TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Faculty of Computer Science Chair of Automata Theory

INTRODUCTION TO NONMONOTONIC REASONING

Anni-Yasmin Turhan

Dresden, WS 2019/20

Section 4

Autoepistemic Logic

Subsection 4.1

Introducing autoepistemic logic

Introducing autoepistemic logic[1]: an example

Idea: formalism to model how an agent forms her own belief sets and how to reason about it.

Example:
Are the Stones playing in Newcastle next week?
—No, because otherwise I would have heard about it.

Observations:

- no definite knowledge that the Stones do not give a concert in Newcastle next week.
- incomplete knowledge and negative answer is rather a conjecture

New knowledge: the Stones are giving a concert in Newcastle next week!

Observations:

- old conclusion by introspection is no longer valid and must be revised—nonmonotonic reasoning!
- long-term knowledge ("If something important is to happen in my city, then I would know about it") has not changed.
- what has changed is that answer is based on fact, not on introspection

[1] autoepistemic: reflection upon self-knowledge

Towards autoepistemic logic

Indicate "believed knowledge" by a modal operator L applied to FOL sentences.
$L\,\varphi$ means intuitively: "I know φ".

Capture the following information:

- Prof Jones is a university professor and thus normally teaches
- If I do not believe that Dr. Jones does not teach, then Dr. Jones does teach

by the modal formula:

$$L\,prof_J \wedge \neg L\,\neg teaches_J \longrightarrow teaches_J$$

The concert example can be captured by:

- $concert \longrightarrow L\,concert$ ("If a concert takes place, then I know about it.")
- $\neg L\,concert$ ("I don't know that a concert takes place.")

Towards syntax and semantics

The L -operator can appear nested in formulae:

$$L\, L\, \varphi \quad \text{or} \quad L\, \neg L\, q \quad \text{or} \quad \neg L\, (p \vee L\, q).$$

Intuitively, the meaning of autoepistemic logic is given in terms of expansions, i.e., collections of pieces of knowledge defining "world views" compatible with and based on the given knowledge.

Expansions are stable,

- if fact φ is in an expansion, then so is $L\, \varphi$
- if fact φ is not in an expansion, then $\neg L\, \varphi$ is in the expansion

Definition 4.1 (Autoepistemic formulae, AE-formula)

Autoepistemic formulae (AE-formulae) are the smallest set satisfying the following:

- each closed FOL formula is an AE-formula
- if φ is an AE-formula, then $L\ \varphi$ is an AE-formula
- if φ and ψ are AE-formulae, then so are the following:
  - $\neg\varphi$
  - $(\varphi \wedge \psi)$
  - $(\varphi \vee \psi)$
  - $(\varphi \longrightarrow \psi)$

The set of all AE-formulae is denoted by *For*.

An autoepistemic theory (AE-theory) is a set of AE-formulae.

Syntax of autoepistemic logic—schema

Sometimes it is convenient to use open FOL formulae in the scope of the L -operator. In such cases the AE-formula reads as a schema, i.e., a collection of ground instances.

E.g.:

$$german(X) \wedge \neg L\, \neg drinksBeer(X) \longrightarrow drinksBeer(X),$$

$$german(bob),\ german(lisa)$$

is read as the autoepistemic theory:

$$german(bob) \wedge \neg L\, \neg drinksBeer(bob) \longrightarrow drinksBeer(bob),$$

$$german(lisa) \wedge \neg L\, \neg drinksBeer(lisa) \longrightarrow drinksBeer(lisa),$$

$$german(bob),\ german(lisa)$$

Sub-formula

Let φ be an AE-formula. The set of subformulae of φ ($sub(\varphi)$) is defined as:

- $sub(\varphi) = \emptyset$ for FOL formula φ
- $sub(\neg\varphi) = sub(\varphi)$
- $sub(\varphi \vee \psi) = sub(\varphi \wedge \psi) = sub(\varphi \longrightarrow \psi) = sub(\varphi) \cup sub(\psi)$
- $sub(L\ \varphi) = \{\varphi\}$

Let T be an AE-theory. The set of subformulae of T is defined as

$$sub(T) = \bigcup_{\varphi \in T} sub(\varphi).$$

Note: $sub()$ does not work recursively; it does not go further into the structure of a formula, after the outmost occurrence of L.

For example:
If $T = \{L\ \neg L\ q, L\ (L\ p \wedge r), \neg L\ r, s\}$, then $sub(T) = \{\neg L\ q, (L\ p \wedge r), r\}$

Degree

The degree of an AE-formula φ ($degree(\varphi)$) is the maximal depth of L -nestings that occurs in φ.

