Nullstellen and Subdirect Representation

Walter Tholen

Dedicated to George Janelidze

Abstract
David Hilbert's solvability criterion for polynomial systems in n variables from the 1890s was linked by Emmy Noether in the 1920s to the decomposition of ideals in commutative rings, which in turn led Garret Birkhoff in the 1940s to his subdirect representation theorem for general algebras. The Hilbert-Noether-Birkhoff linkage was brought to light in the late 1990s in talks by Bill Lawvere. The aim of this article is to analyze this linkage in the most elementary terms and then, based on our work of the 1980s, to present a general categorical framework for Birkhoff's theorem.

Introduction
The first purpose of this article is to exhibit in elementary algebraic terms the linkage between Hilbert's celebrated *Nullstellensatz* [6] for systems of polynomial equations and Birkhoff's *Subdirect Representation Theorem* [2] for general algebras, as facilitated by Noether's work [15] on the decomposition of ideals in rings. Hence, in Section 1 we give a brief tour of the development of the *Nullstellensatz* through Hilbert, Noether and Birkhoff. We do so not from a strictly historical perspective; rather, in today's language we present six versions of the theorem as marked by these three great mathematicians and show how their proofs are interrelated, referring to them as the *HNB Theorems*.

A seventh and an eighth version are given in Section 3, after a discussion of the categorical notion of subdirect irreducibility in Section 2. We illuminate the notion by examples, both traditional and unconventional, in particular in comma categories. We touch upon the dual notion only briefly, but refer the reader to substantial recent work by Matias Menni [14] on Lawvere's concept of cohesion in this context. With a suitable notion of finitariness we formulate the all-encompassing seventh version of the HNB Theorem without recourse to any limits or colimits in the ambient category. The morphism version of it leads to atypical factorizations of morphisms, in the sense that even in standard categories like that of sets one obtains factorizations of maps in a constructive manner (without recourse to choice) which, however, may not be obtained in a functorial way.

Acknowledgement I am indebted to Bill Lawvere for his suggestion to study the topic of this paper and his subsequent interest in and comment on a preliminary (but extended) version of this work, presented in part at The European Category Theory Meeting in Haute Bodeux (Belgium) in 2003. I am also grateful to László Márki for his careful reading of an earlier version of this paper and for detecting an inaccuracy that may be traced back to Birkhoff's original paper (see 2.3(4) below).

1 From Hilbert via Noether to Birkhoff

1.1 Hilbert

In 1892 David Hilbert proved his famous *Nullstellensatz* (see [6]) which we record here in the following form:

HNB Theorem (Version 1) *Any system of polynomial equations*

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_r(x_1, \ldots, x_n) &= 0 \end{aligned} \tag{$*$}$$

in n variables with coefficients in a field k has a common solution $a = (a_1, \ldots, a_n)$ in F^n for any algebraically-closed extension field F of k, unless *there are polynomials $g_i \in k[x_1, \ldots, x_n]$ with*

$$\sum_{i=1}^{r} g_i f_i = 1.$$

It is clear that this last provision is needed since otherwise

$$0 = \sum_{i=1}^{r} g_i(a) f_i(a) = 1$$

for any common solution a of $(*)$. Of course algebraic closedness of F is also indispensable (as the case $n = 1 = r$ shows).

Let us briefly recall the abstract machinery which bridges algebra with geometry. Given *any* set $P \subseteq k[x_1, \ldots, x_n]$ of polynomials one defines the *k-algebraic set* (or *affine k-variety*) of P as the set

$$S(P) = \{a \in F^n \mid \forall f \in P : f(a) = 0\}$$

of common zeroes in F^n to the polynomials of P, for a given extension field F of k. Conversely, given any subset $X \subseteq F^n$, consider the set

$$J(X) = \{f \in k[x_1, \ldots, x_n] \mid \forall a \in X : f(a) = 0\}$$

of polynomials in $k[x_1, \ldots, x_n]$ which have the given points as zeroes. One obtains order-reversing maps

$$\begin{aligned} P &\longmapsto S(P) \\ J(X) &\longleftarrow X \end{aligned}$$

which are adjoint to each other:

$$P \subseteq J(X) \iff X \subseteq S(P),$$

i.e., they form a Galois correspondence. Like any such correspondence, they establish a bijection between the *closed* elements on each side, namely:

— the k-algebraic sets $X \subseteq F^n$, i.e. those X with $X = S(J(X))$, and

— the sets $P = J(X)$ for some $X \subseteq F^n$, i.e. those P with $P = J(S(P))$.

Consequently, one studies k-algebraic sets with the help of this correspondence once one has a good characterisation of the closed sets $P \subseteq k[x_1, \ldots, x_n]$. As a necessary condition, the sets $P = J(X)$ are ideals in the ring $k[x_1, \ldots, x_r]$ that, moreover, are *radical*, i.e., $P = \sqrt{P}$, with

$$\sqrt{P} = \{f \in k[x_1, \ldots, x_n] \mid \exists m \geq 1 : f^m \in P\}.$$

An advanced formulation of the *Nullstellensatz* says that the closed sets P are precisely the radical ideals:

HNB Theorem (Version 2) *For any proper ideal $P \trianglelefteq k[x_1, \ldots, x_n]$ and any algebraically-closed extension field F of k, the ideal $J(S(P))$ given by the k-algebraic set $S(P) \subseteq F^n$ is precisely the radical of P. Hence, $f(a) = 0$ for every common zero $a \in F^n$ of all polynomials in P if and only if $f^m \in P$ for some $m \geq 1$.*

Of course, this formulation entails the seemingly more elementary Version 1. The polynomials $f_1, \ldots, f_r$ of $(*)$ generate the ideal $P = (f_1, \ldots, f_r)$, which is proper since, by hypothesis, $1 \notin P$. In fact, $1 \notin \sqrt{P}$, so that Version 2 gives that $\sqrt{P} = J(S(P))$ is proper, i.e., not the whole ring $k[x_1, \ldots, x_n] = J(\emptyset)$, trivially. But equally trivially one has $S(k[x_1, \ldots, x_n]) = \emptyset$, so that we can't have $S(P) = \emptyset$ since $\sqrt{P} \neq k[x_1, \ldots, x_n]$! Hence, $S(P) \neq \emptyset$. □

Most algebra books show that, conversely, the advanced formulation of the *Nullstellensatz* can be derived from the elementary one and then proceed to give a proof of the elementary version. The reduction is based on a beautiful trick credited to Rabinowitsch and is described already in van der Waerden's classical book on Algebra (see [19]). Here we will elaborate on a more "structural" proof (see 1.2 below). Let us also point out that replacing the finite systems of polynomials $f_1, \ldots, f_r$ by arbitrary sets $P \subseteq k[x_1, \ldots, x_n]$ only seemingly adds generality since the set $S(P) = S(J(S(P)))$ can be described using a finite system of generators of the ideal $J(S(P))$, thanks to:

Hilbert's Basis Theorem *Every ideal in $k[x_1, \ldots, x_n]$ is finitely generated, that is: the ring $k[x_1, \ldots, x_n]$ is Noetherian, for every commutative Noetherian ring k.*

The proof does not require any advanced tools and can be found in any sufficiently sophisticated algebra book (see [7], p. 391).

1.2 Noether

In order to prove Version 2 of the HNB Theorem "directly", we must show

$$J(S(P)) = \sqrt{P}$$

for any proper ideal $P \trianglelefteq [x_1, \ldots, x_n]$, with $S(P) \subseteq F^n$ for an algebraically closed extension field F of k. Since the inclusion "$\supseteq$" is trivial, it suffices to find for any polynomial $f \notin \sqrt{P}$ a point $a \in F^n$ with $a \in S(P)$ but $f(a) \neq 0$. The existence of such a point would be guaranteed once we had proved the following version:

HNB Theorem (Version 3) *Let A be a finitely-generated commutative k-algebra with unit and let F be an algebraically closed extension field of the field k. Then, for any non-nilpotent element $u \in A$ there is a k-homomorphism $\varphi : A \to F$ with $\varphi(u) \neq 0$.*

In fact, with Version 3 at our disposal, we would consider

$$A = k[x_1, \ldots, x_n]/\sqrt{P}$$

which has no non-zero nilpotent elements since $\sqrt{P}$ is a radical ideal, put $u = \pi(f)$ (where $\pi : k[x_1, \ldots, x_n] \to A$ is the projection and $f \notin \sqrt{P}$), get $\varphi : A \to F$ with $\varphi(u) \neq 0$, and put

$$a = (a_1, \ldots, a_n) := (\varphi(\pi(x_1)), \ldots, \varphi(\pi(x_n))).$$

Then

$$p(a) = p(\varphi\pi(x_1), \ldots, \varphi\pi(x_n)) = \varphi\pi(p(x_1, \ldots, x_n)) = \varphi(\pi(p)) = 0$$

for all $p \in P$, while

$$f(a) = f(\varphi\pi(x_1), \ldots, \varphi\pi(x_n)) = \varphi\pi(f(x_1, \ldots, x_n)) = \varphi(u) \neq 0,$$

as desired. □

We now claim that a key ingredient to the proof of Version 3 is:

HNB Theorem (Version 4) *Let u be a non-nilpotent element of the finitely-generated commutative unital k-algebra A. Then (depending on u) there exists an extension field K of k which, as a unital k-algebra, is finitely generated, and a k-homomorphism $\chi : A \to K$ with $\chi(u) \neq 0$.*

Indeed, Version 4 easily implies Version 3 if we make use of two basic but important facts of Algebra (see [1], pp. 57 and 90), which are only needed to "channel" the variable fields K of Version 4 into the fixed field F of Version 3.

