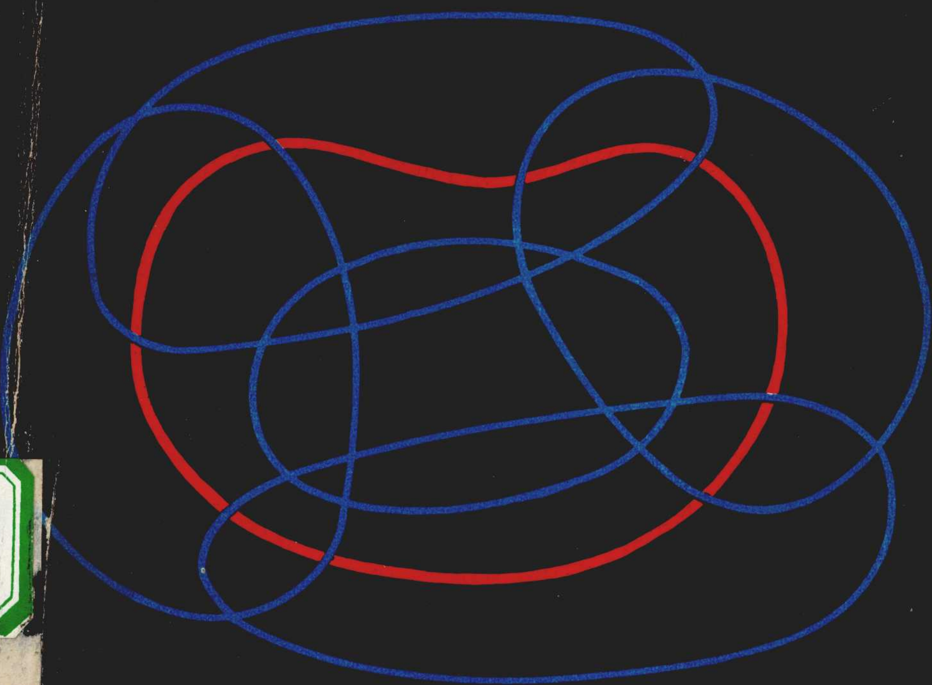
INTRODUCTION TO METRIC AND TOPOLOGICAL SPACES

W A SUTHERLAND



W. A. SUTHERLAND

Introduction to metric and topological spaces

CLARENDON PRESS · OXFORD
1975

Preface

ONE of the ways in which topology has influenced other branches of mathematics in the past few decades is by putting the study of continuity and convergence into a general setting. This book introduces metric and topological spaces by describing some of that influence. The aim is to move gradually from familiar real analysis to abstract topological spaces; the main topics in the abstract setting are related back to familiar ground as far as possible. Apart from the language of metric and topological spaces, the topics discussed are compactness, connectedness, and completeness. These form part of the central core of general topology which is now used in several branches of mathematics. The emphasis is on *introduction;* the book is not comprehensive even within this central core, and algebraic and geometric topology are not mentioned at all. Since the approach is via analysis, it is hoped to add to the reader's insight on some basic theorems there (for example, it can be helpful to some students to see the Heine–Borel theorem and its implications for continuous functions placed in a more general context).

The stage at which a student of mathematics should see this process of generalization, and the degree of generality he should see, are both controversial. I have tried to write a book which students can read quite soon after they have had a course on analysis of real-valued functions of one real variable, not necessarily including uniform convergence.

The first chapter reviews real numbers, sequences, and continuity for real-valued functions of one real variable. Most readers will find nothing new there, but we shall continually refer back to it. With continuity as the motivating concept, the setting is generalized to metric spaces in Chapter 2 and to topological spaces in Chapter 3. The pay-off begins in Chapter 5 with the study of compactness, and continues in later chapters on connectedness and completeness. In order to introduce uniform convergence, Chapter 8 reverts to the traditional approach for real-valued functions of a real variable before interpreting this as convergence in the sup metric.

Most of the methods of presentation used are the common property of many mathematicians, but I wish to acknowledge that the way of introducing compactness is influenced by Hewitt (1960). It is also a pleasure to acknowledge the influence of many teachers, colleagues, and ex-students on this book, and to thank Peter Strain of the Open University for helpful comments and the staff of the Clarendon Press for their encouragement during the writing.

