Truth-Tracking with Non-Expert Information Sources

Joe Singleton and Richard Booth

singletonj1@cardiff.ac.uk

Overview

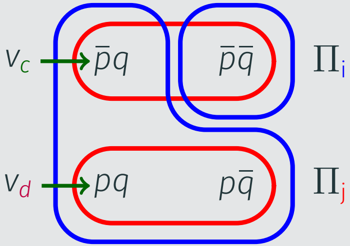
- **Problem:** what can be learned from non-expert information sources?
- We aim to learn both:
  - the **true facts** of the world
  - the **true level of expertise** of the sources
- We adapt the learning framework from recent work combining formal learning theory, belief revision and epistemic logic
- Main results:
  - description of **what can be learned**
  - characterisation of **truth-tracking learning methods**

Logical framework for expertise

Basic framework

- $\mathcal{P}$: finite set of propositional variables (e.g. p, q, ...)
- $\mathcal{S}$: finite set of sources (e.g. i, j, ...)
- $\mathcal{C}$: finite set of cases (e.g. c, d, ...)
- We model expertise with **partitions** of states
  - Sources cannot distinguish states in the same cell
- A **world** is a pair $W = (\{\Pi_i\}_{i \in \mathcal{S}}, \{v_c\}_{c \in \mathcal{C}})$, where
  - Each Π_i is a partition of the set of propositional valuations
  - Each v_c is a valuation

Example



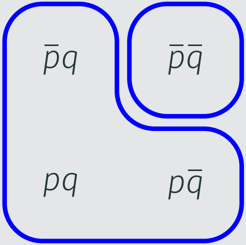
Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states

Example



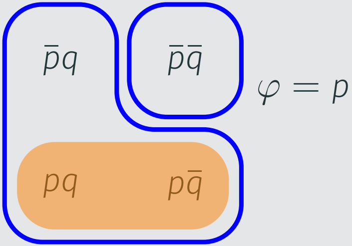
Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states

Example



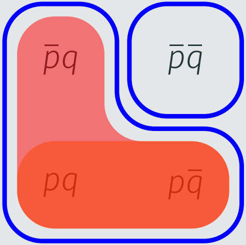
Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states

Example



$\varphi = p$

$W \not\models \mathsf{E}_i p$ ✗

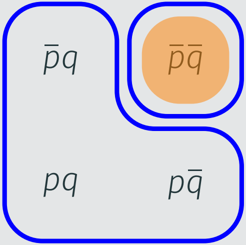
Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states

Example



$$\psi = \neg p \land \neg q$$

$$W \models \mathsf{E}_i\psi$$ ✅

Expertise and soundness

- i has expertise on φ if i can always determine the correct value of φ:

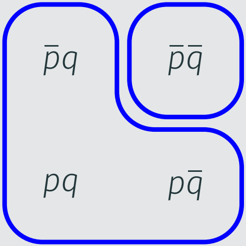
$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states
- φ is sound for i if φ is true *up to lack of expertise* of i

$$W, c \models \mathsf{S}_i\varphi \iff \Pi_i[v_c] \cap \mathrm{mods}(\varphi) \neq \emptyset$$

  - true state indistinguishable from some φ state

Example



Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

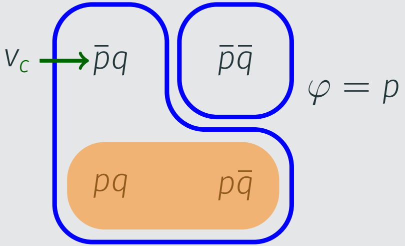
$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

  - φ states always distinguishable from $\neg\varphi$ states
- φ is **sound** for i if φ is true *up to lack of expertise* of i

$$W, c \models \mathsf{S}_i\varphi \iff \Pi_i[v_c] \cap \mathrm{mods}(\varphi) \neq \emptyset$$

  - true state indistinguishable from some φ state

Example



Expertise and soundness

- i **has expertise** on φ if i can always determine the correct value of φ:

$$W \models \mathsf{E}_i\varphi \iff (u \in \mathrm{mods}(\varphi) \implies \Pi_i[u] \subseteq \mathrm{mods}(\varphi))$$

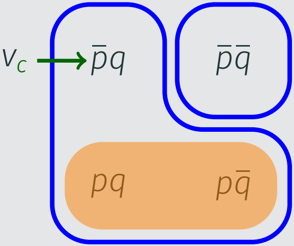
  - φ states always distinguishable from $\neg\varphi$ states

