Diagonal Arguments and Cartesian Closed Categories

by

F. William Lawvere

The similarity between the famous arguments of Cantor, Rüssell, Gödel and Tarski. is well-known, and suggests that these arguments should all be special cases of a single theorem about a suitable kind of abstract structure. We offer here a fixed-point theorem in cartesian closed categories which seems to play this role. Cartesian closed categories seem also to serve as a common abstraction of type theory and propositional logic, but the author's discussion at the Seattle conference of the development of that observation will be in part described elsewhere ["Adjointness in Foundations", to appear in Dialectica, and "Equality in Hyperdoctrines and the Comprehension Schema as an Adjoint Functor", to appear in the Proceedings of the AMS Symposium on Applications of Category theory].

1. By a cartesian closed category is meant a category $\mathbb{C}$ equipped with the following three kinds of right adjoints: a right adjoint 1 to the unique

$$\mathbb{C} \to \mathbf{1},$$

a right adjoint $\times$ to the diagonal functor

$$\mathbb{C} \to \mathbb{C}\times\mathbb{C},$$

and for each object A in $\mathbb{C}$, a right adjoint $(\ )^A$ to the functor

$$\mathbb{C} \xrightarrow{A\times(\ )} \mathbb{C}.$$

The adjunction transformations for these adjoint situations, also assumed given, will be denoted by δ,π in the case of products and by λ_A,ϵ_A in the case of exponentiation by A. Thus for each X one has

$$X \xrightarrow{X\lambda_A} (A\times X)^A$$

and for each Y one has

$$A\times Y^A \xrightarrow{Y\epsilon_A} Y.$$

Given $f{:}A\times X \to Y$, the composite morphism

$$X \xrightarrow{X\lambda_A} (A\times X)^A \xrightarrow{f^A} Y^A$$

will be called the "λ-transform" of the morphism f. A morphism $h:X \to Y^A$ is the λ-transform of f iff the diagram

$$\begin{tikzcd}
A\times X \arrow[d, "A\times h"'] \arrow[dr, "f"] & \\
A\times Y^A \arrow[r, "Y\epsilon_A"'] & Y
\end{tikzcd}$$

is commutative, showing in particular that f can be uniquely recovered from its λ-transform. Taking the case $X = 1$, one has that every $f:A \to Y$ gives rise to a unique $\ulcorner f \urcorner :1 \to Y^A$ and that every $1 \to Y^A$ is of that form for a unique f. Since for every $a:1 \to A$ one has (dropping the indices A,Y on ε when they are clear)

$$\langle a, \ulcorner f \urcorner \rangle \epsilon = a.f,$$

one calls ε the "evaluation" natural transformation; note however that we do <u>not</u> assume in general that f is determined by the knowledge of all its "values" a.f.

Although we do not make use of it in this paper, the usefulness of cartesian closed categories as algebraic versions of type theory can be further illustrated by assuming that the coproduct

$$2 = 1+1$$

also exists in $\mathbb{C}$. It then follows (using the closed structure), that for every object A

$$A\times 2 = A+A$$

and so in particular that 2 is Boolean-algebra-object in $\mathbb{C}$, i.e. that among the morphisms

$$2\times 2\times \ldots \times 2 \to 2$$

in $\mathbb{C}$ there are well determined morphisms corresponding to all the finitary (two-valued) truth tables, and that these satisfy all the commutative diagrams expressing the axioms of Boolean algebra. Equivalently, for each X the set

$$P_{\mathbb{C}}(X) = \mathbb{C}(X,2)$$

of "$\mathbb{C}$-attributes of type X" becomes canonically an actual Boolean algebra, and varying X along any morphism of $\mathbb{C}$ induces contravariantly a Boolean homomorphism of attribute algebras. The morphisms $1 \to 2$ form $P_{\mathbb{C}}(1)$ the Boolean algebra of "truth-values"; among these are the two coproduct injections which play the roles of "true" and "false". For any "constant of type X" $x:1 \to X$ and any attribute φ of type X, $x.\varphi$ is then a truth-value. Now noting that

