Complexity of the interpretability logic IL

Luka Mikec\textsuperscript{1}  Fedor Pakhomov\textsuperscript{2}  Mladen Vuković\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, Faculty of Science, University of Zagreb, Croatia

\textsuperscript{2}Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia; Research University Higher School of Economics, Moscow, Russia

Dubrovnik, 2018
Interpretability

- Let $T_1$ and $T_2$ be some first order theories given through their sets of axioms (*not* sets of theorems).
- Roughly, translation of $T_2$ in $T_1$ is a pair $(f, U)$ where:
  - $f(R(\vec{x})) = f(R)(\vec{x})$;
  - $f(A \rightarrow B) = f(A) \rightarrow f(B)$ etc.;
  - $f(\forall x F) = \forall x (U(x) \rightarrow f(F))$ etc.;
- Thus, translation mostly preserves structure.
- $T_1$ interprets $T_2$ ($T_1 \triangleright T_2$) if
  \[ T_2 \vdash F \Rightarrow T_1 \vdash f(F), \]
  for all sentences $F \in \mathcal{L}(T_2)$. 
Interpretability

▶ In particular, we can study interpretability between finite extensions of a given theory.
▶ For example, since $PA \nvdash \Diamond \top$, we have:

$$PA + \Diamond_{PA} \top \succ PA,$$

where $\Diamond_{PA} \top$ formalizes consistency of $PA$ within $PA$.
▶ Furthermore, we can ask which interpretabilities can be proven within the base theory.

$$T + A \succ T + B \implies T \vdash \text{Int}(\neg A \neg, \neg B \neg)$$
Interpretability logics

▶ The language of interpretability logics is given by

\[ A ::= p | \bot | A \rightarrow A | \Box A | A \triangleright A, \]

where \( p \) is a propositional variable.

▶ Let \( T \) be a formal theory, and \( \text{Int}(\neg A \neg, \neg B \neg) \) a sentence formalizing \( T + A \triangleright T + B \).

▶ Arithmetical interpretation \( \ast \) assigns sentences to modal formulas, such that:
  ▶ \((A \rightarrow B)^{\ast} = A^{\ast} \rightarrow B^{\ast}\) etc.;
  ▶ \(p^{\ast}\) is a sentence (fixed by \( \ast \));
  ▶ \((\Box A)^{\ast} = \text{Pr}_T(A^{\ast})\);
  ▶ \((A \triangleright B)^{\ast} = \text{Int}_T(A^{\ast}, B^{\ast})\).
Interpretability logics

- Given a formal theory $T$,

$$A \in IL(T) :\iff T \vdash A^*.$$

- Research focuses on sufficiently strong theories, able to deal with syntax ("sequential theories").

- Interpretability logics of sequential theories contain the basic interpretability logic IL.
Basic interpretability logic **IL**

- **Basic interpretability logic IL:**
  
  propositionally valid formulas (in the new language);
  
  **K** \( \square(A \rightarrow B) \rightarrow (\square A \rightarrow \square B) \);
  
  **L"ob** \( \square(\square A \rightarrow A) \rightarrow \square A \);
  
  **J1** \( \square(A \rightarrow B) \rightarrow A \triangleright B \);
  
  **J2** \( (A \triangleright B) \land (B \triangleright C) \rightarrow A \triangleright C \);
  
  **J3** \( (A \triangleright C) \land (B \triangleright C) \rightarrow A \lor B \triangleright C \);
  
  **J4** \( A \triangleright B \rightarrow (\lozenge A \rightarrow \lozenge B) \);
  
  **J5** \( \lozenge A \triangleright A \).

  - rules: modus ponens and necessitation \( A/\square A \).

  (parentheses priority: \( \neg, \square, \lozenge; \land, \lor; \triangleright; \rightarrow, \leftrightarrow \))

- \( \square A \) is provably equivalent with \( \neg A \triangleright \bot \), and \( \lozenge A \) is defined as \( \neg \square \neg A \).

