Deeply-Polarized Type Theory as a Generalized Algebraic Theory

Jacob Neumann
University of Nottingham, United Kingdom

Abstract

One of the long-recognized benefits of *categories with families* (CwFs) as a model theory for dependent type theory is their presentation as a generalized algebraic theory. Recent developments in the semantics of type theory make use of *second-order* generalized algebraic theories (SOGATs), which allow for specification of type theories in a higher-order abstract syntax that makes variable binding and stability under substitution implicit. Moreover, by interpreting such SOGATs in presheaf categories, these second-order theories can readily be translated back to first-order theories.

We highlight the phenomenon of *deep polarization*, which arises in the semantics of directed type theory. Directed type theories—variants of Martin-Löf type theory designed for synthetic reasoning about (higher) categories—often adopt a 'polarized' typing discipline of positive and negative types in order to axiomatize co- and contra-variance. Deep polarization—the extension of this polarity into the variable binding and substitution of the language—is difficult to express in a higher-order abstract syntax that has made these implicit. We show how to resolve this problem, and give a SOGAT presentation of deeply-polarized type theory.

2012 ACM Subject Classification Theory of computation → Type theory

Keywords and phrases semantics, directed type theory, homotopy type theory, category theory, generalized algebraic theory

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction

Categories with families (CwFs)—a notion of 'model' for Martin-Löf type theory [23, 24] introduced by Dybjer [13]—enjoy some considerable advantages. First, they are *generalized algebraic theories (GATs)* in the sense of Cartmell [11]. This means that all the data constituting a CwF can be laid out explicitly as first-order operations constrained by various equations, making CwFs particularly amenable to computer formalization. Moreover, CwFs enjoy the advantage of being expressed in terms of very standard category-theoretic constructions (such as presheaves and categories of elements), so specialist knowledge in topos theory or higher category theory is not necessary to comprehend constructions involving CwFs. Finally, as we discuss later, CwFs are highly-modular, in that they provide a flexible framework for making metatheoretic arguments about type theories equipped with a variety of different constructs (see [12] for a wider survey of the possibilities for modelling type theory with CwFs).

In the present work, we begin to develop a confluence between two active areas of research in the semantics of dependent type theory. First, we extend the notion of CwF to encompass constructions important to *directed type theory*. The goal of directed type theory is to develop a type theory that can serve as a language for *synthetic (higher) category theory*, analogous to how homotopy type theory [32] serves as a synthetic language for higher groupoids [33]. Several approaches to directed type theory (particularly [22, 6]) build such a language atop a *polarized* type theory, that is, a type theory with modalities axiomatizing the phenomena of co- and contra-variance.

Second, we incorporate cutting-edge research into the semantics of type theory, namely the theory of *second-order generalized algebraic theories*, or SOGATs [30]. SOGATs serve as a more convenient language for articulating type theories, because they constitute a *higher-order abstract syntax (HOAS)* [27], where much of the cumbersome bureaucracy of expressing elaborate type theories in terms of CwFs can be made implicit. We can reason metatheoretically about this syntax, thanks to semantics (also due to Hofmann [18]) which interpret HOAS in presheaf categories. Moreover, such second-order generalized algebraic theories can be systematically translated into first-order generalized algebraic theories—particularly CwFs with additional structure—which capture the same constructs.

1.1 Contribution and Organization

We explore how SOGATs and the presheaf semantics of HOAS can be leveraged to study polarized type theory, anticipating similar studies of both directed type theory and the modal type theory of [15].

In Section 2, we articulate the phenomenon of *deep polarization*, a modality on the contexts, substitutions, and types of several standard models of type theory. In this section, we give the first of several semantic presentations of deeply-polarized type theory, *(concretely-)polarized CwFs*.

In Section 3, we rehearse the dependently-typed analogue of Hofmann's presheaf semantics of higher-order abstract syntax, and the process by which the theories written in such syntax—*second-order* generalized algebraic theories—are elaborated to obtain *first-order* generalized algebraic theories like Π-CwFs. To our knowledge, this material has not been given a single, detailed, elementary exposition of this kind.

Finally, in Section 4, we modify our notion of polarized CwF to be amenable to expression in a higher-order abstract syntax, arriving at the notion of *abstractly-polarized CwFs*. We then give a second-order GAT articulating deep polarization, which, when unfolded via the procedure from Section 3, yields the theory of abstractly-polarized CwFs. This raises several interesting questions, which hopefully will help inform the development of a SOGAT presentation of multi-modal type theory.

1.2 Related Work

Our analysis in Section 2 assumes familiarity with several standard models of dependent type theory, such as the set model (introduced as a first example of CwFs in [13]); the groupoid model [19]; and the setoid model [16, 3, 4, 10]. Two formalizations of the setoid model—the AGDA formalization of [4], and the COQ formalization of [10]—provided important insights behind our notion of "abstractly-polarized CwF", as did contemplation of polarization in the theory of Awodey's *natural models* [7].

The first study of what we call "deep polarization" was in the directed type theory of Harper and Licata [22], and continued (with some modification) by Nuyts [26]. Subsequent directed type theories only incorporate what we call "shallow polarization" (e.g. North [25]), or did not adopt a modal typing discipline for co- and contra-variance at all (e.g. Riehl–Shulman [28] and Licata–Weaver [2]). The *category model of directed type theory* as a deeply-polarized type theory comes from the work of Altenkirch–Sestini [29], elaborated further by the present author and Altenkirch in [6]. Neither develop a general-purpose polarized model theory of which the category model is an instance—we do so here.

The theory of SOGATs stems from the work of Uemura [31, 30], and is further developed in [9] and [8]; our presentation of *local representability* draws significantly on the latter. The idea of interpreting higher-order abstract syntax in presheaf models was introduced in [18] for the simply-typed case. Doing the same for dependent type theories makes essential use of the presheaf CwFs of [17, Sect. 4], as well as the construction of universes in presheaf models sketched in [20]. A general statement of the process for obtaining a first-order theory from a second-order theory is still forthcoming, but [5] applies this technique towards the study of internal parametricity.

1.3 Metatheory and Notation

We adopt an informal type theory as our metatheory. For instance, we use the type-theoretic notation $x\colon X$ to indicate that x is an element of some set X. We generally assume the *uniqueness of identity proofs* (particularly in Section 3), but this assumption maybe ought to be dropped for studying the *univalent* structures of Section 2. We use AGDA-style notation for dependent products, writing $(x\colon A) \to B(a)$ instead of $\prod_{x\colon A} B(a)$. We will also adopt the convention that *propositions* are *subsingleton types*, types with at most one element. For instance, a relation R on some set X will be understood as a function taking $x, x'\colon X$ and returning the *subsingleton type* $R(x, x')$, whose element (if there is one) we think of as a "witness" that x is R-related to x'. This has the advantage of making it trivial to think of preorders as instances of categories (i.e. a preorder is just a category with subsingleton hom-sets), and likewise setoids as instances of groupoids.

We make significant use of category-theory concepts. We'll denote the collection of objects of a category $\mathbb{C}$ as $|\mathbb{C}|$ and the set of morphisms between I and J as $\mathbb{C}\,[I, J]$. We make use of standard categories, such as Set, Preord, and Cat (ignoring its 2-categorical aspects). We'll write $\int F$ for the category of elements of some presheaf F. The notion of a *dependent natural transformation* will be important for us: given a presheaf $F\colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}$ and $G\colon (\int F)^{\mathrm{op}} \to \mathsf{Set}$, we'll write

$$\int_{I\colon \mathbb{C}} (\phi\colon F(I)) \to G(I, \phi)$$

for the set of transformations α, whose I-component is a dependent function $\alpha_I\colon (\phi\colon F(I)) \to G(I, \phi)$ satisfying a dependent version of naturality: for all $i\colon \mathbb{C}\,[J, I]$ and $\phi\colon F(I)$, it is the case that $G\ i\ (\alpha_I \phi) = \alpha_J (F\ i\ \phi)$. If G doesn't actually depend on ϕ, i.e. $G\colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}$, then this is just the usual set of natural transformations $F \to G$.