Fact 1. *Any extension field K of k which, as a k-algebra, is finitely generated, is algebraic over k.*

Fact 2. *Let K be an algebraic extension field of k. Then any embedding $k \to F$ into an algebraically-closed field F can be extended to an embedding $K \to F$.*

Under the hypothesis of Version 3, obtain K and χ from Version 4, where K is an algebraic extension of k by Fact 1. From Fact 2 one obtains an extension

$\psi : K \to F$ of the given embedding $k \to F$, and $\varphi := \psi\chi$ will now do the job for Version 3. □

Let us now "reformulate" Version 4 in the spirit of Emmy Noether's work:

HNB Theorem (Version 5) *Let u be a non-nilpotent element in a commutative ring A. Then there exists a prime ideal Q in A with $u \notin Q$.*

PROOF. Since u is not nilpotent, the zero ideal in A does not meet the set $S = \{u^n \mid n \geq 1\}$. Zorn's Lemma allows us to find an ideal Q of A which is maximal w.r.t. the property $Q \cap S = \emptyset$. Such an ideal is indeed prime: for any elements $a, b \in A$ outside Q, the ideals $(a) + Q$, $(b) + Q$ must meet S, hence $u^k \in (a) + Q$ and $u^l \in (b) + Q$ for some $k, l \geq 1$; hence $u^{k+l} \in (ab) + Q$, which is possible only if $ab \notin Q$. □

One can now proceed to derive Version 4 from Version 5, as follows. For the k-algebra A as in Version 4 and $u \in A$, $u \neq 0$, obtain Q as in Version 5 and consider the projection $\sigma : A \to A/Q$ and the embedding $A/Q \hookrightarrow L$ of the integral domain A/Q into its field of fractions L. In the intermediate domain $R = (A/Q)[\sigma(u)^{-1}]$ we can choose a maximal ideal $M \neq R$ and form the projection $\tau : R \to K := R/M$. The field K is still finitely generated as a k-algebra, and $\tau(\sigma(u)) \neq 0$ since the invertible element $\sigma(u)$ of R cannot belong to M. Hence, with $\iota : A/Q \hookrightarrow R$, $\chi := \tau\iota\sigma$ does the job for Version 4. □

We remark that, trivially, the assumption of non-nilpotency of u is a necessary condition for the assertion of Version 5: having a prime ideal Q in A with $u \notin Q$, one deduces $u^n \notin Q$ for all $n \geq 1$, in particular $u^n \neq 0$.

Let us also record here a "more structural" formulation of Version 5:

Corollary *P is a radical ideal of a commutative ring R if and only if P is the intersection of a set of prime ideals in R.*

PROOF. "if" is trivial (see above), and for "only if" apply Version 5 to $A := R/P$. □

The Corollary represents an infinitary version of one of Noether's [15] famous decomposition theorems for ideals, only the existence part of which we record here.

Noether's Irreducible Decomposition Theorem *Every ideal P in a commutative Noetherian ring R is the intersection of finitely many irreducible ideals.*

Here we call an ideal $Q \trianglelefteq R$ *irreducible* if $Q \neq R$ and if $Q = I \cap J$ for ideals $I, J \trianglelefteq R$ is possible only if $Q = I$ or $Q = J$; briefly: if Q can be presented as the intersection of finitely many ideals only if Q is one of the intersecting ideals (observe the case of an empty intersection!).

PROOF. Assume that the set of ideals in R which fail to be finite intersections of irreducible ideals is not empty. Since R is Noetherian, there is a maximal (w.r.t. $\subseteq$) element in this set, call it Q. Neither can we have $Q = R$ (since R is an empty intersection of irreducible ideals), nor can Q be irreducible. Hence $Q = I \cap J$ with $Q \neq I \trianglelefteq R$ and $Q \neq J \trianglelefteq R$. But by the maximality of Q, both

I and J are finite intersections of ideals, whence also Q is one: contradiction. □

Noether's *Irreducible Decomposition Theorem* is one of the two pillars needed to prove her *Primary Decomposition Theorem* (the other being the fact that irreducible ideals are primary); Lasker [11] proved it in 1904 for polynomial rings and Noether [15] in 1920 in full generality. We also note that Noether discusses the relationships of her ideal-theoretical results with the *Nullstellensatz* especially in Section 10 of her paper [15]. It is evident that the analysis of the algebraic fundamentals of the *Nullstellensatz* was one of the main goals of her paper.

1.3 Birkhoff

In his famous 1944 paper [2] Garrett Birkhoff recognized that, by applying Noether's lattice-theoretic arguments for ideals more generally to congruence relations of general algebras, one is able to establish key elements of her theory in this much more general context. Here, by a *general algebra* A (which he referred to as just "algebra" or "abstract algebra", but *not* "universal algebra" – a term that he reserves for the *study* of such algebras) is just a set with a small collection of *finitary* operations $\omega : A^{n_\omega} \to A$. Any family P_i $(i \in I)$ of congruence relations on A (= equivalence relations on A which are subalgebras of $A \times A$) with

$$\bigcap_{i\in I} P_i = \Delta_A$$

(where the identity relation Δ_A is the bottom element in the lattice $\mathsf{Con}\,A$ of congruences on A) corresponds to an embedding

$$s : A \to \prod_{i\in I} S_i \qquad (**)$$

with $p_i s[A] = S_i$ for all $i \in I$ (where p_i is a product projection): having P_i $(i \in I)$ consider $S_i = A/P_i$ and s with $p_i s = \pi_i : A \to S_i$ the canonical projection; having S_i and s, take P_i to be the congruence relation induced by the homomorphism $p_i s$. We call $(**)$ a *subdirect representation* of A. (Birkhoff calls $(**)$ a "subdirect union", but note that during the acceptance process the original title of his paper got changed from "Subdirect products in universal algebra" to "Subdirect unions in universal algebra", a term favoured by McCoy whom Birkhoff mentions repeatedly.) Of special interest are those algebras A for which any subdirect representation $(**)$ is trivial, in the sense that $p_i s : A \xrightarrow{\sim} S_i$ is an isomorphism for at least one index i; they are called *subdirectly irreducible (sdi)*. We formulate his main result, commonly known as *Subdirect Representation Theorem*, here as:

HNB Theorem (Version 6) *Every general algebra A has a subdirect representation $(**)$ with all algebras S_i subdirectly irreducible.*

Proof (Sketch). For every pair (x, y) of distinct elements in A one uses Zorn's Lemma to find a congruence $P_{x,y}$ on A which is maximal with the property that $(x, y) \notin P_{x,y}$ (finiteness of the operations is crucial here). Then, trivially,

$$\bigcap_{x \neq y} P_{x,y} = \Delta_A,$$

and one can invoke the machinery outlined earlier and, using maximality of $P_{x,y}$, show that $S_{x,y} = A/P_{x,y}$ is subdirectly irreducible. In fact, $S_{x,y}$ *has a least non-identical congruence relation, a property which characterizes subdirect irreducibility.* □

We shall repeat this argument in greater detail and generality in Section 3 where we give Version 7 of the HNB-Theorem. Here we discuss only the question to which extent Version 6 generalizes Version 5, if any. To this end first we record a proposition mentioned by Birkhoff with credit to McCoy:

Proposition (Birkhoff [2] and McCoy [13]) *A subdirectly irreducible commutative ring without non-zero nilpotent elements is a field (and conversely).*

Let us recall now that Version 5 produces for a commutative ring A without nilpotent elements a subdirect representation

$$A \rightarrowtail \prod_{i \in I} D_i$$

with all $D_i = A/Q_i$ integral domains, but we cannot be sure about their subdirect irreducibility. On the other hand, Version 6 produces the subdirect presentation $(**)$, but the sdi rings S_i occurring in there may fail to be integral domains, unless they have no nilpotent elements. Hence, strictly speaking, Version 5 and 6 (restricted to commutative rings) are logically incomparable.

