Bayesian Pragmatics

Daniel Lassiter

Stanford Linguistics

Jelinek Memorial Workhop
Charles University, Prague
July 9, 2014

Rich language understanding requires resolving massive uncertainty about context and interlocutors' goals and beliefs.

- How and why do these factors influence interpretation?
- How can interpretation be mostly easy and successful when it ought to be really hard and error-prone?

Answer: listeners construct rich interpretations by combining

- domain-specific linguistic knowledge
- knowledge of the world and their interlocutors

via domain-general probabilistic inference and social reasoning mechanisms.

Toward computational psychosemantics

in which Montague and Grice meet the Rev. Bayes

- Bayesian inference in brief
- Bayes/Grice: Interpretation as intention recognition and as latent variable estimation
- Iterated-reasoning architecture for language understanding
- Three kinds of pragmatic inference
  - Inferring speaker beliefs, desires on the basis of observed utterances [implicature]
  - Inferring literal meaning [ambiguity resolution]
  - Inferring values of free variables [context-sensitive meaning]

Caveats and hat tips

The framework is currently under development with input from lots of people including (but not limited to):

Mike Frank, Noah Goodman, Andreas Stuhlmüller, Leon Bergen, Roger Levy, Judith Degen, Adam Vogel, Chris Potts, Percy Liang, Gerhard Jäger, Michael Franke, ...

Today's story emerges largely from my collaborations with Noah Goodman (Lassiter & Goodman '13, Goodman & Lassiter '14; usual disclaimers apply).

Formal models of information & inference

Deductive paradigm

- An info state I_A is an unstructured set of possibilities, representing the total of what agent A fully believes
- All-or-nothing inference: p follows from I_A iff true everywhere
- Learning q contracts the live possibilities: $I_A \underset{\text{learn } q}{\Longrightarrow} I_A \cap q$

Formal models of information & inference

Deductive paradigm

- An info state I_A is an unstructured set of possibilities, representing the total of what agent A fully believes
- All-or-nothing inference: p follows from I_A iff true everywhere
- Learning q contracts the live possibilities: $I_A \underset{\text{learn } q}{\Longrightarrow} I_A \cap q$

Bayesian paradigm

- An info state P_A is an measure on a set of possibilities $L \subseteq W$, representing agent A's graded beliefs
- Learning q updates the measure without changing the domain:

$$P_A(prop) \underset{\text{learn } q}{\Longrightarrow} P_A(prop|q) = \frac{P_A(prop \ \& \ q)}{P_A(q)}$$

- Similar notion of entailment; richer, graded inference

Cognitive interpretation (Tenenbaum et al. '11)

Generative models represent common-sense theories of causal relations between features of the world.

- generate predictions about future events
- Test against evidence; refine model

Simple example:

- Fires, incense, asteroids can cause smoke
- So, if you see smoke one of these things might be causing it
- Observations influence both $P(\text{cause})$ and $P(\text{smoke}|\text{cause})$

Priors matter, as they should.

Applications in many areas of cognitive science: learning, reasoning, categorization, language, vision, motor control, ...

Bayes & Grice

(Grice '89, Lewis '69, Clark '96)

Themes from Grice:

- literal meaning underdetermines communicated content
- the rest requires inferring speakers' latent intentions
- pragmatic enrichments are inferences from chosen action together with rationality, cooperativity assumptions
  - We're fixing a car, and you hand me a tool I don't recognize
  - 'Give me that tool' when we both know the most useful tool
  - 'Al met a woman' when it's relevant if he also married her

Implicit in Grice: Ls maintain a model of S's action planning and use it to generate smart inferences

- L model of S generates predictions about what S will do
- pragmatic inferences emerge from choice to say u rather than saying or doing something else

Listeners actively model speakers' motivations (Clark '75)

Nixon, not long before he was deposed, was quoted as saying at a news conference, "I am not a crook." We all saw immediately that Nixon shouldn't have said what he said. He wanted to assure everyone that he was an honest man, but the wording he used was to deny that he was a crook. Why should he deny that? He must have believed that his audience was entertaining the possibility that he was a crook, and he was trying to disabuse them of this belief. But in so doing, he was tacitly acknowledging that people were entertaining this possibility, and this was something he had never acknowledged before in public. Here, then, was a public admission that he was in trouble, and this signaled a change in his public posture. My inferences about Nixon's utterance stopped about there, but I am sure that the knowledgeable White House press corps went on drawing further inferences. In any event we all took this utterance a long way.

Listeners actively model speakers' motivations (Clark '75)

Nixon, not long before he was deposed, was quoted as saying at a news conference, "I am not a crook." We all saw immediately that Nixon shouldn't have said what he said. He wanted to assure everyone that he was an honest man, but the wording he used was to deny that he was a crook. Why should he deny that? He must have believed that his audience was entertaining the possibility that he was a crook, and he was trying to disabuse them of this belief. But in so doing, he was tacitly acknowledging that people were entertaining this possibility, and this was something he had never acknowledged before in public. Here, then, was a public admission that he was in trouble, and this signaled a change in his public posture. My inferences about Nixon's utterance stopped about there, but I am sure that the knowledgeable White House press corps went on drawing further inferences. In any event we all took this utterance a long way.