2 The Γ-Cube and PCwFs

We begin with the following observation: the set model, the setoid model, and the groupoid model are all instances of a common pattern. Namely, if we let "structure" refer generically to either sets, setoids, or groupoids, then the "structure model" is given by:

- contexts Δ, Γ are *structures*;
- a substitution $\sigma\colon \mathsf{Sub}\ \Delta\ \Gamma$ is a *structure morphism* from Δ to Γ;
- a type $A\colon \mathsf{Ty}\ \Gamma$ is a Γ-indexed *family of structures*;
- the context extension $\Gamma \triangleright A$ is the *total structure of the Grothendieck construction*; and
- a term $t\colon \mathsf{Tm}(\Gamma, A)$ is a section of the *projection structure morphism* $p_A\colon \mathsf{Sub}\ (\Gamma \triangleright A)\ \Gamma$.

These follow a common pattern because we can view setoids as a generalization of sets (by dropping the assumption of anti-symmetry), and groupoids as a generalization of setoids (a setoid is just a groupoid with subsingleton hom-sets). So we have the following chain of generalizations:

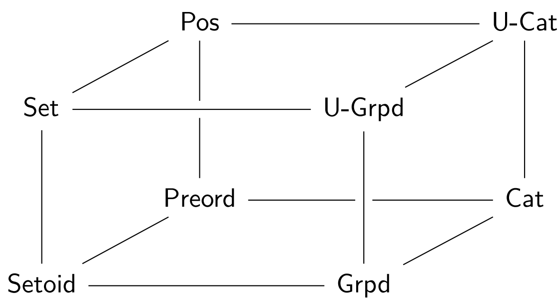
$$\mathsf{Set} \longrightarrow \mathsf{Setoid} \longrightarrow \mathsf{Grpd}.$$

We can take this further: as [6] notes, if we're just considering the basic CwF structure (i.e. not further type-formers like Π or identity types), then there's no need for the 'invertability' assumption of groupoids—we can generalize this to *categories*,

$$\mathsf{Set} \longrightarrow \mathsf{Setoid} \longrightarrow \mathsf{Grpd} \longrightarrow \mathsf{Cat}.$$

The *category model* is just the above schema, with "structure" replaced by "category", "structure morphism" by "functor", etc. This appears to be as far as we can generalize: it's not clear how to make sense of essential elements—particularly *indexed families of structures* and *the total structure of the Grothendieck construction*—for notions of "structure" *more* general than categories.

However, observe that these three steps of generalization—dropping anti-symmetry, asserting *proof relevance*,[1] and dropping symmetry—are completely orthogonal. Thus, we get the following *cube*, whose corners are all categories of structures which can serve as the contexts in a model of type theory. We therefore dub it "the Γ cube".



Here, U-Cat refers to *univalent categories*[2]—categories for whom the notion of equality between objects coincides with isomorphism (this principle—a truncated form of Voevodsky's univalence axiom—is the proof-relevant analogue of anti-symmetry); U-Grpd are such categories whose morphisms are also invertible.[3] We're not aware of any systematic study of these two models of type theory, but we include the structures here for the sake of completeness.

Setting the semantics of type theory aside for a moment, there is much to be said about the inter-relationships of these eight categories and their arrangement in this diagram. The left face and the right face correspond approximately) to the 0- and 1-*truncation levels* of homotopy type theory [21] a setoid $(X, \sim)$, for instance, can have nontrivial "0-dimensional structure" in that X may have multiple elements, but has trivial "1-dimensional structure" because $x \sim x'$ cannot have multiple witnesses. A groupoid can have both nontrivial 0-dimensional structure (its objects), and nontrivial 1-dimensional structure (its hom-sets). Likewise for the other three left-right pairs. Moreover, there is a *reflective subcategory* relationship between subsequent truncation levels: categories can be reflected to preorders (and groupoids into setoids, etc.) by truncating their hom-sets to subsingletons; this operation

[1] We can think of a groupoid as a *proof-relevant* setoid: in a setoid, the equivalence relation $\sim$ takes two elements of the setoid and returns a proposition, the proposition that those two elements are $\sim$-related. Groupoids are the same, except replace 'proposition' with 'set', i.e. there is a *set of witnesses* that the two objects are "related", and this set can potentially have multiple inhabitants.

[2] Defined as "saturated categories" in [1], and referred to merely as "categories" throughout that work and in [32, Chap. 9]

[3] Alternatively, "univalent groupoid" can be taken to mean the 1-truncated types of homotopy type theory [32], which, as noted, can be understood as *synthetic groupoids*. Developing a similar theory of univalent synthetic *categories* is a central motivation for directed (homotopy) type theory.

is left adjoint to the inclusion $\mathsf{Preord} \hookrightarrow \mathsf{Cat}$.[4] Finally, let us note that we could extend the cube further to the right, adding a face for 2-categories and 2-groupoids, 3-categories and 3-groupoids, and so on; this, however, would take us too far afield for the present purpose.

The up–down dimension—the presence or absence of antisymmetry/univalence—is also well-studied. For instance, a given preorder X can be "completed" into an equivalent *poset* $\downarrow P$: its collection of down-sets, ordered by subset inclusion. The proof-relevant analogue of this construction is the *Rezk completion* [1, Thm. 8.5], which, for each category C, obtains a weakly-equivalent *univalent* category as a subcategory of the category of presheaves on C.

But it will be the back–forth dimension that will occupy our attention here. Throughout, we will refer to the back face of the Γ-cube as "polarized" and the front face as "neutral". Each of the back-face categories come equipped with an "opposite" operation: for any category Γ, we have the opposite category Γ^{op}, which has the same objects as Γ but all morphisms flipped around. This extends to a functor $(\_)^{\mathrm{op}} : \mathsf{Cat} \to \mathsf{Cat}$. We say that Cat is polarized because, in general, Γ and Γ^{op} are distinct categories with (perhaps) quite different properties. Not so for Grpd: every groupoid is *self-dual*, i.e. isomorphic to its opposite, and thus there is no real point to considering the opposite operation on groupoids. This distinction between polarized and neutral will prove crucial to our study.

Let's return to considering these categories of structures as models of type theory. Here is the key question: what additional type-theoretic constructs do polarized structures model? Throughout, we'll use the category model (and its relationship to the groupoid model) as paradigmatic—the other examples will be instances of this relationship. So how can we use the opposite endofunctor to define further structure on the category model? Well, the most obvious way is as an operation on *contexts*: since contexts are categories, we can take the opposite of any context. Moreover, since the opposite operation is an endofunctor on Cat, i.e. it has a morphism part as well, we can have a corresponding operation on substitutions. This gives us the following rules.

$$\frac{\Gamma : \mathsf{Con}}{\Gamma^- : \mathsf{Con}} \qquad (\mathsf{Con}\text{-Neg})$$

$$\frac{\sigma : \mathsf{Sub}\ \Delta\ \Gamma}{\sigma^- : \mathsf{Sub}\ \Delta^-\ \Gamma^-} \qquad (\mathsf{Sub}\text{-Neg})$$

These are subject to the equations $(\Gamma^-)^- = \Gamma$ and $(\sigma^-)^- = \sigma$, since the opposite operation is self-inverse. Furthermore, note that the empty context, which we denote $\bullet$ and interpret as the single-object category with only the identity morphism, is definitionally self-dual; thus we also include the rule that $\bullet^- = \bullet$. We will endeavor to interpret the *meaning* of these operations more clearly in just a moment, but for now we continue to see where the semantics leads.

The next place where the opposite operation can be incorporated into the type theory is in the definition of *types*. In the category model, a type $A : \mathsf{Ty}\ \Gamma$ is a Γ-indexed family of categories, that is, a functor $\Gamma \to \mathsf{Cat}$. Given such an A, we can post-compose with $(\_)^{\mathrm{op}}$, to obtain another type in context Γ:

$$\Gamma \xrightarrow{A} \mathsf{Cat} \xrightarrow{(\_)^{\mathrm{op}}} \mathsf{Cat}.$$

[4] For the case of univalent groupoids and sets (treated synthetically in homotopy type theory as 1-types and 0-types, respectively), this is the 0-truncation modality.

As with the context- and substitution-negation operations, we'll indicate this operation with a superscript minus-sign.

$$\frac{A\colon \mathsf{Ty}\ \Gamma}{A^-\colon \mathsf{Ty}\ \Gamma} \qquad \text{(Ty-Neg)}$$

Note that A^- is still a type in context Γ, not Γ^-—this is because the $(\_)^{\mathrm{op}}$ functor is *covariant*. Once again, we'll assert the law that $(A^-)^- = A$.