W. A. S.

Oxford, 1974

Introduction

IN this book we are going to generalize some theorems about convergence and continuity which are probably familiar to the reader in the case of sequences of real numbers and real-valued functions of one real variable. An example of the kind of result we shall be aiming to generalize is the following: *if f is a real-valued function which is defined and continuous on the closed interval $[a, b]$ in the real line, then f is bounded on $[a, b]$, i.e., there exists a real number K such that $|f(x)| \leq K$ for all x in $[a, b]$*. Several such theorems about real-valued functions of a real variable are true and useful in a more general framework, after suitable minor changes of wording. For example if we suppose that f is a real-valued function of two real variables which is defined and continuous in a rectangle $[a, b] \times [c, d]$, then it is still true that f is bounded on this rectangle. Once we have seen that the result generalizes from one to two real variables, it is natural to suspect that it is true for any finite number of real variables, and then to go a step further by asking just how general a situation can the theorem be formulated for, and how generally is it true? These questions lead us to metric and topological spaces.

Before going on to study such questions, it is fair to ask: what is the point of generalization? One answer is simply that it saves time, or at least avoids tedious repetition. For example, if we can show by a single proof that a certain result holds for functions of n real variables, where n is any positive integer, this is better than proving it separately for one real variable, two real variables, three real variables, etc. In the same vein, generalization often gives a unified mental grasp of several results which otherwise might just seem vaguely similar, and in addition to the satisfaction involved, this more efficient organization of material helps some people's understanding. Another gain is that generalization often illuminates the proof of a theorem, because to see how generally a given result can be proved, one has to notice exactly what properties or hypotheses are used at each stage in the proof. A motive which is perhaps more sophisticated is the urge to prove any given result in the appropriate context; what this means is partly a matter of taste, and it is not easy to explain before it has been done.

Against these motives, we should be aware of some dangers in generalization. Most mathematicians would agree that it can be carried to an excessive and useless extent. Just when this stage is reached is a matter of controversy, but the potential reader is warned that some mathematicians would say 'Enough, no more (at least as far as analysis is concerned)' when we get into metric spaces. Also, there is an initial barrier of unfamiliarity to be overcome in moving to a more general framework, with its new language; the extent to which the pay-off is worthwhile is likely to vary from one student to another.

As a consequence of introducing abstractions gradually, the theorem density is low. The title of theorem is reserved for substantial results, which have significance in a broad range of mathematics.

Some paragraphs and exercises are marked † to denote that they require some knowledge of abstract algebra, and others are marked * to denote that they are tentatively thought to be more severe than the rest. There are hints at the end of the book for all such exercises and some others.

At several points in the text, suggestions are made about helpful diagrams which can be drawn. Readers are strongly urged to draw their own diagrams wherever possible, even when no specific suggestions are made.

A previous course in real analysis is regarded as a prerequisite for reading this book. What this means is an introduction (including rigorous proofs) to continuity, differential (and preferably also integral) calculus for real-valued functions of one real variable, and to convergence of real number sequences and series. This material is contained for example in Burkill (1970) or Spivak (1973) (names followed by dates in parentheses refer to the bibliography at the end of the book). Chapter 1 is intended only as a review of some of the real analysis to which the rest of the book most frequently refers. It does not include special functions and facts from calculus which we sometimes use for illustration. The experience of abstraction gained from a previous course in, say, linear algebra would help the reader in a general way to follow the abstraction of metric and topological spaces. However, the student is likely to be the best judge of whether she/he is ready, or wants, to read this book.

Notation and terminology

WE use the logical symbols $\Rightarrow$ and $\Leftrightarrow$ or iff, meaning *implies* and *if and only if*. Throughout, we use the language of sets and maps. Since most introductions to algebra and analysis contain a survey of this language, we merely list notation.

