Probabilistic Multi-agent Only-believing

Qihui Feng, Gerhard Lakemeyer
RWTH Aachen University

May 13, 2024

Overview

Introduction

Knowledge and Belief:

- $\boldsymbol{K}\big(\textit{fair}(\text{Coin}) \wedge \neg \textit{fair}(\text{Die})\big)$
- $\boldsymbol{B}(\textit{fair}(\text{Coin})\colon 0.8)$

Levesque proposed **only-knowing** to precisely capture the (non-)beliefs:

- $\boldsymbol{O}(\textit{fair}(\text{Coin})) \models \neg \boldsymbol{K}(\textit{fair}(\text{Die})) \wedge \neg \boldsymbol{K}(\neg \textit{fair}(\text{Die}))$
- $\boldsymbol{O}(\textit{fair}(\text{Coin})) \models \neg \boldsymbol{B}(\textit{fair}(\text{Die})\colon r)$ for any $r \in [0, 1]$

Introduction (cont'd)

Research on only-knowing:

- Probabilistic only-believing: The logic $\mathcal{OBL}$
- Projection reasoning: $\boldsymbol{O}(KB_1) \rightarrow [action]\boldsymbol{O}(KB_2)$
- Multi-agent only-knowing:
  - Previous works in both propositional and first-order cases
  - No first-order account faithfully follows Levesque's principle of only-knowing.

Introduction (cont'd)

Levesque's notion of only-knowing: Given the only-knowing of the agent, any subjective formula will either be inferred or disproved.

- $\boldsymbol{O}(\mathrm{KB}) \models \boldsymbol{K}\beta$ iff $\mathrm{KB} \models \beta$; $\boldsymbol{O}(\mathrm{KB}) \models \neg\boldsymbol{K}\beta$ iff $\mathrm{KB} \not\models \beta$.

Desiderata for multi-agent extension:

i) non-beliefs on irrelevant items:
$\boldsymbol{O}_a(\neg \mathit{fair}(\mathrm{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\mathrm{Coin}))) \models \neg\boldsymbol{K}_a\boldsymbol{K}_b(\mathit{fair}(\mathrm{Die}))$

ii) non-beliefs on mental states with deeper nesting
$\boldsymbol{O}_a(\neg \mathit{fair}(\mathrm{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\mathrm{Coin}))) \models \neg\boldsymbol{K}_a\boldsymbol{K}_b\boldsymbol{K}_a(\mathit{fair}(\mathrm{Coin}))$

To model only-knowing up to all depths is semantically difficult.

Introduction (cont'd)

To model only-knowing up to depth k?

- Modality $\boldsymbol{O}_a^{(k)}$: agent a's only-knowing(believing) up to depth k.

The new desiderata: for $\boldsymbol{K}_a\beta$ with depth no more than k,

- Either $\boldsymbol{O}_a^{(k)}\alpha \models \boldsymbol{K}_a\beta$ or $\boldsymbol{O}_a^{(k)}\alpha \models \neg\boldsymbol{K}_a\beta$?

Overview

Logic $\mathcal{OBL}_m$

A first-order modal logic with equality and featured with:

- A finite set of agents, e.g. $Ag = \{a, b\}$;
- Modalities for belief and only-believing for each agent.

Syntax

- Standard FO formulae;
- $\boldsymbol{B}_i(\alpha : r)$: α is believed by agent i with degree r (written $\boldsymbol{K}_i\alpha$ if $r = 1$).
- $\boldsymbol{O}_i^{(k)}(\alpha : r)$: all that agent i believes up to depth k is α with degree r.
  - $\boldsymbol{O}_a^{(1)}(fair(\text{Coin}))$: agent a knows $fair(\text{Coin})$ and nothing else about the world.
  - $\boldsymbol{O}_a^{(2)}(fair(\text{Coin}))$: agent a knows $fair(\text{Coin})$ and nothing else about the world, and nothing about Bob's beliefs about the world.

Semantics (Knowledge and Beliefs)

A model is a tuple (w, e_a, e_b) with world w and epistemic states e_a and e_b.
A **world** $w \in \mathcal{W}$ is a set of ground atoms.

