On Gabbay's proof of the Craig interpolation theorem for intuitionistic predicate logic

by Michael Makkai*
(McGill University)

Abstract. Using the framework of categorical logic, we analyze and streamline D. M. Gabbay's proof in Gabbay [1969] mentioned in the title. In the process, we find an apparently new and interesting fact about the relation of coherent and intuitionistic logic.

§1. Introduction.

For any logical system with entailment relation $\vdash$, by saying that the Craig interpolation theorem holds we mean the following:

for any sentences φ_1, φ_2 over the respective vocabularies $\mathcal{L}_1$, $\mathcal{L}_2$, if $\varphi_1 \vdash \varphi_2$, then there is a sentence φ over $\mathcal{L}_1 \cap \mathcal{L}_2$, the common part of the given vocabularies, such that $\varphi_1 \vdash \varphi \vdash \varphi_2$.

The interpolation theorem for classical predicate logic was found by W. Craig [1959]; his proof used a proof-theoretic approach. S. Maehara [1960/61] gave the well-known proof through Gentzen's sequent calculus. K. Schütte [1962] extended the theorem to intuitionistic logic with a similar proof-theoretical proof. (As J. Bell and Machover [1977] point out (see Chapter 9, §13), the theorem for classical logic can be deduced from the intuitionistic version by a use of the double negation interpretation.)

The model theory of classical logic gives several illuminating proofs of the interpolation theorem for classical logic. It is customary (see Chang and Keisler [1973]) to reduce the result

---

*Supported by NSERC Canada and FCAR Quebec

to Robinson's Consistency lemma, originally used by A. Robinson [1956] for a proof of Beth's definability theorem (Beth [1953]), a consequence of Craig interpolation. For the proof of the latter, Robinson used the method of alternating chains of model-extensions. Gabbay's proof of interpolation for intuitionistic logic in Gabbay [1969] exploits the same method. Gabbay's proof is based on Kripke's semantic analysis of intuitionistic logic in Kripke [1965].

A very elegant and conceptually satisfactory proof of interpolation for intuitionistic logic is given by A. M. Pitts [1983]; he uses locale-theoretic methods, and his work is inspired by the ground-breaking work Joyal and Tierney [1984] (available in preprint form earlier) that introduced locales into topos theory (and thus logic). Pitts' proof is constructive as opposed to Gabbay's which is model-theoretic, and uses non-constructive tools such as the compactness theorem. On the other hand, as we intend to show, Gabbay's proof shows something stronger than the interpolation theorem (we have nothing to say here on the question whether Pitts' methods, or other constructive methods, could yield the same additional conclusion).

Let us recall some basic concepts of categorical logic, intended to capture the notion of theory in intuitionistic, and coherent (see below), predicate logic. A *Heyting category* is a category in which

1. all finite limits exist;

2. the poset $\mathrm{Sub}(A)$ of subobject of every object A is a lattice (with 0 and 1);

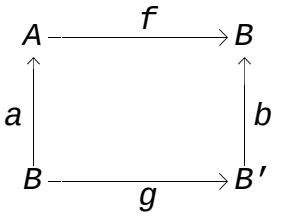
3. for any arrow $f: A \to B$, the map $f^*: \mathrm{Sub}(B) \to \mathrm{Sub}(A)$ defined as taking pullback (inverse image) is a lattice homomorphism;

4. f^* as in 3. has both left and right adjoints

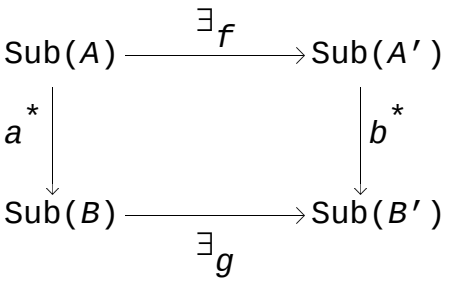
$$\mathrm{Sub}(A) \underset{\forall_f}{\overset{\exists_f}{\underset{\longleftarrow}{\overset{\longrightarrow}{f^*}}}} \mathrm{Sub}(B) \; : \; \exists_f \dashv f^* \dashv \forall_f ;$$

and

5. $\exists_f$ (and $\forall_f$) are stable under pullback (Beck-Chevalley condition): if



is a pullback square, then



commutes.

(The analogous condition for $\forall_f$'s follows.)

The notion of Heyting category embodies that of a theory in multi-sorted intuitionistic predicate logic with equality. This correspondence is explained in Makkai and Reyes [1977], and although Heyting categories are not explicitly introduced in *loc.cit.*, all the categorical operations involved are. In particular, given a theory $(\mathcal{L},\mathcal{T})$ in intuitionistic logic, with vocabulary $\mathcal{L}$ and set of axioms $\mathcal{T}$, we have the corresponding Heyting category $\mathtt{LT}(\mathcal{L},\mathcal{T})$, or $\mathtt{LT}_{\mathtt{Heyting}}(\mathcal{L},\mathcal{T})$ more explicitly, with the letters $\mathtt{L}$ and $\mathtt{T}$ referring to the fact that the construction of $\mathtt{LT}(\mathcal{L},\mathcal{T})$ is like the Lindenbaum-Tarski construction, originally of a Boolean algebra, based on formulas of the theory taken as a basis; see Chapter 8 of *loc.cit.*

We may, for the purposes of the Craig interpolation theorem essentially equivalently, use Heyting pretoposes instead as it is done in Pitts [1983]. In fact, every Heyting pretopos is a Heyting category (although not conversely); however, Heyting categories correspond more directly to the usual symbolic logical notion of theory.

A *Heyting functor* is one between Heyting categories preserving in the usual sense the operations posited in Heyting categories. The various notions of *model* used in intuitionistic logic are *the same*, modulo the correspondence of categories and theories, as Heyting functors with a specific kind of target category. E.g., Kripke models ("modified Kripke structures" in

Gabbay [1981]) of $(\mathcal{L},\mathcal{T})$ correspond to Heyting functors from $\mathtt{LT}(\mathcal{L},\mathcal{T})$ into $\mathtt{Set}^{\mathrm{P}}$, with any poset P ($\mathtt{Set}^{P}$ is the category of functors from P (as a category) into $\mathtt{Set}$, the category of (small) sets); $\mathtt{Set}^{P}$ is a topos, in particular, a Heyting category).

Heyting categories and functors, with all natural transformations between the latter as 2-arrows, form a 2-category called $\mathcal{HEYTING}$; restricting objects to small ones gives the 2-category $\mathcal{H}eyting$. The 2-categorical aspects will play a minor role only; most of the time it is enough to keep in mind the ordinary categories involved.

In fact, the 2-category of Heyting categories that is "good" from an algebraic point of view is the one in which only isomorphism 2-arrows are allowed; by denoting the result of discarding 2-arrows that are not isomorphism by postfixing an asterisk, we are talking about the 2-categories $\mathcal{HEYTING}^{*}$, $\mathcal{H}eyting^{*}$.