If an object a belongs to a set A we write $a \in A$, and if not we write $a \notin A$. If A is a subset of B (perhaps equal to B) we write $A \subset B$ or $B \supset A$. The subset of elements of A possessing some property P is written $\{a \in A : P(a)\}$. A finite set is sometimes specified by listing its elements, say $\{a_1, a_2, \ldots, a_n\}$. Intersection and union of sets are denoted by $\bigcap$ and $\bigcup$. The empty set is written $\varnothing$. Given two sets A and B, the set $\{b \in B : b \notin A\}$ is written $B-A$. Thus in particular if $A \subset B$, the complement of A in B is $B-A$. If for each i in some set I we are given a subset A_i of a set S, then we denote by $\bigcup_{i \in I} A_i$, $\bigcap_{i \in I} A_i$ (or just $\bigcup_I A_i$, $\bigcap_I A_i$) the union and intersection over all i in I; in this situation, I is called an *indexing* set. We use De Morgan's laws, which in the above notation assert

$$S-\bigcup_I A_i = \bigcap_I (S-A_i), \qquad S-\bigcap_I A_i = \bigcup_I (S-A_i).$$

The Cartesian product $A \times B$ of sets A, B is the set of all ordered pairs (a, b) where $a \in A$, $b \in B$. This generalizes easily to the product of any finite number of sets; in particular we use A^n to denote the set of ordered n-tuples of elements from A. We shall not consider infinite products of sets. Occasionally we use 2^A to denote the set of all subsets of a set A.

A map or function (the terms are used interchangably) between sets A, B is written $f: A \to B$. We call A the *domain* of f, and we avoid calling B anything. We think of f as assigning to each a in A an element $f(a)$ in B, although logically it is preferable to define a map as a pair of sets A, B together with a certain type of subset of $A \times B$ (intuitively the graph of f). Persisting with our way of thinking of f, we define the *graph of f* to be $G_f = \{(a, b) \in A \times B : f(a) = b\}$. For any subset $C \subset A$, the (*direct*) *image* $f(C)$ of C under f is $\{b \in B : b = f(c)$ for some c in $C\}$ and for any subset $D \subset B$, the *inverse image* $f^{-1}(D)$ of D under f is $\{a \in A : f(a) \in D\}$. We use the formulae

$$f\Big(\bigcap_I C_i\Big) \subset \bigcap_I f(C_i), \qquad f\Big(\bigcup_I C_i\Big) = \bigcup_I f(C_i),$$

$$f^{-1}\Big(\bigcap_I D_i\Big) = \bigcap_I f^{-1}(D_i), \qquad f^{-1}\Big(\bigcup_I D_i\Big) = \bigcup_I f^{-1}(D_i),$$

$$f^{-1}(B - D_i) = A - f^{-1}(D_i),$$

where $f : A \to B$ is a map and $C_i \subset A$, $D_i \subset B$ for every i in I. Note that equality does not necessarily hold in the first of these. We call f *injective* if $f(a) = f(a') \Rightarrow a = a'$, since the term one–one is a little ambiguous. We should therefore call $f : A \to B$ surjective if $f(A) = B$, but instead we call such an f *onto*. If f has both these properties we call it a *one–one correspondence*. In the special case when a map $f : A \to B$ is injective we may define an inverse function $f^{-1} : f(A) \to A$. When f is injective and $C \subset f(A)$ there is therefore a second meaning of $f^{-1}(C)$. Fortunately this determines the same subset of A as the previous meaning, but we emphasize that the inverse image set $f^{-1}(C)$ is defined for any $C \subset B$, whether f is injective or not, whereas the inverse map f^{-1} is defined only if f is injective, and even when f is injective the direct image $f^{-1}(C)$ of C under f^{-1} is defined only if $C \subset f(A)$.

Given a map $f : A \to B$ and a subset $C \subset A$ we write $f \mid C : C \to B$ for the *restriction* of f to C, defined by $(f \mid C)(c) = f(c)$ for all c in C. In this situation we also refer to f as an *extension* of $f \mid C$ to A. Given maps $f : A \to B$, $g : B \to C$, the *composition* $g \circ f : A \to C$ is defined by $g \circ f(a) = g(f(a))$ for any a in A. In particular if $i : D \to A$ denotes the inclusion map of a subset D in A, and $f : A \to B$ is any map, then $f \circ i = f \mid D$. Special maps used include the identity map $f : A \to A$ of any set A, given by $f(a) = a$ for all a in A, and constant maps $f : A \to B$, given by $f(x) = b$ for every x in A and some fixed b in B.

We shall occasionally assume that the terms *equivalence relation* and *countable set* are understood.

