Chapter 7
Herbrand Method

§7.1 Introduction

In propositional logic, it is possible to determine satisfiability, validity, and logical entailment for a language by looking at the set of all interpretations for the logical constants of that language, i.e. its truth table. Although the truth table for a propositional language can be large, it is always finite; and so there is a simple procedure for checking whether a sentence is satisfiable or valid or whether a sentence logically entails another.

In relational logic, we also have the notion of an interpretation; and this raises the question of whether there is an approach to checking satisfiability, validity, and logical entailment for relational logic that is analogous to the truth table method for propositional logic.

Consider the table below shows all of the interpretations for a relational language with two object constants and one unary relation constant using a universe of discourse consisting of just two objects, ∘ and •. In this case, the set of all possible interpretations is finite.

$\forall$	a	b	p
{∘, •}	∘	∘	{}
{∘, •}	∘	∘	{∘}
{∘, •}	∘	∘	{•}
{∘, •}	∘	∘	{∘, •}
{∘, •}	∘	•	{}
{∘, •}	∘	•	{∘}
{∘, •}	∘	•	{•}
{∘, •}	∘	•	{∘, •}
{∘, •}	•	∘	{}
{∘, •}	•	∘	{∘}
{∘, •}	•	∘	{•}
{∘, •}	•	∘	{∘, •}
{∘, •}	•	•	{}
{∘, •}	•	•	{∘}
{∘, •}	•	•	{•}
{∘, •}	•	•	{∘, •}

Unfortunately, using just this table to check for logical entailment and so forth is not adequate since it covers just one possible universe of discourse. There are many other possibilities. The problem with relational logic is that, in general, the set of possible interpretations for a language is necessarily infinite. In fact, there is not even a systematic way of enumerating all of the possibilities.

The good news is that is not lost. At least in some cases. The trick is to use a set of special models, called *Herbrand models*, to check satisfiability, validity, logical entailment

and so forth. It can be shown that in some cases checking these models alone is sufficient. Furthermore, in some cases, there are only finitely many Herbrand models; and so the process of checking for these conditions is algorithmic.

Sadly, this trick does not always work. Sometimes, Herbrand models alone do not suffice. And, sometimes, even though Herbrand models suffice, they are infinitely large; and so checking satisfaction takes forever.

In order to understand when the trick works and when it does not, we look at the method on increasingly complex subsets of relational logic. First, we examine universal logic, i.e. relational logic without functions or explicit quantifiers. (Free variables are assumed to be universally quantified, but there are no embedded universal quantifiers and no existential quantifiers.) After looking at universal logic, we switch to existential logic but still avoid functions. Then, we look at functional logic but avoid quantifiers.

§7.2 Universal Logic

The *Herbrand Universe* for relational language is the set of all ground terms in the language. For universal logic, this is equivalent to the set of all object constants. If there are no object constants, then we include an arbitrary object constant, say a. For example, in a language with just two object constants a and b, the Herbrand universe is the set $\{a, b\}$.

A *Herbrand interpretation* for a universal language is an interpretation in which (1) the universe of discourse is the Herbrand Universe for the language and (2) each object constant maps into itself.

For example, the interpretation shown below is a Herbrand interpretation for a universal language with objects constants a and b and binary relation constant r.

$$\forall^i = \{a, b\}$$
$$a^i = a$$
$$b^i = b$$
$$r^i = \{\langle a, a\rangle, \langle a, b\rangle\}$$

Note that the object constants are interpreted as themselves, as required by the definition of Herbrand interpretation. The interpretation of the relation constant r is arbitrary; any other binary relation on the Herbrand universe would be acceptable.

A *Herbrand variable assignment* is a variable assignment in which the universe is the Herbrand Universe for the language.

The variable assignment shown below is a Herbrand variable assignment for the universal relational language described above. Each of the variables is mapped into an object constant of the language.

$$x^i = a$$
$$y^i = b$$
$$z^i = b$$

One interesting thing to note about Herbrand interpretations for universal logic is that there are only finitely many of them. Given a finite set of constants, the universe of discourse is finite. The interpretation of the object constants is fixed; and there are only finitely many relations that can be formed from a finite universe of discourse.

The other interesting thing about Herbrand interpretations for universal logic is that they alone are sufficient for checking satisfiability, validity, and logical entailment, because of the Herbrand Theorem.

Universal Herbrand Theorem: *If a set of sentences in a universal language has a model, then it has a Herbrand model.*

For an model i, we can construct a corresponding Herbrand model h as follows. The Herbrand interpretation for each object constant is itself. The Herbrand interpretation h for each relation constant ρ is the set of all tuples of object constants $\tau_1, \ldots, \tau_n$ such that interpretation i satisfies the sentence $\rho(\tau_1, \ldots, \tau_n)$. The Herbrand theorem assures is that the resulting interpretation h satisfies the same sentences as i.

