VARIABLE SET THEORY

MICHAEL BARR, COLIN MCLARTY, CHARLES WELLS

1. Introduction

AUTHOR'S NOTE. This paper was originally written in 1986 during a brief time when *scientific American* had published a number of math articles. By the time it was finished and submitted, this window had apparently closed. At any rate this paper was turned down and, as far as we know, no further mathematics appeared. We have recently decided that the paper might have sufficient interest to at least be posted. With light editing, here is the paper as originally written.

The concept of set, originated about 100 years ago by Georg Cantor and formalized early in this century by Zermelo and Fraenkel, von Neumann, Gödel and Bernays, is a fundamental tool in modern mathematics. It clarifies difficult points in advanced mathematics and mathematicians routinely define new concepts and theories using the language of sets. The idea of set developed in this way concerns what we will call classical sets. The classical approach is based on a rigid concept of membership which does not always faithfully reflect practice either inside or outside of mathematics.

This article describes a new, more general notion of set; variable sets, in which the membership varies with respect to a parameter such as time. The theoretical basis for the concept of variable set is topos theory. Topos theory was developed by A. Grothendieck of the Institut des Hautes Études Scientifiques in France (he is now at the University of Grenoble) in the late 1950's for use in algebraic geometry; in 1969, working at Dalhousie University in Halifax, Nova Scotia, F. W. Lawvere, now at SUNY at Buffalo, and M. Tierney, now at Rutgers University, generalized the concept of topos and worked out the way topos theory could be used to explicate the notion of variable set. The axioms for topos theory were later simplified by Lawvere and Tierney and by C. J. Mikkelsen, then a student at the University of Aarhus, Denmark.

We now turn to two examples which will illustrates various aspects of classical set theory.

1.1. THE CIRCLE Our first example concerns a difficulty faced by nineteenth century algebraic geometers. These mathematicians study the curves and surfaces defined as the solutions of polynomial equations in several variables. For example, the unit circle is the set of all solutions in the x-y plane to the equation

$$x^2 + y^2 = 1 \tag{(*)}$$

[©] Michael Barr, Colin McLarty, Charles Wells, . Permission to copy for private use granted.

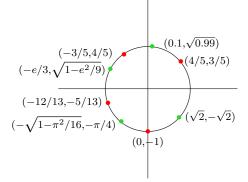


Figure 1: Graph of the unit circle $x^2 + y^2 = 1$ in the x-y plane. Four rational solutions are marked with red dots; four real solutions that are not rational are marked with green dots. There are infinitely many rational solutions, but they do not form a continuum. Every point on the circle shown is a real solution.

The set of solutions of this equation depends on the type of number that one understands x and y to refer to. This set is thus a variable set whose domain of variation consists of types of numbers.

For example, if x and y are understood to denote integers, then there are only four solutions to this equation, namely x = 0, $y = \pm 1$ and $x = \pm 1$, y = 0. If, however, we think of x and y as denoting rational numbers (i.e. fractional numbers such as 1/2 and 7/12, but not such numbers as the square root of 2 or *i*, we then find infinitely many points such as $(\pm 3/5, \pm 4/5)$ on the circle. Even so, we know that the set of rational solutions do not fill out a continuum, as illustrated in Figure 1 above.

The study of rational solutions to algebraic equations is interesting in its own right. Rational solutions to the equation of the circle correspond in a fairly obvious way to triples of integers (a, b, c) for which $a^2 + b^2 = c^2$, called Pythagorean triples since they correspond to right triangles with sides of integer length. Similarly, rational solutions to the analogous equation of higher degree, $x^n + y^n = 1$, correspond to integer solutions to the equation $a^n + b^n = c^n$. The celebrated last "theorem" of Fermat asserts that no non-trivial solutions to this equation exist.

When we allow all real numbers (all numbers with decimal expansions) as type of x and y, the set of solutions is the familiar one dimensional continuum that we ordinarily think of as the circle. Allowing complex numbers (numbers of the form $a + b\sqrt{-1}$) gives a two dimensional continuum as the solution set. There are many more abstract types of numbers and each gives rise to its own circle.

In the classical set theoretic framework, there is thus an integer circle, a rational circle, a real circle, a complex circle and myriads of others; in the setting of variable sets, there is just one circle, the solution of (*). Algebraic geometers originally took this approach but this retarded the subject because the practitioners were not able to put their practice on the same foundation (that is of Cantorian set theory) as the rest of mathematics. Later, when it was put on a firm theoretical basis using the tools of Cantorian set theory, the resulting theory was difficult to understand.

