Comprehensive Analysis of Bayesian Inference in Goal and Belief Attribution

Dog-Man

1 Introduction

This document provides an in-depth analysis of the mathematical and technical components presented in the given images, focusing on Bayesian inference, probabilistic modeling, and related concepts in the context of goal and belief attribution to agents.

2 Quantitatively Interpreting Belief Statements

2.1 Posterior Computation

The process begins with computing the posterior $P(g, b_{0:T}|a_{1:T})$, where g represents the agent's goal, $b_{0:T}$ is the belief history, and $a_{1:T}$ is the series of observed actions. This posterior is crucial for understanding the agent's internal state given their observable behavior.

2.2 Two-Step Interpretation Process

The interpretation of belief statements follows a two-step process:

1. **Translation**: We translate the statement σ from natural language into a formal representation ψ. This step is crucial for bridging the gap between human-understandable statements and machine-processable logic.

2. **Evaluation**: We evaluate the expected truth value of the translated sentence ψ with respect to the inferred distribution over belief states b_T. This step quantifies the degree of belief in the statement.

2.3 Formal Language Representation

To represent statements like "The player believes that there is a red key in box 1", we extend the Planning Domain Definition Language (PDDL) with a `believes` operator. The formal representation becomes:

```
(believes player (exists (?k - key)
                  (and (iscolor ?k red)
                       (inside ?k box1))))
```

This representation allows for a precise and unambiguous encoding of belief statements.

2.4 Belief Operator

The `believes` operator is parameterized by an agent x, such that (believes x ϕ) indicates that agent x believes ϕ. This corresponds to a restricted fragment of epistemic first-order logic, providing a powerful yet tractable way to represent beliefs.

2.5 Translation Process

To perform the translation from natural language to formal representations, we leverage large language models (LLMs) as general-purpose semantic parsers. The process involves:

1. Creating a dataset of example translations ε from natural language to first-order epistemic sentences.
2. Prompting the language model with these examples and the belief statement σ.
3. Inducing a distribution $P(\psi|\sigma, \varepsilon)$ over possible interpretations ψ of σ.

This approach allows for flexible and context-aware translations of natural language belief statements.

2.6 Evaluation of Observer's Credence

To evaluate the observer's credence in $\psi :=$ (believes x ϕ), we:

1. Extract the sentence ϕ attributed to agent x.

2. Compute the expected value of ϕ being true in the belief state b_T:

$$P(\psi|a_{1:T}) = \mathbb{E}_{b_T \sim P(b_T|a_{1:T})}[\phi \text{ is true in } b_T]$$

This expectation is computed as:

$$P(\psi|a_{1:T}) = \sum_{i=1}^{N} w_i \cdot [\phi \text{ is true in } b_T^i] / \sum_{j=1}^{N} w_j$$

where w_i are weights associated with sampled belief states.

2.7 Prior Influence

A notable aspect of this formulation is that $P(\psi|a_{1:T})$ depends on the prior over belief states $P(b_0) = P(s_0)$. This highlights the importance of initial assumptions in Bayesian inference.

2.8 Experimental Priors

Two possibilities for priors are considered:

1. $U_{s_0}(s_0)$ is uniform over the set of possible initial states $s_0 \in S_0$.
2. $U_\psi(s_0)$ induces a uniform prior $P(\psi) = 0.5$ over the statement being true.

In the second case, $P(\psi|a_{1:T})$ becomes a kind of (normalized) likelihood $L(\psi|a_{1:T}) = \frac{P(a_{1:T}|\psi)}{P(a_{1:T}|\psi)+P(a_{1:T}|\neg\psi)}$, which rates how much more evidence there is for the statement ψ as opposed to its negation $\neg\psi$.

3 Joint Inference of Goals and Beliefs

3.1 Bayesian Inference Model

Following Baker et al. (2017), we model observers as performing joint inference over the agent's goal g and belief history $b_{0:T}$ given a series of observed actions $a_{1:T}$. The corresponding posterior factorizes as follows:

$$P(g, b_{0:T}|a_{1:T}) \propto \sum_{s_{0:T}} P(g)P(s_0)P(b_0|s_0) \prod_{t=1}^{T} P(a_t|b_{t-1}, g)P(s_t|s_{t-1}, a_t)P(b_t|s_t)$$

3.2 Perfect Knowledge Assumption

In our setting, we assume the agent has perfect knowledge of the environment, such that $b_t \equiv s_t$. This simplifies the inference to:

$$P(g, s_{0:T}|a_{1:T}) \propto P(g)P(s_0)\prod_{t=1}^{T} P(a_t|s_{t-1}, g)P(s_t|s_{t-1}, a_t)$$

3.3 Particle Filtering Algorithm

To compute this distribution, we initialize a set of weighted samples $\{(g^i, s_0^i, w^i)\}_{i=1}^N$ by enumerating over all possible combinations of goals and initial states. The weight w^i is initialized to $P(g^i)P(s_0^i)$.

We then sequentially update each sample i, multiplying w^i by the likelihood $P(a_t|s_{t-1}^i, g^i)$ of observing action a_t, and simulating the next state s_t that results from a_t:

$$w^i \leftarrow w^i \cdot P(a_t|s_{t-1}^i, g^i)$$
$$s_t \sim P(s_t|s_{t-1}^i, a_t)$$

Each weight w^i represents the unnormalized probability of the pair $(g^i, s_{0:t}^i)$ at step t. Normalizing the weights gives us the probability $P(g^i, s_{0:t}^i|a_{1:t}) = w^i / \sum_{j=1}^N w_j$.

