Metatheory and Proof Theory

Knowledge representation and reasoning are fundamental in AI. In the last two lectures, we looked briefly at knowledge representation. Now, we look at reasoning.

We need to know what it means to say that one wff W 'follows from' some set of wffs Φ. This definition will have something to do with the semantics of FOPL, and we discuss it in the first section below (Metatheory). But, we will see that the definition doesn't give a basis on which we could build a practical, automatable way of doing reasoning. We make a start on doing this in the second section (Proof theory).

1 Logical Metatheory

Recall the notion of a model, M:

$$M = \langle \mathcal{U}, \mathcal{I} \rangle$$

comprising a universe of discourse $\mathcal{U}$ and an interpretation function $\mathcal{I}$. Recall also that the semantic value or denotation of a term, atom or wff α with respect to a model is written:

$$[\![\alpha]\!]^M$$

Once models are defined, we can define satisfiability, unsatisfiability, validity and, most importantly, logical consequence.

1.1 Satisfiability, Unsatisfiability and Validity

Let W be a wff and M be a model.

- A wff W is *satisfiable* iff there exists some model M such that $[\![W]\!]^M = \textbf{true}$.
  Examples of satisfiable wffs are:
  $$p(a, b)$$
  $$p(b, a)$$
  $$\forall x(q(x) \Rightarrow r(x))$$
- A wff W is *unsatisfiable* iff there exists no model such that $[\![W]\!]^M = \textbf{true}$.
  Examples of unsatisfiable wffs are:
  $$p(a, b) \wedge \neg p(a, b)$$
  $$\neg(p(b, a) \Rightarrow p(b, a))$$
- A wff W is *valid* iff, for every model M, $[\![W]\!]^M = \textbf{true}$. (Other terminology: if W is valid, other authors might say W is *unfalsifiable* or W is a *tautology*.)
  Examples of valid wffs are:
  $$p(a, b) \vee \neg p(a, b)$$
  $$p(b, a) \Rightarrow p(b, a)$$

1.2 Logical Consequence

A set of wffs Φ has wff W as a *logical consequence* iff every model that makes Φ true also makes W true. (Or, if you want it in different words, there is no model that makes all the elements of Φ true but makes W false.)

We write this

$$\Phi \models W$$

(Other terminology: if W is a logical consequence of Φ, other authors might say that W is an entailment of Φ or that Φ entails W or that W logically follows from Φ or that W is logically implied by Φ.)

Note that the definition doesn't say anything about models in which one or more of the wffs in Φ are false. These models are not relevant to deciding whether $\Phi \models W$.

An example of a logical consequence:

$$\{\forall x(p(x) \Rightarrow q(x)), p(a)\} \models q(a)$$

It does not matter what p, q and a are interpreted as, *if* the members of the set $\{\forall x(p(x) \Rightarrow q(x)), p(a)\}$ *are* true in M, then $q(a)$ *will* be true in M.

An informal demonstration of this follows. Suppose in one model p denotes the set of men, q denotes the set of mortals and a denotes Socrates. Then, if it is the case that all men are mortals and Socrates is a man (i.e. if $\{\forall x(p(x) \Rightarrow q(x)), p(a)\}$ is true) then it follows that Socrates is mortal (i.e. $q(a)$ is true). Suppose in another model p denotes the set of elephants, q denotes the set of grey things and a denotes Clyde. Then, if it is the case that all elephants are grey and Clyde is an elephant (i.e. if $\{\forall x(p(x) \Rightarrow q(x)), p(a)\}$ is true) then it follows that Clyde is grey (i.e. $q(a)$ is true). Suppose in yet another model p denotes the set of elephants, q denotes the set of yellow things and a denotes Clyde. Then, if it is the case that all elephants are yellow (never mind that it isn't in our world: *if* it *is* the case...) and Clyde is an elephant, then it follows that Clyde is yellow. And so on.

Note how the definition of logical consequence refers to *all* models which make Φ true. This suggests that in order to determine whether W is a logical consequence of a set of wffs we would have to check all interpretations of the wffs for all universes. This is possible in some logics. For example, we can do it for propositional logic (i.e. logic without variables and functions). In propositional logic, checking all interpretations is the same as constructing a truth-table since each row of the table is an interpretation. For example, we can show that

$$\{p \vee q, q\} \models (\neg p \wedge q) \vee p$$

p	q	$p \vee q$	q	$(\neg p \wedge q) \vee p$
T	T	T	T	T
T	F	T	F	T
F	T	T	T	T
F	F	F	F	F

$(\neg p \wedge q) \vee p$ is a logical consequence of $\{p \vee q, q\}$ because it is true in all cases (interpretations) in which $p \vee q$ and q are true (i.e. rows 1 and 3).