1.2. The set of businesses We will now consider a different sort of example which doesn't arise from a mathematical context and which illustrates several aspects of variable sets that do not appear in the previous example. In addition, it illustrates the fact, that notions of variable set theory arise naturally in common usage. Let B denote the set of businesses incorporated in a given state. From the point of view of classical set theory, this is not well defined since businesses go in and out of existence at different times. As a variable set, it is well defined: it varies over intervals of time. We must note explicitly that the intervals of time used do not include the endpoints of the intervals; such intervals are called open intervals. For a given interval of time, the elements of B are the businesses which exist during that whole interval of time. For example, the elements of B for August 6, 1985 are the. businesses which existed over the whole of the day. A business incorporated on that day or a business which was disbanded on that day is not a member of B for that day. During the minute from 12 noon to 12:01 P.M. (between those two times but, not including those exact times), the elements of B are those businesses which existed during the whole of the minute, from noon to 12:01 but not necessarily including noon or 12:01.

It is important to understand that B is not defined at a given instant in time, but rather for each open interval of time. This is contrary to the way we think of scientific functions such as position, which are conceived of as being defined at each instant of time. We do not have to decide, for example, whether a business which is incorporated at 12 noon is an element of B at 12 noon - the concept of element of B at 12 noon is simply meaningless. Such a business exists over any interval beginning after noon and ending before the business ceases to exist. This mirrors the situation in real life; we don't know what happens at a point in time, but only over short intervals of time. See Figure 2.

There is one feature of this example we would like to emphasize. That is, a business that exists over an interval certainly exists over any subinterval included in the interval. The reader should note two things: first that a business existing over a subinterval need not exist over the containing interval, and second that because of mergers and divestitures two or more businesses in some interval may correspond to a single business on a subinterval [see Figure 2]. A careful description of these and related properties gives the mathematical concept of "sheaf", to which we return briefly below.

2. Axioms for variable set theory

Cantor's notion of set theory, codified into axioms in the early years of this century, takes a rigid notion of membership as a primitive undefined term. When Cantor developed set theory, he thought that a sentence which defined a type of thing also defined the set of those things. However, there cannot be a set consisting of all sets S that satisfy sentence $S \notin S$; the supposition that there be such a set leads directly to Russell's well known paradox (due to Bertrand Russell) that such a set could neither be a member of itself nor

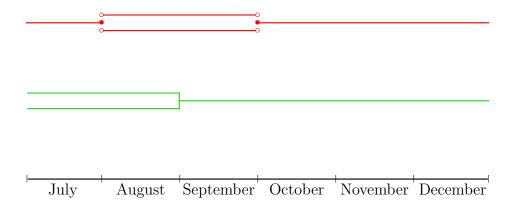


Figure 2: This picture illustrates two variable sets each describing the set of businesses in some very small state. The black line at the bottom shows part of the time line over which they vary. The set colored in red has one element during July, two during August and September and one element after September 30. Over the interval from June to December, the set has no actual elements, even though it is not empty at any time in that interval. The reason is that there is no business which has an existence during that whole interval. This subtlety illustrates the technical distinction between internal and external properties in a variable set. We say that the set has no external elements over that time interval because no business exists for the whole interval, while internally it has elements over the interval because it does over every sufficiently small subinterval. The set shown in green is even more subtle. In the interval July through August there are two businesses, one indicated by the upper branch and the other by the lower branch. But there is a merger in July, so that in the subinterval consisting of August there is only one business. Thus there are four businesses (externally) in the interval July through January: one following the upper branch on both ends, one following the upper branch on the left and the lower branch on the right and so on. Internally, we describe the set by its behavior on all sufficiently small intervals. On any sufficiently small interval, there are one or two businesses. On the other hand, any interval containing the merger in July, however small, will contain a subinterval (on the left) over which there are two businesses and another subinterval (on the right) over which there is only one and on such an interval, we cannot say internally that there is either one or two. We can say internally that there is at least one business and that there are not three or more.

fail to be a member of itself. The way set theorists avoided this difficulty was to allow a set to exist only insofar as its existence is authorized by an axiom. For example, one axiom says there is an empty set and another one says that given sets A and B, there is a set whose elements are those of A and of B. The most commonly used axioms are due to Zermelo and Fraenkel and are complicated. The universe of classical sets consists of all the sets allowed by the Zermelo-Fraenkel axioms. (It can be shown, by the way, that this universe of classical sets is not a set.)