So far, there is no apparent connection between context-negation and type-negation. Moreover, it's not clear how to actually construct terms of type A^-; we have just asserted a bald type operation with no rules for making use of it. This brings us to the keystone of this 'negative type theory': negative *context extension*. Recall that the ordinary (i.e. positive) context extension was given as the total space of the co-fibration obtained by the Grothendieck construction: for $A\colon \mathsf{Ty}\ \Gamma$,

$$\left|\Gamma \rhd^+ A\right| := \sum_{\gamma\,:\,|\Gamma|} |A(\gamma)|$$

$$(\Gamma \rhd^+ A)\ [(\gamma_0, a_0), (\gamma_1, a_1)] := \sum_{\gamma_{01}\,:\,\Gamma\,[\gamma_0,\gamma_1]} (A\ \gamma_1)\ [A\ \gamma_{01}\ a_0, a_1].$$

This is the *covariant* Grothendieck construction, since A is a covariant functor $\Gamma \to \mathsf{Cat}$. But now we can discuss contravariant Cat-valued functors as well: a contravariant functor $A\colon \Gamma^{\mathrm{op}} \to \mathsf{Cat}$ is the same thing as a type in context Γ^-. Given such a type, we can form the *negative context extension* $\Gamma \rhd^- A\colon \mathsf{Con}$ as follows.

$$\left|\Gamma \rhd^- A\right| := \sum_{\gamma\,:\,|\Gamma|} |A(\gamma)|$$

$$(\Gamma \rhd^- A)\ [(\gamma_0, a_0), (\gamma_1, a_1)] := \sum_{\gamma_{01}\,:\,\Gamma\,[\gamma_0,\gamma_1]} (A\ \gamma_0)\ [a_0, A\ \gamma_{01}\ a_1].$$

Here, since $A\colon \Gamma^{\mathrm{op}} \to \mathsf{Cat}$, we have that $A(\gamma_{01})$ is a functor from $A(\gamma_1)$ to $A(\gamma_0)$, hence why we can apply it to $a_1\colon |A(\gamma_1)|$ to obtain an object of $A(\gamma_0)$. We describe this as the "keystone" of the negative polarity because it ties together (Con-Neg), (Sub-Neg), and (Ty-Neg) in the following rule. For any contexts Γ, Δ and any $A\colon \mathsf{Ty}\ \Gamma^-$, we have a bijection

$$\mathsf{Sub}\ \Delta\ (\Gamma \rhd^- A) \cong \sum_{\sigma\,:\,\mathsf{Sub}\ \Delta\ \Gamma} \mathsf{Tm}(\Delta^-, A[\sigma^-]^-) \qquad \text{(LocalRep-Neg)}$$

natural in Δ. The name (LocalRep-Neg) is short for "negative *local representability*"— we'll expound the theory of local representability more in subsequent sections. Note that this equation, without all the minus signs, is the condition usually assumed to hold between the Ty and Tm presheaves and the context extension operation; often the right-to-left direction is spelled out explicitly as a 'pairing' operation satisfying a universal property (see e.g. [13, Defn. 1] or [17, Sect. 3.1]). As we'll explore in the next section, we can view this as a second, parallel CwF structure on the came category of contexts, in addition to the usual, *positive* one.

But let us mention one more significant property satisfied by the category model, preorder model, etc., which we'll call the *distribution law*. So far, we have said nothing to connect the two context extension operations, though they obviously are closely-related. To see how to remedy this, we consider the question: what is $(\Gamma \rhd^+ A)^-$? Can we calculate what this category is, in terms of other operations? Of course, both $\Gamma \rhd^{+} A$ and its opposite share the same set of objects; but what of the morphisms? Well, consider the following calculation.

$$\begin{aligned}
(\Gamma \rhd^{+} A)^{-}\,[(\gamma_0, a_0), (\gamma_1, a_1)] &= (\Gamma \rhd^{+} A)\,[(\gamma_1, a_1), (\gamma_0, a_0)] \\
&= \sum_{\gamma_{10} : \Gamma\,[\gamma_1, \gamma_0]} (A(\gamma_0))\,[A\ \gamma_{10}\ a_1, a_0] \\
&= \sum_{\gamma_{10} : \Gamma^{-}\,[\gamma_0, \gamma_1]} (A(\gamma_0))^{-}\,[a_0, A\ \gamma_{10}\ a_1] \\
&= (\Gamma^{-} \rhd^{-} A^{-})\,[(\gamma_0, a_0), (\gamma_1, a_1)]
\end{aligned}$$

In order for the last line to make sense, we need that A^{-} is a type in context $(\Gamma^{-})^{-}$. But the latter is, of course, just Γ, so, since $A \colon \mathsf{Ty}\ \Gamma$, we have $A^{-} \colon \mathsf{Ty}\ \Gamma$ by (Ty-Neg). Therefore, we have one final law which holds in the polarized models of the Γ-cube:

$$(\Gamma \rhd^{s} A)^{-} = \Gamma^{-} \rhd^{-s} A^{-} \tag{Distr}$$

Here, and throughout, we'll use s as a metavariable for either polarity, $+$ or $-$, and $-s$ is the opposite polarity. So this law also covers the claim that $(\Gamma \rhd^{-} A)^{-} = \Gamma^{-} \rhd^{+} A^{-}$, for when $A \colon \mathsf{Ty}\ \Gamma^{-}$. These laws connect the two context extensions—the two arise from opposite constructions, and therefore it is little surprise that they can be expressed in terms of each other and the negation operations on contexts and types.

With that, we're ready to abstractly state *what kind of thing* our four polarized models are.

▶ **Definition 1** (PCwF). *A* ***polarized category with families (PCwF)*** *consists of the following*

- *A category* Con *(whose hom-sets are denoted* Sub*) with terminal object* $\bullet$
- *A presheaf* $\mathsf{Ty} \colon \mathsf{Con}^{\mathrm{op}} \to \mathsf{Set}$
- *A presheaf* $\mathsf{Tm} \colon (\int \mathsf{Ty})^{\mathrm{op}} \to \mathsf{Set}$
- *An endofunctor* $(\_)^{-} \colon \mathsf{Con} \to \mathsf{Con}$ *such that* $(\Gamma^{-})^{-} = \Gamma$ *and* $(\sigma^{-})^{-} = \sigma$ *for all* Γ *and* σ
- *For each context* Γ*, a function* $(\_)^{-} \colon \mathsf{Ty}\ \Gamma \to \mathsf{Ty}\ \Gamma$ *such that* $(A^{-})^{-} = A$ *for all* A*, and such that* $(\_)^{-}$ *is stable under substitution:*

  $$(A^{-})[\sigma] = (A[\sigma])^{-}$$

- *Context extension operations*

  $$\_ \rhd^{s} \_ \colon (\Gamma \colon \mathsf{Con}) \to \mathsf{Ty}(\Gamma^{s}) \to \mathsf{Con}$$

  such that:
  - $\mathsf{Sub}\ \Delta\ (\Gamma \rhd^{s} A) \cong \sum_{\sigma \colon \mathsf{Sub}\ \Delta\ \Gamma} \mathsf{Tm}(\Delta^{s}, A[\sigma^{s}]^{s})$
  - $(\Gamma \rhd^{s} A)^{-} = \Gamma^{-} \rhd^{-s} A^{-}$.

A PCwF is a model of type theory equipped with a negative polarity. As mentioned in the introduction, several authors (particularly [25]) define versions of directed type theory which include the type-negation operation, but *not* context- nor substitution-negation, nor negative context extension. We adopt the term *shallowly-polarized* for such theories, as opposed to the kind of type theory outlined in Definition 1, which we call *deeply-polarized*. We borrow the "deep" and "shallow" terminology from the theory of domain-specific languages (e.g. [14]), though somewhat loosely. A shallowly-polarized type theory just treats negation as a type annotation, whereas a deeply-polarized type theory extends the polarization into the basic mechanics of the type theory, i.e. contexts, substitutions, and the assumption of free variables (context extension). Therefore, it seems that a shallowly-polarized type theory could be shallowly embedded into an unpolarized host theory, whereas deeply-polarized type theory would require a deep embedding. We won't attempt to make this point precise here—we just use the terminology to establish intuition for these different kinds of polarization.

