Herbrand's Theorem

Herbrand's theorem constitutes a method for reducing the question of whether a single formula has a model to the satisfiability of a potentially infinite set of propositional formulas.

Example Let $L = \{R\}$, where R is a ternary relation symbol. Let A be the formula

$$A: \forall x \exists y \exists z \Big((\neg Rxyz \leftrightarrow Rxzy) \wedge (\neg Rxyz \leftrightarrow Rzyx) \wedge (\neg Rxyz \leftrightarrow Ryxz) \Big)$$

We wish to decide if A has a model, i.e., if A is true in some L-structure.

We fix an L-structure $\mathcal{M} = \langle M, R^{\mathcal{M}} \rangle$, which we suppose is a model of A.

Fix also some element $a \in M$.

Since $\mathcal{M} \models A$, we can substitute a for x in A and see ~~that~~ there are $a_0, a_1 \in M$ s.t.

$$\mathcal{M} \models (\neg Raa_0a_1 \longleftrightarrow Raa_1a_0) \wedge$$
$$(\neg Raa_0a_1 \longleftrightarrow Ra_1a_0a) \wedge$$
$$(\neg Raa_0a_1 \longleftrightarrow Ra_0aa_1).$$

Again, substituting a_0 for x, there are $a_{00}, a_{01} \in M$ s.t. $\mathcal{M} \models (\neg Ra_0a_{00}a_{01} \longleftrightarrow Ra_0a_{01}a_{00}) \wedge$

$$(\neg Ra_0a_{00}a_{01} \longleftrightarrow Ra_{01}a_{00}a_0) \wedge$$
$$(\neg Ra_0a_{00}a_{01} \longleftrightarrow Ra_{00}a_0a_{01})$$

Similarly, we can do the same for a_1 and find $a_{10}, a_{11} \in M$ with properties as above.

In general, repeating this construction, we can construct $a_s \in M$ for every finite binary sequence s, i.e., s is a finite sequence of 0's and 1's, such that for any such s,

$$\mathcal{M} \models (\neg Ra_sa_{s0}a_{s1} \longleftrightarrow Ra_sa_{s1}a_{s0}) \wedge$$
$$(\neg Ra_sa_{s0}a_{s1} \longleftrightarrow Ra_{s1}a_{s0}a_s) \wedge$$
$$(\neg Ra_sa_{s0}a_{s1} \longleftrightarrow Ra_{s0}a_sa_{s1})$$

Here a is identified with $a_\emptyset$, where $\emptyset$ is the empty sequence.

We can now use the information about a hypothetical model $\mathcal{M} \models R$ to actually construct a model $\mathcal{N} \models R$.

We let b_s, s a binary sequence, be new points and set

$$N = \{ b_s \mid s \text{ a binary seq.} \}.$$

So to make $\mathcal{N} = \langle N, \ldots \rangle$ into an L-structure, we need to interpret R in $\mathcal{N}$. That is, for all binary seq. s, t, u we need to decide whether

$$(b_s, b_t, b_u) \in R^{\mathcal{N}}.$$

So for every s, t, u, let $P_{s,t,u}$ be a new propositional variable. We seek a valuation v satisfying the following formulas:

$$\neg P_{s,s0,s1} \longleftrightarrow P_{s,s1,s0}$$

$$\neg P_{s,s0,s1} \longleftrightarrow P_{s1,s0,s}$$

$$\neg P_{s,s0,s1} \longleftrightarrow P_{s0,s,s1}$$

We thus see that A has a model if and only if there is such a valuation.

In the above case, we can let

$$v(P_{s,s0,s1}) = T$$

and $$v(P_{s,t,u}) = F \quad \text{for all other } (s,t,u).$$

Definition A formula A is *propositionally satisfiable* if $\neg A$ is not a tautology of first order logic. Similarly, $\{A_1, \ldots, A_n\}$ is *propositionally satisfiable* if $A_1 \wedge A_2 \wedge \ldots \wedge A_n$ is propositionally satisfiable. Finally, an infinite set of formulas is *prop. sat.* if all of its finite subsets are prop. sat.