As an example of this construction, consider a language with object constants a and b and a single binary relation r.

Now, consider the interpretation i shown below. There are just two elements in the universe of discourse. The constant a maps to $\circ$; b maps to $\bullet$; c also maps to $\bullet$; and r is an arbitrary binary relation in terms of $\circ$ and $\bullet$.

$$\begin{aligned}
\forall^i &= \{\circ, \bullet\} \\
a^i &= \circ \\
b^i &= \bullet \\
r^i &= \{\langle \circ, \bullet \rangle, \langle \bullet, \bullet \rangle\}
\end{aligned}$$

The corresponding Herbrand interpretation is shown below. The constants a and b maps to themselves. In order to get the interpretation of r, we consider each relational sentence involving r and the constants a and b. For each, we check whether it is satisfied by i. If so, we include the ocrreponding pair of constants in our interpretation. If not, we do not include them. In this case, this leads trivially to the interpretation shown.

$$\begin{aligned}
\forall^h &= \{a, b\} \\
a^h &= a \\
b^h &= b \\
r^h &= \{\langle a, b \rangle, \langle b, b \rangle\}
\end{aligned}$$

As a more interesting example, consider the following interpretation for the same language. In this case, both object constants refer to the same object, viz. $\bullet$; and one of the elements in the universe of discourse, viz.$\circ$, has no corresponding object constant.

$$\forall^j = \{\circ, \bullet\}$$
$$a^j = \bullet$$
$$b^j = \bullet$$
$$r^j = \{\langle\circ, \bullet\rangle, \langle\bullet, \bullet\rangle\}$$

The corresponding Herbrand interpretation is shown below. The constants a and b map to themselves. In order to get the interpretation of r, we consider each relational sentence involving r and the constants a and b. For each, we check whether it is satisfied by i. If so, we include the ocrreponding pair of constants in our interpretation. If not, we do not include them. In this case, this leads trivially to the interpretation shown.

$$\forall^h = \{a, b\}$$
$$a^h = a$$
$$b^h = b$$
$$r^h = \{\langle a, a\rangle, \langle b, b\rangle\}$$

In the first example, the "structure" of the interpretation and its corresponding Herbrand interpretation are very similar. In this second example, the structure is very different. In interpretation j, both object constants refer to the same object whereas in the corresponding Herbrand interpretation, they refer to different objects. In the Herbrand interpretation, there is an object constant that refers to every element in the universe of discourse, whereas in interpretation j, there is an object $\circ$ with no corresponding object constant. These differences might lead one to think that the two interpretations do not satisfy the same sentences. The Herbrand theorem assures us that, despite these differences, they satisfy exactly the same sentences, provided that we restrict our attention to universal logic. The reader is encouraged to try out the interpretations on various sentences to confirm that this is indeed correct.

The importance of the Herbrand theorem is that it gives us a way of checking logical entailment in a finite amount of time. This is called the Herbrand Method. Starting with a set of premises and a possible conclusion, we loop over Herbrand interpretations, cross out each interpretation that does *not* satisfy the premises. Then, we check whether each remaining interpretation satisfies the conclusion. If so, the premises relationally entail the conclusion. Otherwise, logical entailment does not hold.

Suppose, for example, we were asked to determine whether the sentences $(p(x) \Rightarrow q(x))$ and $(p(a) \vee p(b))$ logically entails $q(a) \vee q(b)$. Remember that in relational logic, free variables are interpreted as universally quantified variables; hence the implication here must be true for all objects in the universe of discourse.

We start by writing out the set of all Herbrand interpretations and crossing out those that do not satisfy the premises. The first table below on the left is the complete set of Herbrand interpretations for this language. The second table results from crossing out those interpretations that do not satisfy $(p(x) \Rightarrow q(x))$. The third table results form crossing out the interpretations that do not satisfy $(p(a) \vee p(b))$. The fourth table is teh result of crossing out all rows that do not satisfy both premises.

