V : Number systems and set theory

Any reasonable framework for mathematics should include the fundamental number systems which arise in the subject:

1. The ***natural numbers*** $\mathbf{N}$ (also known as the ***nonnegative integers***).
2. The (***signed***) ***integers*** $\mathbf{Z}$ obtained by adjoining negative numbers to $\mathbf{N}$.
3. The ***rational numbers*** $\mathbf{Q}$ obtained by adjoining reciprocals of nonzero integers to $\mathbf{Z}$.
4. The ***real numbers*** $\mathbf{R}$, which should include fundamental constructions like n^{th} roots of positive rational numbers for an arbitrary integer $n > 1$, and also all "infinite decimals" of the form $b_1 \cdot 10^{-1} + b_2 \cdot 10^{-2} + \ldots + b_k \cdot 10^{-k} + \ldots$ where each b_i belongs to $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Up to this point we have tacitly assumed that such number systems are at our disposal. However, in both the naïve and axiomatic approaches to set theory it is eventually necessary to say more about them.

The naïve approach. In naïve set theory it is necessary to do two things. First, one must describe the properties that the set – theoretic versions of these number systems should satisfy. Second, something should be said to justify our describing such systems as **THE** natural numbers, **THE** integers, **THE** rational numbers, and **THE** real numbers. This usage suggests that we have completely unambiguous descriptions of the number systems in terms of their algebraic and other properties. One way of stating this is that

> ***any system satisfying all the conditions for one of the standard systems*** $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{Q}$ ***or*** $\mathbf{R}$ ***should be the same as*** $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{Q}$ ***or*** $\mathbf{R}$ ***for all mathematical purposes*****,**

with some explicit means for mechanical translation from the given system to the appropriate standard model. In less formal terms, it we have any systems **X** which satisfy all the fundamental properties of one of the systems $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{Q}$ or $\mathbf{R}$, then **X** is essentially a ***mathematical clone*** of the appropriate number system.

There are good theoretical and philosophical reasons for asking such questions about the essential uniqueness of the number systems, but these question also have some important practical implications for the development of mathematics. If there would be two systems that satisfy the basic properties of $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{Q}$ or $\mathbf{R}$ but differ from a standard model in some significant fashion, then clearly we might get different versions of mathematics depending upon which example is chosen. To illustrate this, suppose we decided to develop a version of the real numbers in which infinite base 10 "decimal expansions" are replaced by expansions with some other number base, say 16 (to conform with the internal arithmetic of some computer) or 60 (as in Babylonian mathematics). We **expect** that everything should work the same regardless of the

numerical base we choose for expressing quantities, but at some point it is necessary to **confirm** that our expectation is fulfilled.

Later in this unit we shall describe precisely the notion of a mathematical clone. For the time being we note that examples of this concept have already been encountered in Section **IV.6** when we talked about whether two partially ordered sets have the same ***order type***. Given two such partially ordered sets, the **1 – 1** order preserving correspondence from one to another can be viewed as a formal mathematical way of saying that either of the partially ordered sets is a clone of the other.

Our coverage in this unit will mainly concern the first item described in the naïve approach; namely, the formal properties of the number systems and the mathematical statements of their uniqueness properties. Later in these notes (and largely for reference purposes) we shall explain why the basic properties describe these number systems in a totally unambiguous manner.

The axiomatic approach. In axiomatic set theory it is necessary to assume the existence of systems with the given properties and to prove these properties describe them unambiguously (the latter proceeds exactly the same as in naïve set theory).

One new issue in the axiomatic approach is the goal of keeping the basic assumptions for set theory as simple as possible. Assuming the existence of four separate but clearly interrelated number systems is a convenient first step, but at some point it is natural to ask if we really need to make such a long list of assumptions in order to set everything up. Aside from possible aesthetic considerations, there is the practical consideration that long lists of assumptions raise questions whether there might be some logical inconsistency; after all, the whole idea of a proof by contradiction is that one makes so many assumptions that the conclusions end up contradicting each other, and it would undermine everything if such contradictions could be derived from the axioms for set theory itself. We shall address some of these issues in the final unit of the notes.

