Modals, conditionals, and probabilistic generative models

Topic 1: intro to probability & generative models; a bit on modality

Dan Lassiter, Stanford Linguistics
Université de Paris VII, 25/11/19

4 lectures: The plan

1. probability, generative models, a bit on epistemic modals
2. indicative conditionals
3. causal models & counterfactuals
4. reasoning about impossibilia

Mondays except #3 –
it'll be Wednesday 11/11,
no meeting Monday 11/9!

Today: Probabilistic generative models

- widespread formalism for cognitive models
- allow us to
  - integrate model-theoretic semantics with probabilistic reasoning — today
  - make empirical, theoretical advances in conditional semantics & reasoning — 2
  - make MTS procedural, with important consequences for counterfactuals & representing impossibilia — 3,4

How we’ll get there …

- probability
  - aside on epistemic modals
- exact and approximate inference
- kinds of generative models
  - (causal) Bayes nets
  - structural equation models
  - probabilistic programs

Probability theory

What is probability?

La théorie de probabilités n'est au fond, que le bon sens réduit au calcul: elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte. -Laplace (1814)

Probability is not really about numbers; it is about the structure of reasoning. -Shafer (1988)

What is probability?

- probability is a logic
- usually built on top of classical logic
  - an enrichment, not a competitor!
- familiar style of semantics, combining possible worlds with degrees

Interpretations of probability

- Frequentist: empirical/long-run proportion
- Propensity/intrinsic chance
- Bayesian: degree of belief

All are legitimate for certain purposes.

For cognitive modeling, Bayesian interpretation is most relevant

intensional propositional logic

Syntax	**Semantics**
For $i \in \mathbb{N},\ p_i \in \mathcal{L}$	$[\![\phi]\!] \subseteq W$
$\phi, \psi \in \mathcal{L} \Rightarrow \neg\phi \in \mathcal{L}$	$[\![\neg\phi]\!] = W - [\![\phi]\!]$
$\Rightarrow \phi \wedge \psi \in \mathcal{L}$	$[\![\phi \wedge \psi]\!] = [\![\phi]\!] \cap [\![\psi]\!]$
$\Rightarrow \phi \vee \psi \in \mathcal{L}$	$[\![\phi \vee \psi]\!] = [\![\phi]\!] \cup [\![\psi]\!]$
$\Rightarrow \phi \rightarrow \psi \in \mathcal{L}$	$[\![\phi \rightarrow \psi]\!] = [\![\neg\phi]\!] \cup [\![\psi]\!]$

Truth: ϕ is true at w iff $w \in [\![\phi]\!]$

ϕ is true (simpliciter) iff $w_@ \in [\![\phi]\!]$

Classical ('Stalnakerian') dynamics

C is a context set ($\approx$ information state).

If someone says "ϕ", choose to update or reject.

Update: $C[\phi] = C \cap [\![\phi]\!]$

$C[\phi]$ entails ψ iff $C[\phi] \subseteq [\![\psi]\!]$

from PL to probability

For sets of worlds substitute probability distributions:

P: $Prop \to [0,1]$, where

1. $Prop \subseteq \wp(W)$
2. $Prop$ is closed under union and complement
3. $\mathrm{P(W)} = 1$
3. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

Read $P(\llbracket\phi\rrbracket)$ as "the degree of belief that ϕ is true"
i.e., that $w_@ \in \llbracket\phi\rrbracket$

(Kolmogorov, 1933)

conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

One could also treat conditional probability as basic and use it to define conjunctive probability:

$$P(A \cap B) = P(A|B) \times P(B)$$

probabilistic dynamics

A core Bayesian assumption:

> For any propositions A and B, your degree of belief $P(B)$, after observing that A is true, should be equal to your conditional degree of belief $P(A|B)$ before you made this observation.

Dynamics of belief are determined by the initial model (‘prior’) and the data received.

probabilistic dynamics

This assumption holds for Stalnakerian update too.

Bayesian update is a generalization:

$$C_1 \underset{\text{observe } \phi}{\Longrightarrow} C_2 = C_1 \cap [\![\phi]\!]$$

$$P_1([\![\psi]\!]) \underset{\text{observe } \phi}{\Longrightarrow} P_2([\![\psi]\!]) = P_1([\![\psi]\!] | [\![\phi]\!])$$

1) Eliminate worlds where observation is false.
2) If using probabilities, renormalize.

random variables

a random variable is a partition on W – equiv., a Groenendijk & Stockhof '84 question meaning.

