YONEDA WITHOUT TEARS

PATRICK STEVENS

https://www.patrickstevens.co.uk/misc/YonedaWithoutTears/YonedaWithoutTears.pdf

1. Introduction

This document will assume that you are familiar with the notion of a category and a functor. Ideally you will also be familiar with the idea of a natural transformation and a hom-set.

This notation varies widely; we will fix one version of it now.

Definition 1. *Fix a locally small category $\mathcal{C}$, and pick an object $A \in \mathcal{C}$. The* hom-set $\mathrm{Hom}(\mathrm{A}, -)$ *is defined to be the set $\{f \in \mathrm{mor}\mathcal{C} : \mathrm{dom} f = A\}$. There is of course a dual:* $\mathrm{Hom}(-, \mathrm{A})$ *is defined to be $\{f \in \mathrm{mor}\mathcal{C} : \mathrm{cod} f = A\}$.*

Definition 2. *Let $F, G : \mathcal{C} \to \mathcal{D}$ be functors. A* natural transformation *from F to G is a selection α, parametrised over each $X \in \mathcal{C}$, of an arrow α_X in $\mathcal{D}$ from $FX \to GX$. We require this selection to be "natural" in that whenever $f : X \to Y$ is an arrow in $\mathcal{C}$, we have*

$$(FX \xrightarrow{Ff} FY \xrightarrow{\alpha_Y} GY) = (FX \xrightarrow{\alpha_X} GX \xrightarrow{Gf} GY)$$

Definition 3. *Let $F, G : \mathcal{C} \to \mathcal{D}$ be functors. Then the set of natural transformations from F to G is denoted* $\mathrm{Nat}[F, G]$.

Then you should be able to parse the statement of the Yoneda lemma, though not necessarily understand it:

Theorem 1 (The Yoneda lemma)**.** *Let $\mathcal{C}$ be a category, and let $G : \mathcal{C} \to \mathbf{Set}$ be a functor. Let A be an object of $\mathcal{C}$. Then*

$$\mathrm{Nat}[\mathrm{Hom}(\mathrm{A}, -), G] \cong GA$$

and moreover the bijection is natural in both G and A.

2. The Relevant Interpretation of "Category"

The main insight comes from asking the question, "how can I better understand what $\mathrm{Hom}(\mathrm{A}, -)$ means?".

There are many ways to understand what a category is, but the relevant one here is that a category is a description of a many-sorted algebraic theory with unary functions between them. An object is a placeholder for a type, and an arrow is a placeholder for a function from that type.

A category is only an abstract description of some types and their interactions. It's a mistake to think of an object of a category viewed this way as "being" $\mathbb{N}$. We only get to think of the type in this way when we instantiate the theory: when we find a model of it somewhere. Since sets are the easiest places to work, we will consider set-models only.

2.1. **Examples.**

- Any category at all describes a theory which has a model that has only one type which is empty: simply interpret all the type-templates (i.e. objects in the category) as being that set, and all arrows become the identity.
- Any category at all describes a theory which has a model that has only one type which has one element in: simply interpret all the type-templates (i.e. objects in the category) as being that set, and all arrows again become the identity.
- Any category with one object (i.e. any monoid when viewed as a category) has a model where there is only one type, and the elements of that type are the elements of the monoid (i.e. the arrows in the category). (Bear this one in mind.)
- The category with three objects A, B, C, with unique non-identity arrows $f, g : A \to B$, $h : B \to C$, $k : A \to C$ has a model where A is represented by $\mathbb{Z}$, B is represented by $\mathbb{N}$, and C is represented by Bool. Indeed, take $f = |\cdot|$ the absolute value function, g the negation function, and h the "is even" function.
- That category also has a model (which is not a set-model) where A is the type of groups, B is also the type of groups, and C is the type of sets. Take f to be the identity, g to be the "construct a group with the same objects but with multiplication reversed", and h to be "take the underlying set of the group".
- That category has a much more boring set-model: take $A = \{1, 2\}$, $B = \{5, 6\}$, and $C = \{1\}$. Let $f : n \mapsto n + 4$, let g be the other bijection, and let h be $n \mapsto 1$.

In general, any category has lots and lots of models.

2.2. **What is a model anyway?** A set-model of the theory is identifying, for each type-description in the category, a set whose elements are representing elements of that type. It also identifies, for each description of a unary predicate on the types, a function between the types; and the models of the unary predicates have to compose in a way that is reflected in the type-description.

This is just a functor from $\mathcal{C} \to \mathbf{Set}$!

So the first key insight is that a functor from $\mathcal{C} \to \mathbf{Set}$ is precisely a model of the theory described by $\mathcal{C}$.