Rich inferences from

- choice to utter *u* instead of anything else S could have said or done
- inferences rely on a model of how S would have behaved if intentions were different

Sampling of psycholinguistic evidence

"Put the cube in the can": Ls rapidly narrow attention to items for which it's possible to fulfill command (Chambers et al. '02)

Sampling of psycholinguistic evidence

"Put the cube in the can": Ls rapidly narrow attention to items for which it's possible to fulfill command (Chambers et al. '02)

"Hand me the cake mix": Ls infer that the cake mix near to them is intended, since otherwise S would pick it up himself — unless S's hands are full (Hanna & Tanenhaus '04)

Sampling of psycholinguistic evidence

"Put the cube in the can": Ls rapidly narrow attention to items for which it's possible to fulfill command (Chambers et al. '02)

"Hand me the cake mix": Ls infer that the cake mix near to them is intended, since otherwise S would pick it up himself — unless S's hands are full (Hanna & Tanenhaus '04)

"What's above the cow with shoes?": Ls assume Ss don't ask questions they know the answer to (Brown-Schmidt et al. '08)

Bayesian action understanding (Baker, Saxe & Tenenbaum '09)

A woman is walking down the street. She suddenly stops, turns around, and runs in the opposite direction. Why did she do that?

- She'd missed her bus and had to take the subway.
- She realized she'd forgotten to turn off the stove.
- There's been an alien invasion, and she saw one coming.

Human inferences in simple cases captured by "inverse planning":

$$P(\text{Belief, Desire} \mid \text{action}) = \frac{P(\text{action} \mid \text{Belief, Desire}) \times P(\text{Belief, Desire})}{\sum_{\text{Bel*, Des*}} P(\text{action} \mid \text{Bel*, Des*}) \times P(\text{Bel*, Des*})}$$

Sampling gloss: "use model to generate predictions about beliefs, desires; on that basis, predict action; throw out incorrect predictions; update beliefs by reference to belief-desire combos that remain."

Bayesian language understanding

Key insight of Grice, Lewis, Clark: language understanding is a special case of action understanding.

Key question (refined from beginning)

How can communication be easy and successful most of the time, when it should be difficult and highly error-prone?

Proposed answer: listeners use powerful, domain-general Bayesian inference mechanisms to combine

- linguistic knowledge (potentially domain-specific)
- predictions generated by social reasoning, in particular active modeling of speakers' linguistic and non-linguistic choices

Given richly structured beliefs, inference from observed utterance allows listeners to update beliefs about many other variables

Recursive interpretation (cf. Lewis '69, Clark '96)

L reasons about what the world is like (including S's intentions), given observed utterance u:

$$P_L(w|u) \propto P_S(u|w) \times P_L(w)$$

S reasons about what to say, given (a) desire that L infer w, and (b) private utterance preferences:

$$P_S(u|w) \propto P_L(w|u) \times P_S(u)$$

Problem: this reasoning will go on forever.

Solution: pick a base case, recurse up to some (low) level.

Recursive interpretation simplified (Franke '08, Frank&Goodman '12)

Pragmatic listener L_1 reasons about w given priors and speaker model.

$$P_{L_1}(w|u) \propto P_{S_1}(u|w) \times P_{L_1}(w)$$

S_1 reasons about what to say, given that they want a literal listener L_0 to infer their intention.

$$P_{S_1}(u|w) \propto P_{L_0}(w|u) \times P_{S_1}(u)$$

Literal listener L_0 simply assumes u is true, without reasoning about S.

$$P_{L_0}(w|u) = P_{L_0}(w|u \text{ is true})$$

Recursive interpretation simplified (Franke '08, Frank&Goodman '12)

Pragmatic listener L_1 reasons about w given priors and speaker model.

$$P_{L_1}(w|u) \propto P_{S_1}(u|w) \times P_{L_1}(w)$$

S_1 reasons about what to say, given that they want a literal listener L_0 to infer their intention:

$$P_{S_1}(u|w) \propto P_{L_0}(w|u) \times P_{S_1}(u)$$

Literal listener L_0 simply assumes u is true, without reasoning about S.

$$P_{L_0}(w|u) = P_{L_0}(w|u \text{ is true})$$

Note: this model encodes a speaker preference for informative utterances.

Pragmatic enrichment: Implicature

- Dad, I met a girl [quantity]
- Mary ate most of your cookies [quantity]
- The candidate is punctual and his wife is friendly [relevance]
- Mrs. X uttered a series of sounds closely corresponding to the tune of "Home sweet home" [manner]

Recursive Bayes generates quantity implicatures automatically

(Franke '08, Frank & Goodman '12, G. & Stuhlmüller '13, Vogel e.a. '13)

- Manner implicatures: see Bergen et al. '12.
- Relevance: seems straightforward, but not yet investigated.

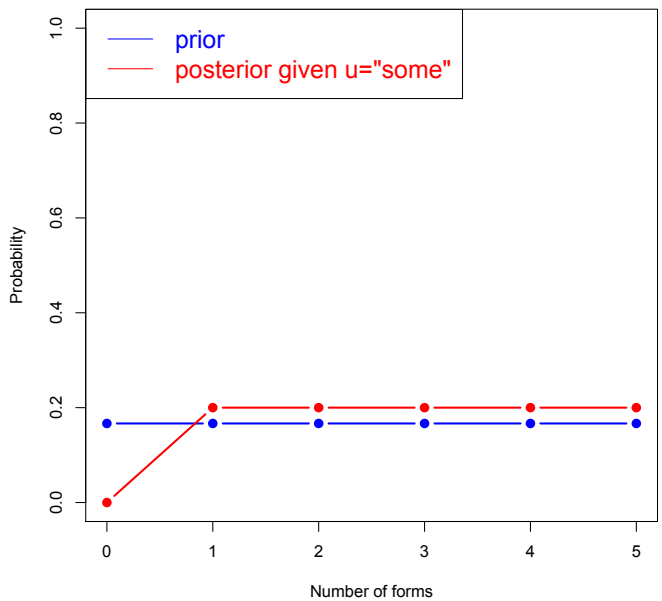
Quantity implicature

A: We had 5 forms left to fill out. How many did you do?