Nevertheless, we regard Birkhoff's Theorem as a variant of the *Nullstellensatz*, noting also that Version 5 does not necessarily precede Version 6. In fact, Version 5 (the proof of which uses Zorn's Lemma [21]) wasn't established yet in this form when Noether wrote her paper. Birkhoff's first aim was to generalize Noether's *Irreducible Decomposition Theorem* to general algebras whose congruence lattices satisfy the ascending chain condition. Then, with the aim of removing this restriction, he formulated his principal theorem as in Version 6, immediately discussing its ring-theoretic repercussions.

The categorical Versions 7 and 8 as presented in Section 3 are straight generalizations of Version 6 but also imply Version 5.

2 Subdirectly irreducible and simple objects

2.1 Subdirect irreducibility

Monomorphisms $f : A \to \prod_{i \in I} S_i$ in a category $\mathcal{A}$ correspond bijectively to jointly monic families $(f_i : A \to S_i)_{i \in I}$ of morphisms in $\mathcal{A}$. Our discussion in 1.3

suggests the following product-free definition of subdirectly irreducible object in a category $\mathcal{A}$.

Definition.

(1) An object A is (*finitely*) *subdirectly irreducible* if every small (finite) jointly monic family $(f_i : A \to S_i)_{i\in I}$ of morphisms with common domain A contains at least one monomorphism $f_{i_0} : A \to S_{i_0}$. We often write *sdi* (*fsdi*) for (finitely) subdirectly irreducible.

(2) A (*finite*) *subdirect representation* of an object A is a small (finite) jointly monic family $(f_i : A \to S_i)_{i\in I}$ of strong epimorphisms f_i in $\mathcal{A}$. We say that A is *subdirectly represented by the objects* S_i $(i \in I)$ in this case. The representation is *trivial* if f_{i_0} is an isomorphism for some $i_0 \in I$.

Remark. (1) Every sdi object is trivially fsdi. The empty family contains no (mono)morphism, hence cannot be jointly monic when A is fsdi. Hence, for every fsdi object A there exists a pair of morphisms $a, b : P \to A$ with $a \neq b$. Any such pair is called a *doubleton* of A. Objects without doubletons are called *preterminal*. In other words then, fsdi *objects, and a fortiori,* sdi *objects are never preterminal*.

(2) If $\mathcal{A}$ has (finite) products, then A is sdi (fsdi) if and only if for every monomorphism $f : A \to \prod_{i\in I} S_i$ into a (finite) product, $p_{i_0} f$ is monic for at least one product projection p_{i_0}.

(3) If $\mathcal{A}$ has (strong epi, mono)-factorizations, then A is sdi (fsdi) if and only if every (finite) subdirect representation of A is trivial.

(4) From (2) and (3) we see that Definition (1) captures Birkhoff's original notion of sdi algebra (see 1.3). Our notion of fsdi object is equivalent to Birkhoff's notion of *weakly irreducible* algebra when $\mathcal{A}$ is a category of general algebras.

Let us first clarify the precise relationship between the two notions introduced by Definition (1). Recall that an (*up-*)*directed poset* is a partially ordered set $(X, \leq)$ with upper bounds to its finite subsets; hence $X \neq \emptyset$, and for all $x, y \in X$ there exists $z \in Z$ with $z \geq x$, $z \geq y$. A (*down-*)*directed* (or inverse) diagram in $\mathcal{A}$ is a functor $D : I^{\mathrm{op}} \to \mathcal{A}$ where I is an up-directed poset (considered as a category). We call the object A in $\mathcal{A}$ *directedly subdirectly irreducible* (*dsdi*) if every monic cone with vertex A over a down-directed diagram in $\mathcal{A}$ contains at least one monomorphism.

Proposition. *In a category with finite products, an object is sdi if and only if it is fsdi and dsdi.*

Proof. Sdi trivially implies fsdi and dsdi. Conversely, for a jointly monic family $(f_i : A \to S_i)_{i\in I}$, consider the induced cone $(f_F : A \to S_F)_{F\subseteq I \text{ finite}}$ over the down-directed diagram given by the finite products

$$S_F = \prod_{i\in F} S_i$$

with "bounding" morphisms $p_F^G : S_G \to S_F$ for $F \subseteq G$ given by projections. This cone is monic, hence it contains a monomorphism f_{F_0} since A is dsdi, and then there must exist $i_0 \in F_0$ with f_{i_0} monic since A is fsdi. □

Remark. Let us call an object A *Artinian* if for every cone of strong epimorphisms with domain A over a down-directed diagram $D : I^{\mathrm{op}} \to \mathcal{A}$ there is $i_0 \in I$ with $D_j \to D_{i_0}$ iso for all $j \geq i_0$. In a category with (strong epi, mono)-factorizations, every Artinian object is dsdi; one concludes that if every object is Artinian, then every fsdi object is sdi.

2.2 Doubletons and congruences

Definition. A pair $(p_1, p_2 : K \to A)$ of morphisms in $\mathcal{A}$ is a *congruence relation* (or simply a *congruence*) on A if for any pair $(q_1, q_2 : L \to A)$ of morphisms with $\mathsf{coker}(p_1, p_2) \subseteq \mathsf{coker}(q_1, q_2)$ there is a uniquely determined morphism $t : L \to K$ with $p_1 t = q_1$, $p_2 t = q_2$; here we write

$$\mathsf{coker}(p_1, p_2) = \{h \mid hp_1 = hp_2\}.$$

Remarks. (1) Every congruence (p_1, p_2) is in fact a *relation*, in the sense that (p_1, p_2) is *jointly monic*: whenever $p_1 t = p_1 s$, $p_2 t = p_2 s$, then $t = s$.

(2) The kernelpair of any morphism is a congruence. Conversely, a congruence is the kernelpair of its coequalizer, provided that the coequalizer exists.

(3) We can set up a category $\mathsf{Con}\, A$ (in fact, a preordered class) whose objects are the congruences on A, and whose morphisms $t : (q_1, q_2) \to (p_1, p_2)$ are as in the Definition above. If the category $\mathcal{A}$ has kernelpairs of regular epimorphisms, then there is, for every object A, a full embedding

$$A \backslash \mathsf{RegEpi} \to \mathsf{Con}\, A,$$

where $A \backslash \mathsf{RegEpi}$ is the comma category of regular epimorphisms with domain A; this is an equivalence of categories if $\mathcal{A}$ has coequalizers of congruences. Hence, congruences (also called *effective equivalence relations*) provide an alternative description of regular epimorphisms, under very mild (co)completeness hypotheses.

(4) $\mathsf{Con}\, A$ has a least element $\Delta_A = (1_A, 1_A)$, and a largest element $\nabla_A = (A \times A \rightrightarrows A)$ whenever the product exists. For any congruence relation $(q_1, q_2 : L \to A)$, the induced morphism $L \to A \times A$ is in fact a regular monomorphism of $\mathcal{A}$.

(5) A morphism $f : A \to B$ is monic if and only if its kernelpair (p_1, p_2) exists and satisfies $p_1 = p_2$; furthermore, any congruence relation (p_1, p_2) on A with $p_1 = p_2$ is isomorphic to Δ_A and is called *discrete*. There is a corresponding fact on ∇_A, provided that $A \times A$ exists: call f *constant* if $fx = fy$ for all $x, y : X \to A$; now, f is constant if and only if its kernelpair exists and is isomorphic to ∇_A.

Remarks (3) and (4) prove:

Corollary. *If the needed products exist, there is a full embedding*

$$\mathsf{Con}\,A \to \mathsf{RegMono}/A \times A.$$

Hence, if $\mathcal{A}$ is wellpowered w.r.t regular monomorphisms and has also kernelpairs of regular monomorphisms, then $\mathcal{A}$ is also cowellpowered w.r.t. regular epimorphisms. □

Recall that a doubleton of an object A is any pair of distinct morphisms with codomain A and common domain. Congruences on A which are doubletons are those which are not isomorphic to $\Delta_A = (1_A, 1_A)$, i.e. which are *non-discrete*.

Definition.

(1) A *generic doubleton* of an object A is a doubleton $a, b : P \to A$ with the property that every morphism $f : A \to B$ with $fa \neq fb$ must be monic.

(2) A is called *simple* if A has a doubleton, and if all of its doubletons are generic.

Remarks. (1) One easily sees that A *is simple if and only if* A *is not preterminal and every morphism with domain* A *is constant or monic.*

(2) Clearly, every simple object is sdi. From the following Proposition it will become clear that the converse statement rarely holds true (see also 2.3).

(3) Let $\mathcal{A}//A$ be the category whose objects are pairs $a, b : P \to A$ of (not necessarily distinct) morphisms in $\mathcal{A}$, and whose morphisms $t : (a, b) \to (c, d)$ are $\mathcal{A}$ morphisms satisfying $ct = a$, $dt = b$. Then the category $\mathsf{Con}\,A$ of congruences on A is a full subcategory of $\mathcal{A}//A$. A *congruence generated by* $a, b : P \to A$ is, by definition, a reflection of (a, b) into $\mathsf{Con}\,A$, and we denote it by $(\overline{a}, \overline{b})$, in case of existence. Hence, it is a congruence on A which comes with a morphism $t : (a, b) \to (\overline{a}, \overline{b})$ such that for any other morphism $s : (a, b) \to (p, q)$ with a congruence (p, q) one has $(\overline{a}, \overline{b}) \leq (p, q)$. One can construct $(\overline{a}, \overline{b})$ as the kernelpair of the coequalizer of (a, b), if these (co)limits exist in $\mathcal{A}$.