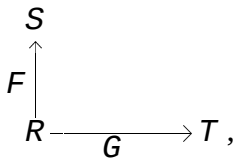
Every Heyting category is a *coherent category*. The latter were called "logical categories", and they play a central role in Makkai and Reyes [1977]. The definition is obtained by deleting reference to $\forall_f$ in 4. (and 5.) in the definition of Heyting category. As it is explained in *loc.cit.*, coherent categories correspond to *coherent theories*. In coherent logic formulas are positive existential (*coherent*). that is, built up of atomic formulas using $\wedge$, $\vee$ and $\exists$, and *axioms* are all of the form $\forall\vec{x}(\varphi(\vec{x}) \longrightarrow \psi(\vec{x}))$ with φ and ψ coherent. In this paper, all the "action" around a Heyting category will take place in the *coherent doctrine*: we will deal with properties of Heyting categories *qua* coherent categories, properties that are not generally true for coherent categories, but hold if the coherent category carries the additional structure of a Heyting category as well. In this way, we "explain" intuitionistic logic via the simpler coherent logic.

A *Boolean* category is a coherent one in which each subobject lattice is complemented (a Boolean algebra); every Boolean category is a Heyting one at the same time. Boolean categories correspond to theories in full classical logic in the usual sense (see *loc.cit.*).

An important point in our story is that the interpolation property of a logical system such as classical logic, intuitionistic logic, and others as well, is equivalent to an exactness property of the (2-) category of the corresponding kind of categories: the 2-category of Boolean categories, resp. Heyting categories. The exactness property in question refers to *pushouts*. For the record,

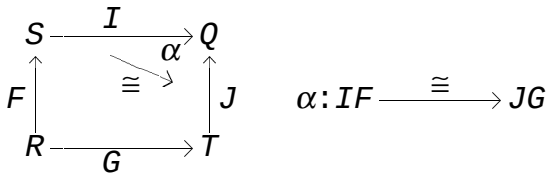
we give the definition of the 2-categorical version of pushout, also called bipushout (see Street [1980]); however, the reader will not go far wrong if he takes the notion to be the ordinary categorical notion, especially in view of the remarks to follow later.

Given a 2-category, and a diagram

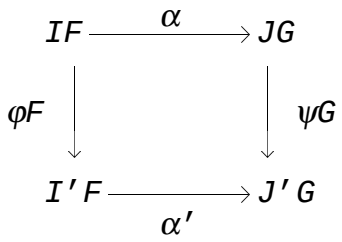


(*)

and any object Q, an extension of the form



is called (temporarily) a Q-*cocone on* (*). Q-cocones on a given triangle (*) and a fixed Q form a category; with another Q-cocone, with data denoted by primed letters, an arrow $(Q, I, J, \alpha) \longrightarrow (Q, I', J', \alpha')$ is a pair $(\varphi: I \longrightarrow I', \psi: J \longrightarrow J')$ such that the square of 2-arrows



commutes; composition of these arrows is defined in the expected way. The category of Q-cocones on the fixed (*) is denoted $\mathtt{Cocone}_Q$ (with reference to (*) suppressed). Note that for any two objects P and Q, a P-cocone $\Gamma = (P, I, J, \alpha)$ and an arrow $H: P \longrightarrow Q$ gives a Q-cocone $\Gamma^*(h) = (Q, HI, HJ, H\alpha)$, and that, in fact, we obtain a functor

$$\Gamma^*: \mathtt{Hom}(P, Q) \longrightarrow \mathtt{Cocone}_Q .$$

We say that Γ is a *pushout of* (*) if for all objects Q, the functor Γ^* is an *equivalence* of categories. Pushouts, when they exist, are determined up to *equivalence*.

The 2-categories $\mathcal{H}eyting^*$, $\mathcal{B}oole^*$, $\mathcal{C}oherent$, are all "good"; they are locally finitely presentable 2-categories (unfortunately, the 2-categorical versions of the well-known concepts of Gabriel and Ulmer [1971] have not been written down in the literature, despite, or maybe because, of the fact that the work to be done is straightforward), and as a consequence, pushouts exist in them. In fact, they can be so constructed that the canonical equivalences Γ^* be surjective on objects (although they still fall short of being isomorphisms), which somewhat simplifies the notation when working with them. A further simplification of the notion of pushout is to require that the structure map $\alpha: IF \longrightarrow JG$ in the pushout-cocone be an identity; in particular that $IF = JG$; in all our examples it is in fact possible to achieve this. When done so, a pushout in the 2-category is almost a pushout in the corresponding ordinary category (without the 2-arrows), except that the uniqueness part of the universal condition holds with the qualification "up to an isomorphism 2-arrow". Note however that the definition of the 2-categorical pushout has 2-dimensional aspects that go beyond the defining properties of the ordinary pushout.

In our proofs, we will never have the need to use the explicit constructions of pushouts; their universal properties will suffice. Nevertheless, it is important to point out that pushouts have a well-known meaning when our categories are given via theories. Suppose we have the intuitionistic theories $(\mathcal{L}_0,\mathcal{R})$, $(\mathcal{L}_1,\mathcal{S})$, $(\mathcal{L}_2,\mathcal{T})$ with $\mathcal{L}_0 \subset \mathcal{L}_1$, $\mathcal{L}_0 \subset \mathcal{L}_2$, $\mathcal{R} \subset \mathcal{S}$ and $\mathcal{R} \subset \mathcal{T}$; moreover, assume that the vocabularies $\mathcal{L}_1$, $\mathcal{L}_2$ have nothing more common than the elements of $\mathcal{L}_0$, $\mathcal{L}_1 \cap \mathcal{L}_2 = \mathcal{L}_0$. We then have the pushout diagram

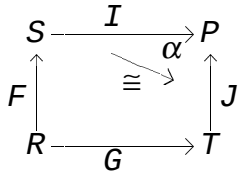
$$\begin{array}{ccc} \mathbb{LT}(\mathcal{L}_1,\mathcal{S}) & \longrightarrow & \mathbb{LT}(\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{S} \cup \mathcal{T}) \\ \uparrow & & \uparrow \\ \mathbb{LT}(\mathcal{L}_1 \cap \mathcal{L}_2, \mathcal{R}) & \longrightarrow & \mathbb{LT}(\mathcal{L}_2,\mathcal{T}) \end{array}$$

in which all arrows are induced by inclusions of vocabularies. This can be seen by applying the universal property of $\mathbb{LT}(\mathcal{L},\mathcal{T})$ as the free Heyting category in which $(\mathcal{L},\mathcal{T})$ is interpreted; see, for the analogous statement of this universal property for the coherent doctrine rather than

the Heyting one, Proposition 8.2.3 (p.247) in Makkai and Reyes [1977].

Consider the following property of any 2-category, the *Strong Amalgamation Property*:

(SAP) If in the pushout



F is conservative, then so is J.

(In any of our "concrete" 2-categories, "conservative" means "reflects isomorphisms"; conservative functors correspond to conservative extensions of theories.)

In the main part of this paper, we will be giving a proof of the

Theorem. Both $\mathcal{B}oole^*$ and $\mathcal{H}eyting^*$ have the SAP.