We'll set aside the study of polarized type theory for a moment, to introduce our other key ingredient: presheaf semantics of higher-order abstract syntax.

3 Presheaf Semantics of Higher-Order Abstract Syntax

We saw in the previous section that the type-negation operation $(\_)^{-} \colon \mathsf{Ty}\,\Gamma \to \mathsf{Ty}\,\Gamma$ had to come equipped with a *stability under substitution* requirement. As the examples of [17, Sect. 3.3] show, we must do this with every type- or term-former we wish to add to the theory. For instance, the Π-type former, which, given $A \colon \mathsf{Ty}\,\Gamma$ and $B \colon \mathsf{Ty}(\Gamma \triangleright A)$, forms the type $\Pi(A, B) \colon \mathsf{Ty}\,\Gamma$, comes with the requirement that, for each $\sigma \colon \mathsf{Sub}\,\Delta\,\Gamma$,

$$\Pi(A, B)[\sigma] = \Pi(A[\sigma], B[\mathsf{q}(\sigma, A)])$$

where $\mathsf{q}(\sigma, A)] \colon \mathsf{Sub}\,(\Delta \triangleright A[\sigma])\,(\Gamma \triangleright A)$ is constructed from the local representability condition. For a fully-featured type theory like Homotopy Type Theory, it can become quite tedious to give such laws for every single construct.

```
Ty : U
Tm : Ty → U*
Π(_,_) : (A : Ty) → (Tm A → Ty) → Ty
lam : ((a : Tm A) → Tm (B a)) ≅ Tm (Π(A,B)) : app
```

Figure 1 Type theory with Π, as a SOGAT

Although such bureaucracy *can* be managed, it will nonetheless be worth the effort to try and automate away these details. We'll do so by passing to a *higher-order abstract syntax (HOAS)*, which abstracts away from explicit substitutions. This makes stability under substitution implicit, so we can focus on giving the appropriate rules for the theory we want. Let's begin with an example: unpolarized type theory with Π-types. A HOAS presentation of such a type theory is given in Figure 1. We'll explain the meaning of these symbols more precisely in a moment, but the important thing to note at this point is the relative simplicity of this presentation. Though we make use of some shorthands (e.g. stipulating the functions lam and app, and insisting they are inverses in just one line), the fact of the matter is that we didn't have to introduce nearly as much *stuff* as in the corresponding first-order presentation, CwFs with Π-types. Namely, we do not explicitly treat contexts and substitutions. Instead of articulating the dependency of B on A in the type $\Pi(A, B)$ using the *object language*—and thereby having to explicitly treat contexts, context extension, substitutions, etc.—we push this work into the *meta-language*, and just ask that B be a meta-language function from terms of A into types. Thus there is no need for substitution laws like the ones above.

This presentation of type theory with Π-types is a *second-order* generalized algebraic theory (a SOGAT), because we allow second-order functions (such as our Π-type former). While this is a simpler and leaner presentation of how the type theory works, we may ultimately want to work with first-order GATs; the model theory of type theories as SOGATs is more complicated, while we already understand well the first-order equivalent: the theory of CwFs. Thankfully, we can view a SOGAT as a specification of a GAT, that is, translate a SOGAT into a GAT capturing the same theory. This is the above-mentioned procedure of "de-SOGAT-ification". To do it, we use Hofmann's [18] presheaf semantics of HOAS to interpret the SOGAT as presheaves, natural transformations, etc. on some unspecified category, and then, using a few clever tricks, elaborate this structure to be able to put it in GAT form. In the present section, we'll sketch this process for *unpolarized* type theory, to prepare for the task (next section) of capturing deeply-polarized type theory as a SOGAT which unfolds to the GAT of PCwFs.

Our first step is to recall the presheaf model of [17, Sect. 4]. For what follows, we'll assume we have two Grothendieck universes in our metatheory, Set_ℓ and $\mathsf{Set}_{\ell+1}$. We'll call sets in the former "small sets" and the latter "large sets", though, as the generic subscript ℓ indicates, this same construction could be performed at every stage of an infinite hierarchy of universes.

▶ **Definition 2** (Presheaf Model). *The **presheaf model** (on $\mathbb{C}$) is the category $\widehat{\mathbb{C}} = \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_{\ell+1}$, endowed with a CwF structure in the following way.*

- $\widehat{\mathsf{Con}} = \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_{\ell+1}$. *A morphism $\sigma : \widehat{\mathsf{Sub}}\ \Delta\ \Gamma$ is a natural transformation of presheaves $\sigma \colon \Delta \to \Gamma$.*
- *The constant-$\mathbf{1}$ presheaf is the terminal object, which we'll denote $\blacklozenge : \widehat{\mathsf{Con}}$. Write $!_\Gamma$ for the unique natural transformation $\widehat{\mathsf{Sub}}\ \blacklozenge\ \Gamma$.*
- *For $\Gamma : \widehat{\mathsf{Con}}$, we define $\widehat{\mathsf{Ty}}(\Gamma)$ as the set of small presheaves on the category of elements of Γ. That is, $\widehat{\mathsf{Ty}}(\Gamma) = (\textstyle\int \Gamma)^{\mathrm{op}} \to \mathsf{Set}_\ell$.*
- *Terms of type A in context Γ are dependent natural transformations from Γ to A:*

$$\widehat{\mathsf{Tm}}(\Gamma, A) = \int_{I:\mathbb{C}} (\phi : \Gamma I) \to A(I, \phi).$$

Note that we decorate the components of this model with hats; this convention will help prevent confusion later on. In addition to the basic CwF structure, presheaf models interpret rich type theories. In particular, presheaf models come equipped with extensional identity types (which we'll denote with $\equiv$) and Π-types. The latter are interpreted in the usual 'Kripke-style', utilizing the dependent Yoneda Lemma. Consider the special case where $A, B \colon \widehat{\mathsf{Ty}}\blacklozenge$, i.e. $A, B \colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_\ell$ (as here, we'll frequently coerce along the isomorphism $\textstyle\int \blacklozenge \cong \mathbb{C}$), then the *function type* $(A \Rightarrow B) \colon \widehat{\mathsf{Ty}}\blacklozenge$ is given by

$$(A \Rightarrow B)\ I = \int_{J\,:\,\mathbb{C}} \mathbb{C}\,[J, I] \times A(J) \to B(J), \qquad (\widehat{\mathbb{C}} \text{ exponentials})$$

i.e. the usual exponential B^A in the presheaf category.

The theory of presheaf models has another feature which will be relevant for our purposes: type universes. As briefly mentioned in [17, Sect. 4] and then elaborated in more detail in [20], we can "lift" the Grothendieck universe Set_ℓ from our metatheory into the theory of the presheaf model, to obtain a type that *classifies types.*

▶ **Proposition 3.** *The presheaf model on $\mathbb{C}$ gives semantics for a large closed type, that is, a $\mathsf{Set}_{\ell+1}$-valued presheaf $\mathbf{U}$ on $\mathbb{C} \cong \textstyle\int \blacklozenge$, such that there is a natural isomorphism*

$$\widehat{\mathsf{Tm}}(\Gamma, \mathbf{U}\,[!_\Gamma]) \cong \widehat{\mathsf{Ty}}(\Gamma). \qquad (\text{Fundamental Property of } \mathbf{U})$$

This is specified as a *large type* to avoid it classifying itself, which would lead to a paradox. Now, observe that if we take a closed type like $\mathbf{U}$ and weaken it into context Γ, then a term of the resulting type $\mathbf{U}\,[!_\Gamma]$ is the same thing as a natural transformation from Γ to $\mathbf{U}$ because $\mathbf{U}$ doesn't actually depend on Γ. Thus the left-hand side of (Fundamental Property of $\mathbf{U}$) could be written as $\widehat{\mathsf{Sub}}\,\Gamma\,\mathbf{U}$—this will be useful later on. Now, in the proof of Proposition 3, included in Appendix A, uses Yoneda-style reasoning to deduce that, if we define $\mathbf{U}(I)$ to be the set $(\mathbb{C}/I)^{\mathrm{op}} \to \mathsf{Set}_\ell$ (that is, small presheaves on the slice category $\mathbb{C}/I$), then we can prove (Fundamental Property of $\mathbf{U}$). So this will be our definition.