We use **N**, **Z**, **Q**, **R**, **C** to denote the sets of natural numbers (or positive integers), integers, rational numbers, real numbers, and complex numbers respectively. We often refer to **R** as the *real line*, and we call the following subsets of **R** *intervals:*

$$
\begin{array}{rl}
\text{(i)} & [a, b] = \{x \in \mathbf{R} : a \leqslant x \leqslant b\}, \\
\text{(ii)} & (a, b) = \{x \in \mathbf{R} : a < x < b\}, \\
\text{(iii)} & (a, b] = \{x \in \mathbf{R} : a < x \leqslant b\}, \\
\text{(iv)} & [a, b) = \{x \in \mathbf{R} : a \leqslant x < b\}, \\
\text{(v)} & (-\infty, b] = \{x \in \mathbf{R} : x \leqslant b\}, \\
\text{(vi)} & (-\infty, b) = \{x \in \mathbf{R} : x < b\}, \\
\text{(vii)} & [a, \infty) = \{x \in \mathbf{R} : x \geqslant a\}, \\
\text{(viii)} & (a, \infty) = \{x \in \mathbf{R} : x > a\}, \\
\text{(ix)} & (-\infty, \infty) = \mathbf{R}.
\end{array}
$$

The intervals in (i), (v), (vii) (and (ix)) are called *closed* intervals; those in (ii), (vi), (viii) (and (ix)) are called *open* intervals, and the others are called *half-open* intervals. When we refer to an interval of types (i)–(iv), it is always to be understood that $b > a$, except for type (i), when on stated occasions we also allow $b = a$. We shall try to avoid the occasional risk of confusing an interval (a, b) in **R** with a point (a, b) in $\mathbf{R}^2$ by stating which of these is meant when there might be any doubt.

Oxford University Press, Ely House, London W. 1

GLASGOW NEW YORK TORONTO MELBOURNE WELLINGTON
CAPE TOWN IBADAN NAIROBI DAR ES SALAAM LUSAKA ADDIS ABABA
DELHI BOMBAY CALCUTTA MADRAS KARACHI LAHORE DACCA
KUALA LUMPUR SINGAPORE HONG KONG TOKYO

CASEBOUND ISBN 0 19 853155 9

PAPERBACK ISBN 0 19 853161 3

FILMSET AT THE UNIVERSITIES PRESS, BELFAST

AND PRINTED BY J. W. ARROWSMITH LTD, BRISTOL 3

Contents

1

Review of some real analysis

1.1 Real numbers

Two popular ways of thinking about the real number system are:

(1) geometrically, as corresponding to all the points on a straight line;

(2) in terms of decimal expansions, where if a number is irrational we think of longer and longer decimal expansions approximating it more and more closely.

Neither of these intuitive ideas is precise enough for our purposes, although each leads to a way of constructing the real numbers from the rational numbers. The second of these ways is described in the appendix. As the reader probably knows, one approach to real numbers is to use a set of axioms. This leaves aside the questions of whether there is any system satisfying the axioms and to what extent such a system is unique. Also, it can hardly be said to explain what the real numbers are to anyone who does not already know in an intuitive sense. But it has the merit of providing quickly a point of departure for analysis.

Many introductions to analysis contain a list of axioms for the real numbers (see, for example, Spivak (1973) or Chapter 1 of Apostol (1957)). A large number of these axioms may be summed up technically by saying that the real numbers form an ordered field. A less precise description is that addition, subtraction, multiplication, and division of real numbers all work in the way that we expect them to, and that the same is true of the way in which inequalities $x<y$ work and interact with addition and multiplication. We shall not review these axioms, but concentrate solely on the so-called completeness axiom. The reasons for this strange behaviour are, first, that this is the axiom which distinguishes the real numbers from the rational numbers (and in a sense analysis from algebra), and secondly that our intuition is unlikely

to let us down on properties deducible from the ordered field axioms, whereas arguments using the completeness axiom tend to be more subtle.

In discussing the completeness axiom we shall proceed as if the real number system were already 'there', and our job is to describe one property of it. (In a strictly axiomatic approach we ought to explain what the completeness axiom says using only the ordered field axioms, and then define the real numbers to be any system satisfying the completeness axiom as well as the ordered field axioms.)

We need some terminology. Let S be a non-empty set of real numbers. An upper bound for S is a number x such that $y \leqslant x$ for all y in S. If an upper bound for S exists we say that S is bounded above. Lower bounds are defined similarly.