(4) The preordered class $\mathsf{Con}\,A$ has small-indexed (finite) infima (= products in the category $\mathsf{Con}\,A$) if the category $\mathcal{A}$ has (finite) products, kernelpairs and coequalizers of congruences: for each of an I-indexed set of congruences one forms its coequalizer $f_i : A \to B_i$ and then the kernelpair of the induced morphism $f : A \to \prod_{i \in I} B_i$. As a consequence we see that $\mathsf{Con}\,A$ has even large-indexed infima if $\mathcal{A}$ is cowellpowered w.r.t. regular epimorphisms and has small-indexed products, kernelpairs and coequalizers of congruences.

Proposition. *Let A be an object of a category $\mathcal{A}$ with kernelpairs, coequalizers of congruences and a terminal object. Then:*

(1) *A is simple if and only if $\mathsf{Con}\,A$ is equivalent to the two-element chain.*

(2) *A is fsdi if and only if* $\mathsf{Con}\,A\backslash\{\Delta_A\}$ *is closed under finite infima in* $\mathsf{Con}\,A$*, provided that* $\mathcal{A}$ *has finite products.*

(3) *For the statements*

i. $\mathsf{Con}\,A\backslash\{\Delta_A\}$ *has a least element,*

ii. *A has a generic doubleton,*

iii. *A is sdi,*

one always has i $\Rightarrow$ ii $\Rightarrow$ iii*, while* ii $\Rightarrow$ i *holds true if* $\mathcal{A}$ *has coequalizers, and* iii $\Rightarrow$ i *if* $\mathcal{A}$ *has (small-indexed) products and is cowellpowered w.r.t. regular epimorphisms.*

We note that the implication ii $\Rightarrow$ iii does not even need the general hypotheses of the Proposition.

Proof. (1) Note that $A \times A$ exists, as the kernelpair of $A \to 1$. Now, if $\mathsf{Con}\,A \cong 2$, A has a doubleton, namely ∇_A. Furthermore, the kernelpair of any morphism $f : A \to B$ which is not monic must be isomorphic to ∇_A, so that f must be constant. Hence, A is simple by Remark (1). Conversely, assuming A to be simple, the coequalizer $f : A \to B$ of any non-discrete congruence (p, q) on A cannot be monic, hence must be constant. Its kernelpair (p, q) is therefore isomorphic to ∇_A.

(2) According to Remark (4), $\mathsf{Con}\,A$ has finite infima. If $\mathsf{Con}\,A\backslash\{\Delta_A\}$ is closed under them, we consider a finite jointly monic family $(f_i : A \to S_i)_{i\in I}$. Then, Δ_A is the infimum of the kernelpairs of the morphisms f_i, at least one of which must be Δ_A as well, by the assumed closure property. Hence, the corresponding morphism f_{i_0} must be monic. Conversely, given a finite family of congruences on A whose infimum is Δ_A, their coequalizers form a jointly monic family to which one applies the hypothesis.

(3) i $\Rightarrow$ ii: Any least non-discrete congruence (p_1, p_2) on A serves as a generic doubleton. Indeed, the kernelpair (q_1, q_2) of any morphism $f : A \to B$ with $fp_1 \neq fp_2$ must be discrete since $q_1 \neq q_2$ would imply $(p_1, p_2) \leq (q_1, q_2)$ and then $fp_1 = fp_2$; hence, f must be monic.

ii $\Rightarrow$ iii: Consider a generic doubleton $a, b : P \to A$ and any monic family $(f_i : A \to S_i)_{i\in I}$. Then $f_{i_0}a \neq f_{i_0}b$ for at least one $i_0 \in I$ and since (a, b) is generic, f_{i_0} must be monic.

ii $\Rightarrow$ i: Starting with a generic doubleton $a, b : P \to A$, one shows that its generated congruence $(\overline{a}, \overline{b})$ is a least element in $\mathsf{Con}\,A\backslash\{\Delta_A\}$. Certainly, $\overline{a} \neq \overline{b}$; furthermore, the coequalizer f of any other non-discrete congruence (q_1, q_2) on A must satisfy $fa = fb$ since otherwise f would be monic and (q_1, q_2) discrete. Hence $(\overline{a}, \overline{b}) \leq (q_1, q_2)$.

iii $\Rightarrow$ i: We may assume $\mathsf{Con}\,A$ to be small. A least element in $\mathsf{Con}\,A\backslash\{\Delta_A\}$ may be constructed as the infimum of all non-discrete congruences on A by Remark (4). Indeed, if this infimum were discrete, then their coequalizers would form a jointly monic family, hence at least one of these would have to be monic, which is impossible. □

Remarks. Let $\mathcal{B}$ be a reflective subcategory of a category $\mathcal{A}$. For any object B of $\mathcal{B}$ one has:

(1) If B is sdi in $\mathcal{A}$, then B is also sdi in $\mathcal{B}$.

(2) If B has a generic doubleton in $\mathcal{A}$, then B has also one in $\mathcal{B}$.

(3) In (1) the converse proposition holds true in *each* of the following cases: 1. all reflexion morphisms are monic, 2. the reflector preserves small jointly monic families, 3. the reflector has a left adjoint.

(4) If strong epimorphisms are regular in $\mathcal{A}$, then for any subdirect representation $(B \xrightarrow{e_i} A)$ in $\mathcal{A}$ one obtains a subdirect representation of B in $\mathcal{B}$ by composing each e_i with a reflexion morphism.

2.3 Some examples

(1) *Sets.* The fsdi objects in Set are precisely the two-element sets, and these are also sdi, even simple, as an easy inspection of the congruence lattices reveals.

The following three examples are mentioned already by Birkhoff [2].

(2) *Distributive lattices and Boolean algebras.* The two-element chain is (up to isomorphism) the only fsdi object in both DLat and Boole, hence sdi and even simple. In fact, any element of an fsdi object in DLat must be 0 or 1 since the meet of the congruences induced by

$$a \wedge - : L \to L, \quad a \vee - : L \to L$$

is trivial, so that one of the two congruences must be trivial. For L in Boole one considers the congruences

$$x \sim y \iff (x \vee y) \wedge (x \wedge y)' \leq a, \qquad x \sim' y \iff (x \vee y) \wedge (x \wedge y)' \leq a'$$

instead, with $(\quad)'$ denoting complementation.

(3) *K-Vector spaces.* In Vec_K the ground field K is (up to isomorphism) the only fsdi object, in fact sdi and simple. Indeed, for any vector space, $\mathsf{Con}\,V \cong \mathsf{Sub}\,V$, and for V to admit non-zero subspaces with trivial intersection V should have dimension at least 2.

(4) *Abelian groups.* Here we find our first example of an fsdi object which is not sdi ($\mathbb{Z}$, the integers), and of an sdi object which is not simple ($\mathbb{Z}_4$, the four-element cyclic group). The fsdi objects of AbGrp include the additive subgroups of the rationals $\mathbb{Q}$; these are the so-called rank 1 torsion-free Abelian groups which are of fundamental importance in the study of torsion-free Abelian groups. However, L. Márki called my attention to the obviously erroneous claim of Lemma 3 in [2] that also the subgroups of $\mathbb{Q}/\mathbb{Z}$ be fsdi; indeed, as a torsion group $\mathbb{Q}/\mathbb{Z}$ is the coproduct of its primary components and thus decomposes

into finite direct products in many non-trivial ways. The sdi objects are precisely the cyclic groups $\mathbb{Z}_{p^n}$ and the quasi-cyclic multiplicative subgroups

$$\mathbb{Z}_{p^\infty} = \{z \in \mathbb{C} \mid \exists k \geq 0 : z^{p^k} = 1\}$$

of the complex numbers (also presentable as the additive subgroups $\mathbb{Z}(p^\infty) = \left\{ \frac{m}{p^k} \middle| m, k \in \mathbb{Z}, k \geq 0 \right\} \leq \mathbb{Q}/\mathbb{Z}$), for some prime p, and $n \geq 1$. Of these, only the cyclic groups $\mathbf{Z}_p$ are simple.

(5) *Commutative rings.* We already saw in Proposition 1.3 that the only sdi objects in CRng without nilpotent elements are the fields. A complete characterization of sdi objects R in CRng was given by McCoy (see [12]), as follows. There are two cases depending on whether the set D of zero divisors in R is the whole ring or not. If $D \neq R$, then R is sdi if and only if

1. D is a maximal ideal,
2. $J = \{x \in R \mid Dx = 0\}$ is a non-zero principal ideal (j),
3. $D = \{y \in R \mid Jy = 0\}$,
4. for $z \in D\backslash J$ there is $w \in D\backslash J$ with $zw = j$.