Now, assume L is a countable language and let A be an L-sentence in prenex form. By adding superflous or "dummy" variables, without changing the logical equivalence class of A, we can suppose that

$A: \quad \forall x_1 \exists x_2 \forall x_3 \exists x_4 \ldots \forall x_{2k-1} \exists x_{2k} B[x_1, \ldots, x_{2k}]$

where B is quantifier free.

Since L is cbbl., the set T of L-terms and the set Θ of all finite sequences $(t_1, \ldots, t_m)$ of L-terms are both cbbl.

We fix an injection

$$\alpha : \Theta \hookrightarrow \mathbb{N}$$

such that

- if the variable x_i occurs in some of $t_1, \ldots, t_m$, then
$$\alpha(t_1, \ldots, t_m) > i$$
- if $j < i$ and $t_1, \ldots, t_i$ are terms, then
$$\alpha(t_1, \ldots, t_j) < \alpha(t_1, \ldots, t_i)$$

<u>Definition</u> A <u>manifestation</u> of the formula A above is any formula of the form:

$$B[t_1, x_{\alpha(t_1)}, t_2, x_{\alpha(t_1,t_2)}, \ldots, t_k, x_{\alpha(t_1,\ldots,t_k)}]$$

where $t_1, \ldots, t_k$ are arbitrary L-terms.

Note that since B is quantifier free, any manifestation of A is a Boolean combination of atomic L-formulas.

<u>Theorem</u> Suppose A is a formula not containing subformulas of the form "$s = t$". Then if the set $\mathbb{M}$ of all manifestations of A is propositionally satisfiable, A has a model.

<u>Proof</u> Let $\mathbb{A}$ be the set of all formulas

$$Rt_1 \dots t_n$$

where R is an n-ary relation symbol and $t_1, \dots, t_n$ are terms. Let also

$$\mathbb{B} = \{ Rt_1 \dots t_n \in \mathbb{A} \mid Rt_1 \dots t_n \text{ is a subformula of a manifestation of } A \}.$$

Suppose $\mathbb{M}$ is propositionally satisfiable. Then, by the compactness theorem for propositional logic, there is a valuation $v : \mathbb{B} \longrightarrow \{T, F\}$ such that $v(B') = T$ for any manifestation

$\mathcal{B}'$ of $\mathcal{A}$. We can extend v arbitrarily to $\mathbb{A}$ by, e.g., setting $v = T$ on $\mathbb{A} \setminus \mathbb{B}$.

Now for any term $t \in \mathcal{T}$, let $\bar{t}$ be a new object. We set $M = \{\bar{t} \mid t \in \mathcal{T}\}$ and define, for any $\bar{t}_1, \dots, \bar{t}_n \in M$ and n-ary relation symbol R,

$$(\bar{t}_1, \dots, \bar{t}_n) \in R^{\mathcal{M}} \iff v(Rt_1 \dots t_n) = T.$$

Also, if f is an n-ary function symbol, set

$$f^{\mathcal{M}}(\bar{t}_1, \dots, \bar{t}_n) = \overline{f(t_1, \dots, t_n)}.$$

We claim now that $\mathcal{M} \models A$.

To see this note first that by choice of v and construction of $\mathcal{M}$, we have for any $\bar{t}_1, \dots, \bar{t}_k \in M$

$$\mathcal{M} \models B[\bar{t}_1, \overline{x_{\alpha(t_1)}}, \bar{t}_2, \overline{x_{\alpha(t_1,t_2)}}, \dots, \bar{t}_k, \overline{x_{\alpha(t_1,\dots,t_k)}}].$$

By descending induction on $i = k, \dots, 1$, we show that for any $\bar{t}_1, \dots, \bar{t}_{i-1} \in M$

$$\mathcal{M} \models \forall x_{2i-1} \exists x_{2i} \dots \forall x_{2k-1} \exists x_{2k}$$

$$B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_{i-1}, \bar{x}_{\alpha(t_1, \dots, t_{i-1})}, x_{2i-1}, \dots, x_{2k}]$$