- φ is **sound** for i if φ is true *up to lack of expertise* of i

$$W, c \models \mathsf{S}_i\varphi \iff \Pi_i[v_c] \cap \mathrm{mods}(\varphi) \neq \emptyset$$

  - true state indistinguishable from some φ state

Example



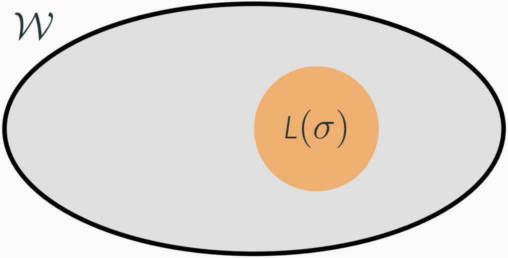
$\varphi = p$

$W, c \models \neg p \wedge \mathsf{S}_i p$ ✅

Learning and truth-tracking

Reports and methods

- We receive **reports** of the form $\langle i, c, \varphi \rangle$
  - "source i reports φ in case c"
- A **learning method** L maps a finite sequence σ to a **conjecture** $L(\sigma) \subseteq \mathcal{W}$, where $\mathcal{W}$ is the set of all worlds



Example

$$L(\sigma) = \{W \mid \forall \langle i, c, \varphi \rangle \in \sigma, W, c \models S_i \varphi\}$$

Streams

- We assume sources report all they consider possible
  - All reports are **sound**: only false due to lack of expertise
  - All sound reports **eventually appear**

Streams

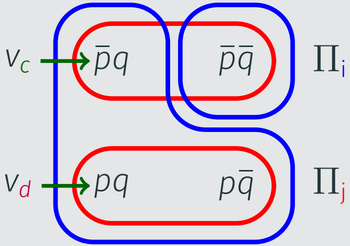
- We assume sources report all they consider possible
    - All reports are **sound**: only false due to lack of expertise
    - All sound reports **eventually appear**
- **Warning:** Strong assumptions! Sources are always honest, and do not distinguish soundness with **beliefs** or **knowledge** ⚠️

Streams (cont'd)

- An infinite sequence of reports ρ is a **stream** for a world W if

$$\langle i, c, \varphi\rangle \in \rho \iff W, c \models \mathsf{S}_i\varphi$$

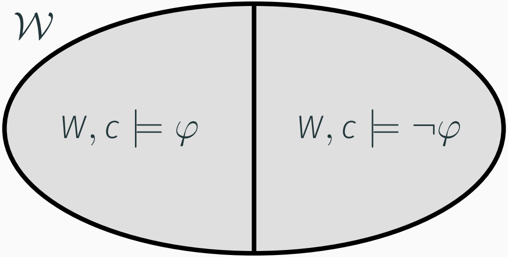
Example



$$\rho = (\langle \mathsf{i}, c, p \lor q\rangle, \langle \mathsf{i}, c, p\rangle, \langle \mathsf{i}, c, q\rangle,$$
$$\langle \mathsf{i}, c, \neg p\rangle, \langle \mathsf{i}, c, \neg q\rangle, \langle \mathsf{i}, c, p \land q\rangle, \ldots)$$

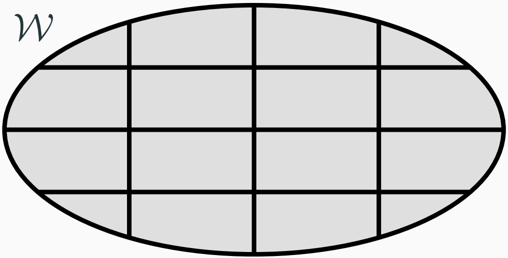
Questions

- We want to design methods L which learn W when fed a stream ρ
- Finding W exactly is too much to ask
- A **question** Q is a partition of $\mathcal{W}$
    - $Q_{\varphi,c}$: does φ hold in case c?