$$\begin{aligned}
\mathbf{rain}? &= [|\textit{is it raining?}|] \\
&= \{\{w|\mathbf{rain}(w)\}, \{w|\neg\mathbf{rain}(w)\}\}
\end{aligned}$$

$$\begin{aligned}
\mathbf{Dan\text{-}hunger} =& [|\textit{How hungry is Dan?}|] \\
=& \{\{w|\neg\mathbf{hungry}(w)(\mathbf{d})\}, \\
& \{w|\mathbf{sorta\text{-}hungry}(w)(\mathbf{d})\}, \\
& \{w|\mathbf{very\text{-}hungry}(w)(\mathbf{d})\}\}
\end{aligned}$$

joint probability

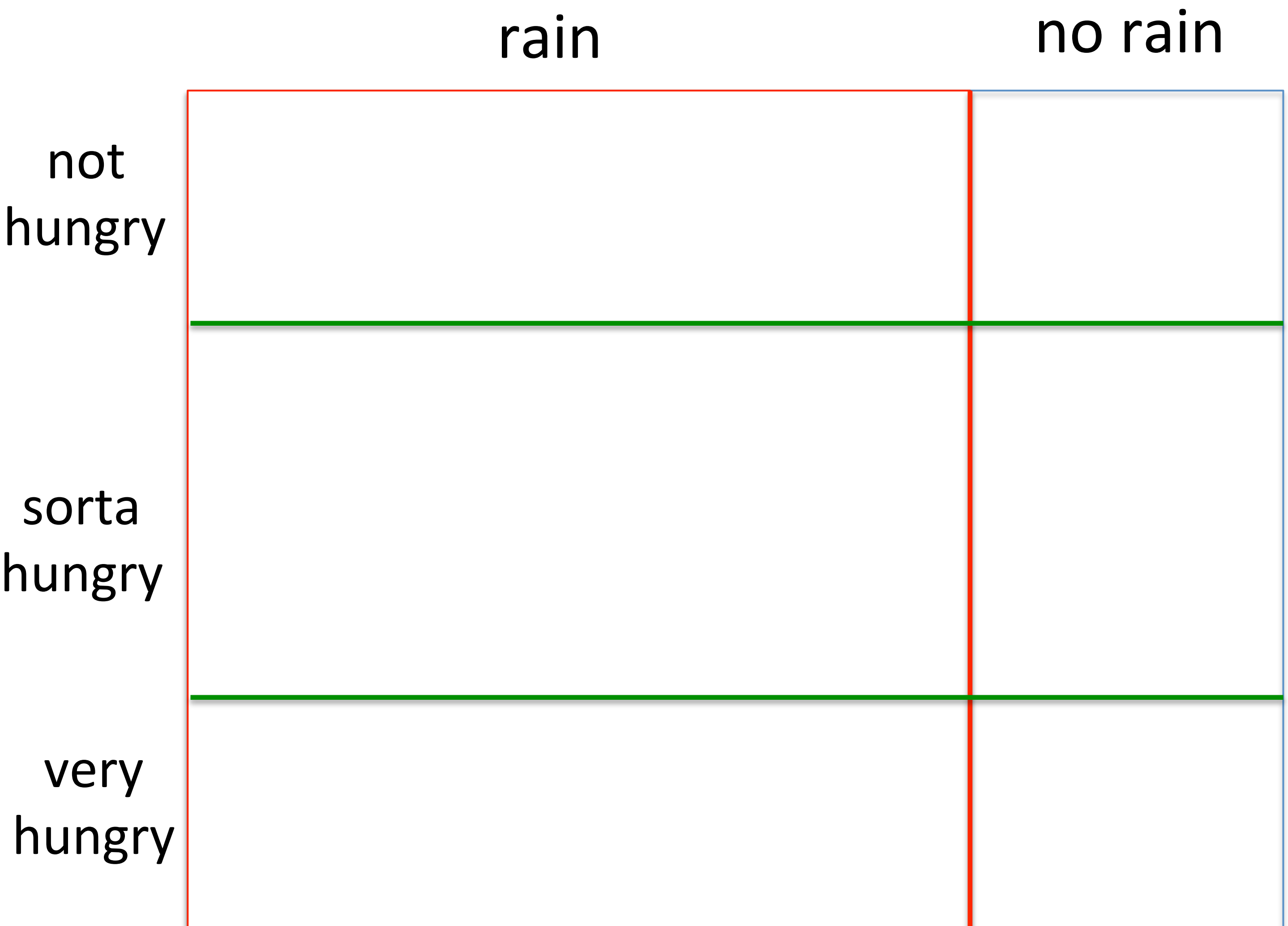
We often use capital letters for RVs, lower-case for specific answers.

P(X=x): prob. that the answer to *X* is *x*

Joint probability: a distribution over all possible combinations of a set of variables.

$$P(X = x \wedge Y = y) \text{ — usu. written — } P(X = x, Y = y)$$

2-RV structured model



A joint distribution determines a number for each cell.