If $D = R$, then R is sdi if and only if

1. there is a prime p such that for every $x \in R$ with $Rx = 0$ there is some k with $p^k x = 0$,
2. $J = \{x \in R \mid Rx = 0, px = 0\}$ is a non-zero principal ideal (j),
3. for every $z \in R$ with $zR \neq 0$ there is $w \in R$ with $zw = j$.

Remarks.

(1) A subdirect representation of a set X by sdi objects is provided by the X-indexed family of maps

$$\delta_x : X \to 2 = \{0, 1\} \quad (x \in X)$$

with $\delta_x(y) = 0$ if and only if $y = x$, provided that $|X| \geq 2$, and by the empty family otherwise.

(2) If $(f_i : X \to S_i)_{i \in I}$ is a subdirect representation by sdi objects in Set and X is finite with n elements, then I has at least $n - 1$ elements.

(3) The projections

$$\pi_p : \mathbb{Z} \to \mathbb{Z}_p = \mathbb{Z}/(p) \quad (p \text{ prime})$$

provide a subdirect representation in CRng by sdi objects, and so does the family

$$\widetilde{\pi}_p : \mathbb{Z} \to \mathbb{Z}_{p^2} = \mathbb{Z}/(p^2) \quad (p \text{ prime})$$

Note in particular that such representations are not unique.

2.4 Comma categories

It is easy to characterize simple and (f)sdi objects in a comma category $\mathcal{A}/T$, based on Proposition 2.2 and the following trivial Lemma.

Lemma. *Let $a : A \to T$ be a morphism in a category $\mathcal{A}$ with kernelpairs and their coequalizers. Then the preordered class $\mathsf{Con}(A, a)$ of congruences of the object (A, a) in $\mathcal{A}/T$ is isomorphic to the preordered subclass of $\mathsf{Con}\, A$ given by those congruences of A below (or isomorphic to) the kernelpair of a.*

Corollary. *Let $a : A \to T$ be a morphism in a complete and weakly cowellpowered category $\mathcal{A}$ with coequalizers, and let $p_1, p_2 : K \to A$ be its kernelpair. Then*

(1) *(A, a) is simple in $\mathcal{A}/T$ if and only if a is not monic and the only congruence of A below (p_1, p_2) is Δ_A.*

(2) *(A, a) is fsdi in $\mathcal{A}/T$ if and only if a is not monic and the intersection of any two non-discrete congruences of A below (p_1, p_2) is non-discrete.*

(3) *(A, a) is sdi in $\mathcal{A}/T$ if and only if there is a least non-discrete congruence of A below (p_1, p_2).* □

Performing the same task in

$$T\backslash\mathcal{A} = (\mathcal{A}^{\mathrm{op}}/T)^{\mathrm{op}}$$

is even easier: given an object $(A, a : T \to A)$ in $T\backslash\mathcal{A}$, every regular epimorphism $f : A \to B$ in $\mathcal{A}$ gives a regular epimorphism $f : (A, a) \to (B, b)$ in $T\backslash\mathcal{A}$, and conversely, provided that the kernelpair of f exists in $\mathcal{A}$ or that $\mathcal{A}$ has finite coproducts. Hence, simple and (f)sdi objects (A, a) in $T\backslash\mathcal{A}$ are given precisely by simple and (f)sdi objects A in $\mathcal{A}$, respectively, under the (co)completeness and smallness (i.e., cowellpoweredness) assumption of Proposition 2.2.

Examples. (1) (A, a) is sdi (even simple) in Set/T if and only if the map $a : A \to T$ has exactly one fibre which is sdi in Set while all other fibres are preterminal; that is: if there is $t_0 \in T$ such that $|a^{-1}t_0| = 2$ and $|a^{-1}t| \leq 1$ for all $t \in T$, $t \neq t_0$.

(2) (A, a) is sdi (even simple) in Vec_K/T if and only if the K-linear map $a : A \to T$ has a kernel of dimension 1.

(3) (A, a) is sdi (simple) in Grp/T if the kernel of the group homomorphism $a : A \to T$ is sdi (simple, respectively) in Grp.

(4) For $\mathcal{A}$ the category of commutative unital rings, $R\backslash\mathcal{A}$ is the category of commutative unital R-algebras. Hence, a commutative unital R-algebra is simple (fsdi, sdi) if and only if it has the respective property as a commutative unital ring.

Remark. For any category $\mathcal{A}$ and $T \in \mathsf{ob}\,\mathcal{A}$ one has:

(1) No object (A, a) of $\mathcal{A}/T$ can have a doubleton if a is monic in $\mathcal{A}$.

(2) If $u, v : R \to A$ is a generic doubleton of A in $\mathcal{A}$, then $u, v : (P, au) \to (A, a)$ is a generic doubleton of (A, a) in $\mathcal{A}/T$ for any a with $au = av$.

2.5 The dual notions: fat and bare points

An object P of a category is a *bare* (*fat*) *point of* $\mathcal{A}$ if P is simple (subdirectly irreducible) in $\mathcal{A}^{\mathrm{op}}$; any monomorphism $P \to A$ in $\mathcal{A}$ is called a *bare* (*fat*; respectively) *point of* A in $\mathcal{A}$. Bare and fat points are characterized by the dual of Proposition 2.2. Denoting by

$$\mathsf{Reg\,Sub}\,A = \mathsf{RegMono}/A$$

the regular subobjects of A in $\mathcal{A}$ we observe that $\mathsf{Reg\,Sub}\,A$ represents the congruences of A in $\mathcal{A}^{\mathrm{op}}$ when $\mathcal{A}$ has cokernelpairs and their equalizers. Hence:

Corollary. *In a cocomplete category $\mathcal{A}$ with equalizers an object*

(1) *P is a bare point of $\mathcal{A}$ if and only if $\mathsf{Reg\,Sub}\,P$ is equivalent to the two-element chain, and*

(2) *P is a fat point of $\mathcal{A}$ if and only if $\mathsf{Reg\,Sub}\,A \backslash \{1_P\}$ has a largest element; equivalently, if there are distinct morphisms $u, v : P \to Q$ such that any morphism $f : A \to P$ with $uf \neq vf$ is epic.*

Examples. (1) In Set the only fat points are the singleton sets, and these are bare.

(2) In Grp, $\mathsf{Reg\,Sub}\,A = \mathsf{Sub}\,A$ since all monomorphisms are regular: this is a nice exercise in group theory. For a group A to admit a largest proper subgroup, A must be cyclic of prime order. Hence, the fat points of Grp are precisely the groups $\mathbb{Z}_p$, p prime, and these are actually bare.

We see now that *the infinite cyclic group $\mathbb{Z}$ does not have a subdirect representation by sdi objects in* $\mathsf{Grp}^{\mathrm{op}}$. In fact, such a representation would have to be given by a family of fat points of $\mathbb{Z}$. Since there is no monomorphism $\mathbb{Z}_p \to \mathbb{Z}$ this family would have to be empty; however, the empty family with codomain $\mathbb{Z}$ is not epic in Grp.

These statements remain unchanged if we trade Grp for AbGrp.

2.6 Presheaves

We characterize the sdi objects of $\mathsf{Set}^{\mathcal{C}^{\mathrm{op}}}$ amongst the representable functors, for any small category $\mathcal{C}$, as follows:

Proposition. *$\mathcal{C}(-, C) : \mathcal{C}^{op} \to \mathsf{Set}$ is sdi in $\mathsf{Set}^{\mathcal{C}^{op}}$ if and only if there exists a doubleton $u, v : P \to C$ in $\mathcal{C}$ with the property that for any other doubleton*

$g, h : D \to C$ in $\mathcal{C}$ one may find morphisms $d_0, d_1, \ldots, d_n : P \to D$, $n \geq 0$, such that

$$\begin{aligned} u \in & \{gd_0, hd_0\}, \\ & \{gd_0, hd_0\} \cap \{gd_1, hd_1\} \neq \emptyset, \\ & \{gd_1, hd_1\} \cap \{gd_2, hd_2\} \neq \emptyset, \\ & \quad \vdots \\ & \{gd_{n-1}, hd_{n-1}\} \cap \{gd_n, hd_n\} \neq \emptyset, \\ v \in & \{gd_n, hd_n\}. \end{aligned}$$

PROOF. Let $\mathcal{C}(-, C)$ be sdi in $\mathsf{Set}^{\mathcal{C}^{\mathrm{op}}}$. According to Proposition 2.2 there exists a generic doubleton $\sigma, \tau : H \to \mathcal{C}(-, C)$, and since he representables form a generating class, there is $\pi : \mathcal{C}(-, P) \to H$ with $\pi\sigma \neq \pi\tau$; by the Yoneda Lemma, we can write $\pi\sigma = \mathcal{C}(-, u)$, $\pi\tau = \mathcal{C}(-v)$ with $u, v : P \to C$ in $\mathcal{C}$. Consider any other doubleton $g, h : D \to C$ in $\mathcal{C}$ and form the coequalizer diagram

$$\mathcal{C}(-, D) \overset{\mathcal{C}(-,g)}{\underset{\mathcal{C}(-,h)}{\rightrightarrows}} \mathcal{C}(-, C) \overset{\alpha}{\to} F.$$

Since α is not monic, for the generic doubleton (σ, τ) we have $\alpha\sigma = \alpha\tau$, hence $\alpha_P \mathcal{C}(P, u) = \alpha_P \mathcal{C}(P, v)$ and then $\alpha_P(u) = \alpha_P(v)$. In other words, u and v are equivalent under the equivalence relation generated by

$$\{(\mathcal{C}(P, g)(d), \mathcal{C}(P, h)(d)) \mid d \in \mathcal{C}(P, C)\}.$$

What this means has been stated explicitly in the Proposition.

