Interpolation Is (Not Always) Easy to Spoil

Andrzej Tarlecki
Institute of Informatics, University of Warsaw, Poland

Abstract

We study a version of the Craig interpolation theorem as formulated in the framework of the theory of institutions. This formulation proved crucial in the development of a number of key results concerning foundations of software specification and formal development. We investigate preservation of interpolation under extensions of institutions by new models and sentences. We point out that some interpolation properties remain stable under such extensions, even if quite arbitrary new models or sentences are permitted. We give complete characterisations of such situations for institution extensions by new models, by new sentences, as well as by new models and sentences, respectively.

2012 ACM Subject Classification Theory of computation → Logic; Theory of computation → Logic and verification

Keywords and phrases interpolation, institutions, institutional abstract model theory, specification theory

Digital Object Identifier 10.4230/LIPIcs.CALCO.2023.8

Acknowledgements Thanks to the anonymous reviewers for a number of useful comments.

1 Introduction

The Craig interpolation theorem [12] states that when an implication $\varphi \Rightarrow \psi$ between premise φ and conclusion ψ holds then there is an interpolant θ built using the symbols the premise and the conclusion have in common that witnesses this implication, that is, such that both $\varphi \Rightarrow \theta$ and $\theta \Rightarrow \psi$ hold. This is one of the fundamental properties of the classical first-order logic, with numerous consequences and links with other key properties developed in the framework of classical model theory [11].

In the area of foundations of system specification and formal development, interpolation proved indispensable for a number of most fundamental features of various approaches. This was perhaps first pointed out in [27], where it was used to ensure composability of subsequent implementation steps (later refined in various forms of the so-called modularisation theorem [42, 41]). In the work on module algebra [3] the interpolation was necessary to obtain crucial distributive laws for their export operator ([31] joined the two threads). The proofs of completeness of proof calculi for consequences of structured specifications rely on interpolation [10, 5] (in fact, no "good" sound and complete such proof calculus may exist without an appropriate interpolation property for the underlying logic [36]). These and further results concerning completeness of various reasoning systems necessary in the process of reliable software development involve interpolation explicitly, but the same idea that showing properties of a union of a number of extensions of a basic theory must rely on some form of interpolation (perhaps disguised as the Robinson consistency [32]) is omnipresent in both practical and foundational aspects of computing.

Applications of logic in computer science face the problem of dealing with numerous logical systems. This follows from the real needs of software development, based on the multitude of application areas as well as of programming paradigms, features and languages. This led to various attempts to abstract away from a specific logical system in use. Such an independence of the foundations for software specification has been successfully achieved by

© Andrzej Tarlecki;
licensed under Creative Commons License CC-BY 4.0

relying on the concept of an institution, introduced by Goguen and Burstall as a formalisation of the concept of a logical system [25]. See for instance [35] for an exhaustive account of such ideas, with further examples in the development of specification formalisms such as CASL [1].

It has been realised quite early that institutions also offer a framework for developing a very abstract version of model theory, going beyond what has been studied within abstract model theory following [2]. This was noted in [37] and expanded in many crucial directions by Diaconescu and his group; his monograph [13] offers an overview of this work, with later developments scattered through numerous articles (see e.g. [16] and references there).

In the institutional model theory the interpolation property is formulated so that it can be studied (and used) for logical systems departing considerably from the first-order logic. This was put forward in [37], but we use here a still more refined formulation of interpolation given in [34, 14]. This formulation uses logical entailment (rather than implication), sets of sentences (rather than individual sentences) and, most crucially, works over arbitrary commutative squares of signature morphisms (rather than over union/intersection squares only), and so caters for instance for the logical systems where one lacks compactness, conjunction and other classical connectives, and even the concept of the set of symbols used in a formula and union/intersection of signatures may not be directly available. The key point of many of the applications mentioned above is the need to abstract away from signature inclusions and deal with interpolation properties with other signature morphisms considered. Subsequent work included development of generic model-theoretic proof techniques to establish interpolation for institutions satisfying a number of structural properties. This led to new interpolation results concerning various logical systems, as well as to studying interpolation in even more general context of non-standard entailment relations [14, 6, 24, 30, 15, 22, 23, 17].

The need for the use of many logical systems leads to the need for establishing their properties, including the interpolation property we study here. Rather than doing this for each system anew, it is desirable to ensure the required properties in the course of systematic construction of new logics, perhaps along the lines aimed at for instance in [29, 28] or [8, 7, 9]. Typically, the new logics are linked with the original ones by institution (co)morphisms [25, 26]. An important line of research was to clarify sufficient conditions on the institution (co)morphisms involved to allow interpolation properties to be "transferred" between the institutions they link [18, 22].

We address a perhaps more basic question that arises in this framework: namely, when interpolation properties can be spoiled by extending a logic by new abstract models or sentences. Looking at the standard formulation, it seems that the answer is always positive. To spoil an interpolant for the premise and the conclusion of a true implication, just add a new model that satisfies the premise but not the interpolant, or the interpolant but not the conclusion, thus spoiling the required implication between the premise and the interpolant, or between the interpolant and the conclusion. This should work, except for the trivial cases when the signature of the premise includes or is included in the signature of the conclusion. At a closer look though, when one considers arbitrary signature morphisms, adding new models for the signature of the premise or for the signature of the conclusion may result in new models for their union signature, and ruin the implication considered.

We explore the consequences of this observation, and give exact characterisations of the situations where interpolation is stable under extensions of institutions. Equivalently, looking at the other side of this coin, we obtain the exact characterisation of the situations where new models or sentences may spoil the interpolation property. More precisely: we consider separately institution extensions where only new models, only new sentences, and both new models and sentences, respectively, are permitted. In each of these three cases complete characterisations are given, formulating necessary and sufficient conditions for a commutative square of signature morphisms under which no such institution extension may spoil interpolation properties over this square.

2 Institutions

2.1 Notational preliminaries

For any function $f\colon X \to Y$, given a set $X' \subseteq X$, $f(X') = \{f(x) \mid x \in X'\} \subseteq Y$ is the image of X' w.r.t. f, and for $Y' \subseteq Y$, $f^{-1}(Y') = \{x \in X \mid f(x) \in Y'\}$ is the coimage of Y' w.r.t. f

Throughout the paper we freely use the basic notions from category theory (*category*, *functor*, *natural transformation*, *pushout*, etc). Composition in any category is denoted by ";" (semicolon) and written in the diagrammatic order. For instance, $f\colon A \to B$ is a *retraction* if for some $g\colon B \to A$ we have $g;f = id_B$, and $f\colon A \to B$ is a *coretraction* if for some $g\colon B \to A$ we have $f;g = id_A$. The collection of objects of any category $\mathbf{K}$ is written as $|\mathbf{K}|$. The category of sets is denoted by $\mathbf{Set}$, and the (quasi-)category of classes by $\mathbf{Class}$.

2.2 Institutions

In the foundations of software specification and development [35] it is standard by now to abstract away from the details of the logical system in use, relying on the formalisation of a logical system as an *institution* [25]. An institution $\mathbf{INS}$ consists of:

- a category $\mathbf{Sign_{INS}}$ of *signatures*;
- a functor $\mathbf{Sen_{INS}}\colon \mathbf{Sign_{INS}} \to \mathbf{Set}$, giving a set $\mathbf{Sen_{INS}}(\Sigma)$ of Σ-*sentences* for each signature $\Sigma \in |\mathbf{Sign_{INS}}|$;
- a functor $\mathbf{Mod_{INS}}\colon \mathbf{Sign}^{op}_{\mathbf{INS}} \to \mathbf{Class}$, giving a class (or a discrete category)[1] $\mathbf{Mod_{INS}}(\Sigma)$ of Σ-*models* for each signature $\Sigma \in |\mathbf{Sign_{INS}}|$; and
- a family $\langle \models_{\mathbf{INS},\Sigma} \subseteq \mathbf{Mod_{INS}}(\Sigma) \times \mathbf{Sen_{INS}}(\Sigma)\rangle_{\Sigma \in |\mathbf{Sign_{INS}}|}$ of *satisfaction relations*

such that for any signature morphism $\sigma\colon \Sigma \to \Sigma'$ the induced translations $\mathbf{Mod_{INS}}(\sigma)$ of models and $\mathbf{Sen_{INS}}(\sigma)$ of sentences preserve the satisfaction relation, that is, for any $\varphi \in \mathbf{Sen_{INS}}(\Sigma)$ and $M' \in \mathbf{Mod_{INS}}(\Sigma')$ the following *satisfaction condition* holds:

$$M' \models_{\mathbf{INS},\Sigma'} \mathbf{Sen_{INS}}(\sigma)(\varphi) \quad \text{iff} \quad \mathbf{Mod_{INS}}(\sigma)(M') \models_{\mathbf{INS},\Sigma} \varphi.$$

The subscripts $\mathbf{INS}$ and Σ are typically omitted. For any signature morphism $\sigma\colon \Sigma \to \Sigma'$, the *translation* $\mathbf{Sen}(\sigma)\colon \mathbf{Sen}(\Sigma) \to \mathbf{Sen}(\Sigma')$ is denoted by $\sigma\colon \mathbf{Sen}(\Sigma) \to \mathbf{Sen}(\Sigma')$, and the *reduct* $\mathbf{Mod}(\sigma)\colon \mathbf{Mod}(\Sigma') \to \mathbf{Mod}(\Sigma)$ by $\_|_\sigma\colon \mathbf{Mod}(\Sigma') \to \mathbf{Mod}(\Sigma)$. For instance, the satisfaction condition may be re-stated as: $M' \models \sigma(\varphi)$ iff $M'|_\sigma \models \varphi$, and given the notation for image and coimage, for $\Phi \subseteq \mathbf{Sen}(\Sigma)$, $\sigma(\Phi) = \{\sigma(\varphi) \mid \varphi \in \Phi\} \subseteq \mathbf{Sen}(\Sigma')$, and for $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma)$, $\mathcal{M}|_\sigma^{-1} = \{M' \in \mathbf{Mod}(\Sigma') \mid M'|_\sigma \in \mathcal{M}\}$. For any signature Σ, the satisfaction relation extends to sets of Σ-sentences and classes of Σ-models. For $\Phi \subseteq \mathbf{Sen}(\Sigma)$, the *class of models of* Φ is $Mod(\Phi) = \{M \in \mathbf{Mod}(\Sigma) \mid M \models \Phi\}$, and for $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma)$, the *theory of* $\mathcal{M}$ is $Th(\mathcal{M}) = \{\varphi \in \mathbf{Sen}(\Sigma) \mid \mathcal{M} \models \varphi\}$. The latter notation is also used for the *theory generated by a set of sentences*: for $\Phi \subseteq \mathbf{Sen}(\Sigma)$, $Th(\Phi) = Th(Mod(\Phi))$.