With this, we can begin to recover the GAT of Π-CwFs from the SOGAT in Figure 1. We interpret the latter in the "host theory" of the presheaf model on some unspecified category $\mathbb{C}$, and then elaborate the presheaf model semantics to obtain a specification of what structure $\mathbb{C}$ must bear. Let's read the lines of Figure 1 one-by-one. To begin, the line $\mathsf{Ty}\colon \mathbf{U}$, interpreted in the presheaf model, says $\mathsf{Ty}\colon \widehat{\mathsf{Tm}}(\blacklozenge, \mathbf{U})$. We can then chain together the following isomorphisms:

$$\begin{aligned}
\widehat{\mathsf{Tm}}(\blacklozenge, \mathbf{U}) &\cong \widehat{\mathsf{Ty}}(\blacklozenge) && \text{(Fundamental Property of } \mathbf{U}\text{)}\\
&= (\textstyle\int \blacklozenge)^{\mathrm{op}} \to \mathsf{Set}_0 && \text{(Defn.)}\\
&\cong \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_\ell.
\end{aligned}$$

Thus, the assertion $\mathsf{Ty}\colon \mathbf{U}$ tells us that Ty is a presheaf on the category $\mathbb{C}$, so we have succeeded in recovering some of the structure of a CwF.

The next line says $\mathsf{Tm}\colon \mathsf{Ty} \to \mathbf{U}^*$; for the moment, just read $\mathbf{U}^*$ as $\mathbf{U}$, so that $\mathsf{Tm}\colon \widehat{\mathsf{Tm}}(\blacklozenge, \mathsf{Ty} \Rightarrow \mathbf{U})$, which can be transformed to be a structure atop $\mathbb{C}$ as follows.

$$\begin{aligned}
\widehat{\mathsf{Tm}}(\blacklozenge, \mathsf{Ty} \Rightarrow \mathbf{U}) &\cong \widehat{\mathsf{Sub}}\,\blacklozenge\,(\mathsf{Ty} \Rightarrow \mathbf{U})\\
&\cong \widehat{\mathsf{Sub}}\,\mathsf{Ty}\,\mathbf{U} && \text{(CCC structure on } \widehat{\mathbb{C}}\text{)}\\
&\cong \widehat{\mathsf{Ty}}(\mathsf{Ty}) && \text{(Fundamental Property of } \mathbf{U}\text{)}\\
&= (\textstyle\int \mathsf{Ty})^{\mathrm{op}} \to \mathsf{Set}_\ell. && \text{(Defn.)}
\end{aligned}$$

In the middle of this calculation, we were treating Ty as a context in the presheaf model: it is, after all, a presheaf on $\mathbb{C}$. But, equivalently, we can understand this as the empty context $\blacklozenge$ extended (using the presheaf model's context extension operation) by a single variable of type Ty, which makes sense, as $\mathsf{Ty}\colon \widehat{\mathsf{Ty}}(\blacklozenge)$ from above.

Now, both of these elaborations have resulted in GAT structure: as Dybjer [13, Section 2.2] showed, the requirements that $\mathbb{C}$ is a category with families $\mathsf{Ty}\colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_\ell$ and $\mathsf{Tm}\colon (\int \mathsf{Ty})^{\mathrm{op}} \to \mathsf{Set}_\ell$ can be expressed as a generalized algebraic theory. We run into issues, however, if we try to do the same for the Π-type former of Figure 1. Given $A\colon \widehat{\mathsf{Tm}}(\blacklozenge, \mathsf{Ty})$, that is, a global section of the Ty presheaf,

$$A\colon \int_{I\,:\,\mathbb{C}} \mathsf{Ty}(I),$$

we can construct the presheaf $\mathsf{Tm}|_A\colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_\ell$ by sending I to $\mathsf{Tm}(I, A_I)$. We use this to begin calculating the type of $\Pi(A, \_)$:

$$\widehat{\mathsf{Tm}}(\blacklozenge, (\mathsf{Tm}|_A \Rightarrow \mathsf{Ty}) \Rightarrow \mathsf{Ty}) \quad\cong\quad \widehat{\mathsf{Sub}}\,(\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})\,\mathsf{Ty}.$$

Now the presheaf $(\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})$, by ($\widehat{\mathbb{C}}$ exponentials), has object part

$$(\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})I = \int_{J\,:\,\mathbb{C}} \mathbb{C}\,[J, I] \times \mathsf{Tm}(J, A_J) \to \mathsf{Ty}(J).$$

The issue is that $\Pi(A, \_)$ remains a second-order function: $\mathsf{Tm}(J, A_J)$ occurs negatively in this expression, and thus we cannot incorporate it into the GAT we have been building.

The reason for our issue is that we did not involve *context extension.* Our hypothesis has been that Figure 1, interpreted in presheaf models and then elaborated, will yield the GAT of Π-CwFs. But the theory of CwFs is incomplete without context extension to tie together contexts, substitutions, types, and terms. And the operations defining Π-types as a type-former atop a CwF structure certainly presupposes context extension. So we need to locate the germ of context extension within our second-order theory. But this raises an immediate question: *how do we talk about context extension when we have no explicit contexts?* It turns out there's an elegant way to smuggle in the logic of context extension, which doesn't force us to axiomatize contexts, substitutions, etc. in the higher object theory (and thereby nullify its advantages as a higher-order abstract syntax), but makes it available to fix our Π-issue. In the simply-typed case, Hofmann observed that the *representability* of presheaves resolved this issue, by allowing one to rewrite negative occurences using the Yoneda Lemma. We'll do the dependently-typed analogue here, using dependent presheaves and *local representability.*

▶ **Definition 4** (Local Representability). *Given $F\colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}$, a presheaf $G\colon (\int F)^{\mathrm{op}} \to \mathsf{Set}$ is called* ***locally representable*** *if, for each $I\colon |\mathbb{C}|$ and $X\colon F(I)$, the restricted presheaf*

$$\begin{aligned} G|_X &: (\mathbb{C}/I)^{\mathrm{op}} \to \mathsf{Set} \\ G|_X(J,i) &= G(J, F\ i\ X) \end{aligned}$$

is representable.

The paradigm example of a locally representable presheaf is the Tm presheaf of any CwF: given a context I and a type $B : \mathsf{Ty}\ I$, the restricted presheaf $\mathsf{Tm}|_B$ is represented by the pair $(I \rhd B, \mathsf{p}_B)\colon \mathsf{Con}/I$. The fact that $\mathsf{Tm}|_B$ is naturally isomorphic to the representable presheaf $\mathsf{Sub}\ (\_)\ (I \rhd B)$ is precisely the isomorphism we referred to as "the local representability" condition before.

So let's see how this solves our issue. Let's return to this expression before, where we got stuck:

$$\int_{J\colon \mathbb{C}} \mathbb{C}\,[J, I] \times \mathsf{Tm}(J, A_J) \to \mathsf{Ty}(J).$$

The naturality of A says that for every $i\colon \mathbb{C}\,[J, I]$, we have that $A_I[i] = A_J$.[5] Therefore, we can introduce a spurious dependence between the two terms to the left of the arrow, and rewrite this equivalently as

$$\int_{J\colon \mathbb{C}} \left(\sum_{i\colon \mathbb{C}\,[J,I]} \mathsf{Tm}(J, A_I[i]) \right) \to \mathsf{Ty}(J).$$

The left-hand side of the arrow ought to look familiar: it is precisely this expression which local representability governs. More precisely, if Tm is locally representable, this means that we have an object $I.A_I$ of $\mathbb{C}$, along with a morphism $\mathsf{p}_{A_I}\colon \mathbb{C}\,[I.A_I, I]$ such that

$$\mathbb{C}\,[J, I.A_I] \cong \sum_{i\colon \mathbb{C}\,[J,I]} \mathsf{Tm}(J, A_I[i])$$

[5] The substitution here is the morphism part of the Ty presheaf already deduced.

naturally in (J, i). If this is so, then our expression for $(\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})\ I$ becomes

$$\int_{J\,:\,\mathbb{C}} \mathbb{C}\,[J, I.A_I] \to \mathsf{Ty}(J),$$

which Yoneda tells us is isomorphic to $\mathsf{Ty}(I.A_I)$. Thus we've eliminated the negative appearance of $\mathsf{Tm}(J, A_J)$ in the argument, and we obtain a description of the Π-type former as a generalized algebraic operation:

$$\begin{aligned}\widehat{\mathsf{Tm}}(\blacklozenge, (\mathsf{Tm}|_A \Rightarrow \mathsf{Ty}) \Rightarrow \mathsf{Ty}) &\cong \widehat{\mathsf{Sub}}\ (\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})\ \mathsf{Ty} \\ &= \int_{I\,:\,\mathbb{C}} (\mathsf{Tm}|_A \Rightarrow \mathsf{Ty})\ I \to \mathsf{Ty}(I) \\ &\cong \int_{I\,:\,\mathbb{C}} \mathsf{Ty}(I.A_I) \to \mathsf{Ty}(I).\end{aligned}$$

This is the shape of the familiar Π-type former in the framework of CwFs: for each I, it turns types in $I.A_I$ into types in I. The naturality condition says that this type-former is stable under substitution,[6] which is exactly the condition we wanted to make implicit in the syntax.