Some more specific objectives

Much of this unit is devoted to summarizing familiar properties of the four basic number systems, so we shall indicate some points that are less elementary and particularly important. In Section 1 the most significant new item is the statement of the Peano Axioms for the natural numbers, and in Section 2 the discussion of finite induction and recursive definitions in the framework of set theory is one of the main topics in the unit. The formulas for counting the numbers of elements in various finite sets in Section 3 start with familiar ideas, and they give systematic rules that are important both for their own sake and for the remaining units of the course. Finally, the description of the real numbers in Section 4 is fundamentally important. Although this description is fairly concise, it contains everything that is needed to justify the standard facts about real numbers and to develop calculus in a mathematically rigorous fashion. The latter development is covered in subsequent courses. Although the justification of the usual expansions for real numbers is also somewhat peripheral to the present course, for the sake of completeness we shall explain how our formal description of the real numbers yields their familiar properties which are used in everyday work, both inside and outside of mathematics.

V. 1 : The natural numbers and integers

(Halmos, §§ 11 – 13; Lipschutz, §§ 2.1, 2.7 – 2.9)

In many respects the positive integers form the most basic number system in all of the mathematical sciences. Some reasons for this are historical or philosophical, but logical considerations are particularly important for the systematic development of mathematics.

Clearly we would like our descriptions of number systems to summarize their basic algebraic properties concise but understandable. In particular, it simplifies things considerably if we can say that addition, subtraction and multiplication are always defined. Since the positive integers are not closed under subtraction, clearly they do not fulfill this condition. Therefore we shall begin by describing the integers, and we shall view the positive integers as a subset of the integers with certain special properties.

The important algebraic properties of the integers split naturally into three classes, two of which are fairly general and one of which is more focused.

Basic rules for addition and multiplication. Formally, these are the conditions defining an abstract type of mathematical system known as a ***commutative ring with unit.***

FIRST AXIOM GROUP FOR THE INTEGERS. *The integers are a set* $\mathbf{Z}$**,** *and they have binary operations* $\mathbf{A : Z \times Z \to Z}$**,** *normally expressed in the form* $\mathbf{A(u, v) = u + v}$**,** *and* $\mathbf{M : Z \times Z \to Z}$**,** *normally expressed in the form* $\mathbf{M(u, v) = uv}$ *or* $\mathbf{u \cdot v}$ *or* $\mathbf{u \times v}$**,** *which satisfy the following algebraic conditions:*

1. (Associative Laws). *For all* **a, b, c** *in* $\mathbf{Z}$**,** $\mathbf{(a + b) + c = a + (b + c)}$ *and* $\mathbf{(ab)c = a(bc)}$**.**
2. (Commutative Laws). *For all* **a, b** *in* $\mathbf{Z}$**,** $\mathbf{a + b = b + a}$ *and* $\mathbf{ab = ba}$**.**
3. (Distributive Law). *For all* **a, b, c** *in* $\mathbf{Z}$**,** $\mathbf{a(b + c) = ab + ac}$**.**
4. (Existence of 0 and 1). *There are distinct elements* **0, 1** *in* $\mathbf{Z}$ *such that for all* **a** *we have* $\mathbf{a + 0 = a}$**,** $\mathbf{a \times 0 = 0}$ *and* $\mathbf{a \times 1 = a}$**.**
5. (Existence of negatives or additive inverses). *For each* **a** *in* $\mathbf{Z}$ *there is an element* $\mathbf{-a}$ *in* $\mathbf{Z}$ *such that* $\mathbf{a + (-a) = 0}$**.**

Notational footnote: The notation $\mathbf{Z}$ for the integers has become fairly standard in mathematical writings, and it is apparently derived from the German word for numbers (*Zahlen*) and/or cyclic (*zyklisch*).