- $w, e_a, e_b \models \mathit{fair}(\text{Coin})$ iff $\mathit{fair}(\text{Coin}) \in w$

Epistemic states are defined inductively:

- 1-**distribution** assigns each world a probability: $\mathcal{W} \to \mathbb{R}_{[0,1]}$;
- 1-**epistemic state** is a set of 1-distributions.

Example

$$w_1 = \{\mathit{fair}(\text{Coin})\} \text{ and } w_2 = \{\mathit{fair}(\text{Die})\},\ d^1(w) = \begin{cases} 0.5 & w \in \{w_1, w_2\} \\ 0 & \mathit{otherwise}. \end{cases}$$

Let $e_a = \{d^1\}$, then $w, e_a, e_b \models \boldsymbol{B}_a(\mathit{fair}(\text{Coin}) : 0.5)$

Semantics (Nested Beliefs)

$\mathcal{E}^1$ denotes the set of all 1-epistemic states. For any $k > 1$,

- k**-distribution** $d^k : (\mathcal{W} \times \mathcal{E}^{k-1}) \to \mathbb{R}_{[0,1]}$
- a k**-epistemic state** is a set of k-distributions

Example

Let $w_1 = \{\mathit{fair}(\text{Coin})\}$, $w_2 = \{\mathit{fair}(\text{Die})\}$, $w_3 = \{\mathit{fair}(\text{Coin}), \mathit{fair}(\text{Die})\}$.

$$\tilde{d}^1(w) = \begin{cases} 0.5 & w \in \{w_1, w_3\} \\ 0 & \mathit{otherwise.} \end{cases} \qquad d^2(w, e_b^1) = \begin{cases} 0.3 & w = w_1, e_b^1 = \{\tilde{d}^1\} \\ 0.7 & w = w_2, e_b^1 = \{\tilde{d}^1\} \\ 0 & \mathit{otherwise.} \end{cases}$$

Let $e_a = \{d^2\}$, then

- $w, e_a, e_b \models \boldsymbol{B}_a(\mathit{fair}(\text{Coin}) : 0.3)$
- $w, e_a, e_b \models \boldsymbol{K}_a(\boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$

Semantics (Only-Believing)

Suppose that $e_a \in \mathcal{E}^2$ (2-epistemic state)

- $w, e_a, e_b \models \boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$ iff for any 2-distribution d,

$$d \in e_a \iff w, \{d\}, e_b \models \boldsymbol{K}_a(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$$

i.e. there is a **unique** $e_a \in \mathcal{E}^2$ which satisfies $\boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$
Every $e_a \in \mathcal{E}^k$ can be **uniquely** "reduced" to an $e'_a \in \mathcal{E}^{k-1}$ s.t.

$$w, e_a, e_b \models \alpha \text{ iff } w, e'_a, e_b \models \alpha \text{ for any } \alpha \text{ not deeper than k-1}$$

For $e_a \in \mathcal{E}^3$, $w, e_a, e_b \models \boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$ iff
e_a reduce to $e'_a \in \mathcal{E}^2$ and $w, e'_a, e_b \models \boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$

Entailment and Validity

Compatibility:

- Formulae like $\boldsymbol{O}_a^{(1)}(\neg \mathit{fair}(\text{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin})))$ are illegal.
- *e compatible* with α: the depth of *e* is not less than the depth of α

We say Σ entails α (written $\Sigma \models \alpha$) iff:

- For each model (w, e_a, e_b) compatible with Σ, α, if $(w, e_a, e_b) \models \sigma$ for all $\sigma \in \Sigma$, then $(w, e_a, e_b) \models \alpha$

We say α is valid iff $\{\} \models \alpha$

Overview

Properties of Knowledge

$\mathcal{OBL}_m$ follows the $\mathrm{KD45_n}$ properties. For agent $i \in Ag$,