Conversely, let us assume the existence of a doubleton $u, v : P \to C$ in $\mathcal{C}$ with the stated property. It suffices to show that $\mathcal{C}(-, u), \mathcal{C}(-, v) : \mathcal{C}(-, P) \to \mathcal{C}(-, C)$ is generic. Indeed, given any natural transformation $\alpha : \mathcal{C}(-, C) \to F$ with $\alpha\mathcal{C}(-, u) \neq \alpha\mathcal{C}(-, v)$, we show that $\alpha_D : \mathcal{C}(D, C) \to FD$ is monic for every object D, as follows: α corresponds to an element $a \in FC$ with $(Fu)(a) \neq (Fv)(a)$; if we had $g \neq h$ in $\mathcal{C}(D, C)$ with $\alpha_D(g) = \alpha_D(h)$, then $(Fg)(a) = (Fh)(a)$, and therefore

$$\begin{aligned} (Fu)(a) &= (Fd_0)(Fg)(a) = (Fd_0)(Fh)(a) \\ &= (Fd_1)(Fg)(a) = (Fd_1)(Fh)(a) \\ &\quad \vdots \\ &= (Fd_n)(Fg)(a) = (Fd_n)(Fh)(a) \\ &= (Fv)(a). \end{aligned} \qquad \square$$

Remark. $\mathsf{Set}^{\mathcal{C}^{\mathrm{op}}}$ may well possess sdi objects which are not representable. Consider, for example, $\mathcal{C}$ to be discrete, with two objects, i.e., $\mathsf{Set}^{\mathcal{C}^{\mathrm{op}}} = \mathsf{Set} \times \mathsf{Set}$. Since

$$\mathsf{Con}(X, Y) \cong \mathsf{Con}\, X \times \mathsf{Con}\, Y,$$

the object (X, Y) is fsdi if and only if ($|X| = 2$ and $|Y| \leq 1$) or ($|X| \leq 1$ and $|Y| = 2$), and these are also simple and therefore sdi. Hence, there are exactly four isomorphism types of fsdi objects, all of which are simple and therefore sdi, but neither of which is representable.

3 The categorical HNB Theorem

3.1 Noether's Irreducible Decomposition Theorem

It is quite evident that, by mimicking the arguments given for ideals of a commutative ring, we should be able to establish Noether's Irreducible Decomposition Theorem for strong quotients of an object in a category, provided that the object is Noetherian in the following sense:

Definition. An object A in a weakly cowellpowered (= cowellpowered with respect to strong epimorphisms) category $\mathcal{A}$ is *Noetherian* if every non-empty set of strong quotients of A contains a maximal element. Hence, in every non-empty set of objects of $A \backslash \mathsf{StrongEpi}$ (the comma category of strong epimorphisms with domain A) we find a strong epimorphism $e : A \to B$ with the property that every strong epimorphism $f : B \to C$ must be an isomorphism.

Theorem. *Every strong quotient object of a Noetherian object A in a weakly cowellpowered category $\mathcal{A}$ with (strong epi, mono)-factorizations admits a finite subdirect representation by fsdi objects.*

Proof. Suppose that the class of quotient objects of A which fail to admit a finite subdirect representation by fsdi objects is not empty, and let $e : A \to B$ be maximal in that class. Then B cannot be fsdi (since otherwise $1_B : B \to B$ would be a suitable representation) and must therefore admit a non-trivial finite subdirect representation $(f_i : B \to S_i)_{i \in I}$ (see Remark (3) of 2.1). Since none of the f_i is an isomorphism, by maximality of B every S_i admits a finite subdirect representation $(g_{ij} : S_i \to T_{ij})_{j \in J_i}$ by fsdi objects T_{ij}. Then

$$f_i g_{ij} : B \to T_{ij} \quad (i \in I, j \in J_i)$$

is a finite subdirect representation of B by fsdi objects, in contradiction to the choice of B. $\square$

3.2 Finitary objects

In order to establish Birkhoff's *Subdirect Representation Theorem* (see 1.3) categorically we should, in view of Theorem 3.1, work in a weakly cowellpowered category with (strong epi, mono)-factorizations. However, it is clear from Example (2) of 2.5 that this is not sufficient: we need to capture the categorical consequences of the fact that Birkhoff deals with *finitary* algebras in order to establish his theorem.

Definition.

(1) Let the up-directed diagram $D : I \to \mathcal{A}$ be given by the morphisms $f_{ij} : A_i \to A_j$ $(i \leq j \in I)$. A *doubleton of* D is a doubleton $a, b : P \to A_{i_0}$ in $\mathcal{A}$ for some $i_0 \in I$ such that $f_{i_0 j} a \neq f_{i_0 j} b$ for all $j \geq i_0$. A cocone $D \to \Delta B$, given by morphisms $h_i : A_i \to B$ $(i \in I)$, is said to *detect* the doubleton if $h_{i_0} a \neq h_{i_0} b$.

(2) An object P of $\mathcal{A}$ is *finitary* if any doubleton with domain P of an up-directed diagram gets detected by at least one of its cocones.

Remarks. (1) Clearly, if some cocone detects a doubleton of an up-directed diagram, so does its colimit (existence granted).

(2) Any cocone $(h_i)_{i \in I}$ of any up-directed diagram $(f_{ij})_{i \leq j}$ with

$$\mathsf{ker}\, h_i = \bigcup_{j \geq i} \mathsf{ker}\, f_{ij} \qquad (+)$$

for all $i \in I$ (see 2.1) detects all doubletons of the diagram. Here we write $\mathsf{ker}\, f = \{(x, y) | fx = fy\}$.

(3) The colimit $(h_i)_{i \in I}$ of every up-directed diagram $(f_{ij})_{i \leq j}$ in Set satisfies $(+)$.

(4) For any doubleton with domain P of an up-directed system in $\mathcal{A}$ with a colimit, the colimit detects the doubleton if the colimit is preserved by $\mathcal{A}(P, -) : \mathcal{A} \to \mathsf{Set}$.

Since *finitely-presentable* objects are precisely those whose representables preserve directed colimits, we obtain:

Proposition.

(1) *Every finitely-presentable object in a category with directed colimits is finitary.*

(2) *Every category with directed colimits preserved by a given faithful right adjoint functor into some small discrete power of* Set *has a generator (= generating set) consisting of finitely-presentable (hence, of finitary) objects.* □

Remark. If P is a finitary object in the category $\mathcal{A}$ with up-directed colimits, then for every $p : P \to T$ in $\mathcal{A}$, (P, p) is a finitary object in $\mathcal{A}/T$.

3.3 HNB categories and theorems

Definition. $\mathcal{A}$ is a *Hilbert-Noether-Birkhoff* (*HNB*) *category* if

(HNB1) $\mathcal{A}$ has (strong epi, mono)-factorizations,

(HNB2) $\mathcal{A}$ is weakly cowellpowered, (HNB3) $\mathcal{A}$ has a generator consisting of finitary objects.

Remarks.

(1) Folklore theorems show that (HNB2) implies (HNB1) if $\mathcal{A}$ has all small connected colimits. (HNB1) comes also for free if $\mathcal{A}$ is wellpowered and has all small connected limits. Note that in the presence of connected limits (HNB3) implies wellpoweredness of $\mathcal{A}$ if the generator of (HNB3) is strong. (A generator $\mathbf{G}$ in $\mathcal{A}$ is *strong* if a monomorphism $u : B \to A$ is an isomorphism whenever every morphism with domain in $\mathbf{G}$ and codomain A factors through u.) Finally, (HNB3) implies (HNB2) if strong epimorphisms in $\mathcal{A}$ are regular.

(2) Every locally finitely-presentable category is HNB (see Proposition 3.2); in particular every quasi-variety of finitary general algebras and every presheaf category.

(3) The category of topological spaces is HNB; in fact, a singleton space provides a one-object finitely-presentable generator.