$$X \times 2^X \xrightarrow[(2)\epsilon_X]{} 2$$

is a "binary operation" we could write it between its arguments, so that we have

$$x \epsilon \ulcorner\varphi\urcorner = x.\varphi,$$

an equality of truth values; thus if we think of $\ulcorner\varphi\urcorner:1 \to 2^X$ as the constant naming the subset of X corresponding to the attribute φ, one sees that the above equation expresses the usual "comprehension" axiom.

Returning to our immediate concern, we define a morphism $g:X \to Z$ to be <u>point-surjective</u> iff for every $z:1 \to Z$ there exists $x:1 \to X$ with $xg = z$. This does not imply that g is necessarily "onto the whole of Z", since there may be few morphisms with domain 1; for example if (as in the next section) X and Z are set-valued functors, then a natural transformation g is point-surjective if every element of the <u>inverse limit</u> of Z comes from an element of the inverse limit of X. In case Z is of the form Y^A, an even weaker notion of surjectivity can be considered, which in fact suffices for our fixed point theorem. Namely

$$X \xrightarrow{g} Y^A$$

will be called <u>weakly point-surjective</u> iff for every $f:A \to Y$ there is x such that for every $a:1 \to A$

$$\langle a, xg \rangle \epsilon = a.f$$

Finally we say that an object Y has the <u>fixed-point property</u> iff for every endomorphism $t:Y \to Y$ there is $y:1 \to Y$ with $y.t = y$.

<u>Theorem</u> In any cartesian closed category, if there exists an object A and a weakly point-surjective morphism

$$A \xrightarrow{g} Y^A$$

then Y has the fixed point property.

Proof: Let $\bar{g}$ be the morphism whose λ-transform is g. Then for any $f:A \to Y$ there is $x:1 \to A$ such that for all $a:1 \to A$

$$\langle a,x \rangle \bar{g} = a.f.$$

Now consider any endomorphism t of Y and let f be the composition

$$\begin{tikzcd} A \arrow[r, "A\delta"] & A\times A \arrow[r, "\bar{g}"] & Y \arrow[r, "t"] & Y; \end{tikzcd}$$

thus there is x such that for all a

$$\langle a,x \rangle \bar{g} = \langle a,a \rangle \bar{g}t$$

since $a(A\delta) = \langle a,a \rangle$. But then $y = \langle x,x \rangle \bar{g}$ is clearly a fixed point for t.

The famed "diagonal argument" is of course just the contrapositive of our theorem. Cantor's theorem then follows with $Y = 2$.

Corollary If there exists $t:Y \to Y$ such that $yt \neq y$ for all $y:1 \to Y$ then for no A does there exist a point-surjective morphism

$$A \to Y^A$$

(or even a weakly point-surjective morphism).

2. Russells Paradox does not presuppose that set theory be formulated as a higher type theory; that is, for A the set-theoretical universe, we do not need 2^A for the argument. In fact we need only apply the proof of our theorem, with $\bar{g}:A\times A \to 2$ as the set-theoretical membership relation, dispensing with g entirely. That is, more generally, our theorem could have been stated and proved in any category with only finite products (no exponentiation) by simply phrasing the notion of (weak) point-surjectivity as a property of a morphism

$$A\times X \to Y;$$

however discovering the latter form (or at least calling it surjectivity!) seems to require thinking of such a morphism as a family of morphisms $A \to Y$ indexed by the elements of X, suggesting that a closed category is the "natural" setting for the theorem.

In fact the more general form of the theorem just alluded to (for categories with products) follows from the cartesian closed version which we have proved, by virtue of the following remark. Notice that it would suffice to assume $\mathbb{C}$ small (just take the full closure under finite products of the two objects A,Y)

Remark Any small category $\mathbb{C}$ can be fully and faithfully embedded in a cartesian closed category in a manner which preserves any products or exponentials which may exist in $\mathbb{C}$.