- (Nec) If $\models \alpha$, then $\models \boldsymbol{K}_i\alpha$
- (K) $\models \boldsymbol{K}_i\alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta) \supset \boldsymbol{K}_i\beta$
- (D)$\models \boldsymbol{K}_i\alpha \supset \neg\boldsymbol{K}_i\neg\alpha$
- (4) $\models \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\boldsymbol{K}_i\alpha$
- (5) $\models \neg\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$
- $\boldsymbol{K}_i\alpha \wedge \neg\alpha$ is satisfiable

Barcan formulae:

- $\models \forall x.\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\forall x.\alpha$
- $\models \exists x.\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\exists x.\alpha$

Properties of Beliefs

The degree of belief follows the laws of probability:

- $\models \boldsymbol{B}_i(\alpha\colon r) \supset \neg \boldsymbol{B}_i(\alpha\colon r')$ for $r' \neq r$
- $\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\neg\alpha\colon 1 - r)$
- $\models \boldsymbol{B}_i(\alpha \wedge \beta\colon r) \wedge \boldsymbol{B}_i(\alpha \wedge \neg\beta\colon r') \supset \boldsymbol{B}_i(\alpha\colon r + r')$

Only-Believing

Modality $\boldsymbol{O}_a^{(k)}$ precisely captures agent *a*'s beliefs and non-beliefs up to depth *k*.

- non-beliefs on irrelevant items:
  $\boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\mathrm{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\mathrm{Coin}))) \models \neg \boldsymbol{K}_a \boldsymbol{K}_b(\mathit{fair}(\mathrm{Die}))$
- non-beliefs on deeper mental states:
  $\boldsymbol{O}_a^{(2)}(\neg \mathit{fair}(\mathrm{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\mathrm{Coin}))) \not\models \neg \boldsymbol{K}_a \boldsymbol{K}_b \boldsymbol{K}_a(\neg \mathit{fair}(\mathrm{Coin}))$
  $\boldsymbol{O}_a^{(3)}(\neg \mathit{fair}(\mathrm{Coin}) \wedge \boldsymbol{K}_b(\mathit{fair}(\mathrm{Coin}))) \models \neg \boldsymbol{K}_a \boldsymbol{K}_b \boldsymbol{K}_a(\neg \mathit{fair}(\mathrm{Coin}))$

For $i \in Ag$, given *i*-objective formulae α and β s.t. the depth of $\boldsymbol{K}_i(\beta)$ not greater than *k*, then $\boldsymbol{O}_i^{(k)}(\alpha)$ entails either $\boldsymbol{K}_i\beta$ or $\neg\boldsymbol{K}_i\beta$.

- $\boldsymbol{O}_i^{(k)}(\alpha) \models \boldsymbol{K}_i\beta$ iff $\alpha \models \beta$; $\boldsymbol{O}_i^{(k)}(\alpha) \models \neg\boldsymbol{K}_i\beta$ iff $\alpha \not\models \beta$

Autoepistemic Reasoning

$\mathcal{OBL}_m$ can represent defaults about another agent's beliefs:

Example

Let KB $= \{\neg \mathit{fair}(\text{Coin})\}$,
$\delta = \forall r.(r \neq 0 \supset \neg \boldsymbol{B}_a(\neg \boldsymbol{K}_b(\mathit{fair}(\text{Coin})) : r)) \supset \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))$
Bob believes *fair*(Coin) unless otherwise (Bob does not believes *fair*(Coin)) is believed (by Alice) with a non-zero degree

- $\boldsymbol{O}_a^{(2)}(\text{KB} \wedge \delta) \models \boldsymbol{K}_a \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))$
- $\boldsymbol{O}_a^{(2)}(\text{KB} \wedge \delta \wedge \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))) \models \boldsymbol{K}_a \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))$
- $\boldsymbol{O}_a^{(2)}(\text{KB} \wedge \delta \wedge \neg \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))) \models \neg \boldsymbol{K}_a \boldsymbol{K}_b(\mathit{fair}(\text{Coin}))$

Conclusion

In this work, we

- propose an logical account for multi-agent only-believing
- prove properties on beliefs and only-believing
- explore the capability of default reasoning about nested beliefs

For future work:

- extend to belief after actions ✓
- develop mechanisms for projection reasoning
- join common beliefs and only-believing

Thank you!