p	q
$\{\}$	$\{\}$
$\{\}$	$\{a\}$
$\{\}$	$\{b\}$
$\{\}$	$\{a,b\}$
$\{a\}$	$\{\}$
$\{a\}$	$\{a\}$
$\{a\}$	$\{b\}$
$\{a\}$	$\{a,b\}$
$\{b\}$	$\{\}$
$\{b\}$	$\{a\}$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$\{a,b\}$	$\{\}$
$\{a,b\}$	$\{a\}$
$\{a,b\}$	$\{b\}$
$\{a,b\}$	$\{a,b\}$

p	q
$\{\}$	$\{\}$
$\{\}$	$\{a\}$
$\{\}$	$\{b\}$
$\{\}$	$\{a,b\}$
$-$	$-$
$\{a\}$	$\{a\}$
$-$	$-$
$\{a\}$	$\{a,b\}$
$-$	$-$
$-$	$-$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$-$	$-$
$-$	$-$
$\{a,b\}$	$\{a,b\}$

p	q
$-$	$-$
$-$	$-$
$-$	$-$
$-$	$-$
$\{a\}$	$\{\}$
$\{a\}$	$\{a\}$
$\{a\}$	$\{b\}$
$\{a\}$	$\{a,b\}$
$\{b\}$	$\{\}$
$\{b\}$	$\{a\}$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$\{a,b\}$	$\{\}$
$\{a,b\}$	$\{a\}$
$\{a,b\}$	$\{b\}$
$\{a,b\}$	$\{a,b\}$

p	q
$-$	$-$
$-$	$-$
$-$	$-$
$-$	$-$
$-$	$-$
$\{a\}$	$\{a\}$
$-$	$-$
$\{a\}$	$\{a,b\}$
$-$	$-$
$-$	$-$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$-$	$-$
$-$	$-$
$\{a,b\}$	$\{a,b\}$

Now, we do the same for the desired conclusion. The two table below show our work. The table on the left is, once again, the full Herbrand table. The table on the right results from crossing out all interpretations that do not satisfy the desired conclusion.

p	q
$\{\}$	$\{\}$
$\{\}$	$\{a\}$
$\{\}$	$\{b\}$
$\{\}$	$\{a,b\}$
$\{a\}$	$\{\}$
$\{a\}$	$\{a\}$
$\{a\}$	$\{b\}$
$\{a\}$	$\{a,b\}$
$\{b\}$	$\{\}$
$\{b\}$	$\{a\}$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$\{a,b\}$	$\{\}$
$\{a,b\}$	$\{a\}$
$\{a,b\}$	$\{b\}$
$\{a,b\}$	$\{a,b\}$

p	q
$-$	$-$
$\{\}$	$\{a\}$
$\{\}$	$\{b\}$
$\{\}$	$\{a,b\}$
$-$	$-$
$\{a\}$	$\{a\}$
$\{a\}$	$\{b\}$
$\{a\}$	$\{a,b\}$
$-$	$-$
$\{b\}$	$\{a\}$
$\{b\}$	$\{b\}$
$\{b\}$	$\{a,b\}$
$-$	$-$
$\{a,b\}$	$\{a\}$
$\{a,b\}$	$\{b\}$
$\{a,b\}$	$\{a,b\}$

Finally, we compare the premise table to the conclusion table and see that every row remaining in the former also remains in the latter. Consequently, we can conclude that the premises in this case logically entail the conclusion.

Although it is always finite, the Herbrand model for a universal language can be very large. If a language has n object constants, then the Herbrand universe is of size n.

This in itself is not so bad. What hurts is the number of intepretations of the language's relation constants. With n constants, one can form n^k k-tuples, leading to 2^{n^k} possible k-ary relations. With m k-ary relation constants, this leads to $(2^{n^k})^m$ possible Herbrand interpretations.

To make this concrete, consider a language with ten object constants and three binary relation constants. In this case, the Herbrand universe has size 10. There are 100 2-tuples, leading to 2^{100} binary relations, leading to 2^{300} possible Herbrand interpretations. For problems of this size, the Herbrand method is impractical.

§7.3 Existential Logic

Existential Logic is the same as universal logic except that it includes quantified sentences. It is called existential logic because the problems that arise are due to the presence of existential quantifiers. Universal quantifiers can also cause problems but only when they are equivalent to existential quantifiers, i.e. when used in a "negative context" such as the target of a negation or the antecedent of an implication or reduction.

The bad news is that the Herbrand theorem does not hold for existential logic.

Existential Herbrand Theorem: *If a set of sentences in an existential language has a model, then it does not necessarily have a Herbrand model.*

To see that this is true, consider the following example. Suppose we have a language with two object constants a and b and a single unary relation constant r. Now, consider the following sentences in existential logic.

$$\neg r(a)$$
$$\neg r(b)$$
$$\exists x.r(x)$$

The interpretation shown below is a model of these sentences. There are three objects in the universe of discourse. The object constants a and b map to two of these objects, and the relation constant r maps to a relation that is true of the third object only.

$$\forall^i = \{\circ, \bullet, \star\}$$
$$a^i = \circ$$
$$b^i = \bullet$$
$$r^i = \{\langle\star\rangle\}$$

The correspondinng Herbrand interpretation in this case is shown below. We have a universe of discourse with just two objects, and the relation corresponding to

$$\forall^h = \{a, b\}$$
$$a^h = a$$
$$b^h = b$$
$$r^h = \{\}$$

Unhappily, this interpretation is not a model of the sentences above. The third sentence requires that there be an object that satisfies r, and this interpretation does not contain any such object.