Choice of RVs determines the model's 'grain': what distinctions can it see?

marginal probability

$$P(X = x) = \sum_y P(X = x \wedge Y = y)$$

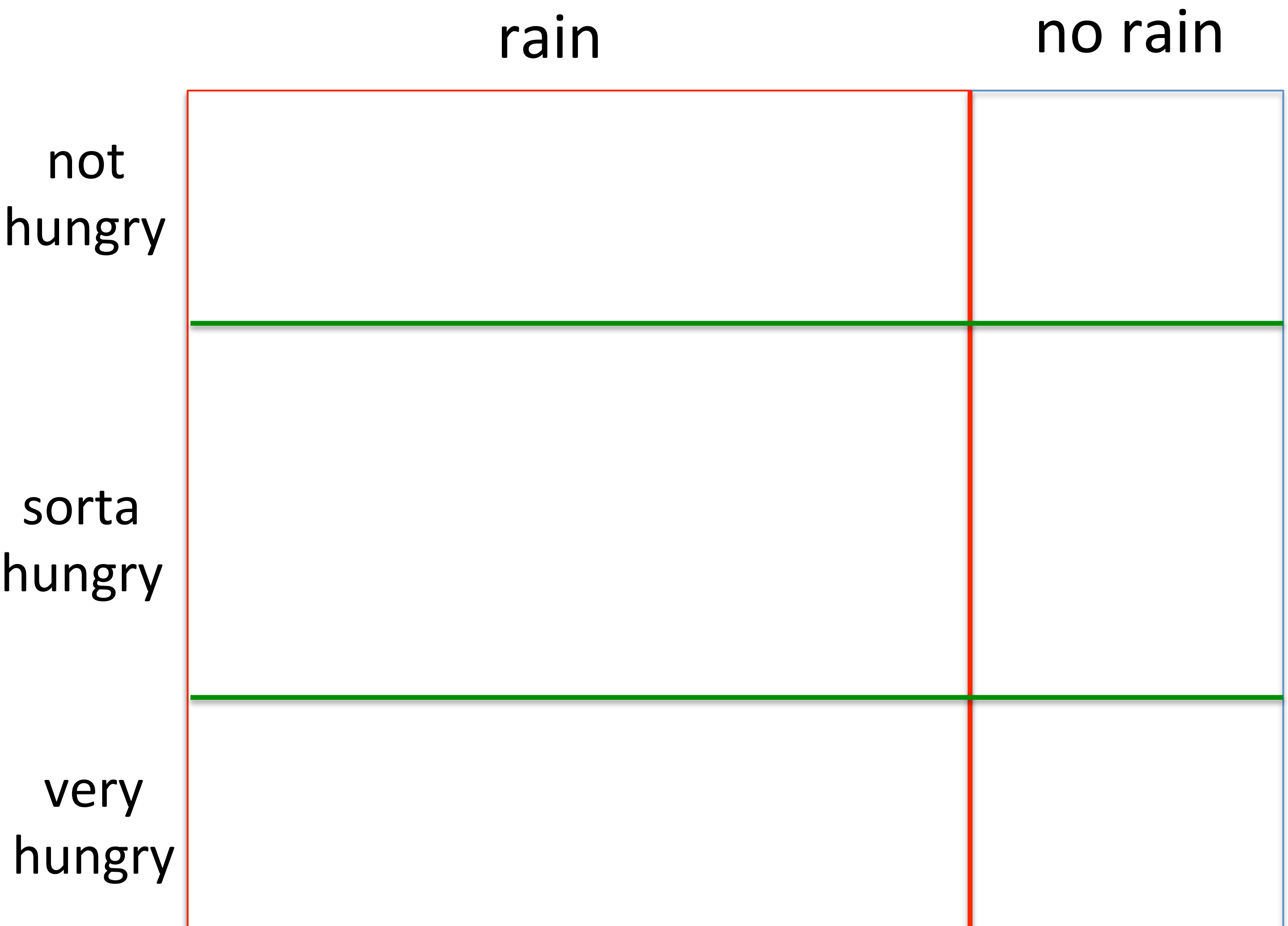
- obvious given that RVs are just partitions
- P(it's raining) is the sum of:
  - P(it's raining and Dan's not hungry)
  - P(it's raining and Dan's kinda hungry)
  - P(it's raining and Dan's very hungry)

independence

$$X \perp\!\!\!\perp Y \Leftrightarrow \forall x \forall y : P(X = x) = P(X = x | Y = y)$$

- X and Y are independent RVs iff:
  - changing P(X) does not affect P(Y)
- Pearl: independence judgments cognitively more basic than probability estimates
  - used to simplify inference in Bayes nets
  - ex.: traffic in LA vs. price of beans in China