EXAMPLES 1.1.1. (a) The set $\mathbf{R}$ of all real numbers has no upper or lower bound.

(b) The set $\mathbf{R}_-$ of all strictly negative real numbers has no lower bound, but for example 0 is an upper bound.

(c) The half-open interval $(0, 1]$ is bounded above and below.

If S has an upper bound u, then S has many upper bounds, since any x satisfying $x \geqslant u$ is another upper bound. This gives the next definition some point.

DEFINITION 1.1.2. Given a non-empty subset S of $\mathbf{R}$ which is bounded above, we call u a *least upper bound for* S if

(a) u is an upper bound for S,
(b) $x \geqslant u$ for any upper bound x of S.

EXAMPLE 1.1.3. In Example 1.1.1(b), 0 is a least upper bound for $\mathbf{R}_-$. For 0 is an upper bound, and it is a *least* upper bound because no strictly negative number x can be an upper bound for $\mathbf{R}_-$, since $\frac{1}{2}x > x$ and $\frac{1}{2}x \in \mathbf{R}_-$. Examples 1.1.1(c) and (b) show that a least upper bound of a set S may or may not be in S.

It follows from Definition 1.1.2 that least upper bounds are unique when they exist. For if u, u' are both least upper bounds for a set S, then since u' is an upper bound it follows that $u \leqslant u'$ by leastness of u (by leastness we mean property 1.1.2(b)). Similarly $u' \leqslant u$, so $u = u'$. Greatest lower bounds are defined similarly. We can now state one form of the completeness axiom for $\mathbf{R}$.

AXIOM 1.1.4. *Any non-empty subset of* **R** *which is bounded above has a least upper bound.*

This axiom is quite subtle, and it is difficult to grasp its full significance until it has been used several times. It corresponds to the intuitive idea that there are no gaps in the real numbers, thought of as the points on a straight line; but the transition from the intuitive idea to the formal statement is not immediately obvious. For some sets of real numbers, such as Examples 1.1.1(b) and (c), it is 'obvious' that a least upper bound exists (strictly speaking, this means that it follows from the ordered field axioms). As a first example of the kind of set to which we really want to apply the completeness axiom, and of its relation to the existence of irrational numbers, let $S=\{x\in\mathbf{Q}:x^2<2\}$. Then intuitively the least upper bound of S is the irrational $\sqrt{2}$ (we write $\sqrt{2}$ for the positive square root of 2). Formally, we need to prove the following assertions:

(a) if u is a least upper bound for S then $u^2=2$,

(b) there is no rational number u such that $u^2=2$.

We shall prove (a) in Example 1.1.8 below. The reader has probably seen a proof of (b); it is included in Exercise 1.5.5 for which hints are given.

For any non-empty subset S of **R** which is bounded above, we call the unique least upper bound *sup* S (sup is short for supremum. Other notation sometimes used is l.u.b. S). Although the completeness axiom was stated in terms of sets bounded above, it is equivalent to the corresponding property for sets bounded below. The next proposition formally states half of this equivalence.

PROPOSITION 1.1.5. *If a non-empty subset* S *of* **R** *is bounded below then it has a greatest lower bound.*

Proof. Let $T=\{x\in\mathbf{R}:-x\in S\}$. The idea of the proof is simply that l is a lower bound for S iff $-l$ is an upper bound for T. The details are left as Exercise 1.5.6.

Just as in the case of least upper bounds, a non-empty subset S of **R** which is bounded below has a unique greatest lower bound called *inf* S (short for infimum) or g.l.b.S.

We next give two applications of the completeness axiom. The proofs use several 'obvious' facts about **R** which in a strictly axiomatic approach would be deduced from the ordered field axioms. We leave the reader to spot these.

PROPOSITION 1.1.6. *The set* $\mathbf{N}$ *of natural numbers is not bounded above.*

Proof. Suppose that $\mathbf{N}$ is bounded above. Then by the completeness axiom there is a real number $u = \sup \mathbf{N}$. For any n in $\mathbf{N}$, $n+1$ is also in $\mathbf{N}$, so $n+1 \leqslant u$. But then $n \leqslant u-1$. Since this is true for all n in $\mathbf{N}$, $u-1$ is an upper bound for $\mathbf{N}$, contradicting the leastness of u. This contradiction shows that $\mathbf{N}$ cannot be bounded above.