Proof: We consider the usual embedding

$$\mathbb{C} \subseteq \mathcal{S}^{\mathbb{C}^{op}}$$

which identifies an object Y with the contravariant set-valued functor

$$X \rightsquigarrow \mathbb{C}(X,Y).$$

By "Yoneda's Lemma" one has for any functor Y and any object A that the value at A of Y

$$AY \xrightarrow{\sim} \mathcal{S}^{\mathbb{C}^{op}}(A,Y)$$

where the right hand side denotes the set of all natural transformations from (the functor corresponding to) A into Y, so that in particular the embedding is full and faithful. It is then also clear that the embedding preserves products (in particular if 1 exists in $\mathbb{C}$ it corresponds to the functor which is constantly the one-element set, which is the 1 of $\mathcal{S}^{\mathbb{C}^{op}}$). For any two functors A,Y the functor

$$C \rightsquigarrow \mathcal{S}^{\mathbb{C}^{op}}(A\times C,Y)$$

plays the role of Y^A. In particular if B^A exists in $\mathbb{C}$ for a pair of objects A,B in $\mathbb{C}$ then

$$(C)B^A \cong \mathbb{C}(C,B^A) \cong \mathbb{C}(A\times C,B) \cong \mathcal{S}^{\mathbb{C}^{op}}(A\times C,B)$$

showing that the embedding preserves exponentiation.

Theorem Let A,Y be any objects in any category with finite products (including the empty product 1); then the following two statements cannot both be true

a) there exists $\bar{g}:A\times A \to Y$ such that for all $f:A \to Y$ there exists $x:1 \to A$ such that for all $a:1 \to A$

$$\langle a,x \rangle \bar{g} = a.f$$

b) there exists $t:Y \to Y$ such that for all $y:1 \to Y$

$$y.t \neq y.$$

Proof: Apply above remark and the proof in the previous section.

Of course the "transcendental" proof just given is somewhat ridiculous, since the incompatibility of a) and b) can be proved directly just as simply as it was proved in the previous section under the more restrictive hypothesis on $\mathbb{C}$. However we wish to take the opportunity to make some further remarks about the above canonical embedding of an arbitrary (small) category into a cartesian closed category $\bar{\mathbb{C}}$ (let the latter denote the smallest full cartesian closed subcategory of $\mathcal{S}^{\mathbb{C}^{op}}$ which contains $\mathbb{C}$). One of the standard ways of embedding a structure into a higher-order structure is to consider "definable" functionals, operators, etc.; however this is difficult to oversee from a simple-minded point of view since it usually requires enumerating all possible definitions. On the other hand in many situations (e.g. functorial semantics of algebraic theories or functorial semantics of elementary theories if the elementary theories are complete) one has come to expect that natural transformations are identical with definable ones or at least a reasonable substitute for definable ones. The latter alternative seems to be at least partly true in the present case. Thus for example we are led to the following definition. If A,B,C,D are objects in a category $\mathbb{C}$ with finite products, a <u>natural operator</u>

$$B^A \xrightarrow{\Phi} D^C$$

shall be simply a natural transformation between the exponential functors of the (functors corresponding to the) given objects in $\mathcal{S}^{\mathbb{C}^{op}}$ (hence in $\bar{\mathbb{C}}$). in particular if $C = 1$ we would call a natural operator a natural functional. Note that 1 will not be a generator for all of $\mathcal{S}^{\mathbb{C}^{op}}$ unless $\mathbb{C} = 1$; however it might conceivably be so for $\bar{\mathbb{C}}$, and we have a partial result in that direction. In fact, in the case that 1 is a generator for $\mathbb{C}$ itself, we can describe in more familiar terms what a natural operator is.