Let T be an AE-theory, then T_n denotes the set of AE-formulae in T with degree less or equal n.

For example: $degree((\neg L\ \neg L\ (p \wedge L\ q))) = 3$.

Kernel

The kernel of an AE-theory T is defined as the set of all FOL formulae that are elements of T (denoted T_0).

For example: if $T = \{p, \neg L\ q, \neg L\ q \longrightarrow s, L\ \neg L\ r, r\}$, then $T_0 = \{p, r\}$.

Definition 4.2 (Normal form)

An AE-formula is in normal form, if it has the form

$$\varphi_1 \wedge \varphi_2 \wedge \cdots \wedge \varphi_n,$$

where each conjunct φ_i ($1 \leq i \leq n$) has the form

$$\beta \vee L\,\gamma_1 \vee \cdots \vee L\,\gamma_{m_i} \vee \neg L\,\delta_1 \vee \cdots \vee \neg L\,\delta_{k_i}$$

with a FOL formula β and AE formulae $\gamma_1, \ldots \gamma_{m_i}, \delta_1, \ldots, \delta_{k_i}$.

Each AE-formula φ can be transformed into an equivalent AE-formula ($nf(\varphi)$) in normal form, such that $degree(\varphi) = degree(nf(\varphi))$.
(Proof: exercise)

Definition 4.3 (AE-interpretation)

An autoepistemic interpretation $\mathcal{I}$ over a signature Σ provides

- a non-empty domain $dom(\mathcal{I})$
- an interpretation $f^{\mathcal{I}}$ for each function symbol $f \in \Sigma$ (as in FOL)
- an interpretation $r^{\mathcal{I}}$ for each predicate symbol $r \in \Sigma$ (as in FOL)
- a truth value $(L\ \varphi)^{\mathcal{I}}$ for every AE-formula $L\ \varphi$.

As in FOL, $\mathcal{I} \models \varphi$ indicates that an AE-interpretation $\mathcal{I}$ satisfies an AE-formula (is an AE-model of) φ.

A formula logically follows from a set M of AE-formulae ($M \models \varphi$) iff φ is valid in all AE-models of M.

For a set of AE-formulae M, $Th(M)$ is the set of AE-formulae that logically follow from M.

Remarks on the semantics of AE formulae

In Def. 4.3 the validity of φ in $\mathcal{I}$ and the validity of $L\,\varphi$ in $\mathcal{I}$ are unrelated:
$L\,\varphi$ is treated as a new atom (a 0-ary predicate) and thus independent of φ.

Intuition:
φ expresses truth of φ, whereas $L\,\varphi$ expresses belief in (/knowledge of) φ.

This choice of semantics admits to "believe in something false" or "not to believe in something true".

The following alternative definition of the semantics captures this observation.

Algebra-based semantics of autoepistemic logics

An algebra with a belief set is a pair $(\mathcal{B}, \mathit{Bel})$, where

- $\mathcal{B}$ is a first order interpretation and
- Bel is a set of AE-formulae.

Validity of AE-formulae in $(\mathcal{B}, \mathit{Bel})$ is defined as:

- $(\mathcal{B}, \mathit{Bel}) \models \varphi$ iff $\mathcal{B} \models \varphi$ for a closed FO formula φ
- $(\mathcal{B}, \mathit{Bel}) \models \neg\varphi$ iff $(\mathcal{B}, \mathit{Bel}) \not\models \varphi$
- $(\mathcal{B}, \mathit{Bel}) \models (\varphi \vee \psi)$ iff $(\mathcal{B}, \mathit{Bel}) \models \varphi$ or $(\mathcal{B}, \mathit{Bel}) \models \psi$
- $(\mathcal{B}, \mathit{Bel}) \models (\varphi \wedge \psi)$ iff $(\mathcal{B}, \mathit{Bel}) \models \varphi$ and $(\mathcal{B}, \mathit{Bel}) \models \psi$
- $(\mathcal{B}, \mathit{Bel}) \models (\varphi \longrightarrow \psi)$ iff $(\mathcal{B}, \mathit{Bel}) \models \varphi$ implies $(\mathcal{B}, \mathit{Bel}) \models \psi$
- $(\mathcal{B}, \mathit{Bel}) \models L\,\varphi$ iff $\varphi \in \mathit{Bel}$.