So all that remains is to explain how we say in the HOAS that Tm is locally representable. This is the reason why Tm is written as $\mathsf{Ty} \to \mathbf{U}^*$. Recall that $\mathbf{U}(I)$ was defined as the set of *all* small presheaves on $\mathbb{C}/I$. Since $\mathbb{C}/I$ is the category of elements of $\mathbf{y}I$, we can understand such a presheaf as a dependent presheaf over $\mathbf{y}I$. It therefore makes sense to speak of local representability for such presheaves, as in the following definition and claim.

▶ **Definition 5.** *Define $\mathbf{U}^* : \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_{\ell+1}$ as the subpresheaf of $\mathbf{U}$ consisting of only those presheaves which are locally representable. That is, $\mathbf{U}^*(I)$ is the set of those presheaves $G\colon (\mathbb{C}/I)^{\mathrm{op}} \to \mathsf{Set}_\ell$ equipped with, for each $J\colon |\mathbb{C}|$ and $i\colon \mathbb{C}\,[J, I]$, an object $J.i$ and morphism $\mathsf{p}_i : \mathbb{C}\,[J.i, J]$ such that*

$$\mathbb{C}\,[K, J.i] \cong \sum_{j\,:\,\mathbb{C}\,[K,J]} G(K, i \circ j)$$

naturally in (K, j).

▶ **Proposition 6.** *There is a natural isomorphism*

$$\widehat{\mathsf{Tm}}(\Gamma, \mathbf{U}^*) \cong \widehat{\mathsf{Ty}}_{l.r.}(\Gamma) \qquad \text{(Fundamental Property of } \mathbf{U}^*\text{)}$$

where $\widehat{\mathsf{Ty}}_{l.r.}(\Gamma)$ is the set of locally representable presheaves $(\textstyle\int \Gamma)^{\mathrm{op}} \to \mathsf{Set}_\ell$.

So then, modifying our calculations from above, the assertion $\mathsf{Tm}\colon \mathsf{Ty} \to \mathbf{U}^*$ ends up meaning that $\mathsf{Tm}\colon \widehat{\mathsf{Ty}}_{l.r.}(\mathsf{Ty})$, and thus that Tm is not just a dependent presheaf on Ty, but a locally representable one, as desired.

We've sketched here the essential ideas, and these can be carried much further. If we spell out the isomorphism given in Figure 1 as terms of two mutually-inverse functions lam and app, we can obtain the λ-abstraction, application, β, and η laws. The former two will be

---

[6] The version given here is only for a "global type" like A. To get the usual statement of the Π-type former and its substitution law, we would need to include the dependence on A as well, i.e. modify Figure 1 to say $\Pi\colon (A : \mathsf{Ty}) \to (\mathsf{Tm}\ A \to \mathsf{Ty}) \to \mathsf{Ty}$. We only avoid doing so here for simplicity of exposition.

natural transformations, whose naturality condition states the stability under substitution condition—just like the Π-type former above. Thus we complete the theory of Π-CwFs as a GAT obtained from the SOGAT in Figure 1. We can further augment this with a huge variety of type- and term-formers: anything which is expressible in the SOGAT language, with the restriction that only elements of $\mathbf{U}^*$ can appear doubly-negative (so that we can rewrite using local representability).

4 Abstract and Concrete Polarization

We now merge these two threads and arrive at our central question: *how can deeply-polarized type theory be treated in a higher-order abstract syntax?* That is, can we write a SOGAT presentation of deeply-polarized type theory, which elaborates to the theory of PCwFs via the procedure given in the previous section? The issue is that there is a contradiction between *deep* and *high*: we said that "deeply-polarized" meant that the polarization acted upon the contexts, substitutions, and context extensions of the theory; but it is precisely these elements which are made implicit when passing to a higher-order abstract syntax. How can we study operations on contexts, in a language which expressly avoids referring to contexts?

As we did with context extension and local representability, we need to find a way to incorporate the logic of deep polarization into the second-order syntax, so that it unfolds to the polarization operations of Definition 1 when we de-SOGAT-ify. Let's revisit (LocalRep-Neg):

$$\mathsf{Sub}\ J\ (I \rhd^- A) \cong \sum_{i:\,\mathsf{Sub}\ J\ I} \mathsf{Tm}(J^-, A[i^-]^-).$$

Here we've switched to I, J for contexts, in anticipation of dealing with presheaves. In this equation, $A\colon \mathsf{Ty}(I^-)$. The first key insight is that we can view $\mathsf{Ty}(I^-)$ as a presheaf in I, the composition of Ty after $(\_)^-$:

$$\begin{aligned}
&\mathsf{Ty}^- &&: \mathsf{Con}^{\mathrm{op}} \to \mathsf{Set}\\
&\mathsf{Ty}^- I &&= \mathsf{Ty}(I^-)\\
&\mathsf{Ty}^-(i\colon \mathsf{Sub}\ J\ I)\colon &&\mathsf{Ty}^-(I) \to \mathsf{Ty}^-(J)\\
&\mathsf{Ty}^-\ i\ A &&= A[i^-].
\end{aligned}$$

Moreover, we can do the same with Tm.

$$\begin{aligned}
&\mathsf{Tm}^- &&: (\textstyle\int \mathsf{Ty}^-)^{\mathrm{op}} \to \mathsf{Set}\\
&\mathsf{Tm}^-(I, A) &&= \mathsf{Tm}(I^-, A^-)\\
&\mathsf{Tm}^-(i) &&: \mathsf{Tm}^-(I, A) \to \mathsf{Tm}^-(J, \mathsf{Ty}^-\ i\ A)\\
&\mathsf{Tm}^-\ i\ t &&= t[i^-].
\end{aligned}$$

Note that the stability of type-negation under substitution is required for this definition to typecheck. Now, it might not be clear why $\mathsf{Tm}^-(I, A)$ was chosen to be $\mathsf{Tm}(I^-, A^-)$ and not $\mathsf{Tm}(I^-, A)$, as either definition would make sense. But if we adopt the former, then (LocalRep-Neg) simplifies nicely:

$$\mathsf{Sub}\ J\ (I \rhd^- A) \cong \sum_{i:\,\mathsf{Sub}\ J\ I} \mathsf{Tm}^-(J, \mathsf{Ty}^-\ i\ A).$$

So, (LocalRep-Neg) says that Tm^- is locally representable, with respect to Ty^-. This makes good on the idea that a PCwF is a category of contexts with two parallel family structures, positive and negative.

Let's encapsulate this structure in a definition.