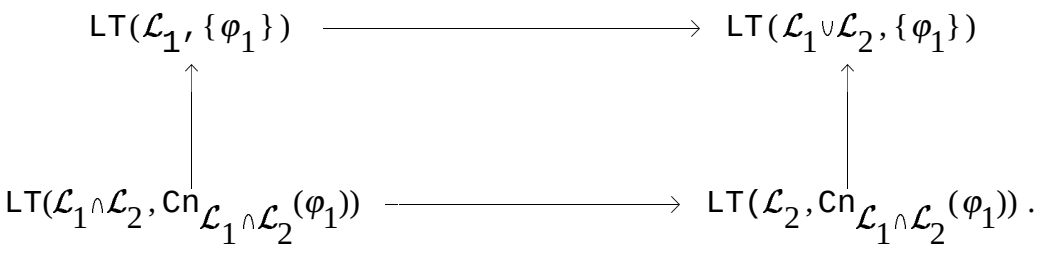
It had long been known that the SAP is equivalent to the Craig interpolation theorem. For one thing, A. Robinson's lemma, which is used in the usual model-theoretical proofs of interpolation, is a special case of SAP for $\mathcal{B}oole$: with reference to the theories from which a pushout diagram is derived, Robinson's lemma says that if $(\mathcal{L}_1 \cap \mathcal{L}_2, \mathcal{R})$ is complete, and $(\mathcal{L}_1, \mathcal{S})$, $(\mathcal{L}_2, \mathcal{T})$ are consistent, then $(\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{S} \cup \mathcal{T})$ is consistent. This follows, since by $(\mathcal{L}_1 \cap \mathcal{L}_2, \mathcal{R})$ being complete and $(\mathcal{L}_1, \mathcal{S})$ consistent, F is conservative; by SAP , J is conservative; since $(\mathcal{L}_2, \mathcal{T})$ is consistent, its conservative extension is consistent.

Nevertheless, I do not know any other place than Pitts' above mentioned paper Pitts [1983] where the SAP is explicitly established, and interpolation is derived from it, for the intuitionistic doctrine. In fact, Pitts goes beyond this, and shows that any pushout square has the interpolation property in an appropriate sense provided SAP holds.

Let me briefly indicate how the interpolation theorem follows from SAP , in a simpler way

than through Robinson's lemma.

With reference to the notation for Craig's theorem above, consider the square



Here, $\mathtt{Cn}_{\mathcal{L}_1 \cap \mathcal{L}_2}(\varphi_1)$ denotes the set of consequences of φ_1 over the vocabulary $\mathcal{L}_1 \cap \mathcal{L}_2$, that is, the set of sentences θ over $\mathcal{L}_1 \cap \mathcal{L}_2$ such that $\varphi_1 \vdash \theta$. The arrows are all induced by inclusions of vocabularies. The diagram is a pushout by the above more general situation: note that $\{\varphi_1\} \cup \mathtt{Cn}_{\mathcal{L}_1 \cap \mathcal{L}_2}(\varphi_1)$ has the same deductive strength as $\{\varphi_1\}$. Now, observe that the left vertical arrow is conservative: if the sentence θ over $\mathcal{L}_1 \cap \mathcal{L}_2$ is provable from φ_1 (that is, has become 1 in the left upper corner), then it *belongs to* $\mathtt{Cn}_{\mathcal{L}_1 \cap \mathcal{L}_2}(\varphi_1)$, thus it is provable from the same set of axioms. By SAP, the right vertical is conservative. Since by the assumption of the interpolation-statement, φ_2 is provable from φ_1, it follows that it is provable from $\mathtt{Cn}_{\mathcal{L}_1 \cap \mathcal{L}_2}(\varphi_1)$, which clearly implies the desired conclusion.

To continue with an account of the necessary background, let us consider a coherent category T, and let $\mathtt{Mod}(T)$ denote the category of all coherent functors $T \longrightarrow \mathtt{Set}$, with all natural transformations as arrows: $\mathtt{Mod}(T) = \mathcal{COHERENT}(T, \mathtt{Set})$. We have the evaluation functor

$$\begin{aligned} \mathtt{e}_T : T &\longrightarrow \mathtt{Set}^{\mathtt{Mod}(T)} \\ A &\longmapsto [X \mapsto M(A)], \end{aligned}$$

and also its variant

$$|\mathtt{e}_T| : T \longrightarrow \mathtt{Set}^{|\mathtt{Mod}(T)|}$$

where in the second case we disregard arrows in $\mathtt{Mod}(T)$ (thus, $\mathtt{Set}^{|\mathtt{Mod}(T)|}$ is merely a (large) Cartesian power of $\mathtt{Set}$). Now, the categorical version of the Gödel completeness

theorem says that, for a small T (as will be assumed throughout), $|\mathbf{e}_T|$ is a conservative functor (see e.g. Makkai [1985], Theorem 2.2.2, p. 254). In fact, this is obviously equivalent to saying that $\mathbf{e}_T$ is conservative. Whereas $\mathbf{Set}^{|\mathbf{Mod}(T)|}$ is a Boolean category, $\mathbf{Set}^{\mathbf{Mod}(T)}$ is not; however, the latter is a Heyting category. A fundamental result of categorical logic is

Joyal's theorem. If T is a (small) Heyting category, then $\mathbf{e}_T$ is a Heyting functor.

Note that $\mathbf{Mod}(T)$ still has the same sense: the category of coherent functors from T to $\mathbf{Set}$! As a consequence of Joyal's and Gödel's theorems, we have a conservative Heyting embedding of T into a very special Heyting category, one of the form $\mathbf{Set}^{\boldsymbol{I}}$ with $\boldsymbol{I}$ a category. Joyal's theorem (combined with Gödel's) is a completeness theorem; in fact, it is essentially equivalent to Kripke's 1965 completeness theorem for intuitionistic logic. From the formulation of Joyal's theorem, it is quite easy to conclude the existence of a conservative Heyting embedding of the form $T \longrightarrow \mathbf{Set}^P$ with a poset P, with possible further conditions on P (regarding its size, or that it be a forest ("tree with many roots")). However, for our purposes, the canonical form of Joyal's theorem is quite sufficient.

For later reference, let us review what is involved in the proof of Joyal's theorem (the proof can be found in Makkai and Reyes [1977], Theorem 6.3.5). In the Heyting category $\mathbf{Set}^{\boldsymbol{I}}$, with $F \xrightarrow{\ g\ } G$ and a subfunctor U of F, the subfunctor $\forall_g U$ of G is given by the prescription

for any $i \in |\boldsymbol{I}|$ and $y \in G(i)$,

$$y \in (\forall_g U)(i) \iff \text{for all arrows } \alpha : i \longrightarrow j \text{ in } \boldsymbol{I},\ \forall x \in Fj\,.\ g_j x = (G\alpha)y \Rightarrow x \in Uj\,.$$