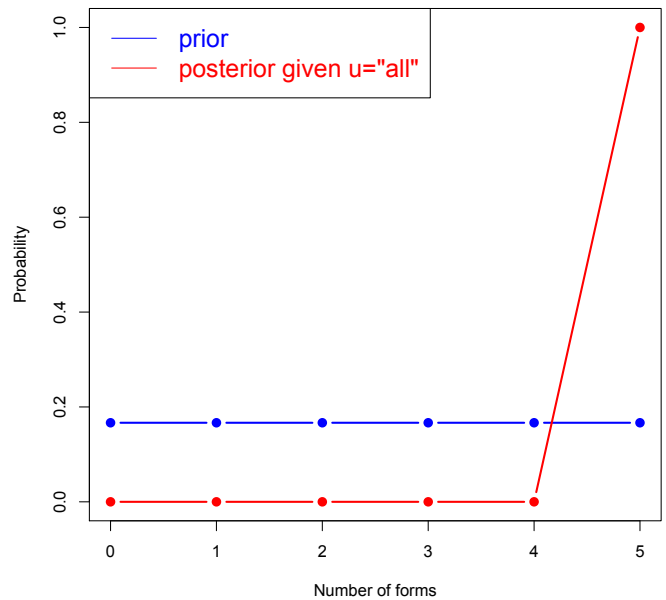
B: I filled out some of them.

Suppose $P_{L_{0/1}}(\text{B filled out } n \text{ forms})$ is uniform for $n \in \{0, ..., 5\}$.

$P_{L_0}(\text{B filled out } n \text{ forms} | u = \text{'some'})$



$P_{L_0}(\text{B filled out } n \text{ forms} | u = \text{'all'})$



Crucial informativity difference when "all" is true.

Informativity effects on speaker model

$$P_{L_1}(w|u) \propto P_{S_1}(u|w) \times P_{L_1}(w)$$
$$P_{S_1}(u|w) \propto P_{L_0}(w|u) \times P_{S_1}(u)$$
$$P_{L_0}(w|u) = P_{L_0}(w|u \text{ is true})$$

Let alternatives be "none", "some", and "all", with uniform prior.

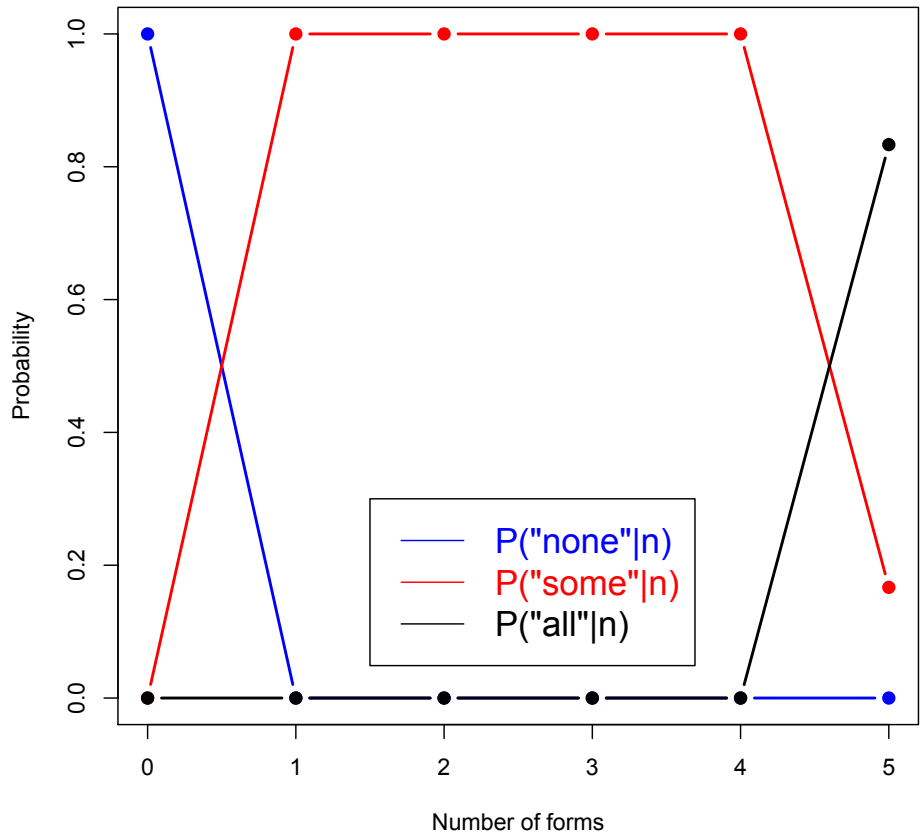
- If $n = 0$, only "none" is true $\Rightarrow$ choose "none"
- If $0 < n < 5$, only "some" is true $\Rightarrow$ choose "some"
- If $n = 5$, "some" and "all" both true. Now,
  - $P_{L_0}(n = 5|\text{"some"}) = 1/5$
  - $P_{L_0}(n = 5|\text{"all"}) = 1$

so $P_{S_1}(\text{"some"}|n = 5) = \frac{1/5}{1/5+1} \approx .17$, $P_{S_1}(\text{"all"}|n = 5) \approx .83$

- S_1 nearly 5 times as likely to say "all" when it is true.

Informativity effects on speaker model

Graphical depiction: S_1 utterance probability by n.