▶ **Definition 7** (Abstractly-Polarized CwF). *An **abstractly-polarized CwF** is a CwF* $(\mathsf{Con}, \mathsf{Sub}, \mathsf{Ty}, \mathsf{Tm}, \ldots)$ *equipped with:*

- *A presheaf* $\mathsf{Ty}^- : \mathsf{Con}^{\mathrm{op}} \to \mathsf{Set}$
- *A locally representable presheaf* $\mathsf{Tm}^- : (\int \mathsf{Ty}^-)^{\mathrm{op}} \to \mathsf{Set}$
- *Natural transformations* $(\_)^- : \mathsf{Ty}^s \to \mathsf{Ty}^s$

such that

- *The* $(\_)^-$ *transformations are both self-inverse*
- $\mathsf{Ty}\ \bullet = \mathsf{Ty}^-\ \bullet$
- *if* $\mathsf{Ty}(J) = \mathsf{Ty}^-(J')$*, then, for all* $B \colon \mathsf{Ty}(J)$,

$$\mathsf{Ty}(J \triangleright^+ B) = \mathsf{Ty}^-(J' \triangleright^- B^-).$$

We call this "abstractly-polarized" because there is no explicit *context-negation* operation: it has been folded into Ty^- and Tm^-. We do, however, still include the *type-negation* operation—the reason for this will be clear shortly. The latter two requirements, which bind together the two structures, will also be useful later. For the sake of comparison, we'll refer to the PCwFs of Definition 1 as "concretely-polarized", since we *do* have the context-negation operation given explicitly. Notice that the structure common to both these definitions—the positive CwF structure and type negation—is what we referred to above as "shallowly-polarized" type theory: abstract and concrete can thus be seen as two means of articulating the extension of shallow polarization to deep polarization.

With this definition, we can return to our main task: expressing deeply-polarized type theory as a SOGAT. The advantage of abstract polarization is that it doesn't refer to explicit operations on contexts. Indeed, it translates nicely into a SOGAT, given in Figure 2.

```
Tyˢ : U
Tmˢ : Tyˢ → U*
(_)⁻ : Tyˢ → Tyˢ
self−inv : (A : Tyˢ) → (A⁻)⁻ ≡ A
Π(_,_) : (A : Ty⁻) → (Tm⁻ A → Ty) → Ty
lam : ((a : Tm⁻ A) → Tm (B a))  ≅  Tm (Π(A,B)) : app
```

Figure 2 Deeply-polarized type theory with Π, as a SOGAT

The calculations go much the same as in the previous section. For instance, in order to resolve the negative appearance of $\mathsf{Tm}^-(A)$ in the argument to the Π-type former, we must make use of the local representability of Tm^- with respect to Ty^-, which, as before, is asserted by $\mathsf{Tm}^- : \mathsf{Ty}^- \to \mathbf{U}^*$. The polarities on the Π-type constructs are taken from the deeply-polarized Π-types of [22] the positive and negative polarities mark the positive and negative occurrences (in the usual sense) within a (dependent) function expression.

This SOGAT *almost* unfolds to give us the GAT of abstractly-polarized CwFs (plus polarized Π-types). The only shortcoming is the final two requirements of Definition 7: that Ty and Ty^- agree on the empty context, and recursively agree across their respective context-extension operations. It's not clear how to assert these in the second-order theory. We could omit these requirements from the definition, but then there would be nothing connecting the two CwF structures together, or either to the type-negation operations. Moreover, as we discuss below, this would complicate the connection between abstract and concrete polarization. So, for the purposes of the present work, we'll simply allow ourselves to assert these equations as part of the de-SOGAT-ification process. Clarifying this situation will be one of the key tasks in developing a SOGAT account of modal type theory.

We conclude this section by considering the circumstances under which abstract and concrete polarization would coincide. As demonstrated above, every concretely-polarized CwF determines an abstract polarization structure, by defining Ty^- and Tm^- in terms of Ty, Tm, and the type- and context-negation operations. To go the other way around, however, requires further assumptions. Namely, we need to be able to do *induction* on contexts, so we can use the components of the abstractly-polarized CwF to concretely define the context-negation operation. To do this, we take the definition of a *contextual CwF* from [12, Defn. 2], and "polarize" it.

▶ **Definition 8** (Polar-Contextual CwF). *An abstractly-polarized CwF is* ***polar-contextual*** *iff there is* $\ell \colon \mathsf{Con} \to \mathbb{N}$*, a length function, such that* $\ell(J) = 0$ *iff* $J = \bullet$*, and* $\ell(J) = n + 1$ *iff exactly one of the following holds:*

- *there is a unique* $I \colon \mathsf{Con}$ *and* $A \colon \mathsf{Ty}(I)$ *such that* $J = I \rhd^+ A$ *and* $\ell(I) = n$*, or*
- *there is a unique* $I \colon \mathsf{Con}$ *and* $A \colon \mathsf{Ty}^-(I)$ *such that* $J = I \rhd^- A$ *and* $\ell(I) = n$*.*

The purpose of this definition is to permit recursive definitions on contexts: if a abstract PCwF is polar-contextual, then an operation on contexts can be defined by specifying its action on the empty context and then recursively on (positively- and negatively-extended contexts (the only purpose served by the length function is to guarantee this is well-founded). Our conjecture is that we can construct the *free* abstract PCwF, a syntax model, and that it will be necessarily polar-contextual. So, the abstract/concrete distinction would disappear in the syntax. However, we leave detailed study of this idea to future work.

Given a polar-contextual PCwF, we then define the context-negation operation as follows.

$$\bullet^- = \bullet$$
$$(I \rhd^+ A)^- = I^- \rhd^- A^-$$
$$(I \rhd^- A)^- = I^- \rhd^+ A^-$$

What we've done here is turn the "Distribution laws" required of concretely-polarized CwFs in Definition 1 into a definition. It is here that we make use of the two problematic requirements from Definition 7: in the second clause of this definition, we take $A \colon \mathsf{Ty}(I)$ on the left-hand side of the equals sign, but $A \colon \mathsf{Ty}^-(I^-)$ on the right. When we were dealing with concrete PCwFs, this was no issue, as Ty^- was $\mathsf{Ty} \circ (\_)^-$ and $(I^-)^- = I$. But here we have to make this an explicit requirement in order for the recursion to carry through.

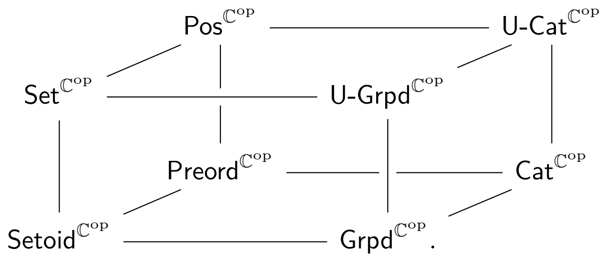
5 Conclusion and Future Work

Here, we have identified *polarization* as an additional dimension of type theory and indicated how deeply-polarized type theories can be treated as both a first- and second-order generalized algebraic theory, the latter serving as a statement of deeply-polarized type theory in a higher-order abstract syntax. We ultimately adopted two first-order GATs describing deeply-polarized type theory, which we termed *concrete* and *abstract*. The former arose more immediately in our choice models and managed to connect the positive and negative CwF structures, but only the latter lent itself to being abstracted to a higher-order theory. It remains for future work to fully understand the relationship between these notions, and resolve definitively whether concrete polarization can similarly be abstracted.

There are numerous other questions to explore in this vein. The most immediate is figuring out how to represent the *unpolarized* type theory within the syntax of polarized type theory. For instance, the category Grpd is both a reflective and coreflective subcategory of Cat: the inclusion/forgetful functor of Grpd $\hookrightarrow$ Cat has left- and right-adjoints, called *localization* and *core*. Core *types* have received particular attention in the directed type theory literature (e.g. [25] uses them to state directed path induction), but treating core as a *deep* operation on contexts is more difficult: while it is possible to define operators Ty^0 and Tm^0 analogous to our notion of abstract polarization, the isomorphism they would satisfy does not have the shape of a *local representability* law like (LocalRep-Neg), so it's unclear how to abstract it to HOAS. A solution to this question might suggest how multi-modal type theories such as [15] could be treated in the HOAS/SOGAT setting.

A presentation of unpolarized and polarized type theory in the same higher-order abstract syntax would likely be a prerequisite for such a treatment of full *directed type theory.* In [6], we develop directed path induction in the category model, but only in *neutral* contexts, i.e. groupoids. We would like to develop adequate machinery to be able to study directed type theory in a higher-order abstract syntax, eventually building a directed analogue to the work of [5].

