by $\Sigma^b_1(L_{\mathrm{PV}})$-LIND.

Theorem 12.1.3 (Buss [106]) *The theory* $S^1_2(\mathrm{PV})$ *is* $\forall\Sigma^b_1(L_{\mathrm{PV}})$*-conservative over* PV_1.

The proof of the theorem relies on Buss's witnessing theorem and we shall not present it here (see [278]).

12.2 Herbrand's Theorem

Herbrand's theorem is a statement about witnessing existential quantifiers in logically valid first-order formulas of a certain syntactic form. Let L be an arbitrary first-order language (possibly empty) and let $A(x, y)$ be a quantifier-free L-formula. The simplest form of Herbrand's theorem says that if $\forall x \exists y A(x, y)$ is logically valid then there are terms $t_1(x), \ldots, t_k(x)$ such that

$$A(x, t_1(x)) \vee \cdots \vee A(x, t_k(x)) \tag{12.2.1}$$

is already logically valid. Note that even if $L = \emptyset$ we have terms: the variables. This can be interpreted as saying that we can compute a witness for y from a given argument x by one of the k terms but not necessarily the same term for all x.

Many results in proof theory have a simple rudimentary version and also a number of more or less (often more rather than less) technically complicated stronger variants. Herbrand's theorem is no exception. The technically more difficult versions are, for example, those describing how to find the terms t_i from *any* first-order proof of $\forall x \exists y A(x, y)$ or formulations for formulas A that are not quantifier-free or are not even in a prenex form (the most cumbersome variant). Fortunately, we will not need these difficult results but only a slight extension of the above informal formulation.

The key difference of the variant we need from the one given above is that the *logical validity* of (12.2.1) is replaced by *propositional validity*. We define a quantifier-free L-formula B to be **propositionally valid** if and only if any assignment of propositional truth values 0, 1 to atomic formulas in B evaluates the whole formula to 1. The only requirement in this assignment is that the same atomic formulas get the same value, but the assignment has a priori no connection with the Tarski truth definition in some L-structure.

Not all logically valid formulas are also propositionally valid. For example, none of the equality axioms

$$x = x, \quad x = y \to y = x, \quad (x = y \wedge y = z) \to x = z \tag{12.2.2}$$

is propositionally valid, and neither are the equality axioms for a relation symbol $R(\bar{x})$ or for a function symbol $f(\bar{x})$ from L:

$$Eq_R(\bar{x}, \bar{y}) : \ \bigwedge_i x_i = y_i \ \to \ R(\bar{x}) \equiv R(\bar{y}) \tag{12.2.3}$$

and

$$Eq_f(\bar{x},\bar{y}) : \bigwedge_i x_i = y_i \rightarrow f(\bar{x}) = f(\bar{y}). \tag{12.2.4}$$

For example, you can give to $x = y$ the value 1 and to $y = x$ the value 0, or to all $x_i = y_i$ the values 1 and to $f(\bar{x}) = f(\bar{y})$ the value 0.

One advantage of Herbrand's theorem over the cut-elimination procedure is that one can give its *complete* proof using only the compactness of propositional logic.

Theorem 12.2.1 (Herbrand's theorem) *Let L be an arbitrary first-order language and let $A(\bar{x},\bar{y})$ be a quantifier-free formula. Assume that $\forall\bar{x}\exists\bar{y}A(\bar{x},\bar{y})$ is logically valid.*

Then there are, for $e \geq 0$ and $k \geq 1$,

- *equality axioms $Eq_j(\bar{u},\bar{v})$, $j \leq e$, of the form (12.2.2), (12.2.3) or (12.2.4) for some symbols of L,*
- *tuples of terms $\bar{r}^i_j(\bar{x})$, $\bar{s}^i_j(\bar{x})$ and $\bar{t}^i(\bar{x})$, for $j \leq e$, $v \leq a$ and $i \leq k$,*

such that

$$(\bigvee_{i,j} \neg Eq_j(\bar{r}^i_j(\bar{x}),\bar{s}^i_j(\bar{x}))) \vee \bigvee_i A(\bar{x},\bar{t}^i(\bar{x})) \tag{12.2.5}$$

is propositionally valid.