2-RV structured model

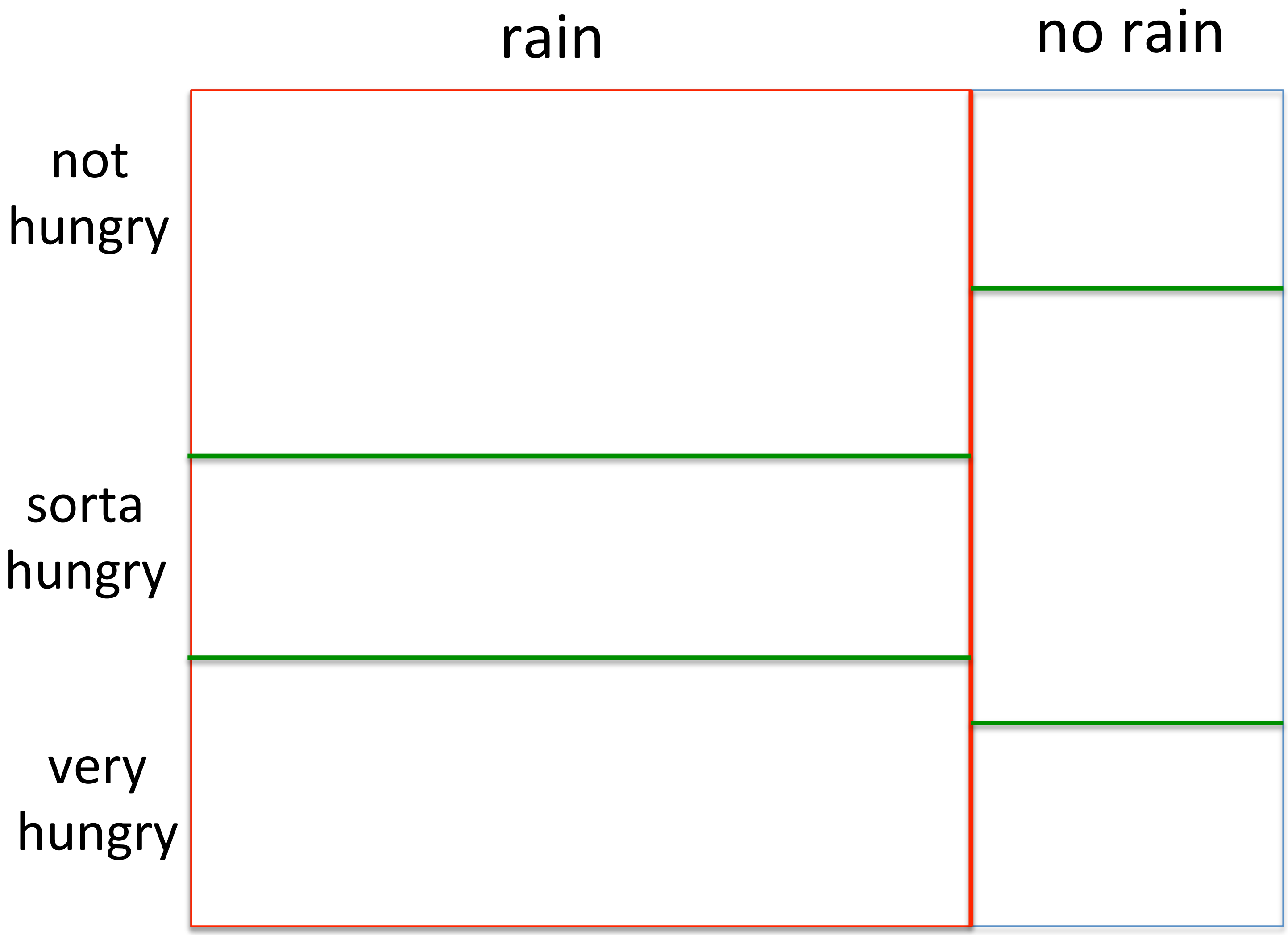


Here, let probability be proportional to area.

rain, **Dan-hunger** independent

- probably, it’s raining
- probably, Dan is sorta hungry

2-RV structured model



rain, **Dan-hunger** not indep.*:* rain reduces appetite

- If rain, Dan's probably not hungry
- If no rain, Dan's probably sorta hungry

inference

Bayes' rule:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Exercise: prove from the definition of conditional probability.

Why does this formula excite Bayesians so?