Relationship between the two semantics

The semantics are equivalent.

1. From a given AE-interpretation $\mathcal{I}$, we define an algebra with a belief set $(\mathcal{B}, \mathit{Bel})$ as follows:
   - the domain of $\mathcal{B}$ and the interpretation of predicate and function symbols are same as in $\mathcal{I}$.
   - $\mathit{Bel} = \{\varphi \mid (L\ \varphi)^{\mathcal{I}} = \mathit{true}\}$
2. From a given algebra with a belief set $(\mathcal{B}, \mathit{Bel})$, we define an AE-interpretation $\mathcal{I}$ as follows:
   - the domain of $\mathcal{I}$ and the interpretation function of predicate and function symbols are same as in $\mathcal{B}$.
   - $(L\ \varphi)^{\mathcal{I}} = \mathit{true}$ iff $\varphi \in \mathit{Bel}$.

Convention: we use the two semantics interchangeably.

By "an AE-interpretation with belief set Bel" we mean $\mathit{Bel} = \{\varphi \mid (L\ \varphi)^{\mathcal{I}} = \mathit{true}\}$.

We define "φ follows from AE-theory T w.r.t. belief set E" (denoted $T \models_E \varphi$) as φ is valid in every AE-model of T with belief set E.

Subsection 4.2
Expansions of autoepistemic theories

Towards expansions — considerations

What knowledge would an agent with introspection have,
given a set of facts (i.e. AE-formulae) T?

The agent's knowledge would be a set E of AE-formulae that

- includes T
- allows introspection
- is grounded in T
  meaning: the knowledge in E must be reconstructable from:
  T, belief in (knowledge of) E, and non-belief in (non-knowledge of) E

Expansion

Let T and E be sets of AE-formulae. We define the following sets

- $L\ E = \{L\ \varphi \mid \varphi \in E\}$
- $\neg L\ E^C = \{\neg L\ \psi \mid \psi \notin E\}$
- $\Omega_T(E) = \{\varphi \mid T \cup L\ E \cup \neg L\ E^C \models \varphi\}$

Definition 4.4 (Expansion)

Let T and E be sets of AE-formulae.

- E is T-sound iff $E \subseteq \Omega_T(E)$
- E is T-complete iff $\Omega_T(E) \subseteq E$
- E is an expansion of T iff $E = \Omega_T(E)$

Intuition:
The agent decides to believe in a set of AE-formulae T.
Based on this, a set of AE-formulae can be deduced from T and the beliefs adopted ($L\ E \cup \neg L\ E^C$). If the deduced set is exactly the set of beliefs E, then E is an expansion.

Alternative characterization of expansions

Observation:
AE-models of $T \cup L\ E \cup \neg L\ E^C$ are just the AE-models of T with belief set E!

Thus we obtain an alternative characterization of expansions.

Corollary 4.5

E is an expansion of an AE-theory T iff $E = \{\varphi \mid T \models_E \varphi\}$.

Example 4.6

Consider the AE-theory T_1:

$$\{german \wedge \neg L\, \neg drinksBeer) \longrightarrow drinksBeer,\ german\}$$

This AE-theory has one expansion.

The formula $\neg drinksBeer$ cannot be derived before $\neg L\, \neg drinksBeer$ is contained in the expansion.

The only expansion of T_1 has the kernel: $Th(\{german, drinksBeer\})$

If we extend T_1 by adding:

$$\{(eatsPizza \wedge \neg L\ drinksBeer) \longrightarrow \neg drinksBeer,\ eatsPizza\}$$

then the theory has two expansions:

- kernel of the first expansion: $\{german, eatsPizza, drinksBeer\}$
- kernel of the second expansion: $\{german, eatsPizza, \neg drinksBeer\}$

Subsection 4.3
Stable sets and their properties

Stable sets — origin

- Stable belief sets were introduced by Robert Stalnaker in the early '80s
- proposed as a formal representation of the epistemic state of an ideally rational agent, with full introspective capabilities.
- assume a propositional language, endowed with a modal operator $\Box\varphi$ interpreted as "φ is believed"
- a set of formulae is a stable set if it is "stable" under classical inference and epistemic introspection
- influenced research on AE logics and nonmonotonic logics in general

Definition 4.7 (stable sets)

Let E be a set of autoepistemic formulae. E is called stable iff

- E is deductively closed, i.e. $E = Th(E)$,
- $\varphi \in E$ implies $L\varphi \in E$, for all AE-formula φ, and
- $\varphi \notin E$ implies $\neg L\varphi \in E$, for all AE-formula φ

Note: Expansions are stable sets by definition.
Thus they inherit all the properties we show for stable sets.