Case $i=k$

Fix $\bar{t}_1, \dots, \bar{t}_{i-1}$ and note that for any $\bar{t}_k \in M$, we have

$$\mathcal{M} \models B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_k, \bar{x}_{\alpha(t_1, \dots, t_k)}],$$

whence

$$\mathcal{M} \models \exists x_{2k} \, B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_k, x_{2k}]$$

and so

$$\mathcal{M} \models \forall x_{2k-1} \exists x_{2k} \, B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_{k-1}, \bar{x}_{\alpha(t_1, \dots, t_{k-1})}, x_{2k-1}, x_{2k}].$$

Induction step: If for any $\bar{t}_1, \dots, \bar{t}_i$ we have

$$\mathcal{M} \models \forall x_{2i+1} \exists x_{2i+2} \dots \forall x_{2k-1} \exists x_{2k}$$

$$B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_i, \bar{x}_{\alpha(t_1, \dots, t_i)}, x_{2i+1}, \dots, x_{2k}],$$

then also for any $\bar{t}_1, \dots, \bar{t}_i \in M$

$$\mathcal{M} \models \exists x_{2i} \forall x_{2i+1} \exists x_{2i+2} \dots \forall x_{2k-1} \exists x_{2k}$$

$$B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \dots, \bar{t}_i, x_{2i}, x_{2i+1}, \dots, x_{2k}]$$

and so for any $\bar{t}_1, \ldots, \bar{t}_{i-1} \in M$

$$\mathcal{M} \models \forall x_{2i-1} \exists x_{2i} \ldots\ldots \forall x_{2k-1} \exists x_{2k}$$

$$B[\bar{t}_1, \bar{x}_{\alpha(t_1)}, \ldots, \bar{t}_{i-1}, \bar{x}_{\alpha(t_1,\ldots,t_{i-1})}, x_{2i-1}, \ldots, x_{2k}].$$

This finishes the induction. So, for $i=1$, we have

$$\mathcal{M} \models \forall x_1 \exists x_2 \ldots \forall x_{2k-1} \exists x_{2k}\, B[x_1, x_2, \ldots, x_{2k}]$$

i.e., $\mathcal{M} \models A$. $\square$

<u>Theorem</u> Suppose A is a formula not containing subformulas of the form "$s=t$" and that the set $\mathbb{M}$ of all manifestations of A is not propositionally satisfiable. Then $\vdash \neg A$.

Lemma Suppose F, G are formulas and w is a variable which is not free in F. Then

$$\{\forall w\,(F \vee G)\} \vdash F \vee \forall w\, G$$

Lemma If B is a formula and w is a variable not appearing in B. Then

$$\vdash \forall w\; B[w/v] \rightarrow \forall v\, B$$

Proof:

$$\underbrace{\forall w\; B[w/v] \longrightarrow \underbrace{\left(B[w/v]\right)[v/w]}_{= B}}_{\text{axiom}}\;,\quad \underbrace{\forall v\left(\forall w\, B[w/v] \longrightarrow B\right)}_{\text{generalization}}\;,$$

$$\underbrace{\forall v\left(\forall w\, B[w/v] \longrightarrow B\right) \longrightarrow \left(\forall w\, B[w/v] \longrightarrow \forall v\, B\right)}_{\text{axiom}}\;,$$

$$\forall w\, B[w/v] \longrightarrow \forall v\, B\,.$$

$\square$

Proof Suppose that the set M of manifestations of A is not propositionally satisfiable.