We shall need the following basic consequences of the preceding algebraic conditions:

Proposition 1. *If* **a** *belongs to a system satisfying the properties listed above, then we have* $\mathbf{(-a)(-b) = ab}$**.**

Proof. The following are special cases of the axioms:

$$0 = a0 = a[b + (-b)] = ab + a(-b)$$

$$0 = 0(-b) = [a + (-a)](-b) = a(-b) + (-a)(-b)$$

The preceding results also show that $ab = -[a(-b)] = (-a)(-b)$.■

Basic rules for ordering. When combined with the previous conditions, these yield a type of mathematical system known as an ***ordered integral domain.***

SECOND AXIOM GROUP FOR THE INTEGERS. *There is a linear ordering on* $\mathbf{Z}$ *such that the following hold:*

1. *If* $a > 0$ *and* $b > 0$, *then* $a + b > 0$ *and* $ab > 0$.
2. *For all* a, b *in* $\mathbf{Z}$, *we have* $a > b$ *if and only if* $a - b > 0$.

Well – ordering of positive elements. This is the assumption that the set $\mathbf{N}$ of nonnegative elements in $\mathbf{Z}$, often called the ***natural numbers,*** is well – ordered with respect to the standard linear ordering.

WELL - ORDERING AXIOM FOR THE POSITIVE INTEGERS. *The set* $\mathbf{N}$ *of all* x *in* $\mathbf{Z}$ *such that* $x \geq 0$ *is well – ordered.*

We shall new derive some basic properties of the integers.

Lemma 2. *If* x *is a nonzero element in a system satisfying the first two groups of axioms, then* x^2 *is positive*.

Proof of Lemma 2. Either x is positive or $-x$ is positive, and in these respective cases it follows that x^2 is positive or $(-x)^2$ is positive. However, the previous proposition implies that $x^2 = (-x)^2$, and thus in either case we know that the square must be positive.■

Lemma 3. *The multiplicative identity* 1 *is positive, and there are no integers* x *for which we have* $0 < x < 1$.

Proof of Lemma 3. First of all, 1 is positive because $1 = 1^2$. Let P be the set of positive elements in $\mathbf{Z}$. By well – ordering it follows that P has a least element m, which must satisfy $m \leq 1$. If strict inequality holds then we have $1 - m > 0$, and therefore we have $m(1 - m) > 0$, which translates to $0 < m^2 < m$, contradicting the minimality of m. Therefore 1 must be the least element of the positive integers.■

We shall need the following elementary but important property of positive integers later in this unit.

Theorem 4. (Long Division Theorem.) *Given two nonnegative integers* a *and* b such that $b > 1$, *there are* ***unique*** *nonnegative integers* q *and* r *such that* $a = bq + r$, *where* $0 \leq r \leq b - 1$.

The numbers **q** and **r** are often called the *integral quotient* and *remainder* respectively.

Proof. We first prove existence. Consider the set of all differences **a – b x,** such that **x** is a nonnegative integer and **a – b x** is nonnegative**.** This set contains **a,** and thus it is nonempty, and as such it has a minimum element **y.** We claim that **y < b;** if this were false, then **y – x** would be another element of the set (it is still nonnegative) and it would be strictly less than **y.** Since **y** is minimal this cannot happen, and therefore we must have **y < b.** This establishes existence.