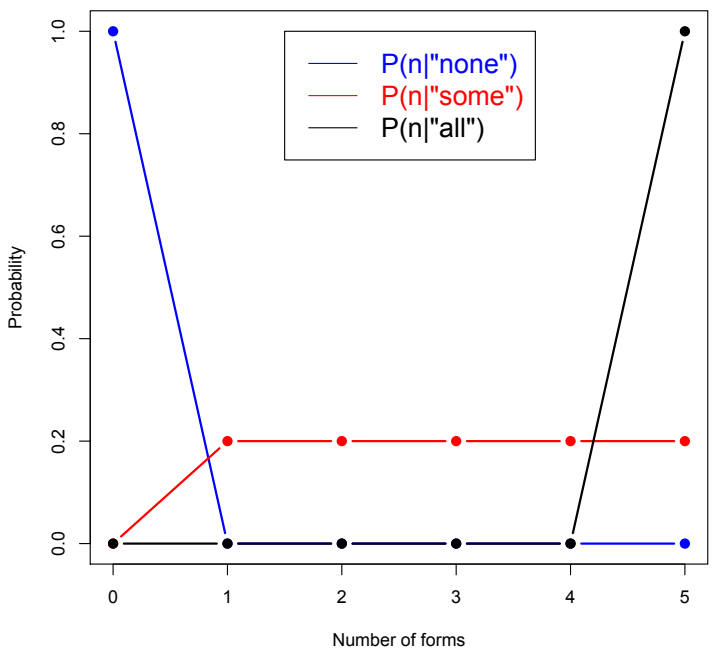


Effects on pragmatic listener

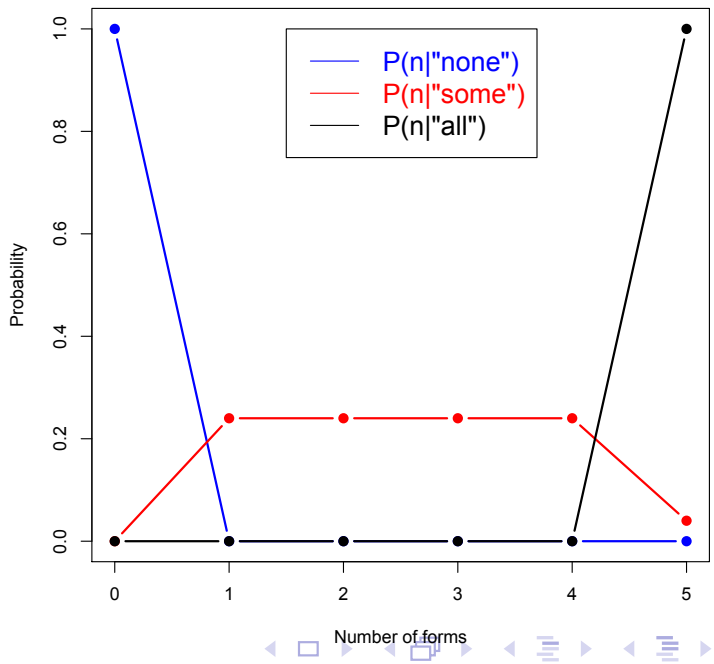
L_1 incorporates this asymmetry and reasons backwards:

- S said "some"
- If "all" were true, he probably would have said "all" instead, since that would be much more informative
- So, "all" probably isn't true: $P_{L_1}(5|\text{"some"}) \approx .04$.

$L_0 : P_{L_0}(\text{B filled out } n \text{ forms}|u)$



$L_1 : P_{L_1}(\text{B filled out } n \text{ forms}|u)$



Ambiguity and underspecification

Bayesian update is only defined for things that can be true or false.

$$P_{L_1}(w|u) \propto P_{S_1}(u|w) \times P_{L_1}(w)$$
$$P_{S_1}(u|w) \propto P_{L_0}(w|u) \times P_{S_1}(u)$$
$$P_{L_0}(w|u) = P_{L_0}(w|u \text{ is true})$$

What if u doesn't pick out a unique proposition?

- "Take AI to the bank" (which kind of bank?)
- "It is big" (What is 'It'? How big is 'big'?)

Ambiguity

Ambiguous utterances have a random choice in their meaning (or, a random choice is required to decide which word was used).

$$[\![bank]\!] = \begin{cases} [\![river\ bank]\!] \text{ with probability } p \\ [\![financial\ bank]\!] \text{ with probability } 1-p \end{cases}$$

- Probabilistic reference projects up through compositional semantics straightforwardly (see Goodman & Lassiter '14).

L_0 conditions on truth of utterance given model. Predictions:

- Plausible interpretations favored, because more likely true [true in many samples from world model]
- Implausible interpretations filtered out as probably false

Ambiguities thus resolved acc. to probability of being true.

- Mary/The otter/The businessman went to the bank.

$\Rightarrow$ World knowledge directly influences interpretation.

Underspecification: Vague scalar adjectives

Scalar adjectives have notoriously context-sensitive meanings.

- I saw a big {baby, football player, tree, skyscraper, planet}.
- Click on a/the large circle. [varying distribution of sizes]

Why do these adjectives mean what they do, relative to a given context? What is the role of the reference class?

How can we communicate any information at all using expressions this semantically flexible?

Theoretical constraints:

- Non-uniform prior on resolved meanings would make interpretation too inflexible
- Context-sensitive interpretations have to emerge from interpretation process and world knowledge

Underspecification: Vague scalar adjectives

Positive-form adjectives compare an object's measure along some scale to a threshold value θ. (Cresswell '76, Kennedy '07, etc.)

$$[\![\textit{Al is tall}]\!]^{\theta} = 1 \text{ iff } \mu_{\textit{height}}(\textbf{Al}) > \theta$$

Common idea: context determines a value for θ, taking into account lexical information, reference classes, ...

- What is "context", and how does it go about doing its job?

Bayesian approach:

- Parametrize model by θ, with uniform prior
- Estimate $P(h, \theta | u)$
- Marginalize out θ to infer Al's height.