In other logics, determining logical consequences cannot in general be done by inspecting models. In particular, it is not possible in FOPL because there may be an infinite number of interpretations (for an infinite number of possibly infinite universes). We clearly need an alternative approach.

2 Proof Theories for FOPL

Given the impossibility of checking infinite numbers of models, we need some apparatus that enables reasoning to take place at a purely syntactic level, before interpretations are considered.

This apparatus is called a *proof theory*, and it is the third component of every logic (alongside its syntax and semantics). (Other terminology: other authors might refer to a proof theory as a (formal) deduction system, an inference system or a logical calculus.) Proof theories can be broadly classified into deduction systems and refutation systems. We look at the former first.

2.1 Informal Presentation

A proof theory enables us to derive conclusions from a set of wffs by *syntactic operations* alone. We manipulate the wffs without reference to their semantics. If the manipulation rules are 'right' (in a sense to be explored later), the new wffs we generate will, in fact, be logical consequences of the original set of wffs. The process is 'mechanical' and thus amenable to automation.

A proof theory comprises a finite set of *inference rules* and a finite set of *logical axiom schemata*.

Inference rules: An inference rule comprises a set of patterns called *conditions* and another pattern called the *conclusion*. Here's an example of a rule called $\Rightarrow$-elimination (although you might know it as Modus Ponens):

$$\frac{W_1, W_1 \Rightarrow W_2}{W_2}$$

Above the line are the conditions; below is the conclusion. If you have wffs that match the conditions, then you can, in a single step, derive a wff that matches the conclusion.

Example. Suppose $\Phi = \{\textit{raining} \Rightarrow (\textit{needMac} \vee \textit{needBrolly}), \textit{raining}\}$. Then we can derive *needMac* $\vee$ *needBrolly*.

Logical axiom schemata: A logical axiom is another name for a valid wff. A logical axiom schemata is a 'template' for a valid wff: if you substitute wffs into the template, you get a valid wff.

Here are examples:

$$W_1 \Rightarrow (W_2 \Rightarrow W_1)$$
$$(W_1 \Rightarrow (W_2 \Rightarrow W_3)) \Rightarrow ((W_1 \Rightarrow W_2) \Rightarrow (W_1 \Rightarrow W_3))$$
$$(W_1 \Rightarrow \neg W_2) \Rightarrow ((W_1 \Rightarrow W_2) \Rightarrow \neg W_2)$$
$$(\neg W_1 \Rightarrow \neg W_2) \Rightarrow ((\neg W_1 \Rightarrow W_2) \Rightarrow W_1)$$

If you substitute *raining* for W_1 and *needMac* $\vee$ *needBrolly* for W_2 in the first logical axiom schemata, you produce the logical axiom (valid wff)

$$\textit{raining} \Rightarrow ((\textit{needMac} \vee \textit{needBrolly}) \Rightarrow \textit{raining})$$

(You can confirm that this is a valid wff by drawing up a truth-table and noting that the wff always evaluates to true. (Exercise!))

Given a set of inference rules, a set of logical axiom schemata and a set of axioms, Φ, a *derivation* of some wff W from Φ is defined to be a finite sequence of wffs $W_1, W_2, \ldots, W_n$ such that $W = W_n$ and for each i $(1 \leq i \leq n)$ either

- W_i is a logical axiom (produced by substituting wffs into the logical axiom schemata), or

- W_i is an axiom (a member of Φ), or
- W_i is the result of applying an inference rule to previous wffs in the sequence.

If there exists a derivation of W from Φ then we say that W is *derivable* from Φ, and we write this as:

$$\Phi \vdash W$$

(Other terminology: if W is derivable from Φ, other authors might say that it is deducible or provable from Φ.)

There are many choices of proof theory. They offer different sets of logical axiom schemata (including none) and different sets of inference rules. We'll look at two example for propositional logic, and then we'll discuss criteria that proof theories can be evaluated against.

2.2 A Hilbert-style proof theory for propositional logic

This proof theory has the following components:

Logical axiom schemata:
Axiom schemata 1: $(\neg W_1 \Rightarrow W_1) \Rightarrow W_1$
Axiom schemata 2: $W_1 \Rightarrow (\neg W_1 \Rightarrow W_2)$
Axiom schemata 3: $(W_1 \Rightarrow W_2) \Rightarrow ((W_2 \Rightarrow W_3) \Rightarrow (W_1 \Rightarrow W_3))$

Inference rules: There is only one rule, $\Rightarrow$-elimination (Modus Ponens):

$$\frac{W_1, W_1 \Rightarrow W_2}{W_2}$$

For your axioms Φ, you must restrict yourself to using $\Rightarrow$ and $\neg$ only. (This is no loss since a wff that contains the other connectives can always be converted into an equivalent wff using only the given three.)