Theorem 4.8

For an AE-theory T and a set of AE-formulae E the following statements are equivalent:

1. *E is an expansion of T*
2. *E is stable, $T \subseteq E$ and is T-sound.*

Proof: blackboard

Lemma 4.9

For a stable set E and an AE-formula φ the following statements are equivalent:

a) $E \models_E \varphi$

b) $E \models \varphi$

c) $\varphi \in E$

For a FOL formula φ, the statements a)-c) are equivalent to

d) $E_0 \models \varphi$

Proof: blackboard

Stable sets are determined by their kernels

Stable sets are uniquely determined by their objective subsets, i.e. their kernels.

Theorem 4.10

For stable sets E and F, $E_0 = F_0$ implies $E = F$.

Proof: blackboard

How can expansions be computed? A first hint

Theorem 4.11

Let T be a first order theory. Then there is a stable set E with $E_0 = T$.

Proof: blackboard

Theorem 4.12 (Orthogonality of stable sets)

Let E and F be different stable sets. Then $E \cup F$ is inconsistent.

Proof: blackboard

Theorem 4.13

If E is a stable set then it is an expansion of E_0.

Proof: blackboard

Subsection 4.4
Computing expansions of AE-theories

To achieve nonmonotonic behavior w.r.t. AE-theories, formulae ("conjectures") can be added to the set of believes that need not be added.

What makes computing expansions difficult?

- nested occurrences of the *L*-operator
- infinitely many conjectures. How to compute *all* expansions?

How to remedy this?

- Nested occurrences of *L* -operator:
  concentrate on potential kernels of expansions (Theorem 4.10).
- establish a Coincidence Lemma:
  "it suffices to consider beliefs or non-beliefs in formulae from $sub(T)$ to determine the expansions of T."

  Only those formulae with *L* -operator play a role in the interpretation of T.

Overview of the computation procedure for expansions

Compute expansions of AE-theories by:

1. partition $sub(T)$ into:
   - $E(+)$: set of beliefs
   - $E(-)$: set of non-beliefs
2. Compute the corresponding kernel $E(0)$ of a potential expansion, using T, beliefs in $E(+)$ and non-beliefs in $E(-)$.
3. Check whether the stable set determined by $E(0)$ is indeed an expansion

Example 4.14

Let $T = \{L\, p \longrightarrow p\}$.
Since $L\, p \longrightarrow p$ is the only AE-formula occurring (at top-level) of T, $sub(T) = \{p\}$.
There are two partitions of $sub(T) = \{p\}$.

$E(+)$	$E(-)$	$E(0)$	$E(+) \subseteq E(0)$?	$E(-) \cap E(0) = \emptyset$?	expansion?
$\{p\}$	$\emptyset$	$Th(\{p\})$	yes	yes	yes
$\emptyset$	$\{p\}$	$Th(\emptyset)$	yes	yes	yes

- $E(0)$: set of first order formulae that follow from $T \cup L\, E(+) \cup \neg L\, E(-)$.
- condition $E(+) \subseteq E(0)$:
  test whether everything that the agent believes in is in $E(0)$.
- condition $E(-) \cap E(0) = \emptyset$:
  ensures that $E(0)$ does not include non-beliefs of the agent

Procedure for computing expansions for AE-theories without L-nestings

Compute expansions no L-nesting (T)

1: $\mathit{Expansions} := \emptyset$
2: **for all** partitions $E(+)$ and $E(-)$ of $\mathit{sub}(T)$ **do**
3: $\quad E(0) := \{\varphi \in \mathit{For}_0 \mid T \cup L\, E(+) \cup \neg L\, E(-) \models \varphi\}$
4: $\quad$ **if** $E(+) \subseteq E$ and $E(-) \cap E = \emptyset$ **then**
5: $\quad\quad \mathit{Expansions} := \mathit{Expansions} \cup \{E(0)\}$
6: $\quad$ **end if**
7: **end for**
8: **return** *Expansions*

Example 4.15

Let $T = \{L\ p \longrightarrow p, \neg L\ \neg L\ p\}$, with $sub(T) = \{p, \neg L\ p\}$.
Now the partitions of $sub(T)$ are no longer first order formulae!