Proof Assume for the sake of contradiction that the conclusion of the theorem is not true. Consider a theory T consisting of

- all instances of all the equality axioms $Eq(\bar{u},\bar{v})$ of the three forms (12.2.2), (12.2.3) and (12.2.4) for all symbols of L, and, for all tuples of terms $\bar{r}(\bar{x}),\bar{s}(\bar{x})$,

$$Eq(\bar{r}(\bar{x}),\bar{s}(\bar{x})),$$

- all instances of $\neg A$ for all tuples of terms $\bar{t}(\bar{x})$,

$$\neg A(\bar{x},\bar{t}(\bar{x})).$$

Claim 1 T is propositionally satisfiable. That is, it is possible to assign to all atomic formulas occurring in T propositional truth values such that all formulas in T become satisfied.

This is the place where we will use the compactness of propositional logic. If T is not propositionally satisfiable, already some finite $T_0 \subseteq T$ is not. But that would mean that a disjunction of the negations of formulas in T_0 is propositionally valid. But such a disjunction is of the form (12.2.5), contradicting our assumption.

Now we define a first-order L-structure in which the sentence $\forall\bar{x}\exists\bar{y}A(\bar{x},\bar{y})$ fails. Let h be a truth assignment to atomic formulas occurring in T that makes all formulas in T true. Let A be the set of all L-terms $w(\bar{x})$. On A define the relation

$$u \sim v \quad \text{if and only if} \quad h(u = v) = 1\ .$$

Owing to the (instances of the) equality axioms (12.2.2) in T it is an equivalence relation. In fact, owing to the (instances of the) equality axioms (12.2.3) and (12.2.4) it is a congruence relation for all symbols of L: the axioms Eq_R and Eq_f hold with $\sim$ in place of $=$.

This determines an L-structure $\mathbf{B}$ with the universe B consisting of all $\sim$-blocks $[u]$ of $u \in A$, $B := A/\sim$, and with L interpreted on B via T:

$$\mathbf{B} \models R([u_1], \ldots, [u_n]) \quad \text{if and only if} \quad h(R(u_1, \ldots, u_n)) = 1$$

and analogously for all function symbols f.

Claim 2 For all quantifier-free L-formulas $C(z_1, \ldots, z_n)$ and all $[u_1], \ldots, [u_n] \in B$, we have

$$\mathbf{B} \models C([u_1], \ldots, [u_n]) \quad \text{if and only if} \quad h(C(u_1, \ldots, u_n)) = 1.$$

The claim is readily established by the logical complexity of C and it implies that

$$\mathbf{B} \models \neg\exists\bar{y} A([x_1], \ldots, y_1, \ldots),$$

contradicting the hypothesis of the theorem. □

From Theorem 12.2.1, we get, without any additional effort, a similar statement for the consequences of **universal theories**, theories all of whose axioms are universal sentences of the form

$$\forall\bar{z} B(\bar{z}),$$

where B is quantifier-free.

Corollary 12.2.2 *Let L be an arbitrary first-order language and let T be a universal L-theory. Let $A(\bar{x}, \bar{y})$ be a quantifier-free formula and assume that $\forall\bar{x}\exists\bar{y} A(\bar{x}, \bar{y})$ is provable in T, i.e. that it is valid in all models of T.*

Then there are $e, a \geq 0$, $k \geq 1$,

- *equality axioms $Eq_j(\bar{u}, \bar{v})$, $j \leq e$, of the form* (12.2.2)*,* (12.2.3) *or* (12.2.4) *for some symbols of L,*
- *axioms $\forall\bar{z} B_u(\bar{z}) \in T$, $u \leq a$,*
- *tuples of terms $\bar{r}^i_j(\bar{x})$, $\bar{s}^i_j(\bar{x})$ and $\bar{w}^i_v(\bar{x})$ and $\bar{t}^{\,i}(\bar{x})$, for $j \leq e$ and $i \leq k$,*

such that

$$(\bigvee_{i,v} \neg B_v(\bar{w}^i_v(\bar{x}))) \vee (\bigvee_{i,j} \neg Eq_j(\bar{r}^i_j(\bar{x}), \bar{s}^i_j(\bar{x}))) \vee \bigvee_i A(\bar{x}, \bar{t}^{\,i}(\bar{x})) \tag{12.2.6}$$

is propositionally valid.