Recall that "1 is a generator for $\mathbb{C}$" simply means that a morphism $f:X \to Y$ in $\mathbb{C}$ is determined by its "values" $x.f:1 \to Y$ for $x:1 \to X$. In that case it is sensible to call the elements of the set $\mathbb{C}(1,X)$ of points of X also the <u>elements of X</u>. Then a function

$$\mathbb{C}(1,X) \to \mathbb{C}(1,Y)$$

is induced by at most one $\mathbb{C}$-morphism $X \to Y$, and in case it is, we say by abuse of language that the function <u>is</u> a morphism of $\mathbb{C}$.

<u>Proposition</u> Suppose that $\mathbb{C}$ is a category with finite products in which 1 is a generator, and that A,B,C,D are objects of $\mathbb{C}$. Then

1) a natural operator

$$B^A \xrightarrow{\Phi} D^C$$

is entirely determined by a single function

$$\mathbb{C}(A,B) \xrightarrow{1\Phi} \mathbb{C}(C,D)$$

and

2) such a function determines a natural operator iff for every object X of $\mathbb{C}$ and for every $\mathbb{C}$-morphism $f{:}A\times X \to B$, the function

$$\mathbb{C}(1,C\times X) \xrightarrow{(f)(X\Phi)} \mathbb{C}(1,D)$$

is a $\mathbb{C}$-morphism, where $(f)(X\Phi)$ is defined by

$$< c,x >\big((f)(X\Phi)\big) = (c)\big((f_x)(1\Phi)\big)$$

for any $c{:}1 \to C$, $x{:}1 = X$, f_x denoting the composition

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & A\times X \arrow[r, "f"] & B. \\ & A\times 1 \arrow[ur, "A\times x"'] & & \end{tikzcd}$$

Proof: We are abusing notations to the extent of identifying a morphism with its λ-transform via the bijections of the form

$$\mathbb{C}(A\times X,B) \cong \bar{\mathbb{C}}(A\times X,B) \cong \bar{\mathbb{C}}(X,B^A).$$

Actually the given operator Φ is a family of functions

$$\mathbb{C}(X,B^A) \xrightarrow{X\Phi} \mathbb{C}(X,D^C)$$

one for each object of $\mathbb{C}$; the "naturalness" condition which this family must satisfy, is, via the abuse, that for every morphism $x{:}X' \to X$ of $\mathbb{C}$, the diagram

$$\begin{tikzcd} \mathbb{C}(A\times X,B) \arrow[r, "X\Phi"] \arrow[d, "x"'] & \mathbb{C}(C\times X,D) \arrow[d, "x"] \\ \mathbb{C}(A\times X',B) \arrow[r, "X'\Phi"'] & \mathbb{C}(C\times X',D) \end{tikzcd}$$

should commute. Now let $X' = 1$. Since 1 is a generator for $\mathbb{C}$, the value of the function $X\Phi$ at a given $f{:}A\times X \to B$ is determined by the knowledge, for each element x of X and element c of C, the result reached in the lower right hand corner by going across then down in the commutative diagram

$$\begin{tikzcd}
\mathbb{C}(A\times X,B) \arrow[r, "X\Phi"] \arrow[d, "x"'] & \mathbb{C}(C\times X,D) \arrow[r, "c"] \arrow[d, "x"] & \mathbb{C}(X,D) \arrow[d, "x"] \\
\mathbb{C}(A,B) \arrow[r, "1\Phi"'] & \mathbb{C}(C,D) \arrow[r, "c"'] & \mathbb{C}(1,D).
\end{tikzcd}$$

But since the same results are obtained by going down then across, all the functions $X\Phi$ are determined by the one function 1Φ, proving the first assertion. The second assertion is then clear, since the definition of $(f)(X\Phi)$ given in the statement of the proposition is just such as to assure naturality of $X\Phi$ provided its values exist.