Each satisfaction relation determines a (semantic) entailment between sets of sentences: $\Phi \subseteq \mathbf{Sen}(\Sigma)$ *entails* $\Psi \subseteq \mathbf{Sen}(\Sigma)$ (or Ψ *is a consequence of* Φ), written $\Phi \models \Psi$, when $\Psi \subseteq Th(\Phi)$. The satisfaction condition implies that the semantic entailment is preserved under

[1] We disregard here model morphisms, irrelevant for the purposes of this paper.

translation along signature morphisms: for any $\sigma\colon \Sigma \to \Sigma'$, if $\Phi \models \Psi$ then $\sigma(\Phi) \models \sigma(\Psi)$. If the opposite implication holds as well, i.e. $\Phi \models \Psi$ iff $\sigma(\Phi) \models \sigma(\Psi)$ for all $\Phi, \Psi \subseteq \mathbf{Sen}(\Sigma)$, we say that $\sigma\colon \Sigma \to \Sigma'$ is *conservative*. It is well-known that if the reduct $\_|_\sigma\colon \mathbf{Mod}(\Sigma') \to \mathbf{Mod}(\Sigma)$ is surjective then $\sigma\colon \Sigma \to \Sigma'$ is conservative.[2]

We typically decorate the names for institution components and for other derived notions by primes, indices, etc, to identify the institution they refer to, and rely on this convention whenever the institution is clear from the context. So, for instance, $\mathbf{Mod}_1$ is the model functor in an institution $\mathbf{INS}_1$, $\models'$ is the satisfaction relation (and entailment) in $\mathbf{INS}'$, etc.

Examples of institutions abound, see e.g. [35, 13]. We just sketch three standard examples.

▶ **Example 1.** The institution **FO** of (many-sorted) first-order logic has signatures that consist of sets of sort names, of operation names with indicated arities and result sorts, and of predicate names with indicated arities. Terms and atomic formulae are defined as usual, and first-order formulae are built using the usual Boolean connectives (including nullary `false`) and quantification. First-order sentences are closed formulae (i.e. formulae with no free occurrences of variables). First-order models consist of many-sorted carrier sets (one set for each sort name), functions to interpret operation names and relations to interpret predicate names, in accordance with their arities and result sorts. Satisfaction of first-order sentences in first-order models is defined as usual. Signature morphisms map sort names to sort names, operation names to operation names and predicate names to predicate names preserving their arities and result sorts. For any such morphism, translation of sentences is defined by renaming sort, operation and predicate names as indicated by the morphism, and model reducts are defined by interpreting the symbols of the source signature as the symbols they are mapped to in the target signature are interpreted in the argument model. This indeed defines an institution [25]. We assume that carrier sets in first-order models are nonempty. The variant of **FO** where empty carrier sets are allowed in models is denoted by $\mathbf{FO}_\emptyset$. Another variant is the institution $\mathbf{FO}_\mathbf{EQ}$ of first-order logic with equality, with a binary equality predicate for each sort, interpreted as the identity relation in all models.

The institution **EQ** of (many-sorted) equational logic may be defined as the restriction of $\mathbf{FO}_\mathbf{EQ}$ to the signatures with no predicates other than equalities (models are usually called algebras then), and sentences limited to universally quantified equalities. $\mathbf{EQ}_\emptyset$ is the variant of **EQ** with empty carriers permitted. See [35, 13] for an explicit definition.

The institution **PL** of propositional logic has finite sets of *propositional variables* as signatures, with signature morphisms being arbitrary functions between those sets. Propositional sentences are built from propositional variables using the usual Boolean connectives (with obvious translations under functions renaming propositional variables). Models over a signature are given as subsets of this signature (consisting of the propositional variables that are satisfied in the model) with reducts w.r.t. signature morphisms given as their coimage. With the usual satisfaction of propositional sentences in such models, the satisfaction condition is easy to check. In fact, the institution **PL** of propositional logic may be viewed as a restriction of the institution of first-order logic to finite signatures with no sort names (and hence no operation names and nullary predicates only).

In the institutions **FO**, **EQ**, and **PL** all injective signature morphisms induce surjective reducts, and so are conservative. This need not be the case for non-injective morphisms. In $\mathbf{FO}_\emptyset$ in $\mathbf{EQ}_\emptyset$, the variants of **FO** and of **EQ** where empty carriers are permitted, not all injective signature morphisms are conservative.

[2] Some authors use “conservative” for signature morphisms that induce surjective reducts [26]. Our more permissive definition seems closer to the standard definition of a conservative theory interpretation [11].

In the above examples all the signatures, sentences and models are quite familiar, and link with many intuitions and implicit assumptions. However, when exploiting the generality of the concept and working with an arbitrary institution, such connotations should be dropped. All the entities involved (signatures, their morphisms, sentences, models, satisfaction relations) are considered entirely abstract, with completely unknown structure and properties. It is perhaps surprising how far one can go with developments of the foundations for software specification [35] and an abstract version of model theory [13] in such an abstract setting.

2.3 Extending institutions by models and sentences

We introduce two basic ways of extending institutions, by adding new "abstract" models, and new "abstract" sentences, respectively. The definitions are shaped after the definition of *constraints* in [25, 35]. The basic observation is that when a new sentence is added over a signature, with some predefined notion of satisfaction in the institution models, it must also be "fitted" to other signatures to mimic its translation along signature morphisms with this signature as a source. Hence, together with each new sentence, we also add its "formal translations" along signature morphisms. The satisfaction of such formal translations is determined by the satisfaction condition. Similarly, when we add a new model over a signature – apart from the model itself, we must also add its "formal reducts".

Consider and arbitrary institution $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle \models_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$.

Suppose that for each signature we are given a set of (new) "sentences" with predefined satisfaction relation in the **INS**-models, which may be organised as a signature-indexed family of sets with relations: $\mathcal{NS} = \langle \mathcal{NS}_\Sigma, \models_\Sigma^{\mathcal{NS}} \subseteq \mathbf{Mod}(\Sigma) \times \mathcal{NS}_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|}$.[3]

The *extension of* **INS** *by sentences* $\mathcal{NS}$ is $\mathbf{INS}^+ = \langle \mathbf{Sign}, \mathbf{Sen}^+, \mathbf{Mod}, \langle \models_\Sigma^+ \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$, where for $\Sigma \in |\mathbf{Sign}|$, $\mathbf{Sen}^+(\Sigma) = \mathbf{Sen}(\Sigma) \cup \{\lceil \tau(\varphi') \rceil \mid \varphi' \in \mathcal{NS}_{\Sigma'}, \tau \colon \Sigma' \to \Sigma\}$.[4] Then for $M \in \mathbf{Mod}(\Sigma)$, $M \models_\Sigma^+ \varphi$ iff $M \models_\Sigma \varphi$ for $\varphi \in \mathbf{Sen}(\Sigma)$, and for $\varphi' \in \mathcal{NS}_{\Sigma'}$, $\tau \colon \Sigma' \to \Sigma$, $M \models_\Sigma^+ \lceil \tau(\varphi') \rceil$ iff $M|_\tau \models_{\Sigma'}^{\mathcal{NS}} \varphi'$. Finally, for $\sigma \colon \Sigma \to \Sigma''$, $\mathbf{Sen}^+(\sigma)(\varphi) = \mathbf{Sen}(\sigma)(\varphi)$ for $\varphi \in \mathbf{Sen}(\Sigma)$, and for $\varphi' \in \mathcal{NS}_{\Sigma'}$, $\tau \colon \Sigma' \to \Sigma$, $\mathbf{Sen}^+(\sigma)(\lceil \tau(\varphi') \rceil) = \lceil (\tau;\sigma)(\varphi') \rceil$.

This defines an institution where for $\Sigma \in |\mathbf{Sign}|$, the new sentences $\varphi \in \mathcal{NS}_\Sigma$ are present as $\lceil id_\Sigma(\varphi) \rceil$. Such an extension does not affect entailments between sets of **INS**-sentences.

Suppose then that for each signature we are given a class of (new) "models" with predefined satisfaction relation for the **INS**-sentences, organised as a signature-indexed family of classes with relations: $\mathcal{NM} = \langle \mathcal{NM}_\Sigma, \models_\Sigma^{\mathcal{NM}} \subseteq \mathcal{NM}_\Sigma \times \mathbf{Sen}(\Sigma) \rangle_{\Sigma \in |\mathbf{Sign}|}$.

The *extension of* **INS** *by models* $\mathcal{NM}$ is $\mathbf{INS}^+ = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}^+, \langle \models_\Sigma^+ \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$, where for $\Sigma \in |\mathbf{Sign}|$, $\mathbf{Mod}^+(\Sigma) = \mathbf{Mod}(\Sigma) \cup \{\lceil M'|_\tau \rceil \mid M' \in \mathcal{NM}_{\Sigma'}, \tau \colon \Sigma \to \Sigma'\}$.[5] Then for $\varphi \in \mathbf{Sen}(\Sigma)$, $M \models_\Sigma^+ \varphi$ iff $M \models_\Sigma \varphi$ for $M \in \mathbf{Mod}(\Sigma)$, and for $M' \in \mathcal{NM}_{\Sigma'}$, $\tau \colon \Sigma \to \Sigma'$, $\lceil M'|_\tau \rceil \models_\Sigma^+ \varphi$ iff $M' \models_{\Sigma'}^{\mathcal{NS}} \tau(\varphi)$. Finally, for $\sigma \colon \Sigma'' \to \Sigma$, $\mathbf{Mod}^+(\sigma)(M) = M|_\sigma$ for $M \in \mathbf{Mod}(\Sigma)$, and for $M' \in \mathcal{NS}_{\Sigma'}$, $\tau \colon \Sigma \to \Sigma'$, $\mathbf{Mod}^+(\sigma)(\lceil M'|_\tau \rceil) = \lceil M'|_{\sigma;\tau} \rceil$.

This defines an institution where for $\Sigma \in |\mathbf{Sign}|$, the new models $M \in \mathcal{NM}_\Sigma$ are present as $\lceil M|_{id_\Sigma} \rceil$. Such an extension may spoil some entailments between sets of **INS**-sentences: for $\Phi, \Psi \subseteq \mathbf{Sen}(\Sigma)$ if $\Phi \models^+ \Psi$ then $\Phi \models \Psi$ but the opposite implication may fail.

---

[3] To avoid any foundational problems below, we may assume that **Sign** is small, or that it is locally small and $\mathcal{NS}_\Sigma \neq \emptyset$ for a set of signatures Σ only.

[4] $\lceil \tau(\varphi') \rceil$ is just our syntax for the sentence $\varphi' \in \mathcal{NS}_{\Sigma'}$ formally "fitted" by $\tau \colon \Sigma' \to \Sigma$ to the signature Σ; we assume that no sentences of the form $\lceil \tau(\varphi') \rceil$ are present in **INS**.

[5] $\lceil M'|_\tau \rceil$ is just our syntax for the model $M' \in \mathcal{NM}_{\Sigma'}$ formally "fitted" by $\tau \colon \Sigma \to \Sigma'$ to the signature Σ; we assume that no models of the form $\lceil M'|_\tau \rceil$ are present in **INS**.