Inference as model inversion:

– Hypotheses *H*: $\{h_1, h_2, \ldots\}$

– Possible evidence *E*: $\{e_1, e_2, \ldots\}$

$$P(H = h_i | E = e) = \frac{P(E = e | H = h_i) \times P(H = h_i)}{P(E = e)}$$

Intuition: use hypotheses to generate predictions about data. Compare to observed data. Re-weight hypotheses to reward success and punish failure.

some terminology

$$\underbrace{P(H = h_i | E = e)}^{\textbf{posterior}} = \frac{\overbrace{P(E = e | H = h_i)}^{\textbf{likelihood}} \times \overbrace{P(H = h_i)}^{\textbf{prior}}}{\underbrace{P(E = e)}_{\textbf{normalizing constant}}}$$

more useful versions

P(e) typically hard to estimate on its own

– how likely were you, a priori, to observe what you did?!?

$$P(e) = \sum_j P(e, h_j)$$
$$= \sum_j P(e|h_j)P(h_j)$$

$$P(H = h_i|e) = \frac{P(e|H = h_i) \times P(H = h_i)}{\sum_j P(e|H = h_j) \times P(H = h_j)}$$

works iff H is a partition!

more useful versions

Frequently you don't need *P(e)* at all:

$$P(h_i|e) \propto P(e|h_i) \times P(h_i)$$

To compare hypotheses,

$$\frac{P(h_i|e)}{P(h_j|e)} = \frac{P(e|h_i)}{P(e|h_j)} \times \frac{P(h_i)}{P(h_j)}$$

example

You see someone coughing. Here are some possible explanations:

- h_1: cold
- h_2: stomachache
- h_3: lung cancer

Which of these seems like the best explanation of their coughing? Why?

example

$$P(\mathbf{cold}|\mathbf{cough}) \propto P(\mathbf{cough}|\mathbf{cold}) \times P(\mathbf{cold})$$
$$P(\mathbf{stomachache}|\mathbf{cough}) \propto P(\mathbf{cough}|\mathbf{stomachache}) \times P(\mathbf{stomachache})$$
$$P(\mathbf{lung\ cancer}|\mathbf{cough}) \propto P(\mathbf{cough}|\mathbf{lung\ cancer}) \times P(\mathbf{lung\ cancer})$$

cold beats **stomachache** in the likelihood
cold beats **lung cancer** in the prior

=> P(**cold|cough**) is greatest
=> both priors and likelihoods important!

A linguistic application: epistemic modals

Modality & probability

Modality is the language of possibility, uncertainty, deliberation:

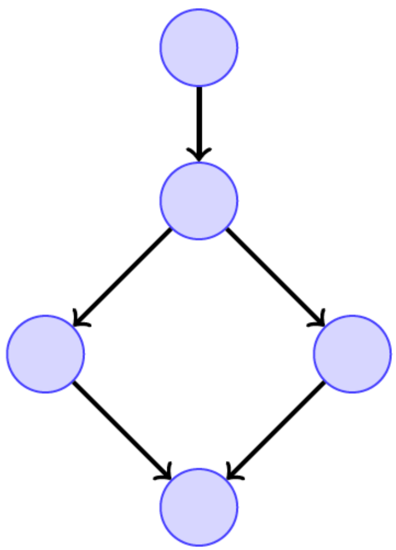
- *likely, certain, possible, must, ...* (epistemic)
- *good, obligatory, must, should ...* (deontic)

Received theories of modal semantics are framed in terms of quantification over a set of best possibilities ("worlds").

My work argues that

- modality is best thought of in terms of scales rather than quantification
- non-maximal possibilities are systematically relevant
- probability plays a crucial role

Lewis-Kratzer semantics



Lewis '73: Rain is better than snow iff the best rain-worlds are ranked above the best snow-worlds.

Kratzer '81: Closely related semantics derived from 'conversational backgrounds', expanded to cover all graded and comparative modalities.

- Dominant framework today

(Portner '09, Kratzer '12, etc.)