3.4 Stochastic Environments

In environments where state transitions $P(s_t|s_{t-1}, a_t)$ are stochastic, this procedure corresponds to a particle filtering algorithm. This allows for efficient inference in complex, dynamic environments.

3.5 Large State Spaces

In settings where there are too many possibilities to enumerate over, we can sample a representative set of goals and states instead. This allows for approximate inference in high-dimensional state spaces.

4 Modeling Goal-Directed Actions

4.1 Boltzmann-Rational Model

We adopt a Boltzmann-rational model of action selection, given the agent's goal g and their belief b_{t-1}:

$$P(a_t|b_{t-1}, g) = \frac{\exp(-\beta \hat{Q}_g(b_{t-1}, a))}{\sum_{a'} \exp(-\beta \hat{Q}_g(b_{t-1}, a'))}$$

Here, $\hat{Q}_g(b_{t-1}, a)$ represents the estimated cost of the optimal plan to achieve goal g after taking action a starting at state s_{t-1}, where s_{t-1} is assumed equal to the agent's belief b_{t-1}.

4.2 Rationality Parameter

β is a "rationality" parameter, controlling how optimal the agent's actions are. A higher β corresponds to more optimal actions. This parameter allows for modeling agents with varying degrees of rationality or decision-making capability.

4.3 Efficient Computation of $\hat{Q}_g(b_{t-1}, a)$

To compute $\hat{Q}_g(b_{t-1}, a)$ efficiently, we leverage recent advances in sequential inverse planning. We use real-time heuristic search as an incremental planning algorithm that rapidly estimates the cost to the goal g from a state s, while reusing past planning results to accelerate planning.

This approach, unlike earlier BToM models which use offline value iteration algorithms, better captures and exploits the online nature of human planning. It allows for more realistic modeling of how humans make decisions in real-time, dynamic environments.

5 Representing Environment and Belief States

5.1 Planning Domain Definition Language (PDDL)

We use the Planning Domain Definition Language (PDDL) to represent both environment states s_t and belief states b_t. PDDL is a first-order language for model-based planning and reasoning.

5.2 PDDL Domain

In a PDDL domain:

- A set of predicates $\mathcal{P}$ describe a set of objects O.
- Each object has a type $\tau \in \mathcal{T}$.

For example, to represent that `key1` is red in color, we write:
`(iscolor key1 red)`
Where:

- `iscolor` $\in \mathcal{P}$ is a predicate
- `key1, red` $\in O$ are objects
- `key, color` $\in \mathcal{T}$ are types

5.3 Environment State Representation

Each environment state s is essentially a list of predicate terms, stating which relations are true or false for every valid combination of objects.

5.4 First-Order Logic Evaluation

Since PDDL is first-order, we can evaluate whether a sentence ϕ is true in a state s. ϕ is compositionally defined in terms of logical operations and (typed) quantifiers over predicates. Predicates can take either objects $o \in O$ or variables ?v as arguments x_i:

$$\phi ::= (P\ x_1\ ...\ x_n) \mid (\text{not}\ \phi) \mid (\text{and}\ \phi_1\ \phi_2) \mid (\text{or}\ \phi_1\ \phi_2) \mid$$
$$(\text{exists}\ (?v - \tau)\ \phi) \mid (\text{forall}\ (?v - \tau)\ \phi)$$

This expressivity allows us to determine the truth value of not just individual relations, but also general queries about whether some property holds for some class of objects.

6 Computational Model

6.1 Generative Model

To explain how human observers attribute goals and beliefs to other agents, we model them as performing approximately Bayesian inference over a generative model of an agent's goals, beliefs, and actions in an environment:

$$
\begin{aligned}
\text{Goal Prior:} \quad & g \sim P(g) \\
\text{State Prior:} \quad & s_0 \sim P(s_0) \\
\text{Belief Update:} \quad & b_t \sim P(b_t|s_t) \\
\text{Action Selection:} \quad & a_t \sim P(a_t|b_{t-1}, g) \\
\text{State Transition:} \quad & s_t \sim P(s_t|s_{t-1}, a_t)
\end{aligned}
$$

Where:

- g is the agent's goal
- s_t is the environment's state at timestep t
- b_t is the agent's belief about the environment at timestep t
- a_t is the agent's action at timestep t

6.2 Complete Knowledge and Veridical Perception

In our setting, we model the agent as having both complete knowledge and veridical perception of the environment, such that their belief at each step t is always equivalent to the environment state $b_t \equiv s_t$.

6.3 Observer Uncertainty

Despite the agent's perfect knowledge, an observer will still have uncertainty over what the agent believes, since they have uncertainty over the initial state s_0.

This model provides a comprehensive framework for understanding how observers infer the goals and beliefs of agents based on their actions, accounting for both the agent's decision-making process and the observer's inference process.

7 Conclusion

This document has provided a detailed explanation of the mathematical and technical components involved in Bayesian inference for goal and belief attribution. We've covered quantitative interpretation of belief statements, joint inference of goals and beliefs, modeling of goal-directed actions, representation of environment and belief states, and the overall computational model. These concepts form the foundation for understanding and modeling how humans attribute mental states to others based on observed behavior.

References

[1] Ying, L., Zhi-Xuan, T., Wong, L., Mansinghka, V., & Tenenbaum, J. (2024). *Grounding Language about Belief in a Bayesian Theory-of-Mind*. arXiv preprint arXiv:2402.10416. https://arxiv.org/abs/2402.10416