(4) If $\mathcal{A}$ has finite products, then $\mathcal{A}$ is HNB if and only if all comma categories $\mathcal{A}/T$ are HNB. In fact, for "only if" observe that in the presence of finite products

— a morphism in $\mathcal{A}/T$ is (strongly) epic if and only if it is (strongly) epic as a morphism in $\mathcal{A}$,

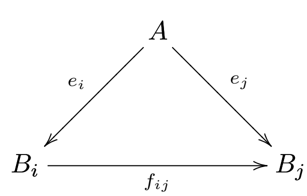
— for a finitary object P in $\mathcal{A}$ every morphism $P \to T$ becomes a finitary object in $\mathcal{A}/T$,

— if $\mathbf{G}$ is a generator of $\mathcal{A}$, then the set of $\mathcal{A}$-morphisms with domain in $\mathbf{G}$ and codomain T becomes a generator of $\mathcal{A}/T$.

Although the notion of HNB category does not entail any existence requirement for limits and colimits, it allows for a proof of a categorical HNB Theorem, as first given in [17] and published in [18]:

HNB Theorem (Version 7) *In an HNB category every object has a subdirect representation by subdirectly irreducible objects.*

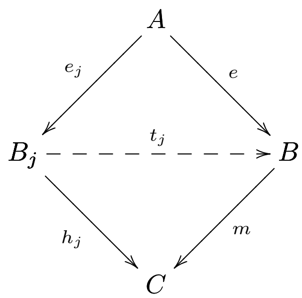
Proof. Let $\mathbf{G}$ be the generator given by (HNB3), and consider any object A. With the help of Zorn's Lemma, for every $\mathbf{G}$-doubleton $x, y : G \to A$ we want to find a maximal element $e_{x,y} : A \to S_{x,y}$ in a representative set $\mathcal{E}_{x,y}$ of all strong epimorphisms e with $ex \neq ey$. Indeed, for any directed diagram



(1)

$(i \leq j \in I)$ in the set $\mathcal{E}_{x,y}$ and any fixed index $i_0 \in I$ $(\neq \emptyset)$, since G is finitary the doubleton $e_{i_0}x, e_{i_0}y$ gets detected by some cocone $h_i : B_i \to C$, hence

$h_{i_0}e_{i_0}x \neq h_{i_0}e_{i_0}y$. Factoring the morphism $h_{i_0}e_{i_0} = h_j e_j$ (StrongEpi, Mono), its strong epi factor e is an upper bound for the directed diagram, which concludes the existence proof for $e_{x,y}$:



(2)

The family $(e_{x,y})_{x,y}$ is, by construction, jointly monic: for any doubleton $a, b : P \to A$ we find $z : G \to P$ with $G \in \mathbf{G}$ and $x := az \neq bz =: y$, hence $e_{x,y}a \neq e_{x,y}b$. Furthermore, the object $S_{x,y}$ is sdi since $e_{x,y}x$, $e_{x,y}y$ is a generic doubleton of $S_{x,y}$. Indeed, the strongly epic factor d of any morphism $f : S_{x,y} \to E$ with $fe_{x,y}x \neq fe_{x,y}y$ satisfies $de_{x,y}x \neq de_{x,y}y$ and must, by maximality of $e_{x,y}$, be an isomorphism, which means that f must be monic. □

Corollary. *An object in an HNB category is sdi if and only if it has a generic doubleton.*

Proof. An object with a generic doubleton in any category is sdi (see Proposition 2.2). Conversely, as just proved, an sdi object in an HNB category has a subdirect representation by objects with generic doubletons. Since this representation must be trivial, the object itself must have a generic doubleton. □

Remarks.

(1) The *subdirect representation rank* $r(A)$ of an object A in an HNB category $\mathcal{A}$ is the least cardinal number $|I|$ occurring in any of the subdirect representations $(f_i : A \to S_i)_{i \in I}$ of A by sdi objects. One shows that if the product $A \times B$ exists in $\mathcal{A}$ and has strong epimorphic projections, then

$$r(A \times B) \leq r(A) + r(B).$$

(2) One concludes from Remark (2) of 2.3 that $r(X) = n - 1$ in Set for every n-element set X, $n \geq 1$.

(3) One easily shows that for a K-vector space V the subdirect representation rank $r(V)$ is its dimension.

3.4 Residually small HNB categories

The categorical impact of the HNB Theorem is especially strong if the category in question has only a set of non-isomorphic sdi objects.

Definition. A category $\mathcal{A}$ is *residually small* if it contains a set of objects such that any sdi object of $\mathcal{A}$ is isomorphic to an object of the set.

The HNB Theorem 3.3 proves the first part of the following Proposition:

Proposition. *Every residually small HNB category has a cogenerator consisting of sdi objects. Conversely, every wellpowered HNB category with a cogenerator is residually small.*

Proof. Given a cogenerator $\mathbf{G}$ of $\mathcal{A}$, for any sdi object S of $\mathcal{A}$ there is a generic doubleton $a, b : P \to S$ and therefore a morphism $f : S \to C$ with $C \in \mathbf{G}$ and $fa \neq fb$, which actually must be monic. Hence, sdi objects are subobjects of objects in $\mathbf{G}$, and there is only a set of them (up to isomorphism) if $\mathcal{A}$ is wellpowered. □

Hence, a residually small HNB category has both a generator and a cogenerator; if one adds to this the existence of *some* limits *or* colimits, then the category has actually all limits and colimits that one can reasonably expect to exist:

Theorem. *For a residually small HNB category $\mathcal{A}$ the following conditions are equivalent:*

- (i) *$\mathcal{A}$ is total (so that its Yoneda embedding has a left adjoint);*
- (ii) *$\mathcal{A}$ is hypercomplete (see [3]);*
- (iii) *$\mathcal{A}$ has all small limits and intersections of large families of monomorphisms;*
- (iv) *$\mathcal{A}$ is cototal;*
- (v) *$\mathcal{A}$ is hypercocomplete;*
- (vi) *$\mathcal{A}$ has all small colimits and cointersections of large families of epimorphisms.*

Proof. For (i) ⇒ (ii) ⇒ (iii) see [3], [9]. (iii) ⇒ (iv) follows from the dual of Day's Theorem [5] (as noted in Corollary 3.6 of [4]) since $\mathcal{A}$ has a cogenerator. (iv) ⇒ (v) ⇒ (vi) ⇒ (i) is dual to (i) ⇒ (ii) ⇒ (iii) ⇒ (iv). □

Corollary. *Let $\mathcal{A}$ be a small-complete, wellpowered and residually-small HNB category, let F be any functor with domain $\mathcal{A}$. Then F has a left adjoint if it preserves all limits, and F has a right adjoint if it preserves all colimits.*

Proof. A total and cototal category is compact and cocompact in the sense of Isbell [8] (see [9]). □

3.5 Examples

(1) *Abelian groups and R-modules.* According to the characterization given in 2.3(4), AbGrp is a residually small HNB category. Alternatively, one may show residual smallness by proving that (the additive group) $\mathbb{Q}/\mathbb{Z}$ is a (single object) cogenerator of AbGrp. More generally, for a unital ring R,

$$\mathsf{hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})$$

can be made into a right R-module by

$$(fr)(s) = f(rs)$$

for all $f : R \to \mathbb{Q}/\mathbb{Z}$, $r \in R$, $s \in R$, and then becomes a cogenerator of the category Mod_R of right R-modules, according to a classical result of module theory. Hence, Mod_R is a residually small HNB category and, as such, total and cototal.

(2) *Groups.* While Grp, like any variety of general algebras, is an HNB category, so that every group admits a subdirect representation by sdi objects, Grp fails to be residually small. In fact, there is even a proper class of simple groups: classically, for every field k the group $PSL(3, k)$ is simple, and there are fields of arbitrary large cardinality. Consequently, Grp fails to have a cogenerator. Grp is cototal but fails to be hypercomplete (see [8], [3]). Hence, in Theorem 3.4 residual smallness is an essential assumption.

Although $\mathsf{Grp}^{\mathrm{op}}$ is residually small (see Example (2) of 2.5), not every group has a subdirect representation in $\mathsf{Grp}^{\mathrm{op}}$. In fact, since Grp fails to have a cogenerator, $\mathsf{Grp}^{\mathrm{op}}$ does not satisfy (HNB3), which is therefore an essential condition for the HNB Theorem (Version 7) to hold. A more refined statement arises from the following example.

(3) *Compact Abelian groups.* The category $\mathcal{A}$ of compact Abelian groups is dually equivalent to AbGrp. Under this equivalence the sphere S^1 in $\mathcal{A}$ corresponds to the infinite cyclic group $\mathbb{Z}$. We saw already in 2.5(2) that $\mathbb{Z}$ does not have a subdirect representation by sdi objects in $\mathsf{Grp}^{\mathrm{op}}$, and neither in $\mathsf{AbGrp}^{\mathrm{op}}$. Hence S^1 has no such representation in $\mathcal{A}$ either, although $\mathcal{A}$ is residually small. This example is due to Wiegandt [20] and shows that the requirement of finitariness for the objects of the generator in (HNB3) is essential for the HNB Theorem to hold. This also shows that $\mathcal{A}$ fails to be locally finitely presentable.