To prove uniqueness, suppose that we have two expressions

$$a = b\,q + r = b\,q' + r',$$

where **q** and **q′** are nonnegative and (say) $0 \leq r \leq r' \leq b - 1$. These conditions imply that $0 \leq r' - r \leq b - 1$, and since

$$b\,(q' - q) = r' - r \leq b - 1$$

it follows that $b\,(q' - q) = 0$. Since **b** is positive this forces $q' - q$ to be equal to **0,** so that $q' = q$. If we substitute this back into the first displayed equation in the paragraph we see that we must also have $r' = r$.■

The Peano Axioms for the natural numbers

There is a very simple and important characterization of $\mathbf{N}$ which is due to G. Peano (1858 – 1932). It depends upon two intuitively clear properties. The first is that zero is the unique nonnegative integer that is smaller than every other nonnegative integer, and the second is that if we are given a nonnegative integer **n,** then **n + 1** is the unique minimal positive integer m such that **m > n.**

Definition. A ***system satisfying the Peano axioms*** is an ordered pair (P, σ) consisting of a set **P** and a function $\sigma: P \to P$ with the following properties [which reflect the nature of σ as a map taking each natural number **m** to its "successor" **m + 1**]:

(1) There is a distinguished element (the *zero element* **0** or 0_N) that is not in the image of σ.

(2) The map σ is $1 - 1$.

(3) If **A** is a subset of **P** such that

(i) $0 \in A$,

(ii) for all $k \in P$, $k \in A$ implies $\sigma(k) \in A$,
then we must have $A = P$.

The third axiom is added to guarantee that **P** is the minimal set satisfying the axioms and containing **0.**

The next result should come as no surprise.

Theorem 5. *If* $\mathbf{N}$ *denotes the natural numbers and* $\sigma : \mathbf{N} \to \mathbf{N}$ *is the function defined by* $\sigma(m) = m + 1$, *then* (P, σ) *satisfies the Peano axioms.*

Proof. The first property follows because $\sigma(x) = 0$ implies $x = -1$, and hence 0 is not in the range of σ. The second follows because $\sigma(x) = \sigma(y)$ means that $x + 1 = y + 1$, and if we subtract 1 from each side we obtain $x = y$. To prove the third, suppose that A is not equal to $\mathbf{N}$. By well – ordering we know that $\mathbf{N} - A$ has a least element m. Since $0 \in A$, we know that $m > 0$. Furthermore, since m is the least element of $\mathbf{N} - A$ then it follows that $m - 1 \in A$. But now if we apply property **(ii)** we conclude that $m = \sigma(m - 1)$ must also lie in A, contradicting our assumption that m does not belong to A. The source of the contradiction is our assumption that A is a proper subset of $\mathbf{N}$, and hence this must be false, so that $A = \mathbf{N}$. ■

Uniqueness of the integers

At the beginning of this unit we indicated that our descriptions of number systems should essentially characterize them uniquely; in other words, we would like to say that if we are given two systems which satisfy our axioms for the integers, then they are the same for all mathematical purposes. This is analogous to the notion of order – isomorphism in Section **IV.6,** and the term ***isomorphism*** is also used to describe the sorts of mathematical equivalences that we shall consider here.

As in the case of partially ordered sets, we shall try to motivate the appropriate concept of isomorphism with an example: If we are given one system which satisfies the given list of properties for the integers, then it is possible to construct a second system by brute force as follows. Let $\mathbf{Z}$ be the original set with operations and order given in the usual manner. Then we can make the set $\mathbf{Z} \times \{0\}$ into a system satisfying the same properties by defining addition by the formula $(x, 0) + (y, 0) = (x + y, 0)$, multiplication by the formula $(x, 0) \cdot (y, 0) = (xy, 0)$, and ordering by the formula $(x, 0) < (y, 0)$ if and only if $x < y$. This may, and in fact ***should***, seem somewhat artificial, for there is an obvious $1 - 1$ correspondence h from $\mathbf{Z}$ to $\mathbf{Z} \times \{0\}$ such that $h(x + y) = h(x) + h(y)$, $h(x \cdot y) = h(x) \cdot h(y)$, and $h(x) < h(y)$ if and only if $x < y$. In other words, the $1 - 1$ correspondence h preserves all the basic structure. A map of this sort is known as an ***isomorphism***. The basic uniqueness result states that any two systems satisfying the listed properties for the integers are related by an isomorphism. Here is the formal statement.