Exercise. *Given that if it is raining then I get wet, and if I get wet then I catch cold, show that if it is raining then I catch cold.*

Finding a derivation is a search process. At each step, there are different options that can be tried. We have to make choices, but if our choices don't work out, we may need to come back and try the other options.

The search space can be huge. In particular, there is an infinite number of wffs we can substitute into each logical axiom schemata.

And this is just a proof theory for *propositional logic*. A similar proof theory for *FOPL*, which is a more expressive logic, has a far worse search space.

2.3 A natural deduction system for propositional logic

In this alternative proof theory, there are no logical axiom schemata. But there are many more inference rules.

$\wedge$-INTRO $\frac{W_1, W_2}{W_1 \wedge W_2}$ $\wedge$-ELIM-LEFT $\frac{W_1 \wedge W_2}{W_1}$ $\wedge$-ELIM-RIGHT $\frac{W_1 \wedge W_2}{W_2}$

$\vee$-INTRO-RIGHT $\frac{W_1}{W_1 \vee W_2}$ $\vee$-INTRO-LEFT $\frac{W_2}{W_1 \vee W_2}$ $\vee$-ELIM $\frac{W_1 \vee W_2, \frac{W_1}{W_3}, \frac{W_2}{W_3}}{W_3}$

$\Rightarrow$-ELIM $\frac{W_1, W_1 \Rightarrow W_2}{W_2}$ $\Rightarrow$-INTRO $\frac{\frac{W_1}{W_2}}{W_1 \Rightarrow W_2}$

$\Leftrightarrow$-ELIM-LEFT $\frac{W_1 \Leftrightarrow W_2, W_1}{W_2}$ $\Leftrightarrow$-ELIM-RIGHT $\frac{W_1 \Leftrightarrow W_2, W_2}{W_1}$ $\Leftrightarrow$-INTRO $\frac{\frac{W_1}{W_2}, \frac{W_2}{W_1}}{W_1 \Leftrightarrow W_2}$

$\neg$-ELIM $\frac{\neg\neg W}{W}$ $\neg$-INTRO $\frac{\frac{W_1}{W_2 \wedge \neg W_2}}{\neg W_1}$

Exercise. *Given that if it is raining then I get wet, and if I get wet then I catch cold, show that if it is raining then I catch cold.*

Again, finding a derivation is a search process. Ths time, we have to choose between numerous possible inference rules. In some cases, we have to choose what wffs to assume.

The upshot again is that we have here another vast search space.

2.4 Soundness and Completeness

A proof theory derives wffs using syntactic operations alone, without reference to the semantics. We could invent a set of logical axiom schemata and inference rules that license even quite bizarre inferences. However, if a proof theory is to be useful, we typically require that the wffs we can derive must tie in with the logical consequences. There are two measures of this: *soundness* and *completeness*. (We came across the word 'completeness' when discussing search. Unfortunately, its use in logic is a related but subtly different use.)

Soundness: A proof theory is *sound* if

$$\Phi \vdash W \text{ implies } \Phi \models W$$

i.e. if all the wffs that can be derived from Φ are also logical consequences of Φ. The proof theory can't derive wffs that aren't logical consequences.

To see an unsound proof theory, imagine a proof theory that contains the following inference rule:

$$\Rightarrow\text{-DUFF}\frac{W_1 \Rightarrow W_2, W_2}{W_1}$$

Wffs derived from this rule are not necessarily logical consequences. For example, we can use this rule to show $\{rain \Rightarrow wet, wet\} \vdash rain$. But we know that $\{rain \Rightarrow wet, wet\} \not\models rain$. (Check it with a truth-table!)

The proof theories given in the previous two subsections are both sound.

Completeness: A proof theory is *complete* if

$$\Phi \models W \text{ implies } \Phi \vdash W$$

i.e. every logical consequence of Φ can be derived from Φ. The proof theory doesn't fail to derive any of the logical consequences.

The proof theories given in the previous two subsections are both complete. To see an example of an incomplete proof theory, imagine dropping one of the logical axiom schemata or one of the inference rules.

Soundness is essential, but completeness might be sacrificed in AI in an effort to reduce the search space and thereby improve the efficiency of an automated proof theory. We might especially sacrifice completeness if doing so loses us only a few 'obscure' inferences.

But before sacrificing these things, there's an alternative. In the worst case, this alternative is no better, but it does seem to give good typical case performance. The alternative, which we look at in the next few lectures, is to use a refutation system on a canonical logical form.