(4) *Residually small varieties of general algebras* (defined by a set Ω of finitary operations and a set of equations). These were first characterized by Taylor [16] who gave a precise upper bound for the maximal size of an sdi algebra S in a residually small variety:

$$|S| \leq 2^k, \quad k = \aleph_0 + |\Omega|.$$

Consequently, in a residually small variety one finds at most 2^{2^k} sdi algebras. Taylor provided examples showing that both bounds are best possible.

It is interesting to note in this context that the number of simple algebras in a variety is not an indicator for residual smallness. There is an example of a variety due to McKenzie which has only two non-isomorphic simple algebras but fails to be residually small.

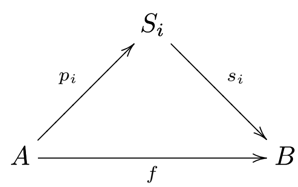
For an extensive table of examples of residually small varieties, see [10].

3.6 Subdirectly irreducible factorization of morphisms

We wish to exhibit more fully the effect of the HNB Theorem (Version 7) for comma categories.

Definition.

(1) A *multifactorization* of a morphism $f : A \to B$ in a category $\mathcal{A}$ is given by a small family of morphisms p_i, s_i $(i \in I)$ such that



(3)

commutes for every $i \in I$. The multifactorization is *monic* if the family $(p_i)_{i\in I}$ is jointly monic, and it is *trivial* if there is some $i_0 \in I$ with p_{i_0} monic.

(2) A morphism f is *sdi* if every monic multifactorization of f is trivial. An *sdi factorization* of f is a monic multifactorization (9) of f with each p_i a strong epimorphism and each s_i an sdi morphism.

Remarks.

(1) A morphism $f : A \to B$ is sdi if and only if (A, f) is an sdi object in $\mathcal{A}/B$. An sdi factorization of f is precisely a subdirect representation of the object (A, f) in $\mathcal{A}/B$ by sdi objects, provided that $\mathcal{A}$ has finite products.

(2) The empty multifactorization of f is monic if and only if f is a monomorphism, i.e., a preterminal object in $\mathcal{A}/B$. Such morphisms have an empty sdi factorization.

(3) A multifactorization of a morphism f may of course be described equivalently by the ordinary factorization of f through the fibred product of the morphisms f_i, or the fibred coproduct of the morphisms p_i.

Since the comma categories of an HNB category with finite products are HNB again (see Remark (4) of 3.3) we can apply the HNB Theorem (Version 7) and obtain:

HNB Theorem (Version 8) *In an HNB category with finite products every morphism has a subdirectly irreducible factorization.*

Of course, existence of finite products granted, Version 8 in fact entails Version 7, since $\mathcal{A}/T \cong \mathcal{A}$ for T terminal in $\mathcal{A}$.

Examples. (1) Sdi morphisms of Set were characterized in Example (1) of 2.4. We can now generalize Remark (1) of 2.3 and establish an sdi factorization of a map $f : X \to T$ by:

$$\begin{array}{ccc} & f[X]+1 & \\ {}^{p_x}\nearrow & & \searrow^{s_x} \\ X & \xrightarrow[f]{} & T \end{array} \tag{4}$$

It is indexed by those $x \in X$ with $|f^{-1}f(x)| \geq 2$, and

$$p_x(y) = \begin{cases} 0 & \text{if} \quad y = x, \\ f(y) & \text{if} \quad y \neq x, \end{cases}$$

$$s_x(t) = \begin{cases} f(x) & \text{if} \quad t = 0, \\ t & \text{if} \quad t \in f[X]; \end{cases}$$

here $1 = \{0\}$, and $0 \notin f[X]$.

It is easy to see that, even in the case $T = 1$ (see Remark 2.3(1)), this construction is not functorial. In fact, it is easy o see that there is no functor that assigns to a set X a product of sdi objects (i.e., a power of 2) which provides a *natural* subdirect representation (so that X is naturally embedded into the product).

(2) Sdi morphisms of Vec_K were described in Example (2) of 4.4 as those K-linear maps whose kernel is isomorphic to the field K. An sdi factorization of $f : V \to W$ in Vec_K is established by fixing a basis $(x_i)_{i \in I}$ of $\mathsf{ker}\, f$ and extending it to a basis $(x_j)_{j \in J}$ of V, with $I \subseteq J$, and then defining a multifactorization of f, as follows:

$$\begin{array}{ccc} & \mathsf{im}\, f \oplus K & \\ {}^{p_i}\nearrow & & \searrow^{s_i} \\ V & \xrightarrow[f]{} & W \end{array} \qquad (i \in I) \tag{5}$$

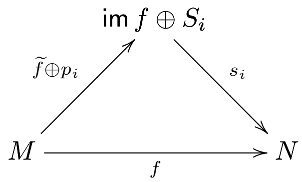
$$p_i(x_j) = \begin{cases} 1 \in K & \text{if } j = i, \\ f(x_i) & \text{if } j \in J, j \neq i, \end{cases}$$

$$s_i\big|_{\mathsf{im}\, f} = \text{ inclusion}, s_i\big|_K = 0. \tag{6}$$

(3) Let $f : M \to N$ be a homomorphism of R-modules, for any ring R, and assume that $\mathsf{ker}\, f$ is a direct summand of M:

$$M = U \oplus \mathsf{ker}\, f$$

for some submodule U of M. Let $(p_i : \mathsf{ker}\, f \to S_i)_{i \in I}$ be a subdirect representation of $\mathsf{ker}\, f$ in Mod_R by sdi objects. Then an sdi factorization of f can be obtained as



(7)

with $\widetilde{f} : U \to \mathsf{im}\, f$ the restriction of f, and with s_i defined as in Example (2).

References

[1] Yuri Bahturin. *Basic structures of modern algebra*. Kluwer Academic Publishers Group, Dordrecht, 1993.

[2] Garrett Birkhoff. Subdirect unions in universal algebra. *Bull. Amer. Math. Soc.*, 50:764–768, 1944.

[3] R. Börger, W. Tholen, M. B. Wischnewsky, and H. Wolff. Compact and hypercomplete categories. *J. Pure Appl. Algebra*, 21:129–144, 1981.

[4] Reinhard Börger and Walter Tholen. Total categories and solid functors. *Canad. J. Math.*, 42:213–229, 1990.

[5] B. J. Day. Further criteria for totality. *Cahiers Topologie Géom. Différentielle Catég.*, 28:77–78, 1987.

[6] David Hilbert. Ueber die vollen Invariantensysteme. *Math. Ann.*, 42:313–373, 1893.

[7] Thomas W. Hungerford. *Algebra*. Holt, Rinehart and Winston, Inc., New York, 1974.

[8] John R. Isbell. Small subcategories and completeness. *Math. Systems Theory*, 2:27–50, 1968.

[9] G. M. Kelly. A survey of totality for enriched and ordinary categories. *Cahiers Topologie Géom. Différentielle Catég.*, 27:109–132, 1986.

[10] E. W. Kiss, L. Márki, P. Pröhle, and W. Tholen. Categorical algebraic properties, a compendium on amalgamation, congruence extension, epimorphisms, residual smallness, and injectivity. *Studia Sci. Math. Hungar.*, 18:79–141, 1983.

[11] E. Lasker. Zur Theorie der Moduln und Ideale. *Math. Ann.*, 60:20–116, 1905.

[12] Neal H. McCoy. Subdirectly irreducible commutative rings. *Duke Math. J.*, 12:381–387, 1945.

[13] Neal H. McCoy. *Rings and ideals*. The Open Court Publishing Company, LaSalle, Ill., 1948.

[14] Matias Menni. An exercise with sufficient cohesion. *Preprint*, 2011. Universidad de La Plata.

[15] Emmy Noether. Idealtheorie in Ringbereichen. *Math. Ann.*, 83:24–66, 1921.

[16] Walter Taylor. Residually small varieties. *Algebra Universalis*, 2:33–53, 1972.

[17] Walter Tholen. Amalgamations in categories. *Seminarbrechte*, 5:121–151, 1979. Fernuniversität Hagen.

[18] Walter Tholen. Injective objects and cogenerating sets. *J. Algebra*, 73:139–155, 1981.

[19] B. L. van der Waerden. *Moderne Algebra II*. Springer, Berlin, 1931.

[20] R. Wiegandt. On subdirect embeddings in categories. *Fund. Math.*, 68:7–12, 1970.

[21] M. Zorn. A remark on method in transfinite algebra. *Bull. Amer. Math. Soc.*, 41:139–155, 1927.