So there are finitely many manifestations $B_1, B_2, \ldots, B_n$ of A s.t. $B_1 \wedge B_2 \wedge \ldots \wedge B_n$ is a propositional antinomy, i.e., s.t.

$$\neg B_1 \vee \neg B_2 \vee \ldots \vee \neg B_n \text{ is a tautology.}$$

$$\text{So} \quad \vdash \neg B_1 \vee \neg B_2 \vee \ldots \vee \neg B_n .$$

Let now $\mathbb{A}$ be the set of all formulas of the form

$$\forall w_{2i+1} \exists w_{2i+2} \ldots \forall w_{2k-1} \exists w_{2k}$$

$$B[t_1, x_{\alpha(t_1)}, t_2, x_{\alpha(t_1,t_2)}, \ldots, t_i, x_{\alpha(t_1,\ldots,t_i)}, w_{2i+1}, \ldots, w_{2k}]$$

where $0 \le i \le k$, $t_1, \ldots, t_i$ are terms, and the variables $w_{2i+1}, \ldots, w_{2k}$ do not appear in any of the terms $t_1, \ldots, t_i, x_{\alpha(t_1)}, \ldots, x_{\alpha(t_1,\ldots,t_i)}$.

Since $B_1, \ldots, B_n \in \mathbb{A}$, we see there is a finite subset $I = \{B_1, \ldots, B_n\} \subseteq \mathbb{A}$ s.t.

$$\vdash \bigvee_{C \in I} \neg C$$

We wish to find another finite subset $J \subseteq A$ s.t. $\vdash \bigvee_{C \in J} \neg C$ and s.t. the number of free variables in $\bigvee_{C \in J} \neg C$ is at least one fewer than in $\bigvee_{C \in I} \neg C$.

Let $C := \forall w_{2i+1} \exists w_{2i+2} \cdots \forall w_{2k-1} \exists w_{2k}$

$$B[t_1, x_{\alpha(t_1)}, \dots, t_i, x_{\alpha(t_1, \dots, t_i)}, w_{2i+1}, \dots, w_{2k}]$$

be a formula of A. Then, by choice of the function α, $\alpha(t_1, \dots, t_i)$ is the largest index of a <u>free</u> variable in C.

Suppose now that

$D := \forall z_{2j+1} \exists z_{2j+2} \cdots \forall z_{2k-1} \exists z_{2k}$

$$B[s_1, x_{\alpha(s_1)}, \dots, s_j, x_{\alpha(s_1, \dots, s_j)}, z_{2j+1}, \dots, z_{2k}]$$

is another formula in A such that

$$\alpha(t_1, \dots, t_i) = \alpha(s_1, \dots, s_j).$$

Then, as α is injective, we have $i = j$ and $t_1 = s_1, \ldots, t_i = s_i$. So C and D differ only in the specific names of their bound variables.

Thus, by the lemma, $\vdash C \longleftrightarrow D$. Therefore, if $C, D \in I$ we get

$$\vdash \bigvee_{E \in I \setminus \{D\}} \neg E$$

by the proof

$$\bigvee_{E \in I} \neg E, \quad C \longleftrightarrow D, \quad \underbrace{(C \longleftrightarrow D) \longrightarrow \Big(\bigvee_{E \in I} \neg E \longrightarrow \bigvee_{E \in I \setminus \{D\}} \neg E \Big)}_{\text{tautology of first order logic}},$$

$$\bigvee_{E \in I \setminus \{D\}} \neg E$$

So, by eliminating duplicates C, D as above, we can suppose that for any distinct $C, D \in I$,

$\gamma(C) :=$ maximal index of a free variable in C

$\neq$ ——————— " ——————— D.

Now choose the $C \in I$ with the largest value of $\alpha(C)$. Then $x_{\alpha(C)}$ is free in C but is not free in any other $D \in I$. Write

$$C: \ \forall w_{2i+1} \exists w_{2i+2} \cdots \forall w_{2k-1} \exists w_{2k}$$
$$B[t_1, x_{\alpha(t_1)}, \ldots, t_i, x_{\alpha(t_1,\ldots,t_i)}, w_{2i+1}, \ldots, w_{2k}].$$

So $\alpha(C) = \alpha(t_1,\ldots,t_i)$.