When using these constructions, we often present new sentences $\mathcal{NS}$ and new models $\mathcal{NM}$ somewhat informally, avoiding much of the notational burden. We disregard the formal distinction between $\varphi \in \mathcal{NS}_\Sigma$ and $\lceil id_\Sigma(\varphi)\rceil$, and between $M \in \mathcal{NM}_\Sigma$ and $\lceil M|_{id_\Sigma}\rceil$. For $\Sigma \in |\mathbf{Sign}|$, we may define the satisfaction relations $\models^{\mathcal{NS}}_\Sigma$ indirectly by defining $Mod^+(\varphi) \subseteq \mathbf{Mod}(\Sigma)$ for $\varphi \in \mathcal{NS}_\Sigma$ (then for $M \in \mathbf{Mod}(\Sigma)$, $M \models^{\mathcal{NS}}_\Sigma \varphi$ iff $M \in Mod^+(\varphi)$), and $\models^{\mathcal{NM}}_\Sigma$ by defining $Th^+(M) \subseteq \mathbf{Sen}(\Sigma)$ for $M \in \mathcal{NM}_\Sigma$ (then for $\varphi \in \mathbf{Sen}(\Sigma)$, $M \models^{\mathcal{NM}}_\Sigma \varphi$ iff $\varphi \in Th^+(M)$).

▶ **Example 2.** We may define an extension of the institution **PL** of propositional logic by sentences, adding for each signature Σ a new sentence $\mathtt{even}_\Sigma$ defined to hold in models that contain an even number of propositional variables. In the resulting extension $\mathbf{PL}^+$ of **PL**, for any $\sigma\colon \Sigma \to \Sigma'$, $\mathbf{Sen}^+(\sigma)(\mathtt{even}_\sigma)$ is $\lceil\sigma(\mathtt{even}_\Sigma)\rceil$, which is distinct from $\mathtt{even}_{\Sigma'}$. Indeed, putting $\mathbf{Sen}^+(\sigma)(\mathtt{even}_\Sigma) = \mathtt{even}_{\Sigma'}$ would violate the satisfaction condition for some σ.

▶ **Example 3.** We may also define an extension of **PL** by models, adding for each signature Σ and Σ-model M, a new model $\widetilde{M}$, where the satisfaction of propositional sentences in $\widetilde{M}$ is defined by interpreting propositional connectives as usual, but the truth of propositional variables is determined separately for each occurrence, from left to right, and after each occurrence the values of all propositional variables are "swapped" (from true to false and vice versa). Thus, for instance the sentence $p \wedge q$ holds in $\widetilde{M}$ if $p \in M$ and $q \notin M$, and $p \vee p$ holds in any model $\widetilde{M}$. In the resulting extension $\mathbf{PL}^+$, for any signature Σ and $M \in \mathbf{Mod}(\Sigma)$, for any $\sigma\colon \Sigma' \to \Sigma$, $\widetilde{M}|_\sigma$ (that is, $\mathbf{Mod}^+(\sigma)(\widetilde{M})$) and $\widetilde{M|_\sigma}$ are distinct Σ'-models, even though they are logically equivalent (satisfy exactly the same propositional sentences).

2.4 Institution morphisms

The are a number of standard notions to capture relationships between different institutions, with institution morphisms [25] and comorphisms [26] perhaps the most common.

For any institutions **INS** and $\mathbf{INS}'$, an *institution morphism* $\mu\colon \mathbf{INS} \to \mathbf{INS}'$ consists of:

- a functor $\mu^{Sign}\colon \mathbf{Sign} \to \mathbf{Sign}'$,
- a natural transformation $\mu^{Sen}\colon \mu^{Sign};\mathbf{Sen}' \to \mathbf{Sen}$, i.e., a family of functions $\mu^{Sen}_\Sigma\colon \mathbf{Sen}'(\mu^{Sign}(\Sigma)) \to \mathbf{Sen}(\Sigma)$ natural in $\Sigma \in |\mathbf{Sign}|$, and
- a natural transformation $\mu^{Mod}\colon \mathbf{Mod} \to (\mu^{Sign})^{op};\mathbf{Mod}'$, i.e., a family of functions $\mu^{Mod}_\Sigma\colon \mathbf{Mod}(\Sigma) \to \mathbf{Mod}'(\mu^{Sign}(\Sigma))$ natural in $\Sigma \in |\mathbf{Sign}|$

such that for all $\Sigma \in |\mathbf{Sign}|$, $\varphi' \in \mathbf{Sen}'(\mu^{Sign}(\Sigma))$ and $M \in \mathbf{Mod}(\Sigma)$, $M \models_\Sigma \mu^{Sen}_\Sigma(\varphi')$ iff $\mu^{Mod}_\Sigma(M) \models'_{\mu^{Sign}(\Sigma)} \varphi'$ (this is referred to as the satisfaction condition for μ). Institution morphisms compose in the obvious, component-wise manner [25].

Semantic entailment is preserved by translation under institution morphisms: for any signature $\Sigma \in |\mathbf{Sign}|$ and sets of sentences $\Phi', \Psi' \subseteq \mathbf{Sen}'(\mu^{Sign}(\Sigma))$, if $\Phi' \models' \Psi'$ then $\mu^{Sen}_\Sigma(\Phi') \models \mu^{Sen}_\Sigma(\Psi')$. If the translation of models $\mu^{Mod}_\Sigma\colon \mathbf{Mod}(\Sigma) \to \mathbf{Mod}'(\mu^{Sign}(\Sigma))$ is surjective then also the opposite implication holds, and $\Phi' \models' \Psi'$ iff $\mu^{Sen}_\Sigma(\Phi') \models \mu^{Sen}_\Sigma(\Psi')$.

For instance, there is an obvious institution morphism from the institution **FO** of first-order logic to the institution **PL** of propositional logic (removing from signatures everything but nullary predicates). For further examples of institution morphisms we refer to [35, 13].

In this paper we deal only with institution morphisms that leave the signature category intact, that is, where the signature functor is the identity. This also allows us to disregard institution comorphisms, since essentially they are the same as institution morphisms then.

An institution morphism $\mu\colon \mathbf{INS} \to \mathbf{INS}'$ is *logically trivial* if it is the identity on signatures and surjective on sentences and models, that is, $\mathbf{Sign}' = \mathbf{Sign}$ and $\mu^{Sign} = id_{\mathbf{Sign}}$, and for all signatures $\Sigma \in |\mathbf{Sign}|$, the functions $\mu^{Sen}_\Sigma\colon \mathbf{Sen}'(\Sigma) \to \mathbf{Sen}(\Sigma)$ and $\mu^{Mod}_\Sigma\colon \mathbf{Mod}(\Sigma) \to \mathbf{Mod}'(\Sigma)$ are surjective. The following fact justifies this terminology:[6]

▶ **Fact 4.** *Logically trivial institution morphisms identify only sentences and models that are logically equivalent.*

Special institution morphisms relate institutions with their extensions by new sentences and by new models, respectively, introduced in Sect. 2.3. If $\mathbf{INS}^+_{\mathcal{NS}}$ is the extension of $\mathbf{INS}$ by new sentences $\mathcal{NS}$ then there is an institution morphism $\mu_{\mathcal{NS}}\colon \mathbf{INS}^+_{\mathcal{NS}} \to \mathbf{INS}$, where $\mu^{Sign}_{\mathcal{NS}}$ and $\mu^{Mod}_{\mathcal{NS}}$ are identities (the former is the identity functor on $\mathbf{Sign}$, the latter is the identity natural transformation on $\mathbf{Mod}\colon \mathbf{Sign}^{op} \to \mathbf{Class}$), and for $\Sigma \in |\mathbf{Sign}|$, $\mu^{Sen}_{\mathcal{NS}}\colon \mathbf{Sen}(\Sigma) \to \mathbf{Sen}^+_{\mathcal{NS}}(\Sigma)$ are inclusions. Similarly, if $\mathbf{INS}^+_{\mathcal{NM}}$ is the extension of $\mathbf{INS}$ by new models $\mathcal{NM}$ then there is an institution morphism $\mu_{\mathcal{NM}}\colon \mathbf{INS} \to \mathbf{INS}^+_{\mathcal{NM}}$, where $\mu^{Sign}_{\mathcal{NM}}$ and $\mu^{Sen}_{\mathcal{NM}}$ are identities, and for $\Sigma \in |\mathbf{Sign}|$, $\mu^{Mod}_{\mathcal{NM}}\colon \mathbf{Mod}(\Sigma) \to \mathbf{Mod}^+_{\mathcal{NM}}(\Sigma)$ are inclusions.

▶ **Fact 5.** *Let $\mathbf{INS}'$ and $\mathbf{INS}''$ be institutions with a common signature category $\mathbf{Sign}$. Consider an institution morphism $\mu\colon \mathbf{INS}' \to \mathbf{INS}''$ with $\mu^{Sign} = id_{\mathbf{Sign}}$. Then for some institution $\mathbf{INS}$, extension $\mathbf{INS}^+_{\mathcal{NS}}$ of $\mathbf{INS}$ by new sentences, extension $\mathbf{INS}^+_{\mathcal{NM}}$ of $\mathbf{INS}$ by new models, and logically trivial institution morphisms $\mu'\colon \mathbf{INS}' \to \mathbf{INS}^+_{\mathcal{NS}}$ and $\mu''\colon \mathbf{INS}^+_{\mathcal{NM}} \to \mathbf{INS}''$ we have $\mu = \mu';\mu_{\mathcal{NS}};\mu_{\mathcal{NM}};\mu''$: $\underbrace{\mathbf{INS}' \xrightarrow{\mu'} \mathbf{INS}^+_{\mathcal{NS}} \xrightarrow{\mu_{\mathcal{NS}}} \mathbf{INS} \xrightarrow{\mu_{\mathcal{NM}}} \mathbf{INS}^+_{\mathcal{NM}} \xrightarrow{\mu''} \mathbf{INS}''}_{\mu}$*

Proof (hint): *Use $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}'', \mathbf{Mod}', \langle \models_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$, where for $\Sigma \in \mathbf{Sign}$, $M' \in \mathbf{Mod}'(\Sigma)$ and $\varphi'' \in \mathbf{Sen}''(\Sigma)$, $M' \models_\Sigma \varphi''$ iff $M' \models'_\Sigma \mu^{Sen}_\Sigma(\varphi')$.*

3 Interpolation

3.1 Classical interpolation

The well-known Craig interpolation theorem [12] states that if an implication between two first-order formulae $\varphi \Rightarrow \psi$ holds then there is a formula θ that uses only the symbols common to φ and ψ such that both $\varphi \Rightarrow \theta$ and $\theta \Rightarrow \psi$ hold; θ is then called an *interpolant* for φ and ψ. This is one of the key properties of first-order logic, with numerous applications, including simpler proofs of similarly famous and important results like the Robinson consistency [32] and Beth definability [4] theorems. The interpolation property has been investigated (and proved or disproved) for many standard extensions (and fragments) of first-order logic [40] as well as for other logical systems, for instance for various modal and intuitionistic logics [21].

The above statement of the interpolation property implicitly involves the following union/intersection square of signatures:

[6] Due to the page limit imposed, proofs are either omitted here or reduced to hints only.

$$\begin{tikzcd}
 & & \boxed{\varphi \Rightarrow \psi} & & \\
 & & \Sigma_p \cup \Sigma_c & & \\
\boxed{\varphi \Rightarrow \theta} & \Sigma_p \arrow[ur] & & \Sigma_c \arrow[ul] & \boxed{\theta \Rightarrow \psi} \\
 & & \Sigma_p \cap \Sigma_c \arrow[ul] \arrow[ur] & & \\
 & & \boxed{\theta} & &
\end{tikzcd}$$

where Σ_p and Σ_c are (first-order) signatures for φ and ψ, respectively, and the arrows indicate signature inclusions.