Underspecification: Vague scalar adjectives

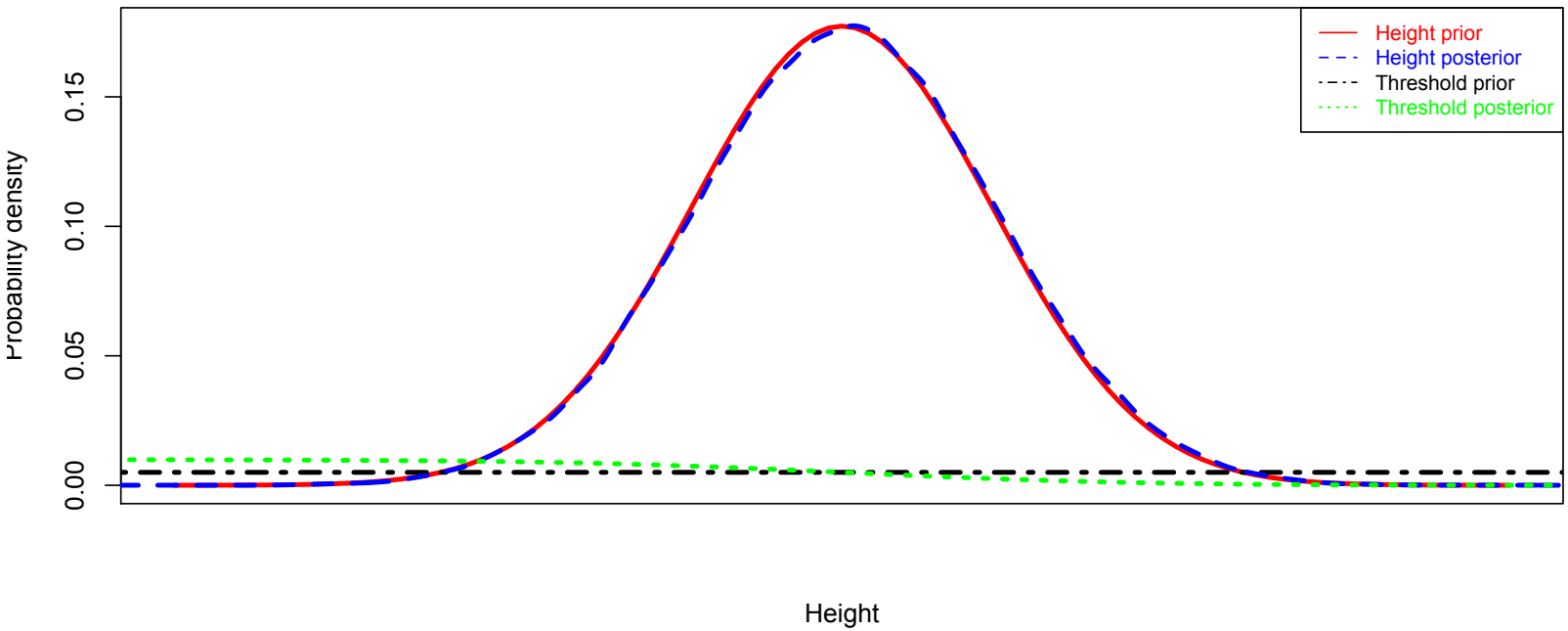
Model with semantic "nuisance" variables at L_0.

$$P_{L_1}(w|u) \propto P_{S_1}(u|w) \times P_{L_1}(w)$$
$$P_{S_1}(u|w) \propto P_{L_0}(w|u) \times P_{S_1}(u)$$
$$P_{L_0}(w|u) = \sum_{\theta} P_{L_0}(w|u \text{ is true relative to } \theta, \theta) \times P_{L_0}(\theta)$$

Note: we're introducing θ at L_0 and eliminating it there as well

Simulation

Disaster: "AI is tall" conveys no information.



- L_0 prefers (implausibly) strong interpretations
- S_1 infers that "tall" is most informative for short people

Variable-passing solution (Lassiter & Goodman '13, cf. Bergen et al. '12)

Solution: instantiate variables at L_1 and pass them down.

Pragmatic listener derives joint inferences about states of the world and interpretations given values for variables.

$$P_{L_1}(w, \theta | u) \propto P_{S_1}(u | w, \theta) \times P_{L_1}(w) \times P_{L_1}(\theta)$$

S_1 chooses u with a preference for informative utterances given θ

$$P_{S_1}(u | w, \theta) \propto P_{L_0}(w | u, \theta) \times P_{S_1}(u | \theta)$$

L_0 conditions on literal meaning, given values for variables.