Let T be a Heyting category, $f : A \longrightarrow B$ an arrow in T, $X \in \mathbf{Sub}(A)$, $\forall_f X \in \mathbf{Sub}(B)$. With $\boldsymbol{I} = \mathbf{Mod}(T)$, $g = \mathbf{e}_T(f)$, $F = \mathbf{e}_T(A)$, $G = \mathbf{e}_T(B)$, and $U = \mathbf{e}_T(X)$, to say that $\mathbf{e}_T$ preserves the particular $\forall_f X$ is to say that the above equivalence holds when $\forall_g U$ on the left is replaced by $\mathbf{e}_T(\forall_f X)$. Let us write $h : M \longrightarrow N$ for $\alpha : i \longrightarrow j$. As usual, one direction (left to right) of the equivalence is automatic. Thus, it remains to show that

for any $M \in \mathrm{Mod}(T)$, and $y \in M(B)$,

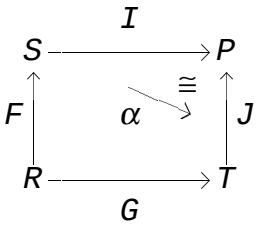
$$y \notin M(\forall_f X) \implies \text{there are } h : M \longrightarrow N \text{ and } x \in N(A) \text{ such that } (Nf)(x) = h_A(y) \text{ and } x \notin N(X).$$

Let us say that the pair (h, x) *witnesses* $y \notin M(\forall_f X)$ if $(Nf)(x) = h_A(y)$ and $x \notin N(X)$. In *loc.cit.*, it is shown, by a straightforward application of the method of diagrams in model theory, that each instance of $y \notin M(\forall_f X)$ is in fact witnessed. We will not repeat the argument here, all the less since similar arguments will have to be made below. However, we make an observation on witnessing to be used later.

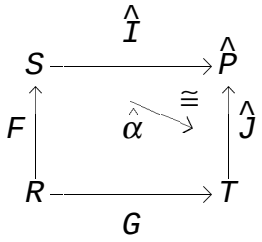
Let us call a map (natural transformation) $k : N \longrightarrow P$ in $\mathrm{Mod}(T)$ (for any coherent T for the moment) *pure* if for any object A in T, any $X \in \mathrm{Sub}(A)$, and any $x \in N(A)$, we have that $k_A(x) \in P(X)$ (if and) only if $x \in N(X)$. The observation is that, with the above notation, (h, x) is a witness for $y \notin M(\forall_f X)$, then so is $(kh, k_A(x))$ provided that k is pure.

§2. The new result and other lemmas.

Given the arrows $F : R \longrightarrow S$, $G : R \longrightarrow T$ in $\mathcal{H}eyting$ (or $\mathcal{B}oole$), we consider the pushout



in $\mathcal{H}eyting$ (or $\mathcal{B}oole$), but also, the pushout



in $\mathcal{C}oherent$. Of course, there is a canonical coherent functor $H:\hat{P}\longrightarrow P$ such that $H\hat{I}=I$, $H\hat{J}=J$ and $H\hat{\alpha}=\alpha$ (H is unique up to isomorphism). Now, if $\hat{J}$ is to be shown to be conservative, we'd better be able to show that $\hat{J}$ is conservative. This we proceed to do now.

Let's pause for a moment. We are claiming that in $\mathcal{C}oherent$, the SAP holds for pushouts of Heyting arrows. To be sure, $\mathcal{C}oherent$ does not satisfy the SAP in general as Pitts [1983] points out (the first example was given by G. E. Reyes); thus, although we deal with a statement formulated in the coherent doctrine, we will have to use the Heyting character of the data involved.

Let $A\in T$, and $X\in \mathrm{Sub}(A)$ with $X\neq 1_A$. We want to show that $\hat{J}X\neq 1_{\hat{J}A}$. Using Gödel completeness, take any $N\in \mathrm{Mod}(T)$ such that $NX\neq NA$. We need a model $L\in \mathrm{Mod}(\hat{P})$ with $L\hat{J}X\neq L\hat{J}A$. But the models L are in an essentially 1-1 correspondence with triples $(M', N', h':M'\mid R\xrightarrow{\cong} N'\mid R)$ with $M'\in \mathrm{Mod}(S)$, $N'\in \mathrm{Mod}(T)$, and with the notation $M'\mid R=MF$, $N'\mid R=NG$; in particular, $L\mid T \underset{\mathrm{def}}{=} LJ=N'$, and thus $L\hat{J}X\neq L\hat{J}A$ means $N'X\neq N'A$. Therefore, if we can make (M', N', h') with a *pure* map $n:N\longrightarrow N'$, we are done! Actually, we first make a careful choice of N with $NX\neq NA$.

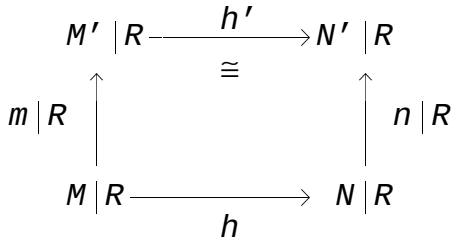
Lemma 1. Suppose that $X\in \mathrm{Sub}(A)$, $X\neq 1_A$ in T . Then there are: $M\in \mathrm{Mod}(S)$, $N\in \mathrm{Mod}(T)$ and a map $h:M\mid R\longrightarrow N\mid R$ such that $NX\neq NA$.

The proof is postponed to Section 3.; it does not need the Heyting character of the data involved.

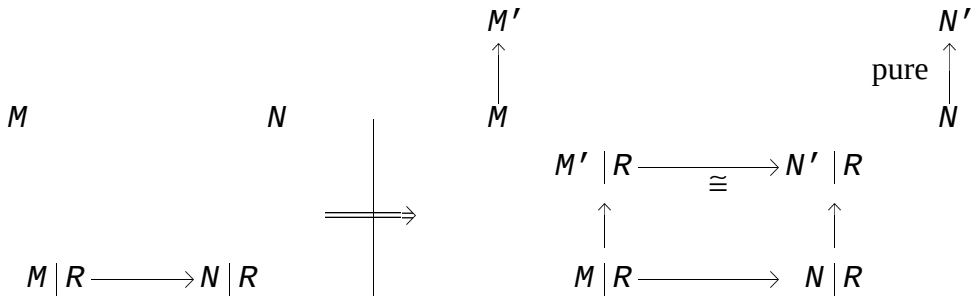
Proposition 2 (Main Lemma). Start with the Heyting functors $F:R\longrightarrow S$, $G:R\longrightarrow T$, without any assumption on conservativeness. Given $M\in \mathtt{Mod}(S)$, $N\in \mathtt{Mod}(T)$ with a map $h:M\,|\,R\longrightarrow N\,|\,R$, there are:

$M'\in \mathtt{Mod}(S)$, $N'\in \mathtt{Mod}(T)$, maps $m:M\longrightarrow M'$, $n:N\longrightarrow N'$

such that n is pure and there is a commutative diagram

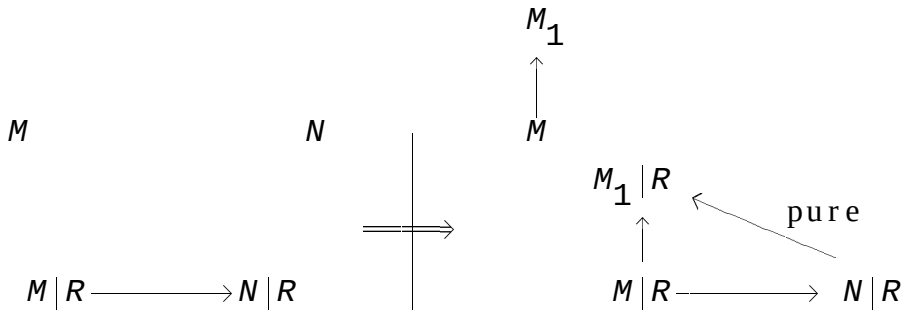


with h' an isomorphism as indicated. Schematically:

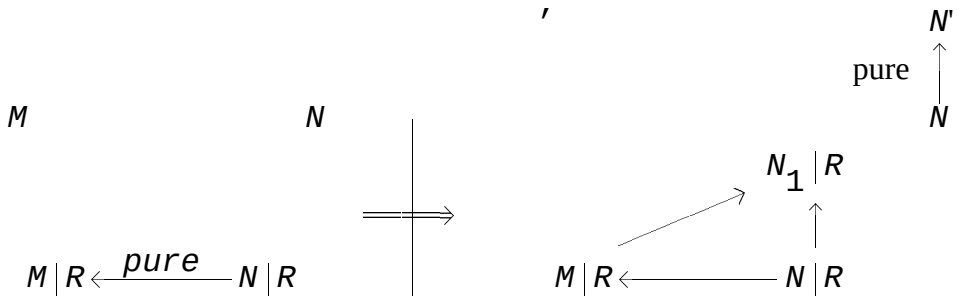


The proposition is proved by the help of the following two lemmas, expressed without words in a similar schematic manner.

Lemma 3.

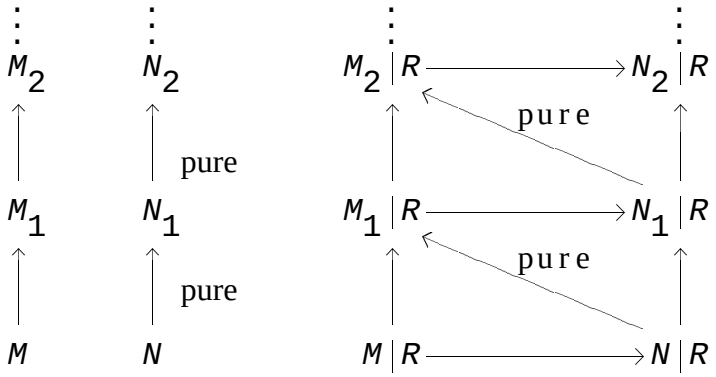


Lemma 4.



The proofs of the last two lemmas are postponed to Section 3.

Proof of Proposition 2 from Lemmas 3 and 4. We build the following diagrams:



We start with the given M, N and $M|R \longrightarrow N|R$. Using Lemma 3, we produce $M \longrightarrow M_1$

and the lowest triangle. Then, applying Lemma 4 to M_1 and N as M and N, we produce $N \xrightarrow{\text{pure}} N_1$ and the next triangle. Etc.

Having the three infinite diagrams, consider $M' = \operatorname{colim}_{i \to \omega} M_i$, $N' = \operatorname{colim}_{i \to \omega} N_i$. We have the canonical colimit coprojections $M \longrightarrow M'$, and $N \longrightarrow N'$; the latter is clearly pure (since all vertical arrows in the diagram of the N's are pure). Finally, the left-to-right horizontal arrows induce a map $M' \longrightarrow N'$, the slanted ones induce one $N' \longrightarrow M'$, and the commutativity of the infinite diagram shows that these latter two arrows are inverses of each other. The isomorphism $M' \longrightarrow N'$ clearly satisfies the commutativity required in Proposition 2.

Let us return to our proof that $\hat{J}$ is conservative provided F is. Using Lemma 1 and Proposition 2, we do have N with $NX \neq NA$, and $(M', N', h' : M' \mid R \xrightarrow{\cong} N' \mid R)$ with a pure map $N \longrightarrow N'$, which, according to what was said above, completes the proof. We have shown:

Proposition 5. Pushouts in $\mathcal{C}oherent$ of pairs of arrows in $\mathcal{H}eyting$ satisfy the SAP.

Let us see that for the case of $\mathcal{B}oole$, this suffices for the proof of the SAP. We claim that the canonical map $H : \hat{P} \longrightarrow P$ is conservative; since the composite of conservative maps is conservative, the desired assertion will follow from Proposition 5. To show the claim, consider the left adjoint Φ to the forgetful functor $\Psi : \mathcal{B}oole^* \longrightarrow \mathcal{C}oherent^*$; for a coherent category T, ΦT is the free Boolean category extension of T. Since Φ is a left adjoint, it preserves colimits; hence it takes a pushout diagram into a pushout diagram. Also, note that if T itself Boolean, then $\Phi T \simeq T$. The reason is that Ψ is a full inclusion (see Mac Lane [1971], Theorem 1 in IV.3, p. 88); the latter statement just says that a coherent functor between Boolean categories is a Boolean functor, which is rather clear.

Our remarks so far imply that in the Boolean case, we have an equvalence $K : P \simeq \Phi \hat{P}$, with $K \circ H = \eta_{\hat{P}}$, where $\eta_{\hat{P}}$ is the canonical unit map of the adjunction. However, $\eta_T : T \longrightarrow \Phi T$ is

always conservative: we have the conservative $e_T : T \longrightarrow \mathbf{Set}^{|\mathbf{Mod}(T)|} = \mathcal{U}$ into a Boolean category; the universal property of η_T implies that there is $L : \Phi T \longrightarrow U$ with $e_T \cong L \circ \eta_T$, from which it follows that η_T is conservative. Our claim follows.

We claim that, in fact

Theorem 6. The canonical coherent functor from the coherent pushout of a pair of Heyting functors to the Heyting pushout of the same pair is always conservative.

This is the new result of this paper. Of course, together with Proposition 5, it implies the SAP for $\mathcal{H}eyting$, and thus, Craig for intuitionistic logic. Note however that Theorem 6 does not follow (directly at least) from the SAP for $\mathcal{H}eyting$, even for pushouts in which one leg or both are conservative.

For the proof of Theorem 6, we show

Proposition 7. With the notation above for the coherent pushout of Heyting arrows, the composite

$$T \xrightarrow{\hat{J}} \hat{P} \xrightarrow{e_{\hat{P}}} \mathbf{Set}^{\mathbf{Mod}(\hat{P})}$$

(and symmetrically,

$$S \xrightarrow{\hat{I}} \hat{P} \xrightarrow{e_{\hat{P}}} \mathbf{Set}^{\mathbf{Mod}(\hat{P})} \quad \text{as well)}$$

is a Heyting functor.