As hinted at in Sect. 1, interpolation proved indispensable for many foundational aspects of computer science, in particular in the area of software specification and development. However, the classical formulation of Craig's interpolation for many applications requires some generalisations, which perhaps do not bring much new insight in the framework of first-order logic, but may be important when other logical systems are considered.

To begin with, the use of implication should be replaced by entailment. Then, we should deal with entailments between sets of sentences, rather than between individual sentences (strictly speaking, this is needed for the premise φ and especially for the interpolant θ – for notational symmetry, we do this for the conclusion ψ as well). Both these generalisations are irrelevant for first-order logic, where implication captures semantic entailment, and a set of sentences in the premise of each single-conclusion entailment may always be replaced by a single sentence (since the logic is compact and has conjunction). However, for instance, working in equational logic we have no implication available, and interpolants cannot be always expressed as a single equation – even though the interpolation property holds if sets of equations are permitted as interpolants [33].

Perhaps most importantly, for instance in applications where parameterised specifications and their "pushout-style" instantiations [42, 20] are involved, we have to go beyond union/intersection squares of signatures and inclusions to relate the signatures. More general classes of signature squares are needed, with non-injective signature morphisms necessary to capture for instance morphisms used to "fit" actual to formal parameters. Typically in applications at least pushouts of signature morphisms are involved, sometimes additionally restricted to indicated classes of morphisms permitted at the "bottom-left" and "bottom-right" of the squares, respectively [42, 41, 5, 13, 30]. However, for the purposes of this paper we will consider interpolation properties for an arbitrary commutative square of signature morphisms.

3.2 Interpolation in an institution

Throughout the rest of this paper, let $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle \models_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$ be an arbitrary institution, and we study interpolation properties over the following commutative square $(*)$ of signature morphisms:[7]

[7] To help memorising the notation: p for *premise*, c for *conclusion*, u for *union* and i for *intersection* (or *interpolant*).

$$\begin{tikzcd}
 & \Sigma_u & \\
\Sigma_p \arrow[ur, "\sigma_{pu}"] & (*) & \Sigma_c \arrow[ul, "\sigma_{cu}"'] \\
 & \Sigma_i \arrow[ul, "\sigma_{ip}"] \arrow[ur, "\sigma_{ic}"'] &
\end{tikzcd}$$

Let $\Phi \subseteq \mathbf{Sen}(\Sigma_p)$ and $\Psi \subseteq \mathbf{Sen}(\Sigma_c)$ be such that $\sigma_{pu}(\Phi) \models_{\Sigma_u} \sigma_{cu}(\Psi)$. An *interpolant* for Φ and Ψ (over diagram $(*)$) is a set $\Theta \subseteq \mathbf{Sen}(\Sigma_i)$ of Σ_i-sentences such that $\Phi \models_{\Sigma_p} \sigma_{ip}(\Theta)$ and $\sigma_{ic}(\Theta) \models_{\Sigma_c} \Psi$.

$$\begin{tikzcd}
 & & \boxed{\sigma_{pu}(\Phi) \models \sigma_{cu}(\Psi)} & & \\
 & & \Sigma_u & & \\
\boxed{\Phi \models \sigma_{ip}(\Theta)} & \Sigma_p \arrow[ur, "\sigma_{pu}"] & & \Sigma_c \arrow[ul, "\sigma_{cu}"'] & \boxed{\sigma_{ic}(\Theta) \models \Psi} \\
 & & \Sigma_i \arrow[ul, "\sigma_{ip}"] \arrow[ur, "\sigma_{ic}"'] & & \\
 & & \boxed{\Theta} & &
\end{tikzcd}$$

To simplify some further statements, if $\sigma_{pu}(\Phi) \not\models_{\Sigma_u} \sigma_{cu}(\Psi)$ then we say that any set $\Theta \subseteq \mathbf{Sen}(\Sigma_i)$ is an interpolant for Φ and Ψ (over diagram $(*)$).

We say that a commutative square $(*)$ of signature morphisms *admits interpolation* if there is an interpolant for every $\Phi \subseteq \mathbf{Sen}(\Sigma_p)$ and $\Psi \subseteq \mathbf{Sen}(\Sigma_c)$ such that $\sigma_{pu}(\Phi) \models \sigma_{cu}(\Psi)$.

▶ **Example 6.** In the institution **FO** of first-order logic, and in any of its variants mentioned in Example 1, if the square $(*)$ is a pushout and at least one of $\sigma_{ip}\colon \Sigma_i \to \Sigma_p$, $\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ is injective on sorts then $(*)$ admits interpolation; otherwise interpolation may fail for $(*)$ (see [6]). In the institution **EQ** of equational logic if the square $(*)$ is a pushout and $\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ is injective then $(*)$ admits interpolation; otherwise interpolation may fail for $(*)$, and in $\mathbf{EQ}_\emptyset$, where empty carriers are permitted, interpolation may fail even for union/intersection squares of signatures (see [39]). In the institution **PL** of propositional logic, all pushouts admit interpolation.

It is well known that the interpolation property of a logical system is fragile. When the logic is extended, when new models or sentences are added, the interpolation property may easily be spoiled. Clearly, this may happen when entirely new signatures are added, with new models and sentences over them. Therefore, in this paper we consider the category of signatures to be fixed, and consider only such extensions of institutions that preserve it.

Throughout the rest of the paper we study how the interpolation property may be spoiled by adding new models or sentences. This will be done from a "local" perspective, for particular commutative squares of signature morphisms, as well as for particular interpolants.

We say that an interpolant $\Theta \subseteq \mathbf{Sen}(\Sigma_i)$ for $\Phi \subseteq \mathbf{Sen}(\Sigma_p)$ and $\Psi \subseteq \mathbf{Sen}(\Sigma_c)$ (over diagram $(*)$) is *stable* under extensions of the institution by models if for every extension $\mathbf{INS}^+$ of **INS** by new models, Θ is an interpolant for Φ and Ψ in $\mathbf{INS}^+$; otherwise we say that the interpolant Θ is *fragile*. Note that adding new sentences cannot spoil a particular interpolant, but may spoil interpolation property for a given diagram.

3.3 Interpolants may be stable

▶ **Lemma 7.** *Consider the diagram* $(*)$ *of signature morphisms.*

1. *If* $\sigma_{ip}\colon \Sigma_i \to \Sigma_p$ *is such that* $\mathbf{Sen}(\sigma_{ip})\colon \mathbf{Sen}(\Sigma_i) \to \mathbf{Sen}(\Sigma_p)$ *is surjective and* $\sigma_{cu}\colon \Sigma_c \to \Sigma_u$ *is conservative then* $(*)$ *admits interpolation.*
2. *If* $\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ *is such that* $\mathbf{Sen}(\sigma_{ic})\colon \mathbf{Sen}(\Sigma_i) \to \mathbf{Sen}(\Sigma_c)$ *is surjective and* $\sigma_{pu}\colon \Sigma_p \to \Sigma_u$ *is conservative then* $(*)$ *admits interpolation.*

Proof (hint): *An interpolant for* $\Phi \subseteq \mathbf{Sen}(\Sigma_p)$ *and* $\Psi \subseteq \mathbf{Sen}(\Sigma_c)$ *is* $\sigma_{ip}^{-1}(\Phi)$ *under* 1., *or* $\sigma_{ic}^{-1}(\Psi)$ *under* 2.

A trivial special case here is when σ_{ip} and σ_{cu}, or σ_{ic} and σ_{pu}, are isomorphisms, which can be further refined as follows:

▶ **Corollary 8.** *The diagram* $(*)$ *of signature morphisms admits interpolation if*

1. $\sigma_{ip}\colon \Sigma_i \to \Sigma_p$ *is a retraction and* $\sigma_{cu}\colon \Sigma_c \to \Sigma_u$ *is a coretraction, or*
2. $\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ *is a retraction and* $\sigma_{pu}\colon \Sigma_p \to \Sigma_u$ *is a coretraction.*

Proof (hint): *The requirements here imply the respective conditions in Lemma 7.*

This shows that if the signature morphisms in $(*)$ satisfy the premises of Cor. 8 then the diagram enjoys a stable interpolation property, which cannot be spoiled by any institution extension that leaves the category of signatures unchanged! No matter how we add new models or sentences, the interpolation property is ensured by the properties of the signature morphisms involved, and the implied properties of the translations of sentences and reducts of models they induce in the institution and in any of its extensions.

The conditions stated in Cor. 8 are in fact quite strong and in many practical situations do not depart too far from the trivial case when Σ_p is (up to isomorphism) included in Σ_c or vice versa. Namely, when the diagram $(*)$ is a pushout then condition 1. implies that $\sigma_{cu}\colon \Sigma_c \to \Sigma_u$ is an isomorphism, and condition 2. implies that $\sigma_{pu}\colon \Sigma_p \to \Sigma_u$ is an isomorphism. Dually, when $(*)$ is a pullback then condition 1. implies that $\sigma_{ip}\colon \Sigma_i \to \Sigma_p$ is an isomorphism, and condition 2. implies that $\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ is an isomorphism.

▶ **Fact 9.** *Let* $\mu\colon \mathbf{INS} \to \mathbf{INS}'$ *be a logically trivial institution morphism. Diagram* $(*)$ *in the category of signatures admits interpolation in* $\mathbf{INS}$ *iff it admits interpolation in* $\mathbf{INS}'$.

Facts 5 and 9 imply that for our study of the fragility of interpolation institution extensions by new models and by new sentences are of primary importance.

4 Spoiling an interpolant by new models

Recall that we study interpolation over a commutative square of signature morphisms $(*)$ in an institution $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle \models_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$. Throughout this section, let $\Phi \subseteq \mathbf{Sen}(\Sigma_p)$ and $\Psi \in \mathbf{Sen}(\Sigma_c)$ be such that $\sigma_{pu}(\Phi) \models \sigma_{cu}(\Psi)$, and let $\Theta \subseteq \mathbf{Sen}(\Sigma_i)$ be an interpolant for Φ and Ψ in $\mathbf{INS}$.

▶ **Lemma 10.** *Suppose that there exists a set of* Σ_p*-sentences* $\Phi^\bullet \supseteq \Phi$ *such that* $\sigma_{ip}(\Theta) \not\subseteq \Phi^\bullet$ *and for all signature morphisms* $\tau\colon \Sigma_u \to \Sigma_p$, *if* $\tau(\sigma_{pu}(\Phi)) \subseteq \Phi^\bullet$ *then* $\tau(\sigma_{cu}(\Psi)) \subseteq \Phi^\bullet$. *Then the interpolant* Θ *for* Φ *and* Ψ *is not stable under extensions of* $\mathbf{INS}$ *by models.*

Proof (hint): *Extend* $\mathbf{INS}$ *by a new* Σ_p*-model* M *with* $Th^+(M) = \Phi^\bullet$. *Then still* $\sigma_{pu}(\Phi) \models^+ \sigma_{cu}(\Psi)$, *but* $\Phi \not\models^+ \sigma_{ip}(\Theta)$.