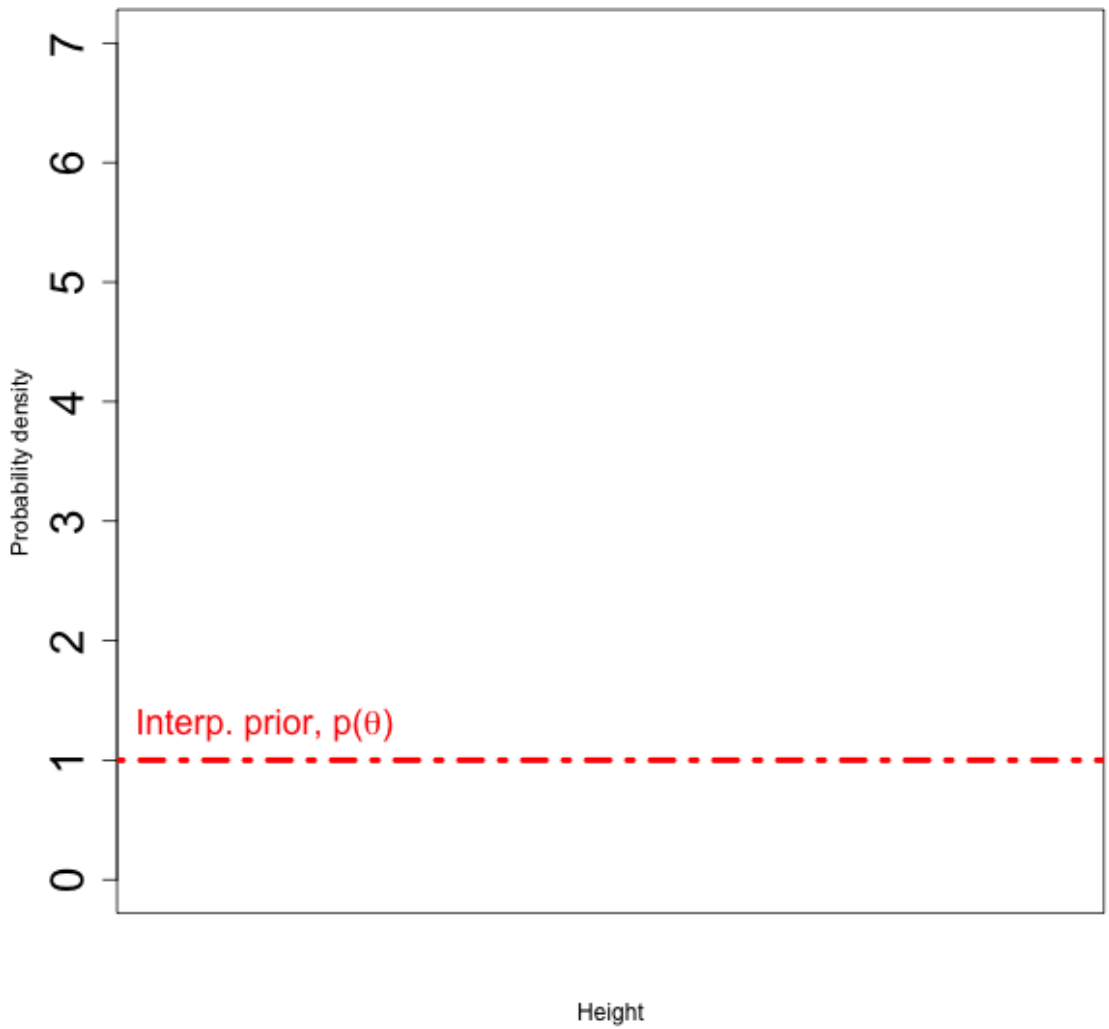
$$P_{L_0}(w | u, \theta) = P_{L_0}(w | u \text{ is true relative to } \theta)$$

Simulation: Gaussian prior ("tall") (Lassiter & Goodman '13)

$$P_{L_1}(h, \theta | u) \propto \qquad \times P_{L_1}(\theta)$$

Let $u =$ "Al is tall".

$P(\theta)$ is uniform: any possible interpretation of *tall* is equally likely *a priori*.



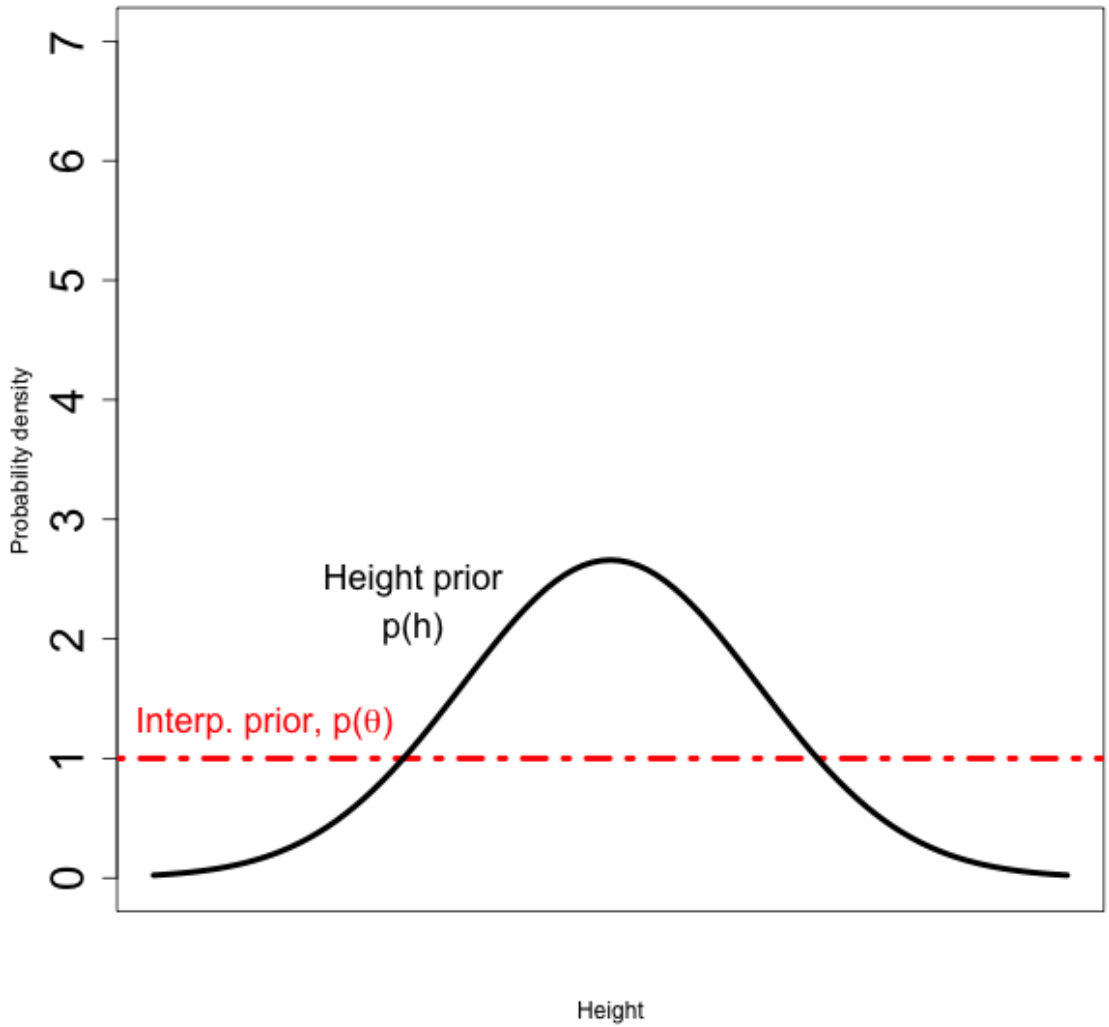
Simulation: Gaussian prior ("tall") (Lassiter & Goodman '13)