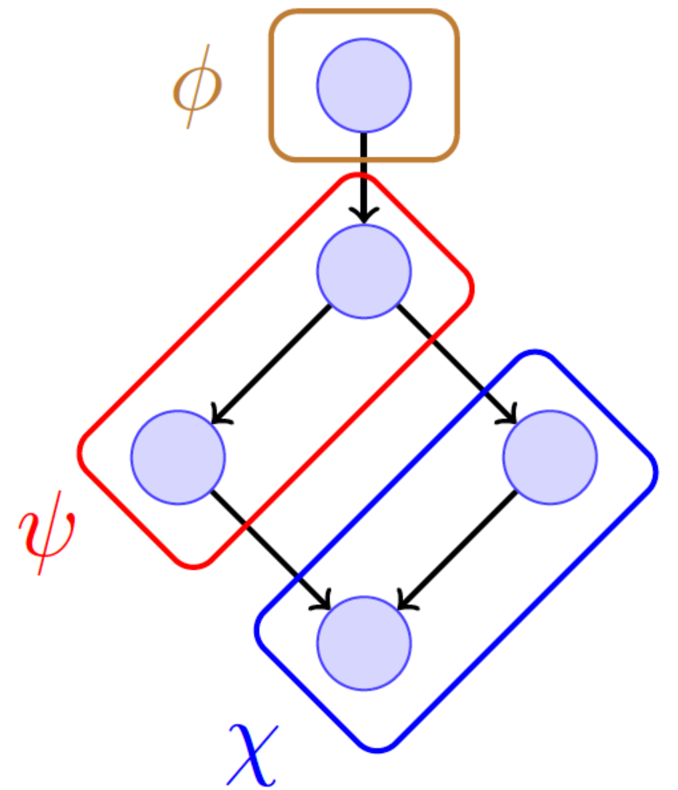
The disjunction problem

What if likelihood = comparative possibility?

Then we validate:

- ϕ is as likely as ψ
- ϕ is as likely as χ
- $\therefore\ \phi$ is as likely as $(\psi \vee \chi)$

Exercise: generate a counter-example.



Probabilistic semantics for epistemic adjectives

An alternative: likelihood is probability.

– fits neatly w/a scalar semantics for GAs

Exercise: show that probabilistic semantics correctly handles your counter-model from previous exercise:

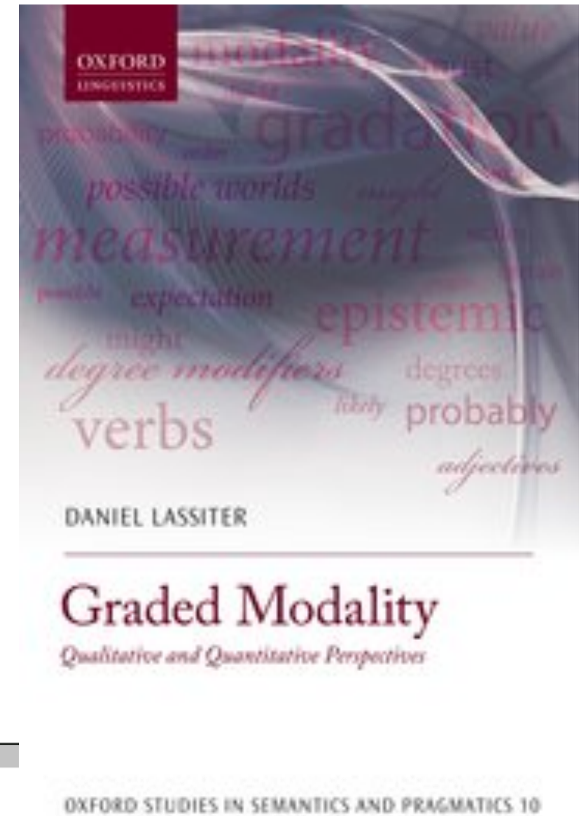
- $\mu_{likely}(\phi) \geq \mu_{likely}(\psi)$
- $\mu_{likely}(\phi) \geq \mu_{likely}(\chi)$
- $\not\models \mu_{likely}(\phi) \geq \mu_{likely}(\psi \vee \chi)$

Key formal difference from comparative possibility?

Other epistemics

Ramifications throughout the epistemic system

- logical relations with *must, might*, *certain*, etc
- make sense of weak *must*



Shameless self-promotion:

Journal of Semantics, 32, 2015: 649–684
doi:10.1093/jos/ffu008
Advance Access publication July 23, 2014

Epistemic Comparison, Models of Uncertainty, and the Disjunction Puzzle

DANIEL LASSITER
Stanford University

Abstract

The best known theory of modality in linguistics (Kratzer 1991, 2012) uses a binary relation on worlds to state truth-conditions for sentences with epistemic auxiliaries, and

Nat Lang Semantics (2016) 24:117–163
DOI 10.1007/s11050-016-9121-8

ORIGINAL PAPER

***Must*, knowledge, and (in)directness**

Daniel Lassiter[1]

Inference & generative models

holistic inference: the good part

probabilistic models faithfully encode many common-sense reasoning patterns.

e.g., explaining away: evidential support is non-monotonic

non-monotonic inference:

- If *x* is a bird, *x* probably flies.
- If *x* is an injured bird, *x* probably doesn't fly.

(see Pearl, 1988)

holistic inference: the bad part

- with *N* worlds we need 2^n-1 numbers
  - unmanageable for even small models
- huge computational cost of inference: update all probabilities after each observation
- is there any hope for a model of knowledge that is both semantically correct and cognitively plausible?

