A Rational Analysis of Persuasion

S. A. Barnett

1 Background

This project will investigate the project of inference within a social context: namely, given that the evidence that an individual receives originates from human communication, how do the inferences drawn from that inference differ from our normative expectations about how such inferences ought to be drawn? In particular, I will focus on the context in which a person hears two different sides of a debate, and must update their beliefs based on the evidence provided each side.

My aim is to provide a computational model of inference over a debate based upon Bayesian principles. This is inspired by the Rational Speech Act model of communication [5], in which language understanding is viewed as a form of Bayesian inference whose sampling procedure is influenced by a theory of mind for the speaker. However, rather than the cooperative principle underpinning the inference model [1], the debate context requires that a pragmatic listener instead models each speaker as aiming to be maximally *persuasive*: I call the principle that can capture this as the *rhetorical principle.* Hence, my project will provide a model that can respect the rhetorical principle in certain contexts.

Previous research has found that listeners in a debate context display two patterns of inference: the *weak evidence effect* [2, 3, 4], and the *strong evidence effect* [6]. The *weak evidence effect* occurs when receiving weak evidence makes people less likely to believe a conclusion relative to the marginal belief. This effect may arise from receivers expecting debaters to produce the strongest evidence for their case. The *strong evidence effect*, on the other hand, occurs when strong evidence does not always lead to stronger inferences. This effect may arise from the possibility that stronger evidence makes the listener believe that the debater was overly biased, causing the listener to discount the evidence.

2 Question

Can weak and strong evidence effects be accounted for by a rhetorical principle?

3 Method

The experiments are run with the *sticks game*. In this game, a sample of N sticks is drawn, with the stick length being modelled as i.i.d. draws from a standard uniform distribution:

$$\{U_n\}_{n=1}^{N} \sim \mathcal{U}[0, 1].$$

Two agents, A_1 and A_2 observe all of the sticks in the sample, and at each time step the agents take it in turn to show a stick to the judge J. The judge must decide whether the sample is 'long': that

is, whether

$$\bar{U} := \frac{1}{N}\sum_{n=1}^{N} U_n \geq 0.5.$$

Experiment 1 The human plays the judge, and observes two time steps of the game for different values of N. At each step, the human is asked to report her confidence that the sample is 'long'. These confidence ratings are contrasted with the conditional probability evaluations (where the data is assumed to be drawn i.i.d.) to test for the weak and strong evidence effects.

Experiment 2 Two different models of the judge are evaluated on the same task: one in which the judge *assumes* that each agent is sampling the longest and shortest sticks (respectively), and one in which the agent has uncertainty over how the agents are biased, or whether they are biased at all. The plot of the posterior belief in the sample length is used to test for the weak and strong evidence effects.

Experiment 3 The agents are biased, and modelled in order to be maximally persuasive to a judge with uncertainty over the bias of the speaker. Concretely, at each step they select the stick that minimizes the KL divergence between having full confidence in the view consistent with the agent's bias, and the posterior determined by the judge in the above model. The intention of this experiment is to model the strong evidence effect in the speakers.

References

[1] H. Paul Grice, Peter Cole, and Jerry Morgan. "Logic and conversation". In: *1975* (1975), pp. 41–58.

[2] Craig R M McKenzie, Susanna M Lee, and Karen K Chen. "When Negative Evidence Increases Confidence: Change in Belief After Hearing Two Sides of a Dispute". In: *Journal of Behavioral Decision Making* 15.1 (2002), p. 17.

[3] Philip M. Fernbach, Adam Darlow, and Steven A. Sloman. "When good evidence goes bad: The weak evidence effect in judgment and decision-making". In: *Cognition* 119.3 (June 2011), pp. 459–467. ISSN: 00100277. DOI: 10.1016/j.cognition.2011.01.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027711000394 (visited on 10/15/2019).

[4] Adam J. L. Harris, Adam Corner, and Ulrike Hahn. "James is polite and punctual (and useless): A Bayesian formalisation of faint praise". In: *Thinking & Reasoning* 19.3 (Sept. 2013), pp. 414–429. ISSN: 1354-6783, 1464-0708. DOI: 10.1080/13546783.2013.801367. URL: https://www.tandfonline.com/doi/full/10.1080/13546783.2013.801367 (visited on 11/12/2019).

[5] Noah D Goodman and Michael C Frank. "Pragmatic language interpretation as probabilistic inference". In: *Trends in cognitive sciences* 20.11 (2016), pp. 818–829.

[6] Amy Perfors, Daniel J Navarro, and Patrick Shafto. "Stronger evidence isn't always better: A role for social inference in evidence selection and interpretation". In: *CogSci* (2018), p. 6.