Since $\vdash \bigvee_{D \in I} \neg D$, we have by generalization

$$\vdash \forall x_{\alpha(t_1,\ldots,t_i)} \left(\bigvee_{D \in I} \neg D \right)$$

and since $x_{\alpha(t_1,\ldots,t_i)}$ is only free in C,

$$(*) \quad \vdash \bigvee_{D \in I \setminus \{C\}} \neg D \ \vee \ \forall x_{\alpha(t_1,\ldots,t_i)} \neg C$$

Let now w_{2i-1} be any variable not occurring in C and let $w_{2i} = x_{\alpha(t_1,\ldots,t_i)}$. Then if

$$C': \ \forall w_{2i-1} \exists w_{2i} \cdots \forall w_{2k-1} \exists w_{2k}$$
$$B[t_1, x_{\alpha(t_1)}, \ldots, t_{i-1}, x_{\alpha(t_1,\ldots,t_{i-1})}, w_{2i-1}, \ldots, w_{2k}]$$

the following is an axiom

$$C' \longrightarrow \exists w_{2i} C$$

($\exists w_{2i} C$ is obtained from C' by universal instantiation)

So, as $\vdash \exists w_{2i} C \longleftrightarrow \neg \forall w_{2i} \neg C$, we obtain

$$\vdash C' \longrightarrow \neg \forall w_{2i} \neg C$$

and thus $\vdash \forall w_{2i} \neg C \rightarrow \neg C'$

i.e.,

$$\vdash \forall x_{\alpha(t_1,\dots,t_i)} \neg C \longrightarrow \neg C'.$$

Combining with (*), we get

$$\vdash \bigvee_{D \in I \setminus \{C\}} \neg D \quad \vee \neg C'.$$

Letting $J = I \setminus \{C\} \cup \{C'\}$, we have found a subset $J \subseteq A$ s.t. $\vdash \bigvee_{D \in J} \neg D$ and $\bigvee_{D \in J} \neg D$ has one less free variable than $\bigvee_{D \in I} \neg D$.

Continuing this by induction, we eventually find $\mathcal{J} \subseteq \mathcal{A}$ finite s.t. $\vdash \bigvee_{D \in \mathcal{I}} \neg D$ and

$\bigvee_{D \in \mathcal{J}} \neg D$ has no free variables. So each $D \in \mathcal{J}$

is of the form

$$\forall w_1 \exists w_2 \ldots \forall w_{2k-1} \exists w_{2k} \; B[w_1, \ldots, w_{2k}] ,$$

where $w_1, \ldots, w_{2k}$ are variables. By the lemma we can change these variables to $x_1, \ldots, x_{2k}$, and obtain

$$A : \forall x_1 \exists x_2 \ldots \forall x_{2k-1} \exists x_{2k} \; B ,$$

whence

$$\vdash \underbrace{\neg A \vee \neg A \vee \ldots \vee \neg A}_{|\mathcal{I}| \text{ many}}$$

i.e.

$$\vdash \neg A . \qquad \square$$

Theorem If B is a formula w/o quantifiers and no subformula of the form "$s = t$" and if

$$A: \ \forall x_1 \exists x_2 \cdots \forall x_{2k-1} \exists x_{2k} B, \qquad (*)$$

where $B = B[x_1, \ldots, x_{2k}]$, then the following are equiv.

(a) A has no model

(b) there are instantiations $B_1, \ldots, B_n$ of A s.t.

$$\neg B_1 \vee \cdots \vee \neg B_n \quad \text{is a tautology},$$

(c) $\vdash \neg A$.

Suppose F is any formula w/o free variables and assume $\models F$, i.e., F is true in any model. Then $\neg F$ has no model. Now, by imitating the construction of a prenex form of $\neg F$, one can find a formula A as in $(*)$ logically equivalent to $\neg F$ and, moreover, s.t. $\vdash \neg A \leftrightarrow F$. So, as $\models F$, $\neg F$, and thus A, has no model, whence $\vdash \neg A$. It follows that $\vdash F$.

<u>Cor</u> For any formula F w/o free variables,

$$\models F \iff \vdash F.$$