The topos axioms also give rules describing exactly what sets one can build, so that a specific topos can be seen as a universe of variable sets: those allowed by the axioms. Here too, a set (i.e. a variable set) is a set only if it one can show it exists on the basis of the axioms. The set building axioms are not, however, based on the membership relation, but on the notion of mapping and the composition of mappings [see Figure (b)]. Just as in classical set theory the notion of membership is taken as undefined, so here the mappings and their composition are undefined terms. These mappings and composition must satisfy the laws of category theory, a branch of mathematics invented during World War II by S. Eilenberg, now at Columbia University, and S. Mac Lane, now of the University of Chicago.

There are two set-builder axioms. The first says that for each equation in a finite number of unknowns built up using mappings, there is a set of solutions of this equation. The reader may wonder how this is done without elements. We will illustrate this by a simple example.

Suppose S and T are two sets and f and g are two mappings from S to T. We are going to describe the solution set of the equation f = g. We define the set S_0 of solutions to the equation f(x) = g(x) together with a function $e: s_0 \longrightarrow S$. The data S_0 and e must have the following property (which can be shown to characterize them uniquely). First $f \circ e = g \circ e$ and second, given any mapping $h: R \longrightarrow S$, if $f \circ h = g \circ h$ then there is a unique mapping $u: R \longrightarrow S_0$ for which $e \circ u = h$. See Figure 3.

The reader may easily see that in classical set theory, S_0 is the set described in familiar set-builder notation as

$$S_0 = \{ x \in S \mid f(x) = g(x) \}$$

The mapping e is the inclusion mapping that takes an element of S which is a solution of the equation to that element itself. The definition of S_0 we have given makes the set-builder notation work in any topos.

The second set-builder axiom says that given any set, we can form the set of all its subsets. Again, one expects subsets to be defined in terms of elements, but one can give an appropriate mapping theoretic definition. These two axioms are enough to allow one to reason about variable sets using many of the standard techniques for reasoning about classical sets—but not all those techniques. Some of the things one can do, for example, is to form the intersection or union of two subsets of a variable set and form the variable set of all mappings between two variable sets.

On the other hand, in most cases it is not possible to form the complement of a subset of a variable set. In classical sets, if S is a set and T is a subset of S, then the complement

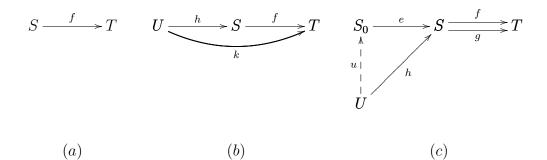


Figure 3: Category theory uses mapping and composition as primitives, rather than the set theoretic membership relation (usually denoted ϵ), and correspondingly uses a notation which vividly displays patterns of mappings. A single arrow (a) says f is a mapping from S to T. A diagram like (b) is usually used to show that h is a function from U to S, k a function from U to T and the composite of f and h, denoted $f \circ h$, is k. The definition of the set of solutions, $E = \{x \in S \mid f(x) = g(x)\}$ is codified in diagram (c): the mappings e, f, g, and h go between sets as indicated, $f \circ e = g \circ e$, and for any h, with $f \circ h = g \circ h$, there is a unique mapping u as shown with $e \circ u = h$.

of T is the subset of S which contains all the elements of S not in T. Consider the set B of businesses discussed previously. For the purposes of this discussion, let us suppose that all businesses are of one of two types: individually owned businesses and corporations. Let C be the subset of corporations. Businesses sometimes convert from individual ownership to corporate status. If a business converts from one status to the other at noon on a given day, then in an interval containing noon it is wrong to say that the business is in the subset C of corporations and it is also wrong to say that it is in the subset of individually owned businesses. It follows from this observation that there is no subset D of B which is disjoint from C and for which the union of C and D is all of B.