What's more, it is easy to see that there can be no model built on the Herbrand universe. Because of the first two sentences, the relation corresponding to r must not be true of these constants, and there is no other object in the universe of discourse to satisfy the third sentence.

The upshot of this is that the Herbrand method does not work for existential logic. It can be used to show that a set of sentences is satisfiable. However, the failure to satisfy all Herbrand interpretations does not mean that there are no interpretations that satisfy the sentences, as this example demonstrates.

§7.4 Functional Logic

Functional logic is the same as universal logic except that it includes functional terms. We are still not permitted to use quantifiers, though as with universal logic free variables are implicitly universally quantified.

Recall that the *Herbrand Universe* for a relational language is the set of all ground terms in the language. For universal logic, this is equivalent to the set of all object constants. For functional logic, we must also include complex terms that can be formed by applying function constants to ground terms. If there are no object constants, then we include an arbitrary object constant, say a.

As an example, consider a universal language with two object constants a and b. In this case, the Herbrand universe consists of just these two constants.

$$\{a, b\}$$

Now, let us add two unary function constants, say f and g. In this case, we must add additional terms, as shown below. Note that the Herbrand universe is infinite in this case, as it is for all functional languages.

$$\{a, b, f(a), f(b), g(a), g(b), f(f(a)), f(f(b)), f(g(a)), f(g(b)), \ldots\}$$

A *Herbrand interpretation* for functional logic is an interpretation in which (1) the universe of discourse is the Herbrand Universe for the language, (2) each object constant maps into itself, and (3) each function constant denotes a function that maps ground terms to themselves.

For example, the interpretation shown below is a Herbrand interpretation with objects constants a and b and unary function constants f and g. Each of the object constants maps into itself, and each function constant is interpreted as a function that maps terms into a new term comprised of that function constant applied to those terms.

$$\forall^i = \{a, b, f(a), f(b), g(a), g(b), f(f(a)), f(f(b)), f(g(a)), f(g(b)), \ldots\}$$
$$a^i = a$$
$$b^i = b$$
$$f^i = \{\langle a \rightarrow f(a), \rangle, \langle b \rightarrow f(b)\,\rangle, \langle f(a) \rightarrow f(f(a))\rangle, \ldots\}$$
$$g^i = \{\langle a \rightarrow g(a), \rangle, \langle b \rightarrow g(b)\,\rangle, \langle g(a) \rightarrow g(f(a))\rangle, \ldots\}$$
$$r^i = \{\langle a, a\rangle, \langle a, b\rangle\}$$

The interpretation shown below is also a Herbrand interpretation for this language. The object constants and function constants are interpreted as before, as required by the definition of Herbrand interpretation. The only thing different, the only thing that can be different is the interpretation of the relation constant r.

$$\forall^i = \{a, b, f(a), f(b), g(a), g(b), f(f(a)), f(f(b)), f(g(a)), f(g(b)), \ldots\}$$
$$a^i = a$$
$$b^i = b$$
$$f^i = \{\langle a \rightarrow f(a), \rangle, \langle b \rightarrow f(b)\,\rangle, \langle f(a) \rightarrow f(f(a))\rangle, \ldots\}$$
$$g^i = \{\langle a \rightarrow g(a), \rangle, \langle b \rightarrow g(b)\,\rangle, \langle g(a) \rightarrow g(f(a))\rangle, \ldots\}$$
$$r^i = \{\langle a, b\rangle, \langle b, a\rangle, \langle f(b), a\rangle, \langle f(b), b\rangle\}$$

As before, we define a *Herbrand variable assignment* as a variable assignment in which the universe is the Herbrand Universe for the language. The only difference is that the universe of discourse now contains more objects.

$$x^i = a$$
$$y^i = b$$
$$z^i = f(g(b))$$

The good news about functional logic is that, unlike existential logic, the Herbrand theorem still holds.

Functional Herbrand Theorem: *If a set of sentences in a functional language has a model, then it has a Herbrand model.*

As with universal logic, many people are surprised by this result. It is possible to have a finite model for a functional language; the universe of discourse of the corresponding Herbrand interpretation is infinitely large; yet the two interpretations satisfy the same sentences.

Unfortunately, in the case of functional logic, the result is not as useful as it is for universal logic. The size of the Herbrand universe for a functional language with is infinite. Consequently, checking the Herbrand interpretations for a language to determine logical entailment is not feasible.