To make the situation perfectly clear, notice that morphisms whose codomain is an exponential object can be discussed even though the exponential object does not exist, just by considering instead morphisms whose domain is a product. There is however then the problem of determining the morphisms whose domain is an exponential, and considering them to be the natural operators is in many contexts the smoothest and most "natural" thing to do. Experts on recursive functions or C^∞ functions between finite-dimensional manifolds may wish to consider the result of taking $\mathbb{C}$ to be these particular categories in the above considerations. They may also wish to consider whether the fixed-point theorem of section one has any applications in those cases.

3. In order to apply the theorem of the previous section to obtain Tarski's theorem concerning the impossibility of defining truth for a theory within the theory itself, we first note briefly how a theory gives rise to a category $\mathbb{C}$ with finite products. Consider two objects A,2 and let the $\mathbb{C}$-morphisms be equivalence classes of (tuples of) formulas or terms of the theory, where two formulas (or terms) are considered equivalent iff their logical equivalence (or equality) is provable in the theory. Thus the morphisms $1 \to A$ are (classes of) constant terms, the morphisms $A\times A \to A$ are (classes of) terms with two free variables, while morphisms $A^n \to 2$ are (classes of) formulas with n free variables so that in particular morphisms $1 \to 2$ are (classes of) sentences of the theory. In particular there is a morphism true: $1 \to 2$ corresponding to the class of sentences provable in the theory and similarly a morphism false: $1 \to 2$ corresponding to the class of sentences whose negation is provabke in

the theory. Morphisms $2^n \to 2$ would include all propositional operations, but we will make no use of that except for the following case:

If the theory is consistent there is a morphism not: $2 \to 2$ such that φ not $\neq \varphi$ for all morphisms $\varphi: 1 \to 2$

In particular we will not need to use the fact that $2 = 1+1$, although that determines the nature of those hom-sets not explicitly spelled out above. Defining composition to correspond to substitution (for example a constant $a: 1 \to A$ composed with a unary formula $\varphi: A \to 2$ composed with not gives the sentence $a\varphi$ not: $1 \to 2$, etc.) we get a category $\mathbb{C}$ with finite products which might be called the Lindenbaum category of the theory. Models of the theory can then be viewed as certain functors $\mathbb{C} \to \mathcal{S}$. We make no use here of the operation in $\mathbb{C}$ induced by quantification in the theory, but the categorical description of this operation will be clear to readers of the two papers cited in the introduction. In our construction above of $\mathbb{C}$ we have tacitly assumed that the theory was a first-order single-sorted one, in which case all objects of $\mathbb{C}$ are isomorphic to those of the form $A^n \times 2^m$, but with trivial modifications we could have started with a higher-order or several-sorted theory with no change of any significance to the arguments below. To make one point somewhat more explicit note that the projection morphisms $A^n \to A$ arise from the variables of the theory.

We then say that <u>satisfaction is definable</u> in the theory iff there is a binary formula $\mathrm{sat}: A \times A \to 2$ in $\mathbb{C}$ such that for every unary formula $\varphi: A \to 2$ there is a constant $c: 1 \to A$ such that for every constant a the following diagram commutes in $\mathbb{C}$

$$\begin{tikzcd}
1 \arrow[r, "a"] \arrow[d, "{\langle a,c \rangle}"'] & A \arrow[d, "\varphi"] \\
A \times A \arrow[r, "\mathrm{sat}"'] & 2
\end{tikzcd}$$

Here we imagine taking for c a Gödel number for (one of the representatives of) φ. The condition would traditionally be expressed by requiring that the sentence

$$a \ \mathrm{sat} \ c \longleftrightarrow a\varphi$$

be provable in the theory, but if $\mathbb{C}$ arises from our construction of the Lindenbaum category this amounts to the same thing.

Combining the above notion with our remark about the meaning of consistency and the theorem of the previous section we have immediately the

Corollary If satisfaction is definable in the theory then the theory is not consistent.