Data types

Some of the set-builder constructions possible in toposes can be applied to data types such as INTEGER or CHAR (characters such as a letter or digit) in high level computer languages such as Pascal. For example, the data type INTEGER itself is the set of solutions of the empty set of equations in one variable of type INTEGER. In Pascal one can construct a data type consisting of pairs of integers; they are the solutions of the empty set of equations in two integer variables. Pascal also allows the construction of a type called SET OF INTEGER, which would be an element of the set of subsets of the integers. However, the powerset capability in Pascal is incomplete: one cannot construct a type "set of pairs of integers". Another capability of Pascal is that it allows construction of a type which is a subrange of another type: thus, one can define a subrange of the INTEGER type which only takes on the values 1, 2, ..., 10. It follows from the axioms for a topos that that sort of construction can be made in a topos as well, but the proof involves ideas beyond the scope of this article.

In Pascal, one cannot construct a data type whose elements are the solutions of an arbitrary nonempty set of equations. As an example, one might want to construct a data type "relatively prime pairs of integers" (in order to implement rational arithmetic, for example). Two integers m and n are relatively prime if the greatest common divisor of m and n is 1. For example, (6, 11) is a relatively prime pair of integers but (6, 8) is not, since 6 and 8 have greatest common divisor 2. Relatively prime pairs of integers can be presented as the solutions of an equation, namely the equation GCD(m, n) = 1, where GCD is the mapping from pairs of integers to integers for which GCD(m, r) is the greatest common divisor of m and n and 1 is the constant mapping which always gives the value 1. It is not possible to define such a data type in Pascal directly as a data type, although it is easy to write a procedure which determines whether an ordered pair is relatively prime. Thus Pascal provides, for its data types, some of the constructive capabilities of a topos, but not all of them.

VARIABLE SET THEORIES WITH DIFFERENT PROPERTIES. Although the axioms for a topos provide a great many techniques and constructions for variable sets, there are still an enormous variety of different variable set theories (toposes) with very different properties. This provides the opportunity to construct variable set theories tailored to specific purposes in mathematics, such as the example of synthetic differential geometry below.

There are two important and essentially different methods of tailoring a variable set theory to a particular mathematical discipline. One is to impose axioms in addition to the ones described above. In most cases in practice these added axioms require that additional properties of classical sets hold for the variable set theory. For example, one could require that all subsets of a variable set have complements. A variable set theory (topos) in which subsets always have complements is called Boolean; a Boolean topos would not, for example, be suitable to use in connection with the set B of businesses discussed previously. The topos used for the circle discussed above, which is called the Zariski topos, is also not Boolean. For example, the complement of the circle in the plane does not exist; for the complement to exist, there would have to be a set of polynomial equations whose solution consisted exactly of the set of points for which $x^2 + y^2 \neq 1$, and one can show that there is no such set of equations.

The other way to obtain a topos suitable for a particular kind of mathematics is to construct the domain of variation, which is technically called a site; the corresponding variable set theory is the category of sheaves over the site. During the early 1940's, Jean Leray (who was at the time a prisoner in occupied France) and later Henri Cartan of the University of Paris originated the study of sheaves. This has been a crucial tool in algebraic geometry and more recently in other areas of geometry. This mathematical technique is used to construct the Zariski topos; the site is based on the different types of numbers. CLASSICAL SETS. The sets of classical (Zermelo-Fraenkel) set theory form a topos with many special properties. in fact, it is possible to force a topos to be essentially the universe of classical sets by requiring that three additional axioms be true.

A topos is two-valued if a set with one element has exactly two subsets. This condition is of course true of classical sets: the two subsets of a one-element set S are S itself and the empty set. A one-element set can be described in a straightforward way using the primitive concepts of mapping and composition.

A topos has the natural numbers if one of its sets is like the classical set of natural numbers (nonnegative whole numbers). The precise specification of how it is like the natural numbers essentially incorporates the principle of mathematical induction. A topos which has the natural numbers also has integers as well as rational and real numbers. In such a topos one can give recursive definitions of functions such as the GCD function previously mentioned.

A topos has choice if in effect it satisfies the classical Axiom of Choice stated using only the primitive concepts of mapping and composition. The classical Axiom of Choice guarantees that given a function $S \longrightarrow T$ with the surjective property that for any element $t \in T$, there is an element $s \in S$ for which f(s) = t, then there is a "choice" function $q: T \longrightarrow S$ for which g(t) = t. This can be stated without elements using the concept of "epimorphism" from category theory.

A topos which satisfies all three of these axioms is in a technical sense essentially the same as the classical sets of Zermelo-Fraenkel. The other properties of classical sets follow from these axioms; for example, the Boolean property that subsets have complements follows from the choice axiom. This theorem, not previously known to logicians, is due to R. Diaconescu, now at Baruch College in New York City.