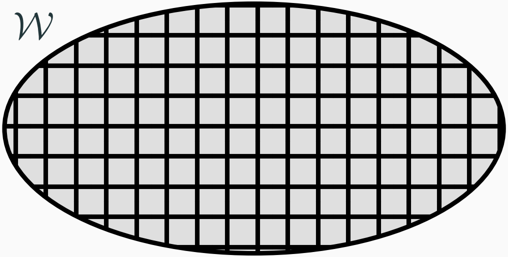
Questions

- We want to design methods L which learn W when fed a stream ρ
- Finding W exactly is too much to ask
- A **question** Q is a partition of $\mathcal{W}$
    - $Q_{\varphi,c}$: does φ hold in case c?
    - Q_{val}: what are the correct valuations? (ignoring partitions)



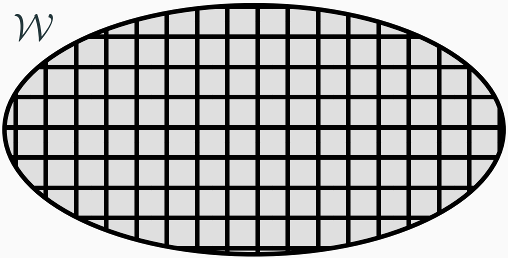
Questions

- We want to design methods L which learn W when fed a stream ρ
- Finding W exactly is too much to ask
- A **question** Q is a partition of $\mathcal{W}$
  - $Q_{\varphi,c}$: does φ hold in case c?
  - Q_{val}: what are the correct valuations? (ignoring partitions)
  - $Q_{\perp} = \{\{W\} \mid W \in \mathcal{W}\}$: what is the actual world?



Questions

- We want to design methods L which learn W when fed a stream ρ
- Finding W exactly is too much to ask
- A **question** Q is a partition of $\mathcal{W}$
    - $Q_{\varphi,c}$: does φ hold in case c?
    - Q_{val}: what are the correct valuations? (ignoring partitions)
    - $Q_{\perp} = \{\{W\} \mid W \in \mathcal{W}\}$: what is the actual world?
    - $Q[W]$ is the **correct answer** at W



Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

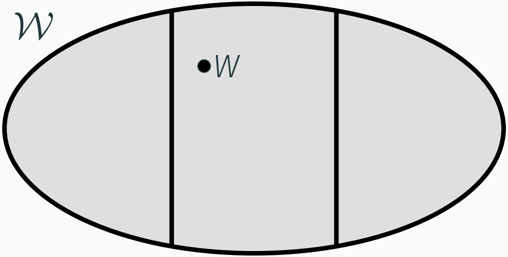
- Q is **solvable** if there is a consistent method L which solves Q

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q

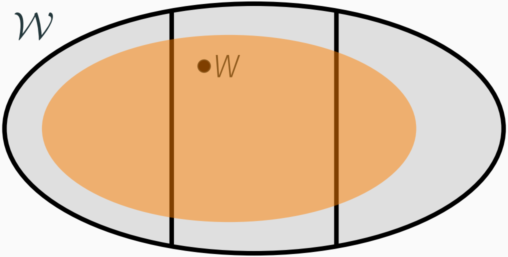


Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



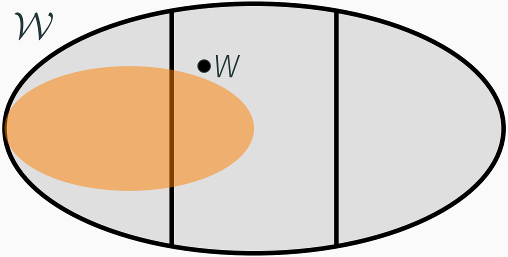
$L(\rho_1)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



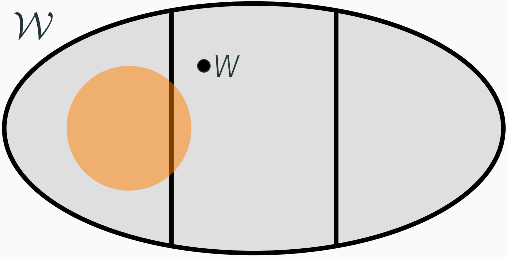
$L(\rho_1, \rho_2)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



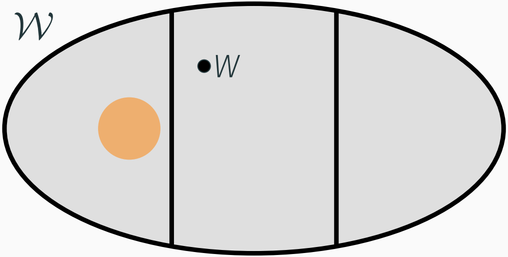
$L(\rho_1, \rho_2, \rho_3)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



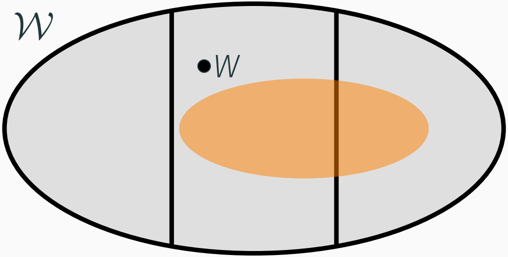
$L(\rho_1, \rho_2, \rho_3, \rho_4)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