In order to show that Truth cannot be defined we first need to say what Truth would mean, which seems to require some further assumptions on the theory, which are however often realizable. Namely we suppose that there is a binary term

$$A\times A \xrightarrow{\text{subst}} A$$

in $\mathbb{C}$ and a ("metamathematical") binary relation

$$\Gamma \subseteq \mathbb{C}(1,A)\times\mathbb{C}(1,2)$$

between constants and sentences for which the following holds.

1) For all $\varphi:A \to 2$ there is $c:1 \to A$ such that for all $a:1 \to A$

$$(a \text{ subst } c)\,\Gamma(a\varphi)$$

For example we could imagine that $d\Gamma\sigma$ means that d is the Gödel number of some one of the sentences which represent σ, and that subst is a binary operation which, when applied to a constant a and to a constant c which happens to be the Gödel number of a unary formula φ, yields the Gödel number of the sentence obtained by substituting a into φ.

Given a binary relation $\Gamma \subseteq \mathbb{C}(1,A)\times\mathbb{C}(1,2)$ we say that Truth (of sentences) is definable in the theory (relative to Γ) provided there is a unary formula Truth:$A \to 2$ such that

2) For all $\sigma:1 \to 2$ and $d:1 \to A$, if $d\Gamma\sigma$ then $d\text{Truth} = \sigma$

Again the traditional formulation would require that the sentence

$$\ulcorner\sigma\urcorner \text{ Truth} \longleftrightarrow \sigma, \text{ for } \ulcorner\sigma\urcorner\Gamma\sigma$$

be provable, but in the Lindenbaum category this just amounts to the equation $\ulcorner\sigma\urcorner\text{Truth} = \sigma$.

Theorem If the theory is consistent and substitution is definable relative to a given binary relation Γ between constants and sentences, then Truth is not definable relative to the same binary relation

Proof: If both 1) and 2) hold then the diagram

$$\begin{tikzcd}
1 \arrow[r, "a"] \arrow[d, "{\langle a,c \rangle}"'] \arrow[dr, "d"] & A \arrow[dr, "\varphi"] & \\
A\times A \arrow[r, "\text{subst}"'] & A \arrow[r, "\text{Truth}"'] & 2
\end{tikzcd}$$

shows that

$$\begin{tikzcd}
A\times A \arrow[r, "\text{subst}"] \arrow[dr, "\text{sat}"'] & A \arrow[d, "\text{Truth}"] \\
& 2
\end{tikzcd}$$

is a definition of satisfaction, contradicting the previous result.

We will also prove an "incompleteness theorem", using the notion of a Provability predicate. Given a binary relation Γ between constants and sentences, we say that <u>Provability is representable in the theory</u> iff there is a unary formula $Pr:A \to 2$ such that

3) Whenever $d\Gamma\sigma$ then

$$dPr = \text{true} \quad \text{iff} \quad \sigma = \text{true}$$

<u>Theorem</u> Suppose that for a given binary relation Γ between constants and sentences of C, substitution is definable and Provability is representable. Then the theory is not complete if it is consistent.

Proof: Suppose on the contrary that $C(1,2) = \{\text{false},\text{true}\}$. Our notion of consistency implies that false $\neq$ true. Condition 3) states that for $d\Gamma\sigma$

a) $\sigma = \text{true}$ implies $dPr = \text{true}$

b) $\sigma \neq \text{true}$ implies $dPr \neq \text{true}$

By completeness b) implies

b') $\sigma = \text{false}$ implies $dPr = \text{false}$

But a) and b') together with completeness mean that whenever $d\Gamma\sigma$

$$\begin{tikzcd}
1 \arrow[d, "d"'] \arrow[dr, "\sigma"] & \\
A \arrow[r, "Pr"'] & 2
\end{tikzcd}$$

is commutative, i.e. that Pr satisfies condition 2) for a Truth-definition, which by our previous theorem yields a contradiction.

Note: Our proposition in section two can be interpreted as a fragment of a general theory developed by Eilenberg and Kelly from an idea of Spanier.