Generative models

We find very similar puzzles in:

- possible-worlds semantics
- formal language theory

Languages: cognitive plausibility depends on representing **grammars**, not stringsets

- ‘infinite use of finite means’

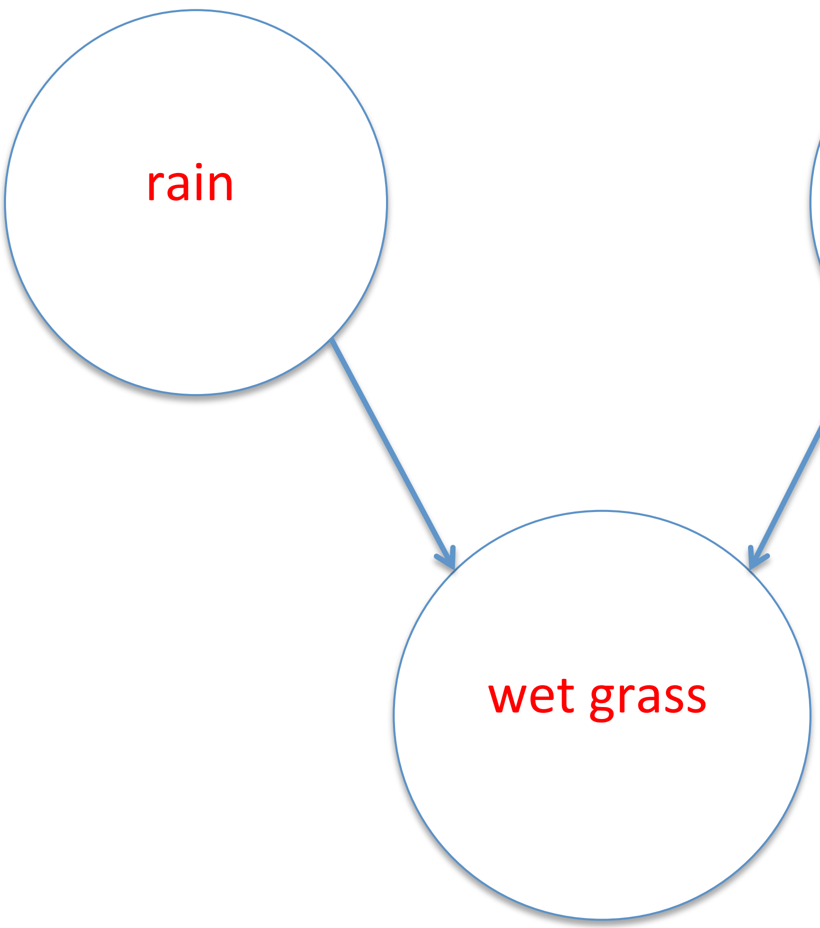
Generative models ~ grammars for distributions

- and for possible-worlds semantics!

Kinds of generative models

- Causal Bayes nets
- Structural equation models
- Probabilistic programs

Causal Bayes nets



wet grass dependent on **rain** and **sprinkler**

rain and **sprinkler** independent (but dependent given **wet grass** !!)

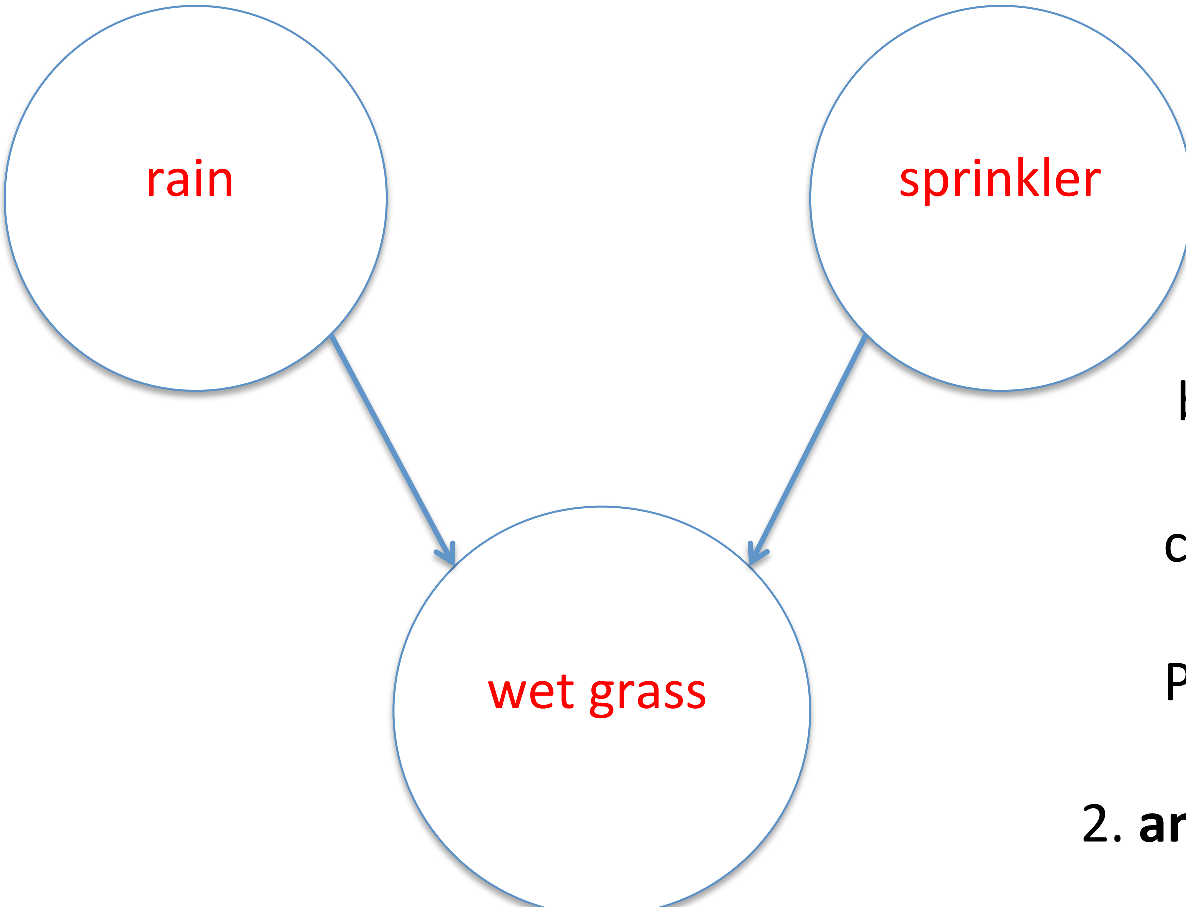
upon observing **wet grass** = 1, update P(V) := P(V|**wet grass** = 1)