- There is no \( T \) such that \( IL(T) = IL \). In fact, we always have \( IL \subset ILW \subset IL(T) \). But, \( IL \) has nice semantics.
Models

- Semantics: extend the usual relational (Kripke) model.
- **IL-frame (Veltman frame):** \( \mathcal{F} = \langle W, R, \{S_w : w \in W\} \rangle \), where:
  1. \( W \neq \emptyset \);
  2. \( R^{-1} \) is well-founded (no \( x_0Rx_1Rx_2R \ldots \) chains);
  3. \( R \) is transitive;
  4. \( S_w \subseteq R(w)^2 \) is reflexive, transitive, contains \( R \cap R(w)^2 \)
     
   \[ wRuRv \text{ implies } uS_wv \];
- **IL-model (Veltman model):** \( \mathcal{M} = \langle W, R, \{S_w : w \in W\}, V \rangle \), where:
  1. \( \langle W, R, \{S_w : w \in W\} \rangle \) is a **IL-frame**;
  2. \( V \subseteq W \times \text{Prop} \) (or \( V : \text{Prop} \rightarrow \mathcal{P}(W) \)).
Models

- Veltman model: $\mathcal{M} = \langle W, R, \{S_w : w \in W\}, V \rangle$.
- $w \Vdash p$ if and only if $wVp$, for $p \in Prop$.
- Logical connectives have classical semantics.
- Truth of a formula $F \triangleright G$ ("$F$ interprets $G$") in a world $w \in \mathcal{M}$:
  \[ w \Vdash F \triangleright G \iff \forall x \in R(w) : x \Vdash F \Rightarrow \exists y \in S_w(x) : y \Vdash G. \]
- Modal soundness and completeness:
  \[ \text{IL} \vdash F \iff \forall F : F \not\vdash F. \]
Extensions and frame conditions

- Some extensions of **IL**:
  - **ILP** \( \text{IL} + A \triangleright B \rightarrow \Box (A \triangleright B) \)
  - **ILM** \( \text{IL} + A \triangleright B \rightarrow A \land \Box C \triangleright B \land \Box C \)
  - **ILM}_0\) \( \text{IL} + A \triangleright B \rightarrow \Diamond A \land \Box C \triangleright B \land \Box C \)
  - **ILW** \( \text{IL} + A \triangleright B \rightarrow A \triangleright B \land \Box \neg A \)
  - **ILW}^*\) \( \text{IL} + A \triangleright B \rightarrow B \land \Box C \triangleright B \land \Box C \land \Box \neg A \)

- These logics are complete w.r.t. certain classes of frames:
  - \( (P) \) \( wRuS_x v \Rightarrow wRv \);
  - \( (M) \) \( wRuS_w v \Rightarrow R(v) \subseteq R(u) \);
  - \( (M)_0 \) \( wRuR_xS_w v \Rightarrow R(v) \subseteq R(u) \);
  - \( (W) \) \( S_w \circ R \) is reverse well-founded for each \( w \);
  - \( (W}^*\) \( (M)_0 \) and \( (W) \).

- **ILW-frame** is **IL**-frame that satisfies \( (W) \) etc.

- Current best guess for **IL(All)** is (a possibly modally incomplete logic) \( \text{ILW} + (R_n)_n + (R^n)_n \). (Joost Joosten)
Complexity

- **IL** conservatively extends **GL** (“provability logic”); **GL** is in PSPACE.
- Closed fragment of **IL** is PSPACE-hard (Bou, Joosten).
- FMP for **IL**: if \( x \vdash F \), then there is finite \( M \) and \( x' \in M \) s.t. \( x' \vdash F \).
- Standard approach: to check if \( \vdash F \), we can (soundness, completeness, FMP) check if there is a finite model of \( \neg F \).
- So, to prove **IL** ∈ PSPACE, it suffices to construct a PSPACE algorithm for checking satisfiability.
Complexity (satisfiability)