Theorem 6. *Suppose that* X *and* Y *are sets with notions of addition, multiplication and ordering which satisfy all the conditions for the integers. Then there is a* ***unique*** $1 - 1$ *correspondence from* h *from* X *to* Y *that is an* ***isomorphism*** *in the appropriate sense:*

> *For all elements* $u, v \in X$ *we have* $h(u + v) = h(u) + h(v)$, $h(u \cdot v) = h(u) \cdot h(v)$, *and* $h(u) < h(v)$ *if and only if* $u < v$. *The map* h *sends the zero and unit of* X *to the zero and unit of* Y *respectively*.

The existence of an isomorphism implies that any reasonable mathematical statement about the addition, multiplication and linear ordering of **X** is also true about **Y** and conversely. A proof of Theorem 6 appears in Unit **VIII.** The proof itself is relatively straightforward and elementary but somewhat tedious; however, it is absolutely necessary to establish such a result if we want to talk about **THE** integers.

V. 2 : Finite induction and recursion

(Halmos, §§ 11 – 13; Lipschutz, §§ 1.11, 4.6, 11.1 – 11.7)

Proofs by ***mathematical induction***, or more precisely by ***finite induction***, play an important role in the mathematical sciences. Furthermore, as noted on page 48 of Halmos,

induction is often used not only to prove things but also to define things,

and because of this we shall describe both the proof definition processes explicitly in this section. Objects defined by induction are often said to be defined ***recursively*** (or by ***finite recursion***)***.*** Examples of recursive definitions arise throughout the mathematical sciences, including set theory itself, and therefore we shall describe the procedure fairly explicitly.

Description of the method

Mathematical induction is often a very powerful technique, but it is really more of a method to provide a formal verification of something that is suspected to be true rather than a tool for making intuitive discoveries, but it is absolutely essential. The use of mathematical induction dates back at least to some work of F. Maurolico (1494 – 1575). There are many situations in discrete mathematics where this method is absolutely essential.

Most of the remaining material on mathematical induction is adapted from the following online references:

http://www.cut-the-knot.org/induction.shtml

http://en.wikipedia.org/wiki/Mathematical_induction

IMPORTANT: The similarity between the phrases **"**mathematical induction**"** and **"**inductive reasoning**"** may suggest that the first concept is a form of the second, but ***this is not the case.*** Inductive reasoning is different from deductive reasoning, but ***mathematical induction is actually a form of deductive reasoning.***

Proofs by mathematical induction involve a sequence of statements, one for each nonnegative integer **n** (sometimes it is impractical to start with **n = 0,** and one can begin instead with an arbitrary integer $\mathbf{n_0}$), and it is convenient to let **P(n)** denote the $\mathbf{n^{th}}$

statement. In the original example from the 16th century, **P(n)** was the familiar formula for the sum of the first **n** odd positive integers:

$$\mathbf{1 + 3 + 5 + \ldots + (2n - 1) = n^2}$$

In this case the first statement **P(1)** is $\mathbf{1 = 1^2}$, the statement **P(2)** is $\mathbf{1 + 3 = 2^2}$, the statement **P(3)** is $\mathbf{1 + 3 + 5 = 3^2}$, and so on.