Lawvere and Tierney used this approach as the basis of a new, more conceptual proof of the independence of the continuum hypothesis from the other axioms of set theory (originally proved by P. Cohen at the University of California at Berkeley). The continuum hypothesis states that no set of real numbers has cardinality both less than that of the whole set of real numbers and greater than that of the set of natural numbers. Cohen showed that this hypothesis was not a consequence of the Zermelo-Fraenkel axioms. It had previously been shown by K. Gödel that the opposite hypothesis is not a consequence either. Thus the Zermelo-Fraenkel axioms do not determine the complete nature of classical sets. [See "NonCantorian Set Theory", by P. Cohen and R. Hersh; Scientific American, December, 1967.]

3. Synthetic Differential Geometry

Differential geometry begins with calculus and develops the techniques to handle curves, surfaces, and higher dimensional spaces. The current set theoretic foundations are somewhat complicated, and raise technical issues which have little to do with geometry. Synthetic differential geometry uses topos theory to define a universe of variable sets where each set has an intrinsic geometry. Specifically, one of the sets, R, is a line comparable to

the real number line in the set theoretic approach. In this universe, every mapping from R to itself is continuous and has a derivative, as distinct from the set of real numbers in set theory which has many discontinuous and geometrically useless mappings to itself.

The axioms for synthetic differential geometry are simple. The topos axioms are assumed. Further axioms require a set R with two selected points, 0 and 1, to serve as origin and unit. Formally, a point on R is a mapping from the one element set to R. Axioms also require an addition and a multiplication of R with the usual properties of associativity, distributivity and commutativity. These axioms make R something like a number line.

The axiom of line type, due to F.W. Lawvere and A. Kock, forces R to have infinitesimals. Define a subset of $D \subseteq R$, by $D = \{x \in R \mid x^2 = 0\}$.

These are the infinitesimals of square zero. Clearly 0 is in D, but D must be larger than 0 because the axiom requires that a mapping f from D to R has both a base value f(0) and a unique slope. Think of D as an infinitesimal interval, bigger than a point but too small to bend. Then for any mapping g from \mathbf{R} to \mathbf{R} and any $x \in \mathbf{R}$ there is an interval similar to D around x and a maps that interval as a line segment with a unique slope. The slope is by definition the derivative of g at x. This formalizes one of the original conceptions of derivative held, at times, by Newton and Leibniz: The derivative of a function at a point was taken to be the slope of the graph over some infinitesimal interval around the point.

The infinitesimals of synthetic differential geometry are unlike those of the nonstandard analysis of the late A. Robinson of Yale University. In particular, the derivative in synthetic differential geometry is exactly the slope of an infinitesimal line segment, which is not true in non-standard analysis.

The resulting topos is not Boolean. For example, $\{0\}$ is a subset of D but has no complement. One can prove $\{0\} \neq D$ and yet there is no $d \in D$ with $d \neq 0$. It follows that there can be no such sets as R and D in classical set theory. Reasoning in this topos must respect this subtlety. By adding a few simple, geometrically motivated axioms one can prove the classical theorems of calculus and differential geometry. Much of this has been done by M. Bunge of McGill University and G. Reyes of the University of Montreal.

On the other hand, one can describe such a topos the other way, by constructing sheaves over a domain of variation. E. Dubuc of the University of Buenos Aires defined a site similar to the one for the Zariski topos but using equations involving a larger class of functions than polynomials (the class.of all differentiable functions). Thus constructing the site presupposes the usual development of calculus in set theory. The axioms described above are all true in the topos, and the usual spaces of differential geometry are included among the variable sets.

4. Alternate approaches to variable set theory

Problems arising from the rigidity of classical set theory became apparent during the 1960's when at least two other approaches were attempted. The first, that of Boolean-

valued models, arose in the process of understanding Cohen's proof that the continuum hypothesis cannot be decided by the Zermelo-Fraenkel axioms and other similar results. Boolean valued models were invented by D. Scott, now at Carnegie-Mellon University, and R. Solovay of University of California at Berkeley. This theory can be totally subsumed as a rather special case of topos theory.

The second approach is that of fuzzy set theory. The relation of fuzzy set theory to topos is more nuanced. Fuzzy sets were invented in 1964 by L. Zadeh at the University of California at Berkeley. The basic idea is to replace the set theoretic membership relation by a fuzzy version of it, one which instead of taking on only the values true and false could take on the value any real number between 0 and 1. The intuitive idea behind this is that an element which is definitely in the set had degree of membership 1 and an element definitely not in it had degree of membership 0. Other degrees of membership represent perhaps uncertain membership although other interpretations are possible.