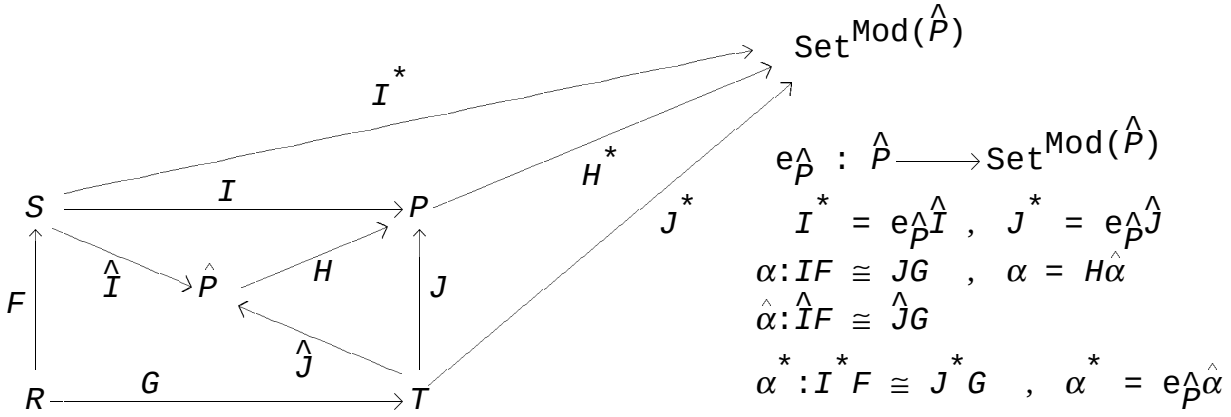
Note that the proposition specializes to Joyal's theorem when both F and G are taken to be identity functors.

Proof. Note that the domain and the codomain of the composite are Heyting categories, although the factors are not necessarily Heyting functors. Looking back at the remarks on the proof of Joyal's theorem, and using the fact that the objects of $\mathbf{Mod}(\hat{P})$ are essentially triples $(M, N, h : M|R \xrightarrow{\cong} N|R)$, a little thought shows that what is to be shown is this: with $f : A \longrightarrow B$ and X as before, given (M, N, h) and $y \in MB$, if $y \notin N(\forall_f X)$, then there is an arrow $(m, n) : (M, N, h) \longrightarrow (M', N', h')$ and some x such that (n, x) is a witness for $y \notin N(\forall_f X)$ in the original sense introduced above. Now, by (the proof of) Joyal's theorem, there is a witness $(n^* : N \longrightarrow N^*, x)$. Apply Proposition 2 for M and N^* as M and N, and $n^*|R \circ h : M|R \longrightarrow N^*|R$ as the arrow h. We obtain $m : M \longrightarrow M'$, $n_1 : N^* \longrightarrow N$ such that n_1 is pure, and the commutative diagram

$$\begin{array}{ccc}
M'|R & \xrightarrow{h'} & N'|R \\
{\scriptstyle m|R}\uparrow & \cong & \uparrow{\scriptstyle n_1|R} \\
M|R & \xrightarrow[n^*|R \circ h]{} & N^*|R \\
{\scriptstyle 1_M|R}\uparrow & & \uparrow{\scriptstyle n^*|R} \\
M|R & \xrightarrow[h]{} & N|R
\end{array} \quad .$$

By a remark made above, since n_1 is pure, $(n_1 n^*, (n_1)_A x)$ is a witness, and thus with $n = n_1 n^*$, m, n and x are as desired.

Proof of Theorem 6. Consider the following diagram:



Here, after the familiar ingredients, I^*, J^* and α^* are defined as shown ($e_{\hat{P}}$ is not drawn). Since by Lemma 7 , I^* and J^* are Heyting functors, by the universal property of P , we have H^* with $IH^* = I^*$, $JH^* = J^*$ and $\alpha H^* = \alpha^*$. Hence, $\hat{I}H^*H = I^*$, $\hat{J}H^*H = J^*$ and $\hat{\alpha}H^*H = \alpha^*$. By the uniqueness part of the universal property of $\hat{P}$, it follows that $H^*H \cong e_{\hat{P}}$. But, by Gödel, $e_{\hat{P}}$ is conservative; hence, so is H .

3. Model theoretical arguments.

The categorical language is superior for the purposes of expressing the important things (stating results, for instance), but not convenient to state certain arguments, e.g. the model-theoretical ones involving the "method of diagrams". In our view (which reverses the historical order), symbolic logic is the auxiliary language, to be introduced to deal in convenient ways with the concepts primarily given in the categorical framework.

With any coherent category T , we associate a particular coherent theory $(\mathcal{L}_T, \underline{T})$, also written simply as $\underline{T}$, the *internal theory* of T ; in Makkai and Reyes [1977], $\mathbb{T}_T$ is written for $(\mathcal{L}_T, T)$; see Chapter 3, Section 5 , p. 128. Here, $\mathcal{L}_T$ has sorts the objects of T , unary sorted operations symbols the arrows of T . The models of $(\mathcal{L}_T, T)$ are *identical* to the models of T (objects of $\mathrm{Mod}(T)$); the (universal algebraic) homomorphisms of models are

the same as the arrows in $\mathtt{Mod}(T)$ (see *loc.cit.*, Theorem 3.5.3.)

With any coherent formula $\varphi(\vec{x})$ with free variables among $\vec{x}$, we have the interpretation of φ; $[\vec{x}\,|\,\varphi] \in \mathtt{Sub}([\vec{x}])$, where $[\vec{x}]$ is the product of the objects that are the respective sorts of the variables in $\vec{x}$ (in [M/R], $[\varphi]$ is written for $[\vec{x}\,|\,\varphi]$). Given any subobject $X \in \mathtt{Sub}(A)$, there is a canonically selected coherent formula $\underline{X}(x)$ such that $[x{\in}A\,|\,\underline{X}(x)] = X$ (here, x is a variable of sort A; " $x{\in}A$ " is just a reminder of this). If $X \in \mathtt{Sub}(A_1{\times}\ldots{\times}A_n)$, we have $\underline{X}(x_1, \ldots, x_n)$ $(= \underline{X}(\vec{x}))$ such that $[\vec{x}{\in}\prod_{i=1}^{n} X_i] \mid \underline{X}(\vec{x})] = X$.

Let's introduce some auxiliary notation. With $F{:}R \longrightarrow S$ a coherent functor, if $\varphi(\vec{x})$ in any formula over $\mathcal{L}_R$ (e.g., in full first order logic, or even in infinitary logic), $\varphi^F(\vec{x})$ denotes its F-translate, that is, the result of substituting FA for A, and Ff for f, for any $f{:}A \longrightarrow B$ in R; the sorting of the variables, both free and bound, is redone accordingly. Thus, $\varphi^F(\vec{x})$ is a formula over $\mathcal{L}_S$. If, in particular, φ is a coherent formula, then $F[\vec{x}\,|\,\varphi] = [\vec{x}\,|\,\varphi^F]$; this is precisely to say that the functor F is coherent. With a little abuse of language, we also write φ^S for φ^F $(F{:}R \longrightarrow S)$.

With $M \in \mathtt{Mod}(R)$, $\mathtt{Diag}^+(M)$ denotes the *positive diagram* of M, that is, the set of all coherent (positive existential) sentences, over the language $\mathcal{L}_R(M)$ containing an individual constant a for each pair $(A{\in}T, a{\in}M(A))$, that are true in M. As it is well-known, the arrows (homomorphisms) out of M are in an essentially one-to-one correspondence with the models of $\mathtt{Diag}^+(M)$. With $G{:}R \longrightarrow T$, say, when writing $\mathtt{Diag}^+(M)^T$, we mean the set of G-translates of all members of $\mathtt{Diag}^+(M)$, with the new individual constants left alone except that their sortings are redone in the obvious way.