$$P_{L_1}(h, \theta | u) \propto \quad \times P_{L_1}(h) \times P_{L_1}(\theta)$$

$P_{L_1}(h)$ = total prob. of worlds in which Al's height is h.

A reasonable prior for heights is a Gaussian.

$P_{L_1}(h)$ is highest for moderate values.

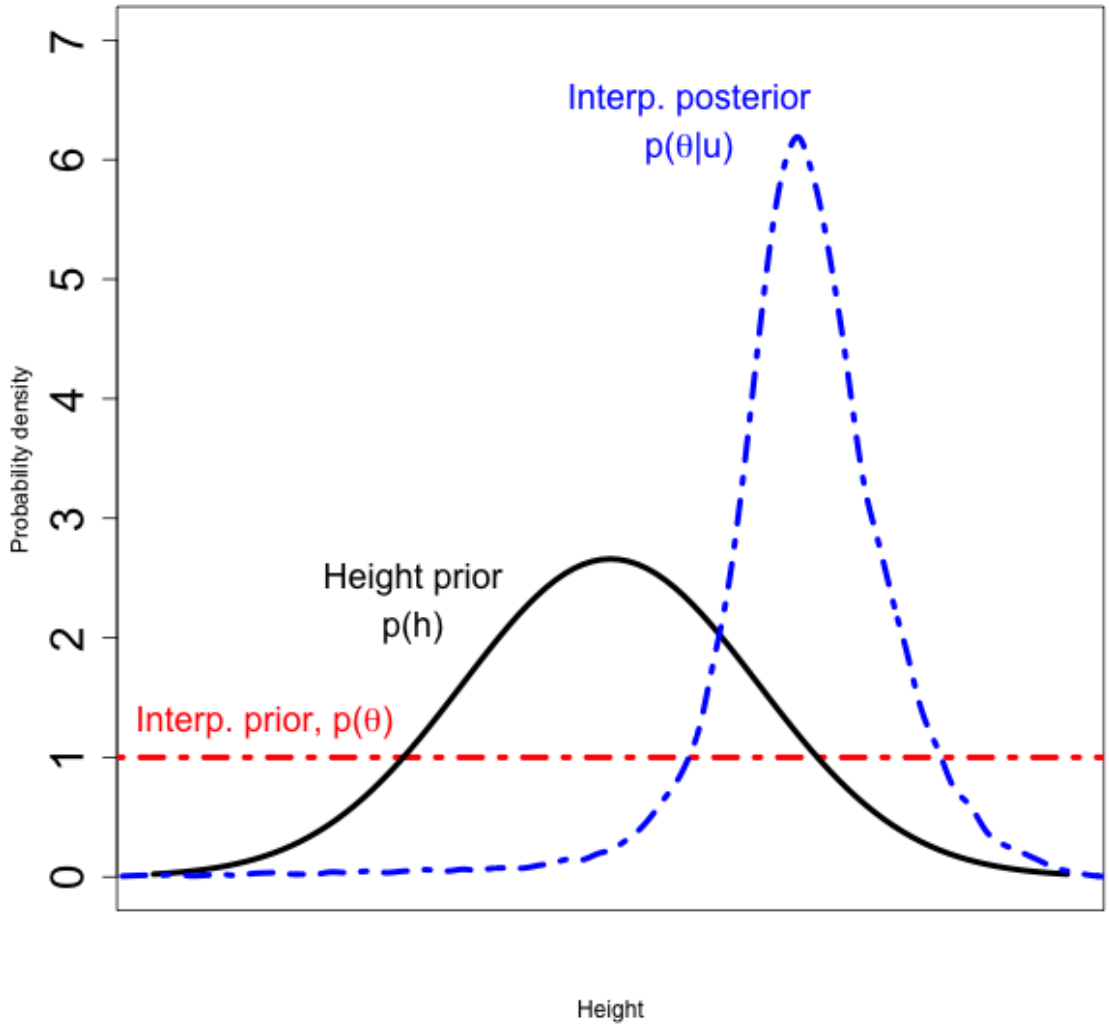


Simulation: Gaussian prior ("tall") (Lassiter & Goodman '13)

$$P_{L_1}(h, \theta | u) \propto P_{S_1}(u | h, \theta) \times P_{L_1}(h) \times P_{L_1}(\theta)$$

$P_{S_1}(u|h,\theta)$ favors higher θ because they make u more informative.