Proof If T proves the formula, already a finite number of axioms $\forall\bar{z} B_v(\bar{z})$, $v \leq a$, from T suffices. Apply Theorem 12.2.1 to the formula

$$\forall\bar{x}\exists\bar{y}, \bar{z}_1, \ldots, \bar{z}_a \, A(\bar{x}, \bar{y}) \vee \bigvee_v \neg B_v(\bar{z}_v).$$

□

We have formulated a version of Herbrand's theorem having propositional validity because that is what we shall use in proving simulations of theories by proof systems. But now we formulate two corollaries of the theorem just in terms of first-order provability in universal theories. These serve as witnessing theorems, and they will be used in some arguments later on.

For simplicity of notation we shall consider just single quantifiers rather than blocks of similar quantifiers (this is without loss of generality).

Corollary 12.2.3 *Let T be a universal theory in a language L and let $\forall x \exists y A(x, y)$, where A is quantifier-free, be provable in T.*

Then there are $k \geq 1$ and L-terms $t_i(x)$, $i \leq k$, such that T proves

$$\bigvee_{i \leq k} A(x, t_i(x)). \tag{12.2.7}$$

Proof First-order logic includes the equality axioms and hence T proves that all instances of all such axioms, as well as of its own axioms, are true. What remains from the disjunction (12.2.6) in Corollary 12.2.2 is given by (12.2.7). □

Let us now assume that our formula is more complex than just $\forall\exists$, say it is a $\forall\exists\forall$-formula, i.e. a formula of the form

$$\forall x \exists y \forall z D(x, y, z), \tag{12.2.8}$$

with D quantifier-free. Let $h(x, y)$ be a new binary function symbol *not* in L. It is often called a **Herbrand function**. Then (12.2.8) is logically valid if and only if

$$\forall x \exists y D(x, y, h(x, y)) \tag{12.2.9}$$

is logically valid; in fact, a theory T in the language L proves (12.2.8) if and only if it proves (12.2.9). It is clear that the validity of the former in an L-structure implies the validity of the latter. But the opposite is also true in the following sense: if (12.2.8) were not true then there would be an a in the structure such that for each b we can find a c there such that $\neg D(a, b, c)$; hence, taking for $h(a, b)$ one such c, interprets the Herbrand function in a way such that (12.2.9) fails.

Combining this reasoning with Corollary 12.2.3 yields the next statement.

Corollary 12.2.4 (The KPT theorem [323]) *Let T be a universal theory in a language L and let $\forall x \exists y \forall z D(x, y, z)$ be provable in T where D is quantifier-free.*

Then there are $k \geq 1$ and L-terms

$$t_1(x), t_2(x, z_1), \ldots, t_k(x, z_1, \ldots, z_{k-1})$$

such that T proves

$$D(x, t_1(x), z_1) \vee D(x, t_2(x, z_1), z_2) \vee \cdots \vee D(x, t_k(x, z_1, \ldots, z_{k-1}), z_k). \tag{12.2.10}$$

Proof Think of T as a theory in the language $L \cup \{h\}$, with h the symbol for the Herbrand function corresponding to the formula. Then the hypothesis of the theorem

implies by Herbrand's theorem that T proves (12.2.9) and, hence, a disjunction of the form

$$\bigvee_{i\le k} D(x, t_i'(x), h(x, t_i'(x))). \tag{12.2.11}$$

Modify this disjunction as follows. Find a subterm s occurring in (12.2.11) that starts with the symbol h and that has the maximum size among all such terms. It must be one of the terms $h(x, t_i'(x))$ sitting at a position z in one of the disjuncts of (12.2.11); say it is $h(x, t_k'(x))$. Replace all its occurrences in the disjunction by a new variable z_k. This maneuver clearly preserves the validity on all structures for $L \cup \{h\}$ that are models of T because we can interpret h arbitrarily. Note that by choosing the maximum-size subterm we know that it does not occur in any t_i' with $i < k$.

Now choose the next to maximum size subterm s of the required form and replace it everywhere by z_{k-1}. The subterm s is either in t_k' in which case we have just simplified t_k' but have not changed anything else, or it may be one of the $h(x, t_i'(x))$, say $h(x, t_{k-1}'(x))$. The subterm s then does not occur in any t_i' for $i < k-1$ but it may still occur in t_k'. This will transform t_k' into a term $t_k''(x, z_{k-1})$ that may depend also on z_{k-1}. Hence the last two disjuncts on the disjunction will look like

$$\cdots \vee D(x, t_{k-1}'(x), z_{k-1}) \vee D(x, t_k''(x, z_{k-1}), z_k)$$

for some term t_k'' with the variables shown.