high probability that at least one enabler is true

(Pearl, 1988)

Demo!

sketch: approx. inference in CBNs



1. **Repeat many times:**

   a. sample a value for nodes with no parents

   P(**rain)**

   P(**sprinkler)**

   b. work downward, sampling values for each node conditional on its parents

   P(**wet grass|rain, sprinkler**)

2. **analyze accepted samples**

Demo!

explaining away

Multiple possible causes leads to the inference pattern **explaining away**.

1. observe that **wet grass** is true:
   => P(**rain**) increases
   => P(**sprinkler**) increases
2. observe that **sprinkler** is true
   => P(**rain**) goes back to prior

Demo!

intransitivity of inference

- if **rain**, infer **wet grass**
- if **wet grass**, infer **sprinkler**
- NOT: if **rain**, infer **sprinkler**

We can’t avoid holistic beliefs; best we can do is exploit independence relationships

exact & approximate inference

A vending machine has one button, producing bagels with probability p and cookies otherwise.

H: the probability p is either .2, .4, .6, or .8, with equal prior probability.

You hit the button 7 times and get

B B B B C B B

What is p?

exact inference

exact calculation

Prior: $\forall h : P(h) \propto 1$

L'hood: $P(\text{seq}|p) = p^{N_B(\text{seq})}(1-p)^{N_C(\text{seq})}$

the observed sequence

$$\forall h : P(h) = 1/|H| = .25$$

$$P(BBBBCBB|p) = p * p * p * p * (1-p) * p * p$$

approximate inference

Monte Carlo approximation

(rejection sampling)

1. repeat many times:
    a. choose *h* according to prior, simulate predictions
    b. accept *h* iff simulated *e* is equal to observed *e*
2. plot/analyze accepted samples

Demo!

Today’s highlights

- Probability as an intensional logic
  - Linguistic application: epistemic modality
- Problems of tractability => generative models
- Sampling is a useful way to think of inference in generative models

Do generative models and sampling have interesting linguistic applications?

Linguistic applications: next 3 lectures

1. indicative conditionals
2. causal models & counterfactuals
3. reasoning about impossibilia

Indicative conditionals

Conditional reasoning as rejection sampling

– enforces Stalnaker's thesis

Background semantics is trivalent

– define a sampler over trivalent sentences

Linguistic advantages:

– avoids Lewis-style triviality results

– semantic treatment of conditional restriction

Connections w/ other ways to avoid triviality

Causal models & counterfactuals

Parenthood in gen. models naturally thought of as causal influence

Counterfactual reasoning as intervention

- connections to Lewis/Stalnaker semantics
- reasons to prefer the causal models approach

Filling a major gap: treatment of complex, quantified antecedents

Reasoning about impossibilia

What if 2 weren't prime?

- doesn't make sense in possible-worlds semantics
- but people understand the question …

Generative models can represent non-causal information, e.g., a theory of arithmetic

- probabilistic programs support interventions
- **lazy computation** means we only compute partial representations

Connections to hyperintensionality

Thanks!

contact:
danlassiter@stanford.edu