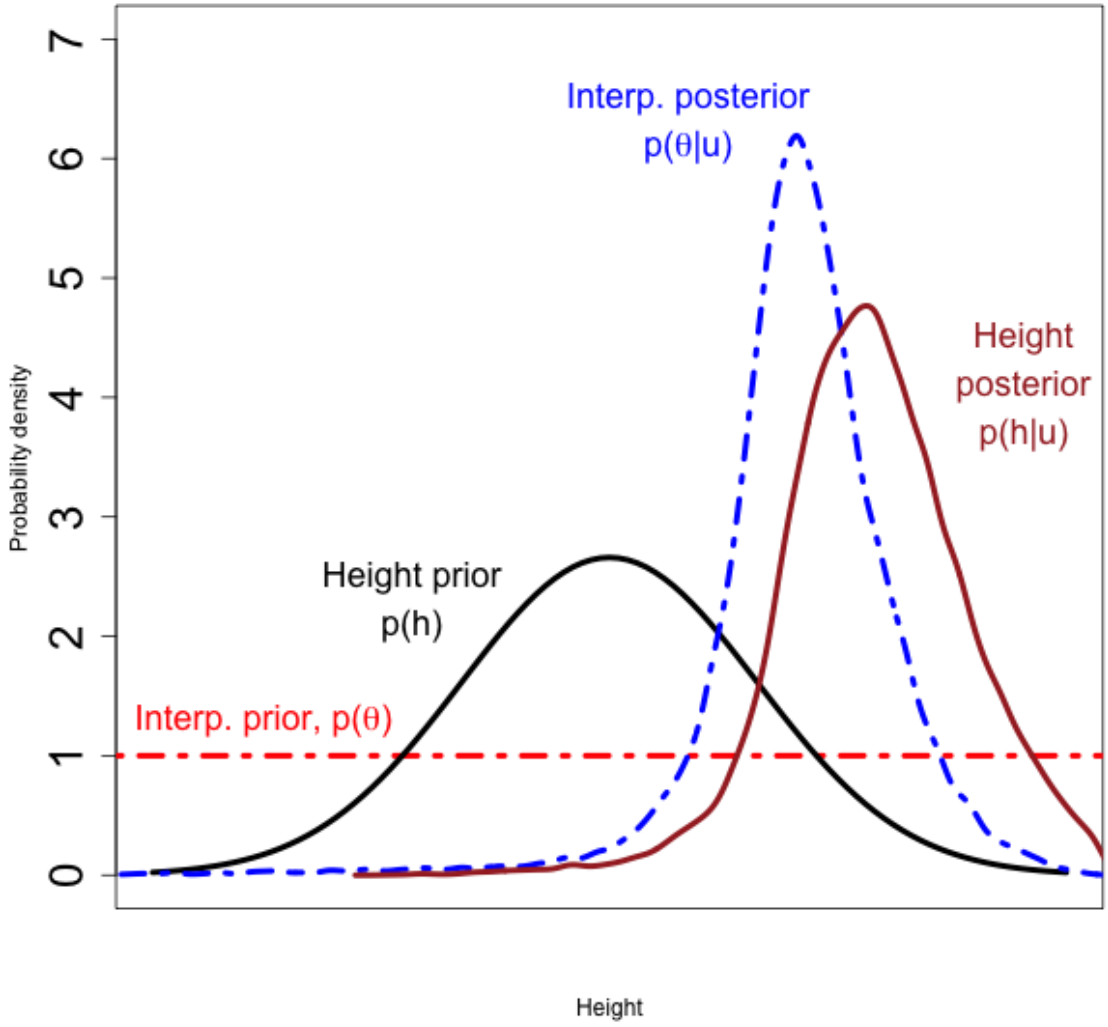
Context-sensitive probabilistic meanings emerge from competition between prior and likelihood terms.



Simulation: Gaussian prior ("tall") (Lassiter & Goodman '13)

Inferred height of Al, given u = "Al is tall".

Note crucial role of statistical priors.



Bayesian interpretation: summary

Implicature: need pragmatic enrichment even with full meanings

Bayesian interpretation requires a proposition at base listener L_0.

When literal meaning doesn't give us one, we infer ...

- and deal with whatever uncertainty remains.

We can resolve uncertainty about literal meaning either at L_0 or L_1.

- L_0: exclusive preference for truthful interpretations $\Rightarrow$ high probability, in some cases logically weak
- L_1: variable passes through speaker model, introducing countervailing preference for informative interpretations

Empirical question which phenomena go which way. (Pronouns, reciprocals, ...?)

Conclusions

Bayesian interpretation offers

- new, unified perspective on important issues in pragmatics
- useful synthesis of work in formal linguistics and cognitive science
- a way to combine benefits of formal models of meaning (precision, productivity) without ignoring pervasive uncertainty and gradation in language

Needed for further progress:

- deepen engagement between computational cognitive science, NLP, formal semantics/pragmatics
- find methods for verification/falsification/fine-tuning of models: behavioral experiments, corpus studies, ...?

Thanks for listening!

Email: danlassiter@stanford.edu

Sorites paradox

1) A 7-foot-tall man is tall.
2) A man who is ϵ shorter than a tall man is also tall.
3) $\therefore$ A 3-foot-tall man is tall.

Translation of (2): If $\mu_{height}(x) > \theta$, then $\mu_{height}(x) - \epsilon > \theta$ as well.

In our simulations, this has high but non-maximal posterior probability: $\approx .98$ with $\epsilon = .01$.

Sorites paradox

1) A 7-foot-tall man is tall.
2) A man who is ϵ shorter than a tall man is also tall.
3) $\therefore$ A 3-foot-tall man is tall.

Probabilistic approach avoids the logical problem of the sorites:

- $P(\text{premise 1}) \approx 1$, $P(\text{premise 2}) \approx .98$, $P(\text{premise 3}) \approx 0$.
- Repeated use of a premise with high but non-maximal probability does not preserve high probability (Kyburg '61)