2.3. **Homomorphisms between models.** We define a certain restricted form of model homomorphism as follows. (Note that this is not quite what is usually meant by a model homomorphism, and I have invented the term "fixed model homomorphism" to describe it.)

A *fixed model homomorphism* α is a function from one model $F : \mathcal{C} \to \mathbf{Set}$ of the theory $\mathcal{C}$ to another model $G : \mathcal{C} \to \mathbf{Set}$, which assigns to each type FA of the model F the corresponding type GA of the model G, in such a way that α respects the predicates $Ff : FA \to FB$ of the model:

$$(FA \xrightarrow{Ff} FB \xrightarrow{\alpha_B} GB) = (FA \xrightarrow{\alpha_A} GA \xrightarrow{Gf})GB$$

Notice that this is a model homomorphism which additionally ensures that FA is always mapped to GA (for any A), so (for example) it won't collapse all the objects FA into a single object in G's image unless G is the trivial model.

You might recognise the definition of a fixed model homomorphism as being the definition of a natural transformation between F and G when viewed as functors.

So the second key insight is that a natural transformation between functors $F : \mathcal{C} \to \mathbf{Set}$ and G is just a fixed homomorphism between the **Set**-models F and G of the theory $\mathcal{C}$.

3. Free models

Throughout mathematics, there is the notion of a free object: an object which somehow has the least possible restriction while still obeying all the rules it has to obey. Can we find a free model of the theory represented by the category $\mathcal{C}$?

Imagine $\mathcal{C}$ has two objects. Then any free model worth its name must have at least two types - otherwise we've definitely lost information in the model. (The theory said there were two types, so our model had better have at least two types or else it's not a good model of the theory.)

Likewise, any free model worth its name had better have *at most* two types, since otherwise we've added extra structure that the theory didn't specify. For concreteness, take our category to be the unique category with two objects and a single arrow from one to the other (and also the identity arrows). Then if our free model had three types in it, that would be very weird (somehow the model would fundamentally require introducing extra things into the universe if we ever wanted to realise the model).

So our free model had better have exactly one type for every object in the category.

Moreover, for the same reasons, every arrow in the category should have exactly one corresponding unary function in the model. (Any fewer, and we've lost the information that the theory should have some particular predicate; any more and we've somehow got a theory with a predicate that can't be realised without adding an extra function to the universe.)

An excellent guess for a free model turns out to be the following (called the *term model*): pick some type (i.e. some object in the category). Let that type have exactly one element, and then chase through all the functions, declaring that everything which is not obviously the same as something we've already made is different. (By analogy with the construction of the free group on some generators: keep piling together the generators, and declare to be different every word which you haven't obviously already made.) Declare that there are no other things in the universe: if we haven't constructed some member of a type this way, then that member can't exist. It's called the term model because we select a type, declare that there is a term of that type, and then see what else is forced to exist.

3.1. **Examples.**

3.1.1. *Simplest example.* For example, in the category above with two objects A and B, and a single arrow f from A to B, we could construct two different models this way. The first is the A-related model: declare that the type A has a single element a, and then all the other things in the universe are those constructed from a. (That is, $\mathrm{id}(a) = a$ which we already know about; $f(a) \in B$ which we don't yet know about so we'll note down that this is a new thing in the universe; and then $\mathrm{id}(f(b))$ which does already exist and is $f(b)$).

The second is the B-related model: declare that the type B has a single element b, and then all the other things in the universe are those constructed from b. (That is, $\mathrm{id}(b) = b$, and then there are no more ways to create elements because there are no other arrows from B.)

So the two term models we have constructed are:

- The one based at A, which has the type A consisting of a single element, and the type B consisting of a single element.
- The one based at B, which has the type A consisting of no elements at all, and the type B consisting of a single element.

3.2. **More complex example.** We'll look at an example with three types: A, B, C with arrows $f, g : A \to B$ and $h : B \to C$, as well as two distinct arrows $hf, hg : A \to C$. One model of this is where $A = \mathbb{N}$, $B = \mathbb{Z}$, $C = \mathrm{Bool}$, and f is the obvious injection $n \mapsto n$, g is the "negate" function $n \mapsto -n$, and h is the "is negative" function.

Then there will be three term models: one based at A, one at B, and one at C.

- At A: we have one element $a \in A$; then two elements of B, namely $f(a)$ and $g(a)$; then two elements of C, namely $hf(a)$ and $hf(b)$.
- At B: we have one element $b \in B$; then only one element of C, namely $h(b)$.
- At C: we have one element $c \in C$ only.