Finally, presheaf models whose base category is itself the category of contexts for a CwF—as we've dealt with here—is often studied as a *two-level type theory (2LTT)*, where the base CwF interprets the "inner type theory" and the presheaf model interprets the "outer type theory". Taking this view, what we've studied here are type theories whose *inner* theory is polarized, but whose outer theory is not. The most apparent way to polarize the outer theory would be to consider presheaves that take values in the polarized categories of the Γ-cube, i.e. study the back face of the "Kripke-fied" Γ-cube:



References

1 Benedikt Ahrens, Krzysztof Kapulkin, and Michael Shulman. Univalent categories and the rezk completion. *Mathematical Structures in Computer Science*, 25(5):1010–1039, 2015.

2 Benedikt Ahrens, Paige Randall North, and Niels van der Weide. Bicategorical type theory: semantics and syntax. *Mathematical Structures in Computer Science*, 33(10), 2023.

3 Thorsten Altenkirch. Extensional equality in intensional type theory. In *Proceedings. 14th Symposium on Logic in Computer Science (Cat. No. PR00158)*, pages 412–420. IEEE, 1999.

4 Thorsten Altenkirch, Simon Boulier, Ambrus Kaposi, Christian Sattler, and Filippo Sestini. Constructing a universe for the setoid model. In *FoSSaCS*, pages 1–21, 2021.

5 Thorsten Altenkirch, Yorgo Chamoun, Ambrus Kaposi, and Michael Shulman. Internal parametricity, without an interval. *Proceedings of the ACM on Programming Languages*, 8(POPL):2340–2369, 2024.

6 Thorsten Altenkirch and Jacob Neumann. The category interpretation of directed type theory. *arXiv preprint*, 2024.

7 Steve Awodey. Natural models of homotopy type theory. *Mathematical Structures in Computer Science*, 28(2):241–286, 2018.

8 Rafaël Bocquet. External univalence for second-order generalized algebraic theories. *arXiv:2211.07487*, 2022.

9 Rafaël Bocquet, Ambrus Kaposi, and Christian Sattler. For the metatheory of type theory, internal sconing is enough. *arXiv preprint arXiv:2302.05190*, 2023.

10 Simon Pierre Boulier. *Extending type theory with syntactic models*. PhD thesis, Ecole nationale supérieure Mines-Télécom Atlantique, 2018.

11 John Cartmell. Generalised algebraic theories and contextual categories. *Annals of pure and applied logic*, 32:209–243, 1986.

12 Simon Castellan, Pierre Clairambault, and Peter Dybjer. Categories with families: Unityped, simply typed, and dependently typed. *Joachim Lambek: The Interplay of Mathematics, Logic, and Linguistics*, pages 135–180, 2021.

13 Peter Dybjer. Internal type theory. In *International Workshop on Types for Proofs and Programs*, pages 120–134. Springer, 1995.

14 Jeremy Gibbons and Nicolas Wu. Folding domain-specific languages: deep and shallow embeddings (functional pearl). In *Proceedings of the 19th ACM SIGPLAN international conference on Functional programming*, pages 339–347, 2014.

15 Daniel Gratzer, GA Kavvos, Andreas Nuyts, and Lars Birkedal. Multimodal dependent type theory. In *Proceedings of the 35th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 492–506, 2020.

16 Martin Hofmann. A simple model for quotient types. In *International Conference on Typed Lambda Calculi and Applications*, pages 216–234. Springer, 1995.

17 Martin Hofmann. Syntax and semantics of dependent types. In *Extensional Constructs in Intensional Type Theory*, pages 13–54. Springer, 1997.

18 Martin Hofmann. Semantical analysis of higher-order abstract syntax. In *Proceedings. 14th Symposium on Logic in Computer Science (Cat. No. PR00158)*, pages 204–213. IEEE, 1999.

19 Martin Hofmann and Thomas Streicher. The groupoid interpretation of type theory. *Twenty-five years of constructive type theory (Venice, 1995)*, 36:83–111, 1995.

20 Martin Hofmann and Thomas Streicher. Lifting grothendieck universes. *Unpublished note*, 199:3, 1999.

21 Nicolai Kraus. *Truncation levels in homotopy type theory*. PhD thesis, University of Nottingham, 2015.

22 Daniel R Licata and Robert Harper. 2-dimensional directed dependent type theory. 2011.

23 Per Martin-Löf. An intuitionistic theory of types: Predicative part. In H.E. Rose and J.C. Shepherdson, editors, *Logic Colloquium '73*, volume 80 of *Studies in Logic and the Foundations of Mathematics*, pages 73–118. Elsevier, 1975.

24 Per Martin-Löf. Constructive mathematics and computer programming. In L. Jonathan Cohen, Jerzy Łoś, Helmut Pfeiffer, and Klaus-Peter Podewski, editors, *Logic, Methodology and Philosophy of Science VI*, volume 104 of *Studies in Logic and the Foundations of Mathematics*, pages 153–175. Elsevier, 1982.

25 Paige Randall North. Towards a directed homotopy type theory. *Electronic Notes in Theoretical Computer Science*, 347:223–239, 2019.

26 Andreas Nuyts. Towards a directed homotopy type theory based on 4 kinds of variance. *Mém. de mast. Katholieke Universiteit Leuven*, 2015.

27 Frank Pfenning and Conal Elliott. Higher-order abstract syntax. *ACM sigplan notices*, 23(7):199–208, 1988.

28 Emily Riehl and Michael Shulman. A type theory for synthetic ∞-categories. *arXiv preprint arXiv:1705.07442*, 2017.

29 Filipo Sestini and Thorsten Altenkirch. Naturality for free!—the category interpretation of directed type theory, 2019. The International Conference on Homotopy Type Theory (HoTT 2019).

30 Taichi Uemura. *Abstract and concrete type theories*. PhD thesis, University of Amsterdam, 2021.

31 Taichi Uemura. A general framework for the semantics of type theory. *Mathematical Structures in Computer Science*, 33(3), mar 2023. URL: `http://dx.doi.org/10.1017/S0960129523000208`, `doi:10.1017/s0960129523000208`.
32 The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.
33 Benno Van Den Berg and Richard Garner. Types are weak ω-groupoids. *Proceedings of the london mathematical society*, 102(2):370–394, 2011.

A Addenda

Here is the proof of Proposition 3:

Proof. To determine what $\mathbf{U} \colon \mathbb{C}^{\mathrm{op}} \to \mathsf{Set}_1$ *must be*, we use Yoneda-style reasoning: if we had such a $\mathbf{U}$, then we could deduce the following.

$$\begin{aligned} \mathbf{U}\ I &\cong \widehat{\mathsf{Sub}}\ (\mathbf{y}I)\ \mathbf{U} && \text{(Yoneda Lemma)} \\ &\cong \widehat{\mathsf{Ty}}(\mathbf{y}I) && \text{(Fundamental Property of } \mathbf{U}\text{)} \\ &= (\mathbb{C}/I)^{\mathrm{op}} \to \mathsf{Set}_0 \end{aligned}$$

Thus it would be a good choice to *define* $\mathbf{U}(I)$ to be the set of small presheaves on the slice category $\mathbb{C}/I$. We can then rearrange the above to see that $\mathbf{U}$ satisfies (Fundamental Property of $\mathbf{U}$) for any *representable* Γ. This then extends to arbitrary Γ, by application of the *co-Yoneda Lemma*, also known as the *density theorem*, which says that every presheaf is the colimit of representables.

$$\begin{aligned} \widehat{\mathsf{Sub}}\ \Gamma\ \mathbf{U} &\cong \widehat{\mathsf{Sub}} \left(\operatorname*{colim}_{(I,\phi)\colon \int \Gamma} \mathbf{y}I \right) \mathbf{U} && \text{(co-Yoneda Lemma)} \\ &\cong \lim_{(I,\phi)\colon \int \Gamma} \widehat{\mathsf{Sub}}\ (\mathbf{y}I)\ \mathbf{U} && \text{(Yoneda preserves limits)} \\ &\cong \lim_{(I,\phi)\colon \int \Gamma} \widehat{\mathsf{Ty}}(\mathbf{y}I) && \text{(above)} \\ &\cong \widehat{\mathsf{Ty}} \left(\operatorname*{colim}_{(I,\phi)\colon \int \Gamma} \mathbf{y}I \right) && (*) \\ &\cong \widehat{\mathsf{Ty}}(\Gamma) && \text{(co-Yoneda Lemma)} \end{aligned}$$

The step marked (*), that $\widehat{\mathsf{Ty}}$ preserves limits, can be proved relatively easily by elementary methods. ◀