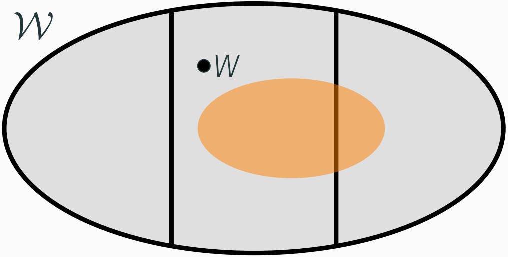
$L(\rho_1 \cdots \rho_5)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



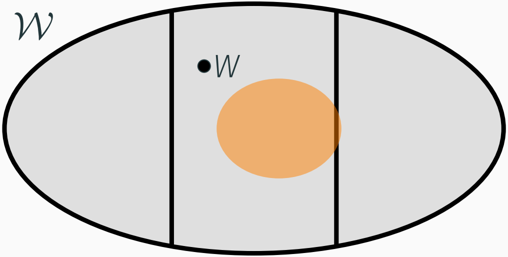
$L(\rho_1 \cdots \rho_6)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



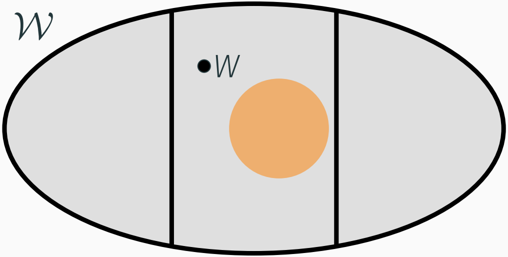
$L(\rho_1 \cdots \rho_7)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



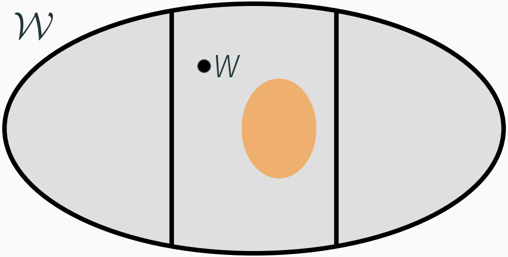
$L(\rho_1 \cdots \rho_8)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



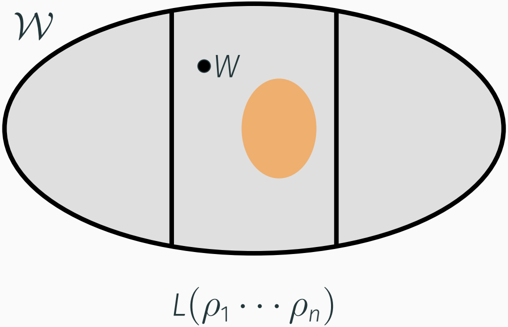
$L(\rho_1 \cdots \rho_9)$

Solvability

- L **solves** Q if given any stream, L eventually finds the correct answer

$$\forall W, \forall \rho \text{ a stream for W}, \exists n \text{ s.t } \forall m \geq n, L(\rho_1 \cdots \rho_m) \subseteq Q[W]$$

- Q is **solvable** if there is a consistent method L which solves Q



What can be learned?

Solvable questions

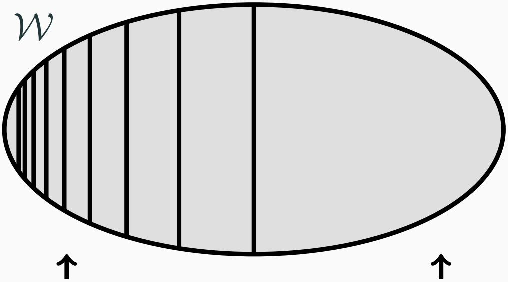
- Which questions are solvable?
- It turns out there is a question Q^* which is the unique **hardest solvable question**

$$W \sim^* W' \iff \forall i \in \mathcal{S}, c \in \mathcal{C}, \Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$$

- Equivalently, W and W' have **exactly the same streams**
- $Q_{\varphi,c}$ is only solvable when φ is a tautology or contradiction ❌
- Q_{val}, $Q_\perp$ not solvable ❌
- **Problem:** if source have no expertise at all, *all* reports are sound. True valuations don't matter!