The method of proof by mathematical induction has two basic steps:

1. Proving that the first statement $\mathbf{P(n_0)}$ is true.
2. Proving that if **P(k)** is true for some value of **k,** then so is the next statement **P(k + 1).**

In effect, *mathematical induction allows one to prove an* ***infinite list of statements***, say **P(1), P(2), P(3), ... ,** *with an argument that has only* ***finitely many steps****.* It may be helpful to visualize this in terms of the domino effect; if you have a long row of dominoes standing on end, you can be sure of two things:

1. The first domino can be pushed over.
2. Whenever a domino falls, then its next neighbor will also fall.

Under these conditions, we know that ***every one of the dominos in the picture below will eventually fall*** if the first one is nudged down in the right direction***.***

Incidentally, there is there is an animated version with Apple iPods at the following online site:

http://www.hemmy.net/2006/04/30/domino-ipod-commercial/

There are some instances where one uses a variant of the principle of mathematical induction stated above; namely, one replaces the assumption in the second step with a stronger hypothesis that **P(m)** is true for ***all*** $\mathbf{m < k + 1}$ and not just for $\mathbf{m = k}$.

Example of a proof by induction. Here is a proof of the summation formula for the first **n** odd integers. The statement **P(1)** merely asserts that $\mathbf{1 = 1^2}$, and hence it is obviously true. Let's assume we know that **P(k)** is also true for some arbitrary **k,** so that we have the equation $\mathbf{1 + 3 + 5 + \ldots + (2k - 1) = k^2}$. The next step in mathematical induction is to derive **P(k+1)** from **P(k)**. To do this, we note that

$$1 + 3 + \ldots + (2k-1) + (2k+1) = [1 + 3 + \ldots + (2k-1)] + (2k+1)$$
$$= k^2 + (2k+1)$$
$$= (k+1)^2$$

which shows that **P(k+1)** is also true because **2k + 1 = 2(k + 1) – 1.** Therefore **P(n)** is true for all **n** and we have proven the general formula by mathematical induction.

Formally, the difference between mathematical induction and inductive reasoning sis that the latter would check the first few statements, say **P(1), P(2), P(3), P(4),** and then conclude that **P(n)** holds for all **n.** *The inductive step* **"P(k)** implies **P(k+1)"** *is missing.* Needless to say, inductive reasoning does not constitute a proof in the strict sense of deductive logic.

Frequently the verification of the first statement in a proof by induction is fairly easy or even trivial, but ***it is absolutely essential to include an explicit statement about the truth of the initial case,*** and also ***it is important to be sure that the inductive step works for every statement in the sequence.*** If these are not done, the final conclusion may be false and in some cases downright absurd.

Example. (Somewhat more difficult than the others) Consider the following defective **"**proof**"** that *a nonempty finite set (purportedly!) contains as many elements as one of its proper subsets*. This is vacuously true for the empty set, so assume it is true for a set with **k** elements. Let **S** be a set with **k + 1** elements**;** we need to show that some proper subset **T** contains the same number of elements as **S.** Let **T** be obtained from **S** by removing one element, and let **U** be obtained from **T** by removing one element. By the induction assumption we know that **#(T) = #(U),** and since we also know **#(S) = #(T) + 1** and **#(T) = #(U) + 1** we conclude that **#(S) = #(T).** This is a ridiculous conclusion, so the point here is to ask, **"***How did this happen?***"** In fact, ***the inductive step we have given is valid for all values of* k *except for the case* k = 0.** However, when **k = 0** it breaks down because **T** must be the empty set, so it is not possible to construct the subset **U** by removing an element from **T.**

Justification of the method

In fact, there are two versions of proof by induction that are used frequently in the mathematical sciences. We shall state and prove both of them.

Theorem 1. (WEAK PRINCIPLE OF FINITE INDUCTION.) *Suppose that for each nonnegative integer* **n** *we are given a statement* **(S_n)** *such that the statements* **(S_n)** *satisfy the following conditions:*

(i) **(S_0)** *is true***.**

(ii) *For all positive integers* **n,** *if* **(S_{n-1})** *is true***,** *then* **(S_n)** *is true***.**

Then *each of the statements* **(S_n)** *is true***.**

Proof: Let **F** be the set of all **n** such that **(S_n)** is false. We claim that **F** is empty; we shall assume the contrary and derive a contradiction.