- Let $\Gamma$ be an adequate set for $A \in \mathcal{L}$: set of subformulas, closed under certain operations (in fact, we use four different adequate set).
- $|\Gamma|$ is polynomial in $|A|$.
- Our algorithm builds models world-by-world (nondeterministically or with backtracking).
- There are functions named (1), (2) and (3).
- (1) only calls (2), which only calls (3), which only calls (1).
Function (1)

- (1) takes $\Delta \subseteq \Gamma$ and checks whether there is a rooted Veltman model of $\Delta$ ($W = \{w\} \cup R(w), w \models \Delta$).
- The starting call will be with $\Delta = \{A\}$.
- (1) looks at all the maximal Boolean consistent $\Delta' \supseteq \Delta$, and returns a positive result if at least one extension is satisfiable.
- Lemma: (1) returns a positive result if and only if $\Delta$ is satisfiable.
Function (2)

- (2) takes a maximal Boolean consistent $\Delta \subseteq \Gamma$ and checks whether there is a rooted Veltman model of $\Delta$.

$$\Delta^+ := \{E \triangleright G \in \Gamma : E \triangleright G \in \Delta\}$$
$$\Delta^- := \{E \triangleright G \in \Gamma : \neg (E \triangleright G) \in \Delta\}$$

- (2) returns a positive answer if the sets $\{\neg (C \triangleright D)\} \cup \Delta^+$ are satisfiable for all $\neg (C \triangleright D) \in \Delta^-$.  

- Lemma: (2) returns a positive result if and only if $\Delta$ is satisfiable. (Proof by merging roots)
Function (3)

- (3) takes a Boolean consistent $\Delta \subseteq \Gamma$ consisting of one negated $\triangleright$-formula $\neg(C \triangleright D)$ and a set of positive $\triangleright$-formulas $\Delta^+$, and checks whether there is a rooted Veltman model of $\Delta$.

- We say that $(N, P)$ is a $(-C \triangleright D, \Delta)$-pair if:
  1. $N, P \subseteq \Gamma$;
  2. $D \in N$;
  3. $\bot \notin P$;
  4. $E \triangleright G \in \Delta^+ \Rightarrow E \in N$ or $G \in P$.

- (3) returns a positive answer if there is a $(-C \triangleright D, \Delta)$-pair $(N, P)$ such that the following holds:
  1. $\{\neg B, B \triangleright \bot \mid B \in N\} \cup \{C, C \triangleright \bot\}$ is satisfiable;
  2. $\{\neg B, B \triangleright \bot \mid B \in N\} \cup \{G, \theta \triangleright \bot\}$ is satisfiable for all $G$ in $P$.

- Lemma: (3) returns a positive result if and only if $\Delta$ is satisfiable. (Proof by joining the models, adding a new root $w$, and adding the $S_w$ where needed – or even make it total).
Wrapping up

- Note that (1) can be calculated in terms of (2) etc.
- Each (1)-(2)-(3) chain adds a new □¬B formula for some \( B \in \Gamma \): so the procedure terminates.
- Algorithm works locally correct: each function does what it is supposed to do assuming the next one does. Correctness follows by induction (starting with leaf nodes in the execution tree).
- \( \text{IL} \) was known to be \( \text{PSPACE} \)-hard (conservatively extends \( \text{GL} \); also \( \text{IL}_0 \)). Thus, \( \text{IL} \) is \( \text{PSPACE} \)-complete.
I believe to have shown that $\text{IL}_W$, $\text{IL}_M$ and $\text{IL}_P$ are also PSPACE-complete.

After checking/finalizing those proofs, the next natural step would be to prove complexity results regarding $\text{IL}_M^0$ and $\text{IL}_W^*$. These logics were only recently proven to have FMP (Luka Mikec, Tin Perkov, Mladen Vuković), but with respect to a much more complex semantics.
L. Mikec, F. Pakhomov, M. Vuković. Complexity of interpretability logics IL. Logic Journal of the IGPL, 2018.

L. Mikec, T. Perkov, M. Vuković. Decidability of interpretability logics $\text{ILM}_0$ and $\text{ILW}^*$. Logic Journal of the IGPL, Volume 25, Issue 5, 1 October 2017, Pages 758–772,