Solvable questions (cont'd)

- **Solution:** investigate what $Q^*[W]$ tells us about W



Q^* very informative $\qquad$ Q^* not very informative

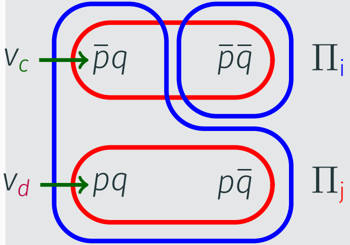
- A property of W is **learnable** if all $W' \in Q^*[W]$ share the same property
  - Any method solving Q^* eventually finds it

Theorem

The true c-valuation is learnable at W iff there is a set Γ st

1. $W, c \models \Gamma$
2. $\mathrm{Cn}(\Gamma)$ *is a maximally consistent set*
3. *For all $\varphi \in \Gamma$, there is $i \in \mathcal{S}$ such that $W \models \mathsf{E}_i \varphi$*

Example



For c, take $\Gamma = \{p \vee q, \neg p\}$:

$W, c \models \Gamma, \quad \mathrm{Cn}(\Gamma) = \mathrm{Cn}(\neg p \wedge q)$

$W \models \mathsf{E}_i(p \vee q), \quad W \models \mathsf{E}_j \neg p,$

✅

For d there is no such Γ!

❌

- Similar result for partitions (omitted)

Truth-tracking methods

A characterisation of truth-tracking

- We have so far only looked solvable questions
- Which methods actually solve them?
- L is truth-tracking if it solves all solvable questions
  - Equivalently, L solves Q^*
- We characterise truth-tracking axiomatically, given three basic properties:
  - **Equivalence:** if $\sigma \equiv \delta$ then $L(\sigma) = L(\delta)$
  - **Repetition:** $L(\sigma \cdots \sigma) = L(\sigma)$
  - **Soundness:** if $W \in L(\sigma)$ then $W, c \models \mathsf{S}_i\varphi$ for all $\langle i, c, \varphi \rangle \in \sigma$

A characterisation of truth-tracking (cont'd)

- Let T_σ be the set of worlds W such that, for all $\langle i, c, \varphi \rangle$:

$$W, c \models \mathsf{S}_i\varphi \iff \exists \psi \equiv \varphi \text{ s.t } \langle i, c, \psi \rangle \in \sigma$$

  - i.e. σ contains all sound reports, up to logical equivalence
- Write $U, c \models \varphi$ iff $W, c \models \varphi$ for all $W \in U$
- **Credulity:** if $T_\sigma, c \not\models \mathsf{S}_i\varphi$ then $L(\sigma), c \models \neg \mathsf{S}_i\varphi$

Theorem

For a method L satisfying ***Equivalence****,* ***Repetition*** *and* ***Soundness****,*

$$\textit{Truth-tracking} \iff \textbf{\textit{Credulity}}$$

Credulity

- **Credulity:** if $T_\sigma, c \not\models \mathsf{S}_i\varphi$ then $L(\sigma), c \models \neg\mathsf{S}_i\varphi$
- More expertise means **fewer sound reports**
- **Credulity** is a principle of **maximal trust**
  - Whenever consistent with T_σ, we should trust i to have expertise to rule out φ
  - Since all sound reports eventually received, mistaken trust can be retracted
- **Consequence:** truth-tracking is not possible deductively; **defeasible reasoning** is required
- Stronger property in terms of expertise directly: if $T_\sigma \not\models \neg\mathsf{E}_i\varphi$ then $L(\sigma) \models \mathsf{E}_i\varphi$

Conclusion

Summary and future work

- Summary:
  - Developed a logical framework to reason about expertise and sound reports
  - Expressed a learning problem in this setting
  - Characterised conditions under which information can be learned
  - Axiomatically characterised truth-tracking learning methods
- Future work:
  - Assumptions on streams are very strong! Can these be lifted?
  - Everything is finite. What results carry over to the infinite case?
  - Bridge with probabilistic reasoning?

Appendix: An example method

- **Intuition**: express credulity with a prior plausibility ordering over worlds
- Conjecture the maximally plausible worlds consistent with soundness statements
- E.g. using the number of partition cells as a measure of expertise:

$$L(\sigma) = \operatorname*{argmax}_{W \in X_\sigma} \sum_{i \in \mathcal{S}} |\Pi_i^W|$$

where $X_\sigma = \{W \mid \forall \langle i, c, \varphi \rangle \in \sigma, W, c \models \mathsf{S}_i \varphi\}$

- This method *is* truth-tracking!