If **F** is nonempty, then there is a least **m** such that **(S_m)** is false, and by the first assumption we know that **m** is positive, so that **m –** 1 is nonnegative. By the minimality assumption on **m** we know that **(S_{m-1})** must be true. Therefore the second condition implies that **(S_m)** is true, yielding a contradiction. The problem arises from our assumption that **F** is nonempty, and therefore the latter set must be empty, which means that each of the statements **(S_n)** is true**.**■

Frequently one needs a version of finite induction with a stronger hypothesis.

Thereom 2. (STRONG PRINCIPLE OF FINITE INDUCTION.) *Suppose that for each nonnegative integer* **n** *we are given a statement* **(S_n)** *such that the statements* **(S_n)** *satisfy the following conditions:*

(i) **(S_0)** *is true***.**

(ii) *For all positive integers* **n,** *if* **(S_k)** *is true for all* **k < n,** *then* **(S_n)** *is true***.**

Then *each of the statements* **(S_n)** *is true***.**

Proof: Let **F** be the set of all **n** such that **(S_n)** is false. We claim that **F** is empty; we shall assume the contrary and derive a contradiction.

If **F** is nonempty, then there is a least **m** such that **(S_m)** is false, and by the first assumption we know that **m** is positive, so that the set of all **k** such that **k < m** is nonempty. By the minimality assumption on **m,** we know **(S_k)** is true for all **k < m.** Therefore the second condition implies that **(S_m)** is true, yielding a contradiction. The problem arises from our assumption that **F** is nonempty, and therefore the latter set must be empty, which means that each of the statements **(S_n)** is true**.**■

One important example of a result whose proof requires the Strong rather than the Weak Principle of Finite Induction is the Fundamental Theorem of Arithmetic (see Rosen, Example 14, p. 250). Another example illustrating the use of the Strong Principle of Finite Induction appears at the end of the next section.

Definition by recursion

The basic idea is fairly simple. We begin to define a function by specifying **f(0),** assume we know how to define **f(x)** for **x < n,** and we use this partial function to find **f(n).** Here is a formal statement of this principle:

Theorem 3. (Recursive Definition Theorem.) *Suppose that* **B** *is a set, and suppose also that for each nonnegative integer* **n** *we have a function* **$H : B^{\{0, \dots, n\}} \to B$,** *let* $\mathbf{N}$ *be the nonnegative integers, and let* **$b_0 \in B$.** *Then there is a unique function* **$f : \mathbf{N} \to B$** *such that* **$f(0) = b_0$** *and for all positive* **n** *we have*

$$f(n) = H\big(f \,\big|\, \{0, \dots, n-1\}\big).$$

Proof. We begin by describing the approach to proving the result. The idea for proving existence is to define a sequence of functions **$g_n : \{0, \dots, n-1\} \to B$** which agree on the overlapping subsets; one then constructs a function **f** whose graph is the union of the

graphs of the partial functions. The uniqueness proof will then reduce to proving uniqueness for the restrictions to each subset $\{0, \dots, n-1\}$.

The function $g_0: \{0\} \to B$ is defined by $g_0(0) = b_0$. Once we are given the function $g_n: \{0, \dots, n-1\} \to B$. we define the function $g_{n+1}: \{0, \dots, n\} \to B$ by $g_{n+1}(k) = g_n(k)$ if $k < n$ and $g_{n+1}(n) = H(g_n)$. Let $G_n \subset \{0, \dots, n-1\} \times B$ be the graph of g_n, and let $G \subset \mathbf{N} \times B$ be the union of the subsets G_n.

We claim that for each $x \in \mathbf{N}$ there is a unique $y \in B$ such that $(x, y) \in G$. If true, then this will imply the existence of a function $f: \mathbf{N} \to B$ whose graph is equal to G. Since G is the union of the graphs G_n, this is equivalent to verifying that for all $n > x$ the elements $g_n(x)$ are all equal; note that $g_n(x)$ is only defined for these values of n. