THE COMPLETENESS THEOREM FOR PROPOSITIONAL LOGIC

ANDREW FOOTE

The key thing I had to realize for the proof of the completeness theorem for propositional logic to “make sense” to me was that interpretations of a language of propositional logic can be thought of as theories.

Normally, an interpretation of a language L of propositional logic is thought of as a function v from the set of the sentence variables in L to $\{0, 1\}$, which is extended to the set of the formulas in L by letting $v(\bot) = 0$ and recursively letting

$$v(\phi \to \psi) = v(\psi)^{v(\phi)} = \begin{cases} 0 & \text{if } v(\phi) = 0 \text{ or } v(\psi) = 1, \\ 1 & \text{if } v(\phi) = 1 \text{ and } v(\psi) = 0 \end{cases}$$

for every pair of sentences ϕ and ψ in L (assuming $\bot$ and $\to$ are the primitive connectives). The value under v of a sentence ϕ is thought of as the truth value of ϕ under v, with 0 standing for falsity and 1 standing for truth.

A theory in L, on the other hand, is thought of as a set of sentences in L, these sentences being the nonlogical axioms of the theory. It’s a completely different type of object from an interpretation.

But if you think about what these formal concepts are trying to get at, they’re quite similar. Both of them are essentially requirements that certain sentences be true. A theory requires its nonlogical axioms to be true. An interpretation requires the sentences true under it to be true, and the sentences false under it to be false, which might appear to be a slightly more elaborate concept, but since a sentence is false iff its negation is true, you know what sentences an interpretation makes false if you know what sentences it makes true. The only substantial difference is that an interpretation is subject to certain restrictions compared to a general theory:

1. An interpretation of L must require every sentence variable in L to be either true or false (and not both).
2. An interpretation of L must require $\bot$ to be false.
3. For every pair of sentences ϕ and ψ in L, an interpretation of L must require $\phi \to \psi$ to be true iff it requires ϕ to be false or requires ψ to be true (or both).

Formally, we can define a function f on the set of the interpretations of L by the rule that for every interpretation v of L, we have

$$f(v) = \{\phi : v(\phi) = 1\} \cup \{\neg\phi : v(\phi) = 0\}.$$

This function f is an injection into the set of the theories in L. But it is not surjective—for every theory M in L, the value $f^{-1}(M)$ exists only if the statements below hold:

1. For every sentence variable A in L, exactly one of A and $\neg A$ is a member of M.
2. $\bot \notin M$.

(3) For every pair of sentences ϕ and ψ in L, we have $\phi \to \psi \in M$ iff $\phi \notin M$ or $\psi \in M$.

An interpretation could reasonably be *defined* as a theory M in L with properties (1)-(3) above.

Now, the completeness theorem says that every syntactically consistent theory in L is semantically consistent, i.e. has a model. A *model* of a theory T in L is an interpretation of L under which every member of T is true, so if we adopt the definition of interpretations as theories, we can say that a model of T is an extension of T which is an interpretation of L. This leads us to the idea that in order to construct a model of an arbitrary syntactically consistent theory T in L, we will have to *extend* T by adding new members.

But how exactly do we extend T? Here, it helps to recall the soundness theorem, which says that for every theory T' in L, every syntactic consequence of T' is a semantic consequence of T'. In particular, if $T' \vdash \bot$, then $T' \models \bot$. Contrapositively, if T' is semantically consistent, then T' is syntactically consistent. Now, every interpretation of L, thought of as a theory in L, is semantically consistent (since it has itself as an extension) and hence, by soundness, syntactically consistent. Therefore one thing we definitely need to do as we add new members to T, in order for it to eventually become an interpretation of L, is preserve the syntactic consistency of T.

In fact, it's not too difficult to see that every theory M in L which can be thought of as an interpretation of L is not only syntactically consistent, but *maximally* syntactically consistent: every proper extension of M is syntactically *in*consistent.

To see this, first, observe that for every sentence ϕ in L, since $\neg\phi$ abbreviates $\phi \to \bot$, property (3) tells us that $\neg\phi \in M$ iff $\phi \notin M$ or $\bot \in M$; and M never contains $\bot$. So we have $\neg\phi \in M$ iff $\phi \notin M$. In other words, M contains exactly one of ϕ and $\neg\phi$.

Now, suppose M' is a proper extension of M. Then it has a member ϕ which is not a member of M. The negation of ϕ must then be a member of M. Since M' extends M, it follows that $\neg\phi \in M'$. But we also have $\phi \in M'$. Since every member of M' is a syntactic consequence of M', it follows that M' is syntactically inconsistent.

Now that we know every interpretation of L is maximally syntactically consistent, the natural next question to ask is whether the converse holds, i.e. every maximally syntactically consistent theory T in L is an interpretation of L. If the converse does hold, then to prove the completeness theorem, all we need to do is extend a syntactically consistent theory to a maximal syntactically consistent theory. As it happens, the converse does hold.

Lemma 1. *For every maximally syntactically consistent theory M in L and every sentence ϕ in L, exactly one of ϕ and $\neg\phi$ is a member of M.*

Proof. Suppose M is a maximally syntactically consistent theory in L and ϕ is a sentence in L.

If M contains both ϕ and $\neg\phi$, then M is syntactically inconsistent. So M contains at most one of ϕ and $\neg\phi$.

If M contains neither ϕ nor $\neg\phi$, then $M \cup \{\phi\}$ and $M \cup \{\neg\phi\}$ are proper extensions of M and hence are syntactically inconsistent, from which it follows by negation introduction and elimination that M syntactically implies both ϕ and $\neg\phi$, and hence is syntactically inconsistent. So M contains at least one of ϕ and $\neg\phi$. □

Theorem 1. *Every maximally syntactically consistent theory M in L is an interpretation of L.*

Proof. Suppose M is a maximally syntactically consistent theory in L.

For every sentence variable A in L, exactly one of A and $\neg A$ is a member of M by the lemma above.

If $\perp \in M$, then M syntactically implies $\perp$ and hence is syntactically inconsistent; so we have $\perp \notin M$.

Suppose ϕ and ψ are sentences in L. We shall prove that $\phi \to \psi \in M$ iff $\phi \notin M$ or $\psi \in M$.

For the forward implication, suppose $\phi \to \psi \in M$. If $\phi \in M$ and $\psi \notin M$, i.e. $\neg\psi \in M$, then M syntactically implies both $\phi \to \psi$ and $\neg\psi$, so by *modus tollens* it follows that $M \vdash \neg\phi$. But we also have $M \vdash \phi$, so M is syntactically inconsistent. This is a contradiction, so one of $\phi \in M$ and $\psi \notin M$ must not hold.

For the backward implication:

(1) First, suppose $\phi \notin M$, i.e. $\neg\phi \in M$. Then $M \vdash \neg\phi$, so $M \vdash \phi \to \psi$.
(2) Second, suppose $\psi \in M$. Then $M \vdash \psi$ and hence $M \vdash \phi \to \psi$.

Either way, we have $M \vdash \phi \to \psi$. Therefore, if $\phi \to \psi \notin M$, i.e. $\neg(\phi \to \psi \in M)$, so that $M \vdash \neg(\phi \to \psi)$, we have that M is syntactically inconsistent, which is a contradiction. So $\phi \to \psi \in M$. □

Now, to complete the proof, we just need to prove that an arbitrary syntactically consistent theory T in L can be extended until it is maximally syntactically consistent. The standard tool for carrying out such proofs is Zorn's lemma. (There are other techniques that can be used, like transfinite induction; if you're not too familiar with proofs like this, using transfinite induction generally makes things clearer. But Zorn's lemma makes the proof more concise, so that's what I'll use here.)

Note that we make use of the "syntactic compactness theorem" here, which says that a theory is syntactically consistent iff each of its finite subsets is syntactically consistent. Unlike the *semantic* compactness theorem, which is most straightforwardly proven as a *consequence* of completeness, the *syntactic* compactness theorem is trivial; it essentially follows from the fact that proofs are finite and hence any proof of $\perp$ only makes use of finitely many axioms.

Theorem 2. *Every syntactically consistent theory has a maximal syntactically consistent extension.*

Proof. Suppose T is a syntactically consistent theory. To prove that T has a maximal syntactically consistent extension, we shall use Zorn's lemma; so suppose $\mathcal{T}$ is a chain of syntactically consistent extensions of T and let T' be the union of the theories in $\mathcal{T}$. To prove that T' is consistent, we shall use the syntactic compactness theorem; so suppose $U = \{\phi_1, \phi_2, \dots, \phi_n\}$ is a finite subset of T. In the case where U is empty, it is certainly syntactically consistent. Otherwise, let T_1, T_2, … and T_n be theories in $\mathcal{T}$ containing ϕ_1, ϕ_2, … and ϕ_n respectively. Then $\{T_1, T_2, \dots, T_n\}$ is a finite chain, so it has a maximum U', which is syntactically consistent and extends U. Therefore U, having a syntactically consistent extension, must be syntactically consistent itself. □