$E(+)$	$E(-)$	$E(0)$	$E(+) \subseteq E$?	$E(-) \cap E = \emptyset$?	expansion?
$\{p, \neg L\ p\}$	$\emptyset$	For_0	yes	yes	yes
$\{p\}$	$\{\neg L\ p\}$	$Th(\{p\})$	yes	yes	yes
$\{\neg L\ p\}$	$\{p\}$	For_0	yes	no	no
$\emptyset$	$\{p, \neg L\ p\}$	$Th(\emptyset)$	yes	no	no

- Line 1: $E(0)$ is inconsistent, since $L\ \neg L\ p$ follows from $L\ E(+)$, but $\neg L\ \neg L\ p \in T$.
- Line 2: $T \cup L\ E(+) \cup \neg L\ E(-) = \{L\ p \longrightarrow p, \neg L\ \neg L\ p, L\ p\}$, thus $E(0) = Th(\{p\})$. Since $p \in E$ and E is stable and consistent, we have $L\ p \in E$ and thus $\neg L\ \notin E$.
- Line 3: $T \cup L\ E(+) \cup \neg L\ E(-)$ contains both $L\ \neg L\ p$ and $\neg L\ \neg L\ p$, thus $E(0) = For_0$
- Line 4: $T \cup L\ E(+) \cup \neg L\ E(-) = \{L\ p \longrightarrow p, \neg L\ \neg L\ p, \neg L\ p\}$. From $p \notin E$ follows $\neg L\ p \in E$ and thus $(E(-) \cap E) \neq \emptyset$

Compute expansions (T)

1: $Expansions := \emptyset$
2: **for all** partitions $E(+)$ and $E(-)$ of $sub(T)$ **do**
3: $\quad E(0) := \{\varphi \in For_0 \mid T \cup L\ E(+) \cup \neg L\ E(-) \models \varphi\}$
4: $\quad$ **Let** E be the unique stable set with kernel $E(0)$
5: $\quad$ **if** $E(+) \subseteq E$ AND $E(-) \cap E = \emptyset$ **then**
6: $\quad\quad Expansions := Expansions \cup \{E(0)\}$
7: $\quad$ **end if**
8: **end for**
9: **return** $Expansions$

Lemma 4.16 (Preservation Lemma)

Let E be a stable set and T an AE-theory.
If $E_0 = \{\varphi \in For_0 \mid T \cup L\ E \cup \neg L\ E^C \models \varphi\}$, then $E = \{\varphi \in For \mid T \cup L\ E \cup \neg L\ E^C \models \varphi\}$.

Proof: blackboard

Lemma 4.17 (Coincidence Lemma)

Let T be an AE-theory. Consider sets of AE-formulae $E(+)$, $E(-)$, $F(+)$, and $F(-)$ with the following properties:

- $sub(T) \subseteq E(+) \cup E(-)$ *and* $E(+) \cap E(-) = \emptyset$ *and*
  $sub(T) \subseteq F(+) \cup F(-)$ *and* $F(+) \cap F(-) = \emptyset$
- $E(+) \cap sub(T) = F(+) \cap sub(T)$
- $E(-) \cap sub(T) = F(-) \cap sub(T)$.

Then the same first order formula follow from $T \cup L\ E(+) \cup \neg L\ E(-)$ as from $T \cup L\ F(+) \cup \neg L\ F(-)$

Proof: blackboard

Theorem 4.18

Let T be an AE-theory and let $sub(T)$ be partitioned into the disjoint sets $E(+)$ and $E(-)$. We consider the following steps:

1. *Compute $E_0 = \{\varphi \in For_0 \mid T \cup L\ E(+) \cup \neg L\ E(-) \models \varphi\}$ and let E be the unique stable set with kernel E_0.*
2. *Check whether $E(+) \subseteq E$ and $E(-) \cap E = \emptyset$.*

Then the following holds:

a) *If the check in Step 2. is positive, then E is an expansion of T.*

b) *Conversely, for every expansion E of T there is a decomposition of $sub(T)$ into $E(+)$ and $E(-)$ such that*
   - *$E(0) = E_0$ and*
   - *the check in Step 2 is positive.*

Proof: blackboard

Computational complexity of reasoning in autoepistemic logics

- For closed FOL formulae in a logic L holds:
  if satisfiability in L is decidable, then so are nonmonotonic reasoning tasks for L.
- deciding whether an AE-theory has a stable expansion: Σ_2^P-complete
- credulous reasoning (using one expansion) is Σ_2^P-complete
  cautious reasoning (using all expansions) is Π_2^P-complete