The negative diagram $\mathtt{Diag}^-(M)$ is defined similarly, with negated positive existential positive sentences. Models of $\mathtt{Diag}^+(M) \cup \mathtt{Diag}^-(M)$ correspond to embeddings, that is, pure maps, out of M.

In this section, we use $\vdash$ in the sense of deducibility in ordinary *classical* predicate logic.

Proof of Lemma 1. Let Σ denote the set of all coherent (positive existential) sentences σ over $\mathcal{L}_R$ such that $\neg\sigma^T$ is a consequence $\underline{T} \cup \{\neg\forall x \underline{X}(x)\}$. We claim that $\underline{S} \cup \{\neg\sigma^S : \sigma \in \Sigma\}$ is consistent. Otherwise, also using that Σ is closed under disjunction, we obtain $\sigma \in \Sigma$ such that $\underline{S} \vdash \sigma^S$. Since F is conservative, $\underline{R} \vdash \sigma$; hence, $\underline{T} \vdash \sigma^T$. But since $\underline{T} \cup \{\neg\forall x \underline{X}(x)\} \vdash \sigma^T$ (definition of Σ), it follows that $\underline{T} \vdash \forall x \underline{X}(x)$, contrary to the hypothesis that $X \neq 1_A$.

Let M be any model of $\underline{S} \cup \{\neg\sigma^S : \sigma \in \Sigma\}$; in particular, $M \in \mathbf{Mod}(T)$. We claim that $\mathbf{Diag}(M|R)^T \cup \underline{T} \cup \{\neg\forall x \underline{X}(x)\}$ is consistent; once proved, the claim establishes the lemma. If the claim fails, there is $\varphi(\vec{\underline{a}}) \in \mathbf{Diag}(M|R)$ with $\underline{T} \cup \{\neg\forall x \underline{X}(x)\} \vdash \neg\varphi(\vec{\underline{a}})$; hence, $\exists\vec{x}\varphi \in \Sigma$; since $M \models (\exists\vec{x}\varphi)^S$, this is a contradiction to the definition of Σ.

Before turning to the proofs of Lemmas 3 and 4, we discuss how the fact that R is a Heyting category is reflected on the "coherent logic" of R.

First of all, with $A \in R$, and X, $Y \in \mathbf{Sub}(A)$, we may consider $\forall_y(X) \in \mathbf{Sub}(Y)$ where $y : Y \longrightarrow A$ is the structure map for $Y \in \mathbf{Sub}(A)$. With z the structure map for $\forall_y(X) \in \mathbf{Sub}(Y)$, we let $X \to Y$ be the subobject of A represented by the monomorphism $y \circ x$. It is easy to see that $X \to Y$ is the "Heyting implication" (relative pseudo-complement) in the sense that for any $Z \in \mathbf{Sub}(A)$, $Z \wedge X \leq Y$ iff $Z \leq X \to Y$. Further, if $f : A \longrightarrow B$ and X, $Y \in \mathbf{Sub}(A)$, then $\forall_f(X \to Y) \in \mathbf{Sub}(B)$ satisfies the following: for any $Z \in \mathbf{Sub}(B)$,

$$Z \leq \forall_f(X \to Y) \iff f^*(Z) \wedge X \leq Y;$$

this is easy to see.

Now, suppose we have the coherent formulas $\varphi(\vec{x}, \vec{y})$, $\psi(\vec{x}, \vec{y})$ over the language $\mathcal{L}_R$; let $A = [\vec{x}]$, $B = [\vec{y}]$, thus $A \times B = [\vec{x}, \vec{y}]$; let $\pi : A \times B \longrightarrow A$ be the projection; let $X = [\vec{x}, \vec{y} \mid \varphi]$, $Y = [\vec{x}, \vec{y} \mid \psi]$. We *define* the subobject $[\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]$ to be $\forall_\pi(X \to Y) \in \mathrm{Sub}(A)$; note that $\forall \vec{y}(\varphi \longrightarrow \psi)$ is not a coherent formula, thus the expression $[\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]$ is not defined in the general context. What would make this definition fully justified is the equivalence

$$? \; : \quad \underline{R} \vdash \forall \vec{x}(\underline{[\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]}(\vec{x}) \longleftrightarrow \forall \vec{y}(\varphi(\vec{x}, \vec{y}) \longrightarrow \psi(\vec{x}, \vec{y}))) \; .$$

This is too much to hope for; however, the left-to-right implication and a suitable weakening of the other implication are true:

$$\underline{R} \vdash \underline{[\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]}(\vec{x}) \longrightarrow \forall \vec{y}(\varphi(\vec{x}, \vec{y}) \longrightarrow \psi(\vec{x}, \vec{y})) \; ; \qquad (1)$$

and for any (coherent/$\mathcal{L}_R$) $\theta(\vec{x})$,

$$\underline{R} \cup \{\theta(\vec{x})\} \vdash \forall \vec{y}(\varphi(\vec{x}, \vec{y}) \longrightarrow \psi(\vec{x}, \vec{y})) \Longrightarrow \underline{R} \cup \{\theta(\vec{x})\} \vdash \underline{[\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]}(\vec{x}) \quad (2)$$

(in the latter, $\vec{x}$ is treated as constants, in the natural way). These assertions are immediately seen by the what was said above.

Furthermore, if $F : R \longrightarrow S$ is Heyting, then, clearly,

$$F([\vec{x} \mid \forall \vec{y}(\varphi \longrightarrow \psi)]) = [\vec{x} \mid (\forall \vec{y}(\varphi \longrightarrow \psi))^S] \; .$$

Proof of Lemma 3. Let f be the given arrow $M \mid R \longrightarrow N \mid R$. In forming $\mathrm{Diag}(M)$ and $\mathrm{Diag}(N \mid R)$, make sure that the individual constants used are disjoint in the two diagrams. Modify $\mathrm{Diag}(N \mid R)$ to $\mathrm{Diag}(N \mid R)^*$ by adding to its language all individual constants in $\mathrm{Diag}(M)$, adding all axioms $\underline{a} = \underline{b}$ whenever $A \in R$ and $f_A(a) = b$, and closing the result under conjunction. It is clear by inspection that the assertion of the lemma is equivalent to saying that

$$\underline{S} \cup \mathtt{Diag}(M) \cup (\mathtt{Diag}^{+}(N|R)^{*})^{S} \cup (\mathtt{Diag}^{-}(N|R))^{S}$$

is consistent. Assume the contrary. Then there are $\theta(\underline{\vec{a}}) \in \mathtt{Diag}(M)$, $\varphi(\underline{\vec{a}}, \underline{\vec{b}}) \in \mathtt{Diag}^{+}(N|R)$, and $\neg\psi(\underline{\vec{b}}) \in \mathtt{Diag}^{-}(N|R)$ such that

$$\underline{S} \vdash \theta(\underline{\vec{a}}) \wedge \varphi^{S}(\underline{\vec{a}}, \underline{\vec{b}}) . \longrightarrow . \psi^{S}(\underline{\vec{b}}) ,$$

that is,

$$\underline{S} \cup \{\theta(\underline{\vec{a}})\} \vdash \forall\vec{y}(\varphi^{S}(\underline{\vec{a}}, \vec{y}) \longrightarrow \psi^{S}(\vec{y})) .$$