There is a certain incoherence (in the technical sense) in fuzzy set theory. It comes about as follows. Classical set theory is based on two predicates, membership being one and equality the other. In fuzzy sets, membership is allowed to be variable but equality is not. To see the problem raised by this omission, consider a fuzzy set S and two fuzzy subsets T and U with the property that the elements whose degree of membership in Tis at most 1/2 are the same as the elements whose degree of membership in U is at most 1/2. Then one would like to say that the degree of equality of T and U is at least 1/2. But fuzzy set theory offers no way of making such an assertion. It turns out that this gap is serious enough to prevent the formation the fuzzy set of fuzzy subsets of a fuzzy set. The result of that is that set theoretic constructions that do not depend on the formation of subsets (the so-called first order theory) can be carried out without problem, but the ones that do require the set of subsets (the so-called higher order constructions) cannot be carried out.

This deficiency of fuzzy sets can be repaired by allowing equality to be fuzzy as well. When this repair is carried out it turns out that the resultant universe of "fuzzy sets with fuzzy equality" is a topos. This topos has essentially the same interval of real numbers between 0 and 1 (the real numbers themselves, not the open subintervals) as its domain of variation. One can easily identify the original fuzzy sets in this topos by a simple criterion. Thus the theory of fuzzy sets, when modified to allow all the constructions needed in mathematics, also turns out to be subsumed as a special case of topos theory. Fuzzy set theory has been generalized in various ways, but those generalizations that look like set theory suffer from the same deficiency and when that deficiency is repaired, the resultant set theory is always that of the variable sets of a topos.

5. Conclusion

Topos theory not only represents a new, different formulation of the foundations of mathematics, it represents a new attitude towards foundations. We mentioned previously that neither the continuum hypothesis nor the negation of that hypothesis follows from the Zermelo-Fraenkel axioms for classical set theory. This means that in effect there is more than one universe of classical sets, at least one in which the continuum hypothesis is true and another in which it is false. In fact, there are many universes of classical sets, since set theorists have discovered' other properties which are independent of the Zermelo-Fraenkel axioms as well. Some mathematicians were disappointed by the discovery that the continuum hypothesis does not follow from the Zermelo-Fraenkel axioms. Those axioms were intended to capture classical set theory and to define it precisely. The fact that more than one universe of classical sets was possible was regarded as a failure of the axiom system.

There was no intention of defining a single universe of classical sets when topos theory was discovered. The axioms for topos theory were known from the beginning, from the very examples that suggested them, to allow an enormous variety of different variable set theories, and the many possibilities allowed by the axioms is a virtue, not a fault. By imposing different additional axioms or by creating a universe of variable sets from the ground up by starting with a particular domain of variation, one obtains different universes within which one can do mathematics suitable to specific applications.

The attitude that an axiomatic framework should allow variation is standard among mathematicians. It is the basis of the axiomatic method, which more than anything else distinguishes twentieth-century mathematics from the mathematics of past centuries. The object or type of data one is studying is specified by statements (axioms) which are required to be true of the object rather than by trying to specify or understand the content or internal nature of the object. Then everything which is deduced about the object (which is thus an abstract object) is deduced from the axioms. This axiomatic method, which is the standard tool-for studying the objects on which mathematicians work, is used in topos theory to study the universe within which the mathematician works, and in the process has the potential for giving the mathematician a much more supple and powerful working environment.

References

- M. Arbib and E. Manes, Arrows, Structures and Functors: The Categorical Imperative, Academic Press, 1975.
- [2] M. Barr and C. Wells, Toposes, Triples and Theories, SpringerVerlag, 1985.
- [3] A. Kock, Synthetic Differential Geometry, London Mathematical Society Lecture Note Series 51, Cambridge University Press, 1991.
- [4] F. W. Lawvere, ed. Toposes, Algebraic Geometry and Logic, Lecture Notes in Math., 274, Springer-Verlag, 1972.

Department of Mathematics and Statistics, McGill University Montreal, Quebec, Canada H3A K6 Department of Mathematics, Cleveland State University Cleveland, Ohio 44115

Department of Mathematics and Statistics, Case Western Reserve University Cleveland, Ohio 44106