The property expressed in Proposition 1.1.6, or some equivalent statement, is called Archimedes' axiom (it is an axiom in some axiom systems for $\mathbf{R}$, although not in ours).

COROLLARY 1.1.7. *Between any two distinct real numbers* x, y *there is a rational number.*

Proof. Suppose $x<y$. Then $y-x>0$, and $1/(y-x)$ exists. By Proposition 1.1.6 there is a positive integer n satisfying $n> 1/(y-x)$, so $1/n<y-x$. Now let $M=\{m \in \mathbf{N} : m/n>x\}$. By Proposition 1.1.6, M is non-empty, otherwise nx would be an upper bound for $\mathbf{N}$. Hence, since $M \subset \mathbf{N}$, M contains a least number v. (This deduction would be fallacious if all we knew was, say, $M \subset \mathbf{R}_+$.) This means $v/n>x$, $(v-1)/n \leqslant x$. Hence $v/n \leqslant x+1/n< x+(y-x)=y$, and v/n is a rational number satisfying $x<v/n<y$ as required.

REMARK. Between any two distinct real numbers there is also an irrational number (see Exercise 1.5.7).

We are now in a position to prove the existence of $\sqrt{2}$.

EXAMPLE 1.1.8. There exists a real number u such that $u^2=2$.

Proof. Let $S=\{x \in \mathbf{Q} : x^2<2\}$ (we could equally well use $\{x \in \mathbf{R} : x^2<2\}$). Then S is non-empty ($1 \in S$). To prove that S is bounded above, there is no need to be fastidious—the existence of some upper bound, however much larger than is strictly necessary, will do. For example, 10 is an upper bound, since if $y>10$ then $y^2>100>2$, and y is not in S. Hence $x \in S \Rightarrow x \leqslant 10$. Now by the completeness axiom, $u = \sup S$ exists. In fact $u \geqslant 1$, since $1 \in S$. We shall prove that $u^2=2$, by showing that each of $u^2>2$, $u^2<2$ leads to a contradiction.

First suppose that $u^2>2$. Then $(u^2-2)/2u>0$, so by Proposition 1.1.6 there exists an n in $\mathbf{N}$ satisfying

$$0<1/n<(u^2-2)/2u.$$

Then

$$(u-1/n)^2=u^2-2u/n+1/n^2>u^2-2u/n>u^2-(u^2-2)=2,$$

so $x\in S\Rightarrow x^2<2<(u-1/n)^2\Rightarrow x<u-1/n$, contradicting the leastness of u.

Secondly suppose that $u^2<2$. Choose an integer n such that $0<1/n\leqslant(2-u^2)/4u$ and $1/n<2u$. Then

$$\begin{aligned}(u+1/n)^2=u^2+2u/n+1/n^2&<u^2+2u/n+2u/n \quad (\text{since } 1/n<2u)\\ &\leqslant u^2+2-u^2 \quad (\text{since } 2-u^2\geqslant 4u/n)\\ &=2.\end{aligned}$$

Hence $u+1/n\in S$, contradicting the fact that u is an upper bound for S. This completes the proof that $u^2=2$.

By similar proofs we could establish the existence of other square roots, cube roots, etc. However, the kind of work needed above can later be done much more efficiently, although the proofs are still based on the completeness axiom.

The completeness axiom can be stated in various other forms. Two of these are mentioned in §1.2 below.

We conclude this brief review of real numbers by recalling two useful inequalities, whose proofs are immediate.

PROPOSITION 1.1.9. $|x+y|\leqslant|x|+|y|$ *for any* x, y *in* $\mathbf{R}$.

COROLLARY 1.1.10. $|x-y|\geqslant\big||x|-|y|\big|$ *for any* x, y *in* $\mathbf{R}$.

1.2 Real sequences

Formally an infinite sequence of real numbers is a map $s:\mathbf{N}\to\mathbf{R}$. This definition is useful for discussing topics such as subsequences and rearrangements without being vague. In practice, however, given such a map s we denote $s(n)$ by s_n and think of the sequence in the traditional way as an infinite ordered string of numbers, using the notation (s_n) or s_1, s_2, $s_2,\ldots$ instead of s for the whole sequence.

It is important to distinguish between a sequence (s_n) and the set of its members, $\{s_n:n\in\mathbf{N}\}$. The latter can easily be finite. For