Thus, by (2),

$$\underline{S} \cup \{\theta(\underline{\vec{a}})\} \vdash \underline{[\vec{x} \mid \forall\vec{y}(\varphi^{S} \longrightarrow \psi^{S})]}(\underline{\vec{a}})$$, which, by $M \models S \cup \{\theta(\underline{\vec{a}})\}$,

implies that

$$M \models \underline{[\vec{x} \mid \forall\vec{y}(\varphi^{S} \longrightarrow \psi^{S})]}(\underline{\vec{a}}) ,$$

that is,

$$M|R \models \underline{[\vec{x} \mid \forall\vec{y}(\varphi \longrightarrow \psi)]}(\underline{\vec{a}}) ,$$

and

$$N|R \models \underline{[\vec{x} \mid \forall\vec{y}(\varphi \longrightarrow \psi)]}(\underline{f\vec{a}}) . \qquad (3)$$

However, $N|R \models \varphi(\underline{f\vec{a}}, \underline{\vec{b}})$, and $N|R \models \neg\psi(\underline{\vec{b}})$, hence

$$N|R \models \neg\forall\vec{y}(\varphi(\underline{f\vec{a}}, \vec{y}) \longrightarrow \psi(\vec{y})) . \qquad (4)$$

(1), (3) and (4) give a contradiction.

Proof of Lemma 4. This is similar; we can be briefer. It suffices to prove that

$$\underline{T} \cup \mathrm{Diag}^{+}(N) \cup \mathrm{Diag}^{-}(N) \cup (\mathrm{Diag}^{+}(M|R)^{*})^{T}$$

is consistent. Otherwise, we have $\theta(\underline{\vec{b}}) \in \mathrm{Diag}^{+}(N)$, $\varphi(\underline{\vec{a}}, \underline{\vec{b}}) \in \mathrm{Diag}^{+}(M|R)^{*}$ and $\psi(\underline{\vec{b}}) \in \mathrm{Diag}^{-}(N)$ such that

$$\underline{T} \cup \{\varphi^{T}(\underline{\vec{a}}, \underline{\vec{b}})\} \vdash \theta(\underline{\vec{b}}) \longrightarrow \psi(\underline{\vec{b}}) \text{ , that is,}$$

$$\underline{T} \cup \{\exists \vec{x} \varphi^{T}(\underline{\vec{x}}, \underline{\vec{b}})\} \vdash \theta(\underline{\vec{b}}) \longrightarrow \psi(\underline{\vec{b}}) \text{ ,}$$

and by (2) ,

$$\underline{T} \cup \{\exists \vec{x} \varphi^{T}(\underline{\vec{x}}, \underline{\vec{b}})\} \vdash \underline{[\vec{y} \mid \theta(\vec{y}) \longrightarrow \psi(\vec{y})]}(\underline{\vec{b}}) \text{ .}$$

Now, since $N|R \longrightarrow M|R$ is pure, and $M|R \vDash \exists \vec{x} \varphi(\vec{x}, \underline{\vec{b}})$, we have that $M|R \vDash \exists \vec{x} \varphi(\vec{x}, \underline{\vec{b}})$, that is, $M \vDash \exists \vec{x} \varphi^{T}(\vec{x}, \underline{\vec{b}})$. Hence, $M \vDash \underline{[\vec{y} \mid \theta(\vec{y}) \longrightarrow \psi(\vec{y})]}(\underline{\vec{b}})$. However, this contradicts (1) and the choice of the formulas θ , ψ .

Let us make some remarks comparing our procedures with those of Gabbay [1969]. In that paper, a version of Robinson's Consistency Lemma is proved, and Craig's theorem is derived from it. The proof of the "Robinson's lemma" is given, essentially, in one piece; no (real) lemmas are formulated; in particular, no statement comparable to Theorem 6 is given. Nevertheless, most of the ingredients of our proofs appear in Gabbay's proof. Instead of *models* (in our sense) of T , "complete and saturated extensions of T " appear; this is indeed an equivalent concept. Lemma 1, Proposition 2 and Proposition 5 do not appear in a recognizable form, but Lemmas 3 and 4 do, in several places (as essentially the same two arguments repeated). Finally, there is a construction of a tree-based Kripke model, with a tree whose nodes are various tuples consisting of formulas, etc; this is avoided in our treatment by the use

of a Joyal-type canonical Kripke model.

References

E. W. Beth [1953]
On Padoa's method in the theory of definition. *Indag. Math.* **15**, 330-339.

J. L. Bell and M. Machover [1977]
A Course in Mathematical Logic. North-Holland, Amsterdam.

C. C. Chang and H. J. Keisler [1973]
Model Theory. North-Holland, Amsterdam.

W. Craig [1957]
Linear reasoning. A new form of the Herbrand-Gentzen theorem. *J. Symbolic Logic* **22**, 250-268.

D. M. Gabbay [1969]
Semantic proof of Craig's interpolation theorem for intuitionistic logic and extensions, Part I. In: *Logic Colloquium'69* (R. O. Gandy and C. M. E. Yates, eds), North-Holland, Amsterdam; pp. 391-402.

D. M. Gabbay [1981]
Semantical Investigations in Heyting's Intuitionistic Logic. D. Reidel Publ. Co., Dordrecht and Boston.

P. Gabriel and F. Ulmer [1971]
Lokal präsentierbare Kategorien. *Lecture Notes in Math*., vol. 221, Springer Verlag.

A. Joyal and M. Tierney [1984]
An extension of the Galois theory of Grothendieck. *Memoirs Amer. Math. Soc.* **51**, no. 309.

S. A. Kripke [1965]
Semantical analysis of intuitionistic logic I. In: *Formal Systems and Recursion Theory* (S. N. Crossley and M. A. E. Dummett, eds). North-Holland, Amsterdam; pp. 92-130.

S. Mac Lane [1971]
Categories for the Working Mathematician. Springer-Verlag, Heidelberg and New York.

S. Maehara [1960/61]
On the interpolation theorem of Craig (in Japanese). *Sugaku* **12**, 235-237.

M. Makkai and G. E. Reyes [1977]
First Order Categorical Logic. *Lecture Notes in Math.*, vol. **611**, Springer-Verlag, Heidelberg and New York.

M. Makkai [1985]
Ultraproducts in categorical logic. In: *Lecture Notes in Math.*, vol. 1130, Springer-Verlag, Heidelberg and New York; pp. 222-309.

A. M. Pitts [1983]
An application of open maps to categorical logic. *J. Pure and Applied Algebra* **29**, 313-326.

A. Robinson [1956]
A result on consistency and its application to the theory of definition. *Indag. Math.* **18**, 47-58.

K. Schütte [1962]
Der Interpolationsatz der intuitionistischen Prädicatenlogik. *Math. Annalen* **148**,192-200.

R. Street [1980]
Fibrations in bicategories. *Cahiers Top. Geom. Diff.* **21**, 111-160.