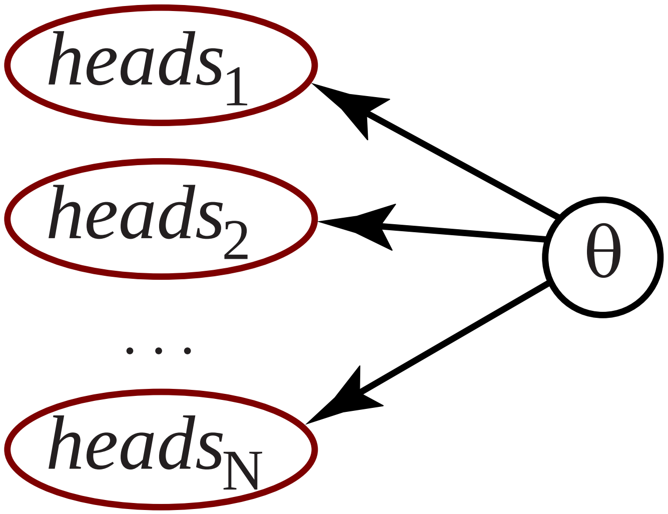
Choose best hypothesis given the evidence:

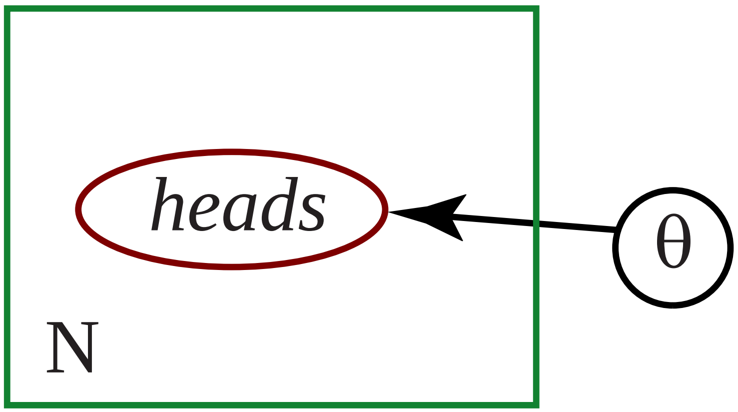
$$
\begin{aligned}
& \arg\max_h P(h|e) \\
= \quad & \arg\max_h \frac{P(e|h)P(h)}{P(e)} \\
= \quad & \arg\max_h P(e|h)P(h) \\
= \quad & \arg\max_h \underbrace{-\log_2 P(e|h)}_{\text{the number of bits to describe the data in terms of the model}} + \underbrace{-\log_2 P(h)}_{\text{the number of bits to describe the model}}
\end{aligned}
$$

Graphical Models for Learning

Idea: model $\longrightarrow$ data

Example: parameter estimation for probability of heads (from [Buntine, JAIR, 94])





Abductive Version of Parameter Learning

$heads(C) \leftarrow$

$\quad turns\_heads(C, \Theta) \wedge prob\_heads(\Theta).$

$tails(C) \leftarrow$

$\quad turns\_tails(C, \Theta) \wedge prob\_heads(\Theta).$

$\forall C \forall \Theta \; \{turns\_heads(C, \Theta), turns\_tails(C, \Theta)\} \in \mathbf{C}$

$\{prob\_heads(\theta) : \theta \in [0, 1]\} \in \mathbf{C}$

$Prob(turns\_heads(C, \theta)) = \theta$

$Prob(turns\_tails(C, \theta)) = 1 - \theta$

$Prob(prob\_heads(\theta)) = 1$ $\longleftarrow$ uniform on [0, 1].

Explaining Data

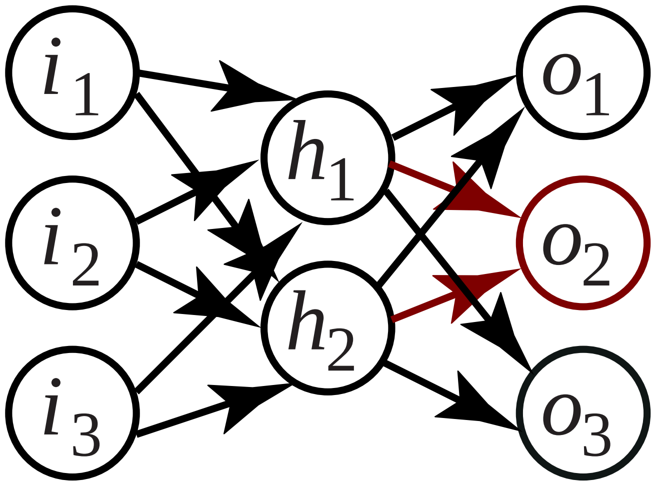
If you observe:

$$heads(c_1), tails(c_2), tails(c_3), heads(c_4), heads(c_5), \ldots$$

For each $\theta \in [0, 1]$ there is an explanation:

$\{prob\_heads(\theta), turns\_heads(c_1, \theta), turns\_tails(c_2, \theta),$
$turns\_tails(c_3, \theta), turns\_heads(c_4, \theta), turns\_heads(c_5, \theta),$
$\ldots\}$

Abductive neural-network learning



$$prop(X, o_2, V) \leftarrow$$

$$prop(X, h_1, V_1) \wedge prop(X, h_2, V_2) \wedge$$

$$param(p_8, P_1) \wedge param(p_{10}, P_2) \wedge \qquad \longleftarrow \text{abducible}$$

$$V = \frac{1}{1 + e^{(V_1 P_1 + V_2 P_2)}}. \qquad \longleftarrow \text{sigmoid additive}$$

Lesson #3

Abductive view of Bayesian learning:

- rules imply data from parameters or possible representations
- parameter values or possible representations abducible
- rules contain logical variables for data items

Evidential versus causal modelling

	Neural Nets	Bayesian Nets
modelling	evidential	causal
	sigmoid additive	linear gaussian
	— related by Bayes theorem	
evidential reasoning	direct	abduction
causal reasoning	none	direct
context changes	fragile	robust
learning	easier	more difficult

Conclusion

- Bayesian reasoning is abduction.
- Bayesian network is a causal modelling language — abduction + probability.
- What logic-based abduction can learn from Bayesians:
  Handle noise, overfitting
  Conditioning: explain everything
  Algorithms: exploit sparse structure, exploit extreme distributions, or stochastic simulation.
- What the Bayesians can learn:
  Richer representations.