Repeat this process as long as there is any occurrence of the symbol h. □

There is a nice interpretation of the disjunction (12.2.3) in terms of a two-player game, the so-called **Student–Teacher game**. Assume that

$$\forall x \exists y \forall z D(x, y, z)$$

is valid in an L-structure (this is usually applied to the standard model, so we may consider that the formula is true). Consider a game between Student and Teacher proceeding in rounds. They both receive some $a \in \{0,1\}^*$, and the task of Student is to find $b \in \{0,1\}^*$ such that $\forall z D(a, b, z)$ is true. They play as follows.

- In the first round Student produces a candidate solution b_1. If $\forall z D(a, b_1, z)$ is true then Teacher says so. Otherwise she gives Student a *counter-example*: some $c_1 \in \{0,1\}^*$ such that $\neg D(a, b_1, c_1)$ holds.
- Generally, before the ith round, $i \ge 2$, Student has suggested solutions $b_1, \ldots, b_{i-1}$ and has received counter-examples $c_1, \ldots, c_{i-1}$. He sends a new candidate solution b_i and Teacher either accepts it or sends her counter-example.

The play may continue for a fixed number of rounds or for an unlimited number, as prearranged. Student wins if and only if he finds a valid solution.

Assume that the disjunction (12.2.11) is valid. Student may use the terms t_i as his strategy: in the first round he sends

$$b_1 := t_1(a_1).$$

If that is incorrect and he gets c_1 as a counter-example, he sends

$$b_2 := t_2(a, c_1)$$

in the second round and similarly in the later rounds. But, because (12.2.11) is valid in the structure, in at most the kth round his answer must be correct. Hence we get

Corollary 12.2.5 *Let T be a universal theory in a language L and let $\forall x \exists y \forall z\, D(x, y, z)$, where D is quantifier-free, be provable in T. Let* **M** *be any model of T.*

Then there are $k \geq 1$ and L-terms $t_i(x, z_1, \ldots, z_{i-1})$, $i \leq k$, such that Student has a winning strategy for the Student–Teacher game, associated with the above formula over **M***, such that he wins in at most k rounds for every a. Moreover, his strategy is computed by the terms $t_1, \ldots, t_k$ as described above.*

12.3 The ‖ . . . ‖ Translation

In Section 1.4 we presented the set of clauses $\mathrm{Def}_C(\bar{x}, \bar{y})$ that define a circuit C: the set is satisfied if and only if $\bar{y}$ is the computation of C on an input $\bar{x}$. We considered there only circuits with one output, the last y-bit. Now we need to extend this notation to circuits which output multiple bits (i.e. strings). It will also be convenient to consider circuits with multiple string inputs (rather than combining one-string inputs). By $\mathrm{Def}_C(\bar{x}_1, \ldots, \bar{x}_r; \bar{y}, \bar{z})$ we denote the set of clauses whose conjunction means that $\bar{y}$ is the computation of C on the inputs $\bar{x}_i$ with output string $\bar{z}$.

The following statement formalizes in resolution the fact that the computations of circuits are uniquely determined by the inputs; it is easily proved by induction on the size of the circuit.

Lemma 12.3.1 *Let C be a size s circuit. Then there are size $O(s)$ resolution derivations of*

$$y_j \equiv u_j \quad \textit{and} \quad z_i \equiv v_i, \textit{ for all } i, j,$$

from the initial clauses

$$\mathrm{Def}_C(\bar{x}_1, \ldots, \bar{x}_r; \bar{y}, \bar{z}) \cup \mathrm{Def}_C(\bar{x}_1, \ldots, \bar{x}_r; , \bar{u}, \bar{v}).$$

Recall that $L_{BA}(\mathrm{PV})$ is the language L_{BA} of bounded arithmetic augmented by all function symbols of L_{PV}; in particular, # is among them. Our aim is to define for all sharply bounded (i.e. Σ_0^b) $L_{BA}(\mathrm{PV})$-formulas $A(x_1, \ldots, x_k)$ a sequence of propositional formulas

$$\|A(x_1, \ldots, x_k)\|^{n_1, \ldots, n_k}$$

with the property that the formula is a tautology if and only if

$$\forall x_1 (|x_1| = n_1) \ldots \forall x_k (|x_k| = n_k)\, A(x_1, \ldots, x_k)$$

is true in **N**.