3.3. **A cyclic example.** Consider the category with two objects A, B only, and arrows $f : A \to B$ and $g : B \to A$ which compose so that $gf = 1_A$ but $fg \neq 1_B$. Then there are two term models:

- At A: there is $a \in A$, then $f(a) \in B$, then $gf(a) \in A$ but that is known to be just a, so we're done.
- At B: there is $b \in B$, then $g(b) \in A$, then $fg(b) \in B$, then $gfg(b) = g(b) \in A$, and we're done.

3.4. **An infinite example.** Consider the category which is the monoid $\mathbb{N}$: namely, a single element A and then an arrow f_i for each $i \in \mathbb{N}$ such that $f_i f_j = f_{i+j}$.

Then there is just one term model, and its elements are $a, f_1(a), f_2(a), \dots$.

3.5. **General definition of the term model.** If you have been thinking categorically, you may have noticed that in every example above, if you simply remove the term from everything when we list elements, we end up with a collection of arrows in the category. For example, in the natural numbers case, we had the elements $a, f_1(a), f_2(a), \dots$; delete the a and we get $\{\mathrm{id}, f_1, f_2, \dots\}$. And this, crucially, is simply a list of the arrows out of A.

In general, the collection of {things which exist in the term model based at an object A} is isomorphic to the set of arrows out of A: that is, it's the hom-set $\mathrm{Hom}(\mathrm{A}, -)$.

So the third key insight is that the term models are precisely the hom-sets of the category.

4. The Yoneda lemma

Now we can understand the Yoneda lemma's statement in the light of these new concepts:

Theorem 2 (The Yoneda lemma, in new terms)**.** *Let $\mathcal{C}$ be an algebraic theory with unary predicates, and let $G : \mathcal{C} \to \mathbf{Set}$ be a model of that theory. Let A be a type in $\mathcal{C}$. Then the collection of fixed model homomorphisms from the term model based at A into the model G is isomorphic to the set of things of type A in the model G. Moreover, the isomorphism is natural in both the type we chose, and the model we chose.*

After a little thought, the existence of the bijection is just obvious. Indeed, a homomorphism from the term model based at A into any other model G is exactly defined by where a goes: nothing exists in the term model except things which are derived by applying arrows to a, so if we've decided where a goes then we've decided where everything in the term model goes. Hence for every fixed model homomorphism from $\mathrm{Hom}(\mathrm{A}, -)$ to G, we can canonically define a member of the concrete type GA which is "where did a end up". Conversely, if we're given an element $x \in GA$ of the instantiation of the type A in model G, we can canonically define a fixed model homomorphism from $\mathrm{Hom}(\mathrm{A}, -)$ by sending our abstract term a to x, and letting all the rest of the elements of the term model get pulled along with it.

4.1. **Naturality.** I'm afraid I don't know of a good way to think about naturality other than just to draw out the diagrams and show they commute; but they're both easy and I can't be bothered to do them right now.

5. Relation to the notion of the free model

It turns out that the Yoneda lemma can be used to prove that the term models together form a "free" collection in some sense. The terse way to say this is that the Yoneda embedding $A \mapsto \mathrm{Hom}(\mathrm{A}, -)$ is full and faithful from $\mathcal{C}^{\mathrm{op}} \to [\mathcal{C}, \mathbf{Set}]$.

In more elementary terms: the conversion from objects to models, given by taking an object and producing the term model, loses no information about the category. If we select two types A and B in the category $\mathcal{C}$, and take a pair of different arrows $f, g : A \to B$, then these two arrows correspond to a pair of fixed model homomorphisms (natural transformations) between the term models $\mathrm{Hom}(\mathrm{A}, -)$ and $\mathrm{Hom}(\mathrm{B}, -)$ (given by "replace a by $f(a)$" and "replace a by $g(a)$" respectively). Moreover, the two homomorphisms really are different from each other.

So given a category (a specification of an algebraic theory), we can produce a specific collection of models for that theory and a specific collection of homomorphisms between the models, such that all the information about the theory can be recovered from the

models. There are loads more models out there, and loads more homomorphisms between those extra models, but if we restrict our attention only to the term models then we recover all the information about the original category. Moreover, remove any of the models or any of the homomorphisms, and we stop being able to pick out the theory we are modelling uniquely.

That is, the collection of term models is "free": they haven't lost us any information about the theory (we can use them to recover the original category entirely), and nor do they contain any extra information (*every* fixed model homomorphism between term models is required, or else we have lost some information about the category).

6. Acknowledgements

This entire document is derived from an answer by Sridhar Ramesh on a Math Overflow answer at `https://mathoverflow.net/a/15143`.