The key property of the set $\Phi^\bullet$ in the above lemma is that it cannot be used to separate $\sigma_{pu}(\Phi)$ from $\sigma_{cu}(\Psi)$ via any morphism $\tau\colon \Sigma_u \to \Sigma_p$. More formally, for any signatures $\Sigma, \Sigma' \in |\mathbf{Sign}|$, we say that $\Upsilon \subseteq \mathbf{Sen}(\Sigma)$ *never separates* $\Phi' \subseteq \mathbf{Sen}(\Sigma')$ *from* $\Psi' \subseteq \mathbf{Sen}(\Sigma')$

when for all morphisms $\tau\colon \Sigma' \to \Sigma$, if $\tau(\Phi') \subseteq \Upsilon$ then $\tau(\Psi') \subseteq \Upsilon$. For any set $\Phi \subseteq \mathbf{Sen}(\Sigma)$, we denote by $[\Phi' \overset{\Sigma'}{\underset{\Sigma}{\leadsto}} \Psi'](\Phi)$ the least set of Σ-sentences that contains Φ and never separates Φ' from Ψ' (it exists since the family of such sets is closed under intersection and is nonempty).

▶ **Corollary 11.** *If $\sigma_{ip}(\Theta) \not\subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\leadsto}} \sigma_{cu}(\Psi)](\Phi)$ then the interpolant Θ for Φ and Ψ is not stable under extensions of* **INS** *by models.*

▶ **Lemma 12.** *Suppose that there exists a set of Σ_c-sentences $\Psi^\circ \subseteq \mathbf{Sen}(\Sigma_c)$ such that $\Psi \cap \Psi^\circ \neq \emptyset$, $\sigma_{ic}(\Theta) \cap \Psi^\circ = \emptyset$ and for all signature morphisms $\tau\colon \Sigma_u \to \Sigma_c$, if $\tau(\sigma_{cu}(\Psi)) \cap \Psi^\circ \neq \emptyset$ then $\tau(\sigma_{pu}(\Phi)) \cap \Psi^\circ \neq \emptyset$. Then the interpolant Θ for Φ and Ψ is not stable under extensions of* **INS** *by models.*
Proof (hint): *Extend* **INS** *by a new Σ_c-model N with $Th^+(N) = \mathbf{Sen}(\Sigma_c) \setminus \Psi^\circ$. Then still $\sigma_{pu}(\Phi) \models^+ \sigma_{cu}(\Psi)$, but $\sigma_{ip}(\Theta) \not\models^+ \Psi$.*

To refine Lemma 12 in the style of Cor. 11, notice that the requirement on $\Psi^\circ \subseteq \mathbf{Set}(\Sigma_c)$ that for $\tau\colon \Sigma_u \to \Sigma_c$, if $\tau(\sigma_{cu}(\Psi)) \cap \Psi^\circ \neq \emptyset$ then $\tau(\sigma_{pu}(\Phi)) \cap \Psi^\circ \neq \emptyset$, means that the set $\mathbf{Sen}(\Sigma_c) \setminus \Psi^\circ$ never separates $\sigma_{pu}(\Phi)$ from $\sigma_{cu}(\Psi)$.

▶ **Corollary 13.** *If $\Psi \not\subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\leadsto}} \sigma_{cu}(\Psi)](\sigma_{ic}(\Theta))$ then the interpolant Θ for Φ and Ψ is not stable under extension of* **INS** *by models.*

Corollaries 11 and 13 present sufficient conditions that ensure that a particular interpolant may be spoiled by an extension of the institution by new models. In fact, these conditions jointly are also necessary:

▶ **Theorem 14.** *The interpolant Θ for Φ and Ψ is stable under extensions of* **INS** *by models if and only if the following conditions hold:*
1. *$\sigma_{ip}(\Theta) \subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\leadsto}} \sigma_{cu}(\Psi)](\Phi)$, and*
2. *$\Psi \subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\leadsto}} \sigma_{cu}(\Psi)](\sigma_{ic}(\Theta))$.*

Proof (hint): *In any extension* $\mathbf{INS}^+$ *of* **INS** *by models such that $\sigma_{pu}(\Phi) \models^+ \sigma_{cu}(\Psi)$, if $\Phi \not\models^+ \sigma_{ip}(\Theta)$ then* 1. *fails, and if $\sigma_{ic}(\Theta) \not\models^+ \Psi$ then* 2. *fails, which proves the "if" part. The "only if" part follows by Corollaries 11 and 13.*

The above theorem gives precise conditions that ensure stability of a particular interpolant under extensions of the institution by new models. Equivalently, this is a precise characterisation of specific interpolation properties that can be spoiled by adding new abstract models. It should be stressed that the conditions in use are purely "syntactic" – they do not refer to the semantic properties of the sets of sentences involved, and depend on a specific syntactic form of the sentences, and the conclusions may change when the sentences considered are replaced by semantically equivalent sentences that are of a different syntactic form.

▶ **Example 15.** Consider a trivial example in the institution **PL** of propositional logic. In the diagram $(*)$, let $\Sigma_p = \{p, r\}$, $\Sigma_c = \{p, q\}$, $\Sigma_u = \Sigma_p \cup \Sigma_c = \{r, p, q\}$, $\Sigma_i = \Sigma_p \cap \Sigma_c = \{p\}$, and the four signature morphisms are inclusions.

Let φ be $r \wedge p$ and ψ be $p \vee q$.[8] Clearly, $\varphi \models \psi$, and φ and ψ have a number of distinct interpolants in **PL**. One such interpolant for φ and ψ is p. Consider the **PL**-model $M = \{r\} \in \mathbf{Mod}_{\mathbf{PL}}(\Sigma_p)$. Let $\mathbf{PL}^+$ be an extension of **PL** by a new Σ_p-model $\widetilde{M}$ (with

[8] When convenient, we write φ for $\{\varphi\}$, relying on the context to impose such identification of a sentence with the one-element set that contains it.

interpretation of propositional sentences "swapping" the valuation of propositional variables, as in Example 3). Then $\widetilde{M} \models^{+} r \wedge p$ while $\widetilde{M} \not\models^{+} p$, and so p is not an interpolant for φ and ψ in $\mathbf{PL}^{+}$. In fact, $\Phi^{\bullet} = \{\varphi \in \mathbf{Sen}_{\mathbf{PL}}(\Sigma_p) \mid \widetilde{M} \models^{+} \varphi\}$ satisfies the premises of Lemma 10.

Moreover, one can easily calculate that $[r \wedge p \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} p \vee q](r \wedge p) = \{r \wedge p, p \vee r, p \vee p\} \subseteq \mathbf{Sen}_{\mathbf{PL}}(\Sigma_p)$ (there are exactly two morphisms from Σ_u to Σ_p that map $r \wedge p$ to $r \wedge p$, they are identities on $\{p, r\}$ and map q to any of the symbols in Σ_p). Thus, by Cor. 11, any interpolant for φ and ψ other than $p \vee p$ may be spoiled by extending $\mathbf{PL}$ by new models.

Indeed, $p \vee p$ is an interpolant for φ and ψ. Since no morphism from Σ_u to Σ_c maps $r \wedge p$ to $p \vee p$, we have $[r \wedge p \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} p \vee q](p \vee p) = \{p \vee p\} \subseteq \mathbf{Sen}_{\mathbf{PL}}(\Sigma_c)$, and so by Cor. 13 in some extension of $\mathbf{PL}$ by new models $p \vee p$ is not an interpolant for φ and ψ. For instance, consider $\mathbf{PL}$-model $N = \{q\} \in \mathbf{Mod}_{\mathbf{PL}}(\Sigma_c)$. Let $\mathbf{PL}^{+}$ be the extension of $\mathbf{PL}$ by a new Σ_c-model $\widetilde{N}$ (with interpretation of propositional sentences "swapping" the valuation of propositional variables, as in Example 3). Then $\widetilde{N} \models^{+} p \vee p$ while $\widetilde{N} \not\models^{+} p \vee q$, and so $p \vee p$ is not an interpolant for φ and ψ in $\mathbf{PL}^{+}$. Summing up: none of the interpolants for φ and ψ in $\mathbf{PL}$ is stable under extensions of $\mathbf{PL}$ by new models.

Let now φ' be $(p \vee r) \wedge (p \vee \neg r)$ and ψ' be $(p \vee q) \wedge (p \vee \neg q)$. Perhaps the most obvious interpolant for φ' and ψ' is p. This interpolant, however, is fragile. Namely,

$$[\varphi' \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \psi']((p \vee r) \wedge (p \vee \neg r)) = \{(p \vee r) \wedge (p \vee \neg r), (p \vee p) \wedge (p \vee \neg p)\} \subseteq \mathbf{Sen}_{\mathbf{PL}}(\Sigma_p).$$

Thus, by Cor. 11, p is not an interpolant for φ' and ψ' in an extension of $\mathbf{PL}$ by new models.

Another interpolant for φ' and ψ' in $\mathbf{PL}$ is $(p \vee p) \wedge (p \vee \neg p)$. Since $(p \vee p) \wedge (p \vee \neg p) \in [\varphi' \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \psi']((p \vee r) \wedge (p \vee \neg r))$, Cor. 11 cannot be used here to conclude that this interpolant gets spoiled in an extension of $\mathbf{PL}$ by new models. Moreover,

$$[\varphi' \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} \psi']((p \vee p) \wedge (p \vee \neg p)) = \{(p \vee p) \wedge (p \vee \neg p), (p \vee q) \wedge (p \vee \neg q)\} \subseteq \mathbf{Sen}_{\mathbf{PL}}(\Sigma_c).$$

Consequently, Cor. 13 does not apply here either, and by Thm. 14 the interpolant $(p \vee p) \wedge (p \vee \neg p)$ for φ' and ψ' in $\mathbf{PL}$ is stable under extensions of $\mathbf{PL}$ by new models.

5 Spoiling interpolation by new models

As in the previous section, consider an institution $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle \models_{\Sigma} \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$, commutative square of signature morphisms $(*)$, and sets of sentences $\Phi \subseteq \mathbf{Sen}(\Sigma_c)$ and $\Psi \in \mathbf{Sen}(\Sigma_c)$ such that $\sigma_{pu}(\Phi) \models \sigma_{cu}(\Psi)$. Theorem 14 gives the exact characterisation of interpolants that are stable under extensions of $\mathbf{INS}$ by new models. Of course, this also characterises interpolants that are fragile. In this section we characterise situations where all interpolants for the premise Φ and conclusion Ψ may be spoiled at once.

▶ **Corollary 16.** *Let $\Phi^{*} = [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\Phi)$ and $\Psi^{*} = [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\sigma_{ic}(\sigma_{ip}^{-1}(\Phi^{*})))$. If $\Psi \not\subseteq \Psi^{*}$ then there is an extension $\mathbf{INS}^{+}$ of $\mathbf{INS}$ by models such that there is no interpolant for Φ and Ψ in $\mathbf{INS}^{+}$.*

Proof (hint): *Extend $\mathbf{INS}$ by a new Σ_p-model M with $Th(M)^{+} = \Phi^{*}$ and a new Σ_c-model N with $Th^{+}(N) = \Psi^{*}$.*

The converse of Cor. 16 does not hold, since the conclusion follows as well when we limit our attention to consequences of Φ, rather than all sentences in $\Phi^{*} = [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\Phi)$.

To avoid repetition, for the rest of this section let $\Theta^* = \sigma_{ip}^{-1}([\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\Phi) \cap Th(\Phi))$ (that is, more explicitly: $\Theta^* = \{\theta \in \mathbf{Sen}(\Sigma_i) \mid \sigma_{ip}(\theta) \in [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\Phi), \Phi \models \sigma_{ip}(\theta)\}$).

▶ **Lemma 17.** *If $\Psi \not\subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\sigma_{ic}(\Theta^*))$ then no interpolant for Φ and Ψ is stable under extensions of* **INS** *by models.*

Proof (hint): *For any interpolant Θ for Φ and Ψ, if $\Theta \not\subseteq \Theta^*$ then Θ is not stable by Cor. 11, and if $\Theta \subseteq \Theta^*$ by Cor. 13.*

The thesis of Lemma 17 seems weaker that that of Cor. 11 – but only superficially so:

▶ **Lemma 18.** *If no interpolant for Φ and Ψ is stable under extensions of* **INS** *by models then in some extension of* **INS** *by models Φ and Ψ have no interpolant at all.*

▶ **Corollary 19.** *If $\Psi \not\subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\sigma_{ic}(\Theta^*))$ then in some extension of* **INS** *by models Φ and Ψ have no interpolant at all.*

▶ **Theorem 20.** *There is an interpolant for Φ and Ψ in every extension of* **INS** *by models if and only if $\Psi \subseteq [\sigma_{pu}(\Phi) \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} \sigma_{cu}(\Psi)](\sigma_{ic}(\Theta^*))$ and $\sigma_{ic}(\Theta^*) \models \Psi$.*

Proof (hint): *For the "if" part: under the assumptions, Θ^* is a stable interpolant for Φ and Ψ. For the "only if" part: any stable interpolant Θ for Φ and Ψ satisfies $\Theta \subseteq \Theta^*$.*

▶ **Example 21.** Recall Example 15. As argued there, every interpolant for $r \wedge p$ and $p \vee q$ in **PL** is fragile. Theorem 20 leads to the same conclusion, of course. Namely, as in Example 15, $[r \wedge p \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} q \vee p](r \wedge p) = \{r \wedge p, p \vee r, p \vee p\}$. Then, using the notation Θ^* defined above for the case at hand, $\Theta^* = \{p \vee p\}$. Recalling from Example 15 again: $[r \wedge p \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} p \vee q](\Theta^*) = \{p \vee p\}$, and so $p \vee q \notin [r \wedge p \overset{\Sigma_u}{\underset{\Sigma_c}{\rightsquigarrow}} p \vee q](\Theta^*)$. Thus, by Thm. 20 and Lemma 18, there is an extension of **PL** by models in which $r \wedge p$ and $p \vee q$ have no interpolant.

As in Example 15, let now φ' be $(p \vee r) \wedge (p \vee \neg r)$ and ψ' be $(p \vee q) \wedge (p \vee \neg q)$, and we get $[\varphi' \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \psi'](\varphi') = \{(p \vee r) \wedge (p \vee \neg r), (p \vee p) \wedge (p \vee \neg p)\}$. Therefore, using the notation Θ^* for the current case, $\Theta^* = \{(p \vee p) \wedge (p \vee \neg p\}$, and then $(p \vee q) \wedge (p \vee \neg q) \in [\varphi' \overset{\Sigma_u}{\underset{\Sigma_p}{\rightsquigarrow}} \psi'](\Theta^*)$. Thus, by Thm. 20, $(p \vee r) \wedge (p \vee \neg r)$ and $(p \vee q) \wedge (p \vee \neg q)$ have an interpolant in every extension of **PL** by models, and indeed, in Example 15 we argued independently that $(p \vee p) \wedge (p \vee \neg p)$ is such an interpolant.

6 Spoiling interpolation by new sentences

As before, in an institution $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle \models_\Sigma \rangle_{\Sigma \in |\mathbf{Sign}|} \rangle$ we study interpolation over a commutative square of signature morphisms $(*)$.

Changes to a logical system that may arise when new sentences are introduced are in no sense dual to those resulting from extending the logical system by new models. In particular, new sentences do not modify the entailments between the sentences of the original system, so we cannot expect that we may spoil interpolants for old sentences. However, new sentences (over the premise and conclusion signatures) may lead to new entailments $\sigma_{pu}(\Phi) \models^+ \sigma_{cu}(\Psi)$ with no interpolant for Φ and Ψ. On the other hand, adding appropriate new sentences (over the interpolant signature) may restore (or establish) the interpolation property.

The first rough idea (see for instance the semantic characterisation of interpolation in [13]) is that to spoil interpolation for the diagram $(*)$, we look for a class $\mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i)$ that is not definable in **INS**, and then build an extension $\mathbf{INS}^+$ of **INS** by new sentences $\varphi \in \mathbf{Sen}^+(\Sigma_p)$ and $\psi \in \mathbf{Sen}^+(\Sigma_c)$ such that $Mod^+(\varphi) = \mathcal{K}|_{\sigma_{ip}}^{-1}$ and $Mod^+(\psi) = \mathcal{K}|_{\sigma_{ic}}^{-1}$. Then $\sigma_{pu}(\varphi) \models^+ \sigma_{cu}(\psi)$, and it may seem that there should be no interpolant for φ and ψ, since such an interpolant would have to define $\mathcal{K}$. However, the latter need not be true in general.

One technical nuance is that a set $\Theta \subseteq \mathbf{Sen}^+(\Sigma_i)$ of sentences may be an interpolant for φ and ψ if $Mod^+(\Theta) \supset \mathcal{K}$ but no model in $Mod^+(\Theta) \setminus \mathcal{K}$ has a σ_{ic}-expansion to a model in $\mathbf{Mod}(\Sigma_c)$. Another technicality is that the requirement that $Mod^+(\varphi) = \mathcal{K}|_{\sigma_{ip}}^{-1}$ may be weakened to $Mod^+(\varphi)|_{\sigma_{ip}} = \mathcal{K}$. At the conclusion side, it is enough to assume that all σ_{ic}-expansions of the models in $\mathcal{K}$ are in $Mod(\psi)$, $\mathcal{K}|_{\sigma_{ic}}^{-1} \subseteq Mod(\psi)$, or equivalently, no model in $\mathcal{K}$ is a σ_{ic}-reduct of a Σ_c-model outside $Mod(\psi)$, $\mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i) \setminus ((\mathbf{Mod}(\Sigma_c) \setminus Mod(\psi))|_{\sigma_{ic}})$. We may also permit a gap between $Mod^+(\varphi)|_{\sigma_{ip}}$ and $\mathbf{Mod}(\Sigma_i) \setminus ((\mathbf{Mod}(\Sigma_c) \setminus Mod(\psi))|_{\sigma_{ic}})$ as long as no definable class separates them.

Most importantly though, adding new sentences over signatures Σ_p and Σ_c may result in adding new Σ_i-sentences (as translations of the added sentences), and some Σ_i-model classes that are not definable in **INS** may become definable in $\mathbf{INS}^+$. The following notion will be used to take care of this: for any signature $\Sigma \in |\mathbf{Sign}|$ and collection $\mathcal{F} = \{\langle \Sigma_j, \mathcal{M}_j \rangle \mid \Sigma_j \in |\mathbf{Sign}|, \mathcal{M}_j \subseteq \mathbf{Mod}(\Sigma_j), j \in \mathcal{J}\}$,[9] we say that a class $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma)$ of Σ-models is *definable in* **INS** *from* $\mathcal{F}$ if for some family of signature morphisms $\tau_l \colon \Sigma_{j_l} \to \Sigma$ where $j_l \in \mathcal{J}$, $l \in \mathcal{L}$, and a set $\Phi \subseteq \mathbf{Sen}(\Sigma)$ of Σ-sentences we have $\mathcal{M} = \bigcap_{l \in \mathcal{L}} \mathcal{M}_{j_l}|_{\tau_l}^{-1} \cap Mod(\Phi)$.

▶ **Theorem 22.** *There is an extension* $\mathbf{INS}^+$ *of* **INS** *by new sentences in which the diagram* $(*)$ *does not admit interpolation if and only if there are classes of models* $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma_p)$ *and* $\mathcal{N} \subseteq \mathbf{Mod}(\Sigma_c)$ *such that*

1. $\mathcal{M}|_{\sigma_{pu}}^{-1} \subseteq \mathcal{N}|_{\sigma_{cu}}^{-1}$ *and*
2. *no class of models* $\mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i)$ *such that* $\mathcal{M}|_{\sigma_{ip}} \subseteq \mathcal{K}$ *and* $\mathcal{K}|_{\sigma_{ic}}^{-1} \subseteq \mathcal{N}$ *is definable in* **INS** *from* $\{\langle \Sigma_p, \mathcal{M} \rangle, \langle \Sigma_c, \mathcal{N} \rangle\}$.

Proof (hint): *For the "if" part, extend* **INS** *by new sentences that define* $\mathcal{M}$ *and* $\mathcal{N}$, *respectively. For the "only if" part, if in an extension* $\mathbf{INS}^+$ *of* **INS** *by new sentences there is no interpolant for* $\Phi^+ \subseteq \mathbf{Sen}^+(\Sigma_p)$ *and* $\Psi^+ \subseteq \mathbf{Sen}^+(\Sigma_c)$ *then* $\mathcal{M} = Mod(\Phi^+)$ *and* $\mathcal{N} = Mod(\Psi^+)$ *satisfy* 1. *and* 2.

▶ **Example 23.** Consider an example in the institution $\mathbf{FO_{EQ}}$ of first-order logic with equality. Let all the signatures in the diagram $(*)$ extend Σ_i, which has sort *Nat*, constant $0 \colon Nat$ and operation $s \colon Nat \to Nat$. In addition, Σ_p has $bop \colon Nat \times Nat \to Nat$ and Σ_c has $\_ + \_ \colon Nat \times Nat \to Nat$. Finally, $\Sigma_u = \Sigma_p \cup \Sigma_c$, and all morphisms in $(*)$ are inclusions.

Let $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma_p)$ be the class of all models with the carrier set freely generated by 0 and s (where each element is the value of exactly one of the terms of the form $s^n(0)$). Let then $\mathcal{N} \subseteq \mathbf{Mod}(\Sigma_c)$ be the class of models that satisfy the following implication:

$$\psi \equiv (\forall x, y{:}Nat.\, x + 0 = x \wedge x + s(y) = s(x + y)) \Rightarrow \forall x, y{:}Nat.\, x + y = y + x$$

Let $\mathbf{FO^+_{EQ}}$ be the extension of $\mathbf{FO_{EQ}}$ by a new Σ_p-sentence φ (and its formal translations) such that $Mod^+(\varphi) = \mathcal{M}$. No Σ_c-sentence is added, since $\mathcal{N}$ is already definable in $\mathbf{FO_{EQ}}$. Clearly, $\mathcal{M}|_{\sigma_{pu}}^{-1} \subseteq \mathcal{N}|_{\sigma_{cu}}^{-1}$, and so $\sigma_{pu}(\varphi) \models^+ \sigma_{cu}(\psi)$.

[9] $\mathcal{J}$ is a set of indices; we introduce such sets of indices whenever convenient.

However, no class of models $\mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i)$ that is definable by first-order sentences excludes non-standard models of natural numbers (with "infinitary" elements). Moreover, there is no signature morphism from Σ_p to Σ_i. Therefore, if $\mathcal{M}|_{\sigma_{ip}} \subseteq \mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i)$ and $\mathcal{K}$ is definable in $\mathbf{FO_{EQ}}$ from $\{\langle\Sigma_p, \mathcal{M}\rangle\}$ then $\mathcal{K}|^{-1}_{\sigma_{ic}} \not\models^+ \psi$ (addition need not commute on "infinitary" arguments). Consequently, φ and ψ have no interpolant in $\mathbf{FO^+_{EQ}}$.

If we remove the binary operation *bop* from Σ_p (and replace it by a unary operation $uop\colon Nat \to Nat$) the situation becomes quite different. We have then a (unique) signature morphism $\tau\colon \Sigma_p \to \Sigma_i$, and the sentence $\lceil\tau(\varphi)\rceil \in \mathbf{Sen}^+(\Sigma_i)$ defines up to isomorphism the standard model of natural numbers, and therefore is an interpolant for φ and ψ.

For institutions like $\mathbf{PL}$, where all classes of models are definable, it might seem that all commutative squares of signature morphisms admit interpolation, and no extension by sentences may spoil this property. However, this need not be the case in general, since for classes of models $\mathcal{M} \subseteq \mathbf{Sen}(\Sigma_p)$ and $\mathcal{N} \subseteq \mathbf{Sen}(\Sigma_c)$ such that $\mathcal{M}|^{-1}_{\sigma_{pu}} \subseteq \mathcal{N}|^{-1}_{\sigma_{cu}}$, the inclusion $\mathcal{M}|_{\sigma_{ip}} \subseteq \mathbf{Mod}(\Sigma_i) \setminus ((\mathbf{Mod}(\Sigma_c) \setminus \mathcal{N})|_{\sigma_{ic}})$ may fail, and then no class $\mathcal{K} \subseteq \mathbf{Mod}(\Sigma_i)$ satisfies $\mathcal{M}|_{\sigma_{ip}} \subseteq \mathcal{K}$ and $\mathcal{K}|^{-1}_{\sigma_{ic}} \subseteq \mathcal{N}$.

The diagram $(*)$ admits *weak amalgamation* if for all models $M \in \mathbf{Mod}(\Sigma_p)$ and $N \in \mathbf{Mod}(\Sigma_c)$ such that $M|_{\sigma_{ip}} = N|_{\sigma_{ic}}$ there is a model $K \in \mathbf{Mod}(\Sigma_u)$ such that $K|_{\sigma_{pu}} = M$ and $K|_{\sigma_{cu}} = N$. The diagram $(*)$ admits *amalgamation* if such a model $K \in \mathbf{Mod}(\Sigma_u)$ is always unique. This is a standard property used extensively in "institutional" foundations of software specifications. Amalgamation (and hence weak amalgamation) holds for pushouts in all the sample institutions and their variants we defined in Example 1; it fails for some non-pushout diagrams though.

▶ **Corollary 24.** *If the diagram $(*)$ does not admit weak amalgamation then it does not admit interpolation in some extension of the institution by new sentences, nor in its further extensions by new sentences.*

Proof (hint): *If $M \in \mathbf{Mod}(\Sigma_p)$ and $N \in \mathbf{Mod}(\Sigma_c)$ give a counterexample to the weak amalgamation for $(*)$ then $\{M\}$ and $\mathbf{Mod}(\Sigma_c) \setminus \{N\}$ satisfy* 1. *and* 2. *in Thm. 22.*

▶ **Theorem 25.** *Assume that in* $\mathbf{INS}$ *each class of Σ_i-models is definable by a set of Σ_i-sentences. Then the diagram $(*)$ admits interpolation in every extension of* $\mathbf{INS}$ *by new sentences if and only if it admits weak amalgamation.*

Proof (hint): *Assuming weak amalgamation for $(*)$, for $\mathcal{M} \subseteq \mathbf{Mod}(\Sigma_p)$ and $\mathcal{N} \subseteq \mathbf{Mod}(\Sigma_c)$, if $\mathcal{M}|^{-1}_{\sigma_{pu}} \subseteq \mathcal{N}|^{-1}_{\sigma_{cu}}$ then $(\mathcal{M}|_{\sigma_{ip}})|^{-1}_{\sigma_{ic}} \subseteq \mathcal{N}$.*

7 Spoiling interpolation by new models and sentences

As so far, we study interpolation over a commuting diagram of signature morphisms $(*)$ in an institution $\mathbf{INS} = \langle \mathbf{Sign}, \mathbf{Sen}, \mathbf{Mod}, \langle\models_\Sigma\rangle_{\Sigma\in|\mathbf{Sign}|}\rangle$. In this section we consider the stability of interpolation under institution extensions by new models and sentences.

An extension of an institution $\mathbf{INS}$ by new models and sentences is an extension $\mathbf{INS}^{++}$ by new sentences of an extension $\mathbf{INS}^{+}$ by new models of the institution $\mathbf{INS}$.

The order of the extensions in the above definition is irrelevant. To see this, suppose that $\mathbf{INS}^{+}$ extends $\mathbf{INS}$ by models $\mathcal{NM} = \langle \mathcal{NM}_\Sigma, \models^{\mathcal{NM}}_\Sigma \subseteq \mathcal{NM}_\Sigma \times \mathbf{Sen}(\Sigma)\rangle_{\Sigma\in|\mathbf{Sign}|}$, and $\mathbf{INS}^{++}$ extends $\mathbf{INS}^{+}$ by sentences $\mathcal{NS} = \langle \mathcal{NS}_\Sigma, \models^{\mathcal{NS}}_\Sigma \subseteq \mathbf{Mod}^+(\Sigma) \times \mathcal{NS}_\Sigma\rangle_{\Sigma\in|\mathbf{Sign}|}$ (see Sect. 2.3 for the definitions and notation). Then define $\mathbf{INS}'$ as the extension of $\mathbf{INS}$ by sentences $\mathcal{NS}' = \langle \mathcal{NS}_\Sigma, \models^{\mathcal{NS}'}_\Sigma \subseteq \mathbf{Mod}(\Sigma) \times \mathcal{NS}_\Sigma\rangle_{\Sigma\in|\mathbf{Sign}|}$, where $M \models^{\mathcal{NS}'}_\Sigma \varphi$ iff $M \models^{\mathcal{NS}}_\Sigma \varphi$ for $\Sigma \in |\mathbf{Sign}|$, $M \in \mathbf{Mod}(\Sigma)$ and $\varphi \in \mathcal{NS}_\Sigma$. Then $\mathbf{INS}^{++}$ coincides with the extension of

INS$'$ by models $\mathcal{NM}' = \langle \mathcal{NM}_\Sigma, \models^{\mathcal{NM}'}_\Sigma \subseteq \mathcal{NM}_\Sigma \times \mathbf{Sen}'(\Sigma)\rangle_{\Sigma \in |\mathbf{Sign}|}$, where for $\Sigma \in |\mathbf{Sign}|$ and $M \in \mathcal{NM}_\Sigma$, $M \models^{\mathcal{NM}'}_\Sigma \varphi$ iff $M \models^{\mathcal{NM}}_\Sigma \varphi$ for $\varphi \in \mathbf{Sen}(\Sigma)$, and for $\tau\colon \Sigma' \to \Sigma$, $\varphi' \in \mathcal{NS}_{\Sigma'}$, $M \models^{\mathcal{NM}'}_\Sigma \lceil\tau(\varphi')\rceil$ iff $\lceil M|_\tau\rceil \models^{\mathcal{NS}}_{\Sigma'} \varphi'$.

Obviously, we have "sinks" of institution morphisms that link institution **INS** and its extension **INS**$^{++}$ by models and sentences, but in general there is no institution morphism between **INS** and **INS**$^{++}$. Their relationship can be captured by another kind of mapping between institutions, where sentences and models translate covariantly [38, 26].

Corollary 8 gives sufficient conditions that ensure that interpolation over a diagram $(*)$ is stable under extensions of the institution by new models and sentences. The key result here is that these conditions are necessary:

▶ **Theorem 26.** *The diagram $(*)$ admits interpolation in all extensions of* **INS** *by new sentences and models if and only if at least one of the following conditions holds:*
1. *$\sigma_{ip}\colon \Sigma_i \to \Sigma_p$ is a retraction and $\sigma_{cu}\colon \Sigma_c \to \Sigma_u$ is a coretraction, or*
2. *$\sigma_{ic}\colon \Sigma_i \to \Sigma_c$ is a retraction and $\sigma_{pu}\colon \Sigma_p \to \Sigma_u$ is a coretraction.*

Proof (hint): *For the "only if" part, let* **INS**$^{++}$ *extend* **INS** *by a new Σ_p-model M and a new Σ_c-model N that do not satisfy any* **INS***-sentences and then by a new Σ_p-sentence φ and a new Σ_c-sentence ψ such that*
- $Mod^{++}(\varphi) = \{M\} \cup \{\lceil N|_{\tau_{pi};\sigma_{ic}}\rceil \mid \tau_{pi}\colon \Sigma_p \to \Sigma_i, \tau_{pi};\sigma_{ip} = id_{\Sigma_p}\}$
- $Mod^{++}(\psi) = \{\lceil M|_{\sigma_{cu};\tau_{up}}\rceil \mid \tau_{up}\colon \Sigma_u \to \Sigma_c, \sigma_{pu};\tau_{up} = id_{\Sigma_p}\} \cup$
$\{\lceil N|_{\tau_{cc}}\rceil \mid \tau_{cc}\colon \Sigma_c \to \Sigma_c, \tau_{cc} \neq id_{\Sigma_c}\}$

If condition 1. *fails then $\sigma_{pu}(\varphi) \models^{++} \sigma_{cu}(\psi)$. If condition* 2. *fails then for any $\Theta \subseteq \mathbf{Sen}^{++}(\Sigma_i)$, if $\varphi \models^{++} \sigma_{ip}(\Theta)$ then $\sigma_{ic}(\Theta) \not\models^{++} \psi$.*

8 Final remarks

In this paper we deal with a general interpolation property, recalling its formulation for an arbitrary logical system formalised as an institution. We study behaviour of interpolation properties over an arbitrary commutative square of signature morphisms under extensions of the institution by new models and sentences. We give an exact characterisation of the situations when a particular interpolant for a premise and a conclusion remains stable under institution extensions by new models (Thm. 14), or looking at this from the other side, when a particular interpolant for a premise and a conclusion is spoiled in some extension of the institution by new models. Another result (Thm. 20) gives sufficient and necessary conditions under which no interpolant for a given premise and conclusion may survive all extensions of the institution by new models, or turning to the positive view, when no extension by new models may spoil the interpolation property for a given premise and conclusion. Then we turn to institution extensions by new sentences, and give an exact characterisation of commutative squares of signature morphisms where adding new sentences may lead to the lack of interpolation (Thm. 22). Incidentally, we clarify here the role of the weak amalgamation property as a necessary condition without which interpolation fails if adding new sentences is permitted (Cor. 24). Finally, we give exact characterisation of commutative squares of signature morphisms where interpolation is ensured for all extensions of the institution by new models and sentences (Thm. 26).

We have carried out our study for the Craig interpolation property. However, in applications a stronger formulation of interpolation is needed: so-called Craig-Robinson (or parameterised) interpolation [19, 13, 35], where the conclusion is required to follow only when an additional "parameter" set of sentences over the signature of the conclusion is added to the premise and, respectively, to the interpolant. In first-order logic Craig-Robinson interpolation can easily be derived from the Craig interpolation property, but in general, in logical systems that lack compactness and standard logical connectives, this need not be the case. We do not treat explicitly Craig-Robinson interpolation here, to avoid extra complication of notation, but the concepts and techniques we use carry over to this case as well, and so the results may easily be adjusted to cover this more general property.

In many applications, the class of signature morphisms and of their commutative squares for which the interpolation property is required may be considerably restricted. Typically, signature pushouts are of the utmost importance, with further restrictions on the classes of morphisms used. In fact, this is often necessary, since many institutions involved (including the many-sorted first-order logic **FO** and equational logic **EQ**) simply do not admit interpolation for arbitrary signature pushouts. It would be interesting to check how such extra requirements on the signature morphisms involved interact with our characterisation theorems.

References

1. Egidio Astesiano, Michel Bidoit, Hélène Kirchner, Bernd Krieg-Brückner, Peter D. Mosses, Donald Sannella, and Andrzej Tarlecki. CASL: the Common Algebraic Specification Language. *Theoretical Computer Science*, 286(2):153–196, 2002. `doi:10.1016/S0304-3975(01)00368-1`.
2. Jon Barwise. Axioms for abstract model theory. *Annals of Mathematical Logic*, 7:221–265, 1974. `doi:10.1016/0003-4843(74)90016-3`.
3. Jan A. Bergstra, Jan Heering, and Paul Klint. Module algebra. *Journal of the Association for Computing Machinery*, 37(2):335–372, 1990. `doi:10.1145/77600.77621`.
4. Evert W. Beth. On Padoa's method in the theory of definition. *Indagationes Mathematicae (Proceedings)*, 56:330–339, 1953. `doi:10.1016/S1385-7258(53)50042-3`.
5. Tomasz Borzyszkowski. Logical systems for structured specifications. *Theoretical Computer Science*, 286(2):197–245, 2002. `doi:10.1016/S0304-3975(01)00317-6`.
6. Tomasz Borzyszkowski. Generalized interpolation in first-order logic. *Fundamenta Informaticae*, 66(3):199–219, 2005. URL: `http://content.iospress.com/articles/fundamenta-informaticae/fi66-3-01`.
7. Carlos Caleiro, Paula Gouveia, and Jaime Ramos. Completeness results for fibred parchments: Beyond the propositional base. In Martin Wirsing, Dirk Pattinson, and Rolf Hennicker, editors, *Recent Trends in Algebraic Development Techniques. Selected Papers from the 16th International Workshop on Algebraic Development Techniques*, volume 2755 of *Lecture Notes in Computer Science*, pages 185–200. Springer, 2003. `doi:10.1007/978-3-540-40020-2_10`.
8. Carlos Caleiro, Paulo Mateus, Jaime Ramos, and Amílcar Sernadas. Combining logics: Parchments revisited. In Maura Cerioli and Gianna Reggio, editors, *Recent Trends in Algebraic Development Techniques. Selected Papers from the 15th Workshop on Algebraic Development Techniques Joint with the CoFI WG Meeting*, volume 2267 of *Lecture Notes in Computer Science*, pages 48–70. Springer, 2001. `doi:10.1007/3-540-45645-7_3`.
9. Carlos Caleiro, Amílcar Sernadas, and Cristina Sernadas. Fibring logics: Past, present and future. In Sergei N. Artëmov, Howard Barringer, Artur S. d'Avila Garcez, Luís C. Lamb, and John Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*, pages 363–388. College Publications, 2005.
10. María Victoria Cengarle. *Formal Specifications with Higher-Order Parameterization*. PhD thesis, Ludwig-Maximilians-Universität München, Institut für Informatik, 1994.
11. Chen-Chung Chang and H. Jerome Keisler. *Model Theory*. North-Holland, third edition, 1990.
12. William Craig. Linear reasoning. A new form of the Herbrand-Gentzen theorem. *The Journal of Symbolic Logic*, 22(3):250–268, 1957. `doi:10.2307/2963593`.
13. Răzvan Diaconescu. *Institution-Independent Model Theory*. Birkhäuser, 2008. `doi:10.1007/978-3-7643-8708-2`.

14 Răzvan Diaconescu. An institution-independent proof of Craig interpolation theorem. *Studia Logica*, 77(1):59–79, 2004. doi:10.1023/B:STUD.0000034185.62660.d6.

15 Răzvan Diaconescu. Interpolation for predefined types. *Mathematical Structures in Computer Science*, 22(1):1–24, 2012. doi:10.1017/S0960129511000430.

16 Răzvan Diaconescu. Three decades of institution theory. In Jean-Yves Béziau, editor, *Universal Logic: An Anthology*, pages 309–322. Birkhäuser, 2012.

17 Răzvan Diaconescu. Generalised graded interpolation. *International Journal of Approximate Reasoning*, 152:236–261, 2023. doi:10.1016/j.ijar.2022.10.018.

18 Răzvan Diaconescu. Borrowing interpolation. *J. Logic and Computation*, 22(3):561–586, 2011. doi:10.1093/logcom/exr007.

19 Theodosis Dimitrakos and Thomas S.E. Maibaum. On a generalised modularization theorem. *Information Processing Letters*, 74(1–2):65–71, 2000. doi:10.1016/S0020-0190(00)00037-5.

20 Hartmut Ehrig, Hans-Jörg Kreowski, James W. Thatcher, Eric G. Wagner, and Jesse B. Wright. Parameter passing in algebraic specification languages. *Theoretical Computer Science*, 28(1–2):45–81, 1984. doi:10.1016/0304-3975(83)90065-8.

21 Dov M. Gabbay and Larisa Maksimova. *Interpolation and Definability: Modal and Intuitionistic Logics*. Oxford University Press, 2005. doi:10.1093/acprof:oso/9780198511748.001.0001.

22 Daniel Găină. Interpolation in logics with constructors. *Theoretical Computer Science*, 474:46–59, 2013. doi:10.1016/j.tcs.2012.12.002.

23 Daniel Găină. Downward Löwenheim–Skolem theorem and interpolation in logics with constructors. *Journal of Logic and Computation*, 27(6):1717–1752, 2015. doi:10.1093/logcom/exv018.

24 Daniel Găină and Andrei Popescu. An institution-independent proof of the Robinson consistency theorem. *Studia Logica*, 85:41–73, 2007. doi:10.1007/s11225-007-9022-4.

25 Joseph A. Goguen and Rodney M. Burstall. Institutions: Abstract model theory for specification and programming. *Journal of the Association for Computing Machinery*, 39(1):95–146, 1992. doi:10.1145/147508.147524.

26 Joseph A. Goguen and Grigore Roşu. Institution morphisms. *Formal Aspects of Computing*, 13(3–5):274–307, 2002. doi:10.1007/s001650200013.

27 Thomas S.E. Maibaum, Martin R. Sadler, and Paolo A.S. Veloso. Logical specification and implementation. In Mathai Joseph and Rudrapatna Shyamasundar, editors, *Foundations of Software Technology and Theoretical Computer Science*, pages 13–30. Springer, 1984. doi:10.1007/3-540-13883-8_62.

28 Till Mossakowski, Wiesław Pawłowski, Donald Sannella, and Andrzej Tarlecki. Parchments for CafeOBJ logics. In Shusaku Iida, José Meseguer, and Kazuhiro Ogata, editors, *Specification, Algebra, and Software - Essays Dedicated to Kokichi Futatsugi*, volume 8373 of *Lecture Notes in Computer Science*, pages 66–91. Springer, 2014. doi:10.1007/978-3-642-54624-2_4.

29 Till Mossakowski, Andrzej Tarlecki, and Wiesław Pawłowski. Combining and representing logical systems using model-theoretic parchments. In Francesco Parisi-Presicce, editor, *Recent Trends in Data Type Specification. Selected Papers from the 12th International Workshop on Specification of Abstract Data Types*, volume 1376 of *Lecture Notes in Computer Science*, pages 349–364. Springer, 1998. doi:10.1007/3-540-64299-4_44.

30 Andrei Popescu, Traian Florin Şerbănuţă, and Grigore Roşu. A semantic approach to interpolation. *Theoretical Computer Science*, 410(12–13):1109–1128, 2009. doi:10.1016/j.tcs.2008.09.038.

31 Gerard R. Renardel de Lavalette. Interpolation in computing science: The semantics of modularization. *Synthese*, 164(3):437–450, 2008. doi:10.1007/s11229-008-9358-y.

32 Abraham Robinson. A result on consistency and its application to the theory of definition. *Indagationes Mathematicae (Proceedings)*, 59:47–58, 1956. doi:10.1016/S1385-7258(56)50008-X.

33 Pieter Hendrik Rodenburg. A simple algebraic proof of the equational interpolation theorem. *Algebra Universalis*, 28:48–51, 1991. doi:10.1007/BF01190411.

34 Grigore Roşu and Joseph A. Goguen. On equational Craig interpolation. *Journal of Universal Computer Science*, 6(1):194–200, 2000. `doi:10.3217/jucs-006-01-0194`.

35 Donald Sannella and Andrzej Tarlecki. *Foundations of Algebraic Specification and Formal Software Development*. Monographs in Theoretical Computer Science. An EATCS Series. Springer, 2012. `doi:10.1007/978-3-642-17336-3`.

36 Donald Sannella and Andrzej Tarlecki. Property-oriented semantics of structured specifications. *Mathematical Structures in Computer Science*, 24(2):e240205, 2014. `doi:10.1017/S0960129513000212`.

37 Andrzej Tarlecki. Bits and pieces of the theory of institutions. In David H. Pitt, Samson Abramsky, Axel Poigné, and David E. Rydeheard, editors, *Proceedings of the Tutorial and Workshop on Category Theory and Computer Programming*, volume 240 of *Lecture Notes in Computer Science*, pages 334–360. Springer, 1986. `doi:10.1007/3-540-17162-2_132`.

38 Andrzej Tarlecki. Towards heterogeneous specifications. In Dov Gabbay and Maarten de Rijke, editors, *Frontiers of Combining Systems 2*, volume 7 of *Studies in Logic and Computation*, pages 337–360. Research Studies Press, 2000.

39 Andrzej Tarlecki. Some nuances of many-sorted universal algebra: A review. *Bulletin of the European Association for Theoretical Computer Science*, 104:89–111, 2011. URL: `http://eatcs.org/beatcs/index.php/beatcs/article/view/121`.

40 Jouko Väänänen. The Craig interpolation theorem in abstract model theory. *Synthese*, 164:401–420, 2008. `doi:10.1007/s11229-008-9357-z`.

41 Paulo A.S. Veloso. On pushout consistency, modularity and interpolation for logical specifications. *Information Processing Letters*, 60(2):59–66, 1996. `doi:10.1016/S0020-0190(96)00146-9`.

42 Paulo A.S. Veloso and Thomas S.E. Maibaum. On the modularization theorem for logical specifications. *Information Processing Letters*, 53(5):287–293, 1995. `doi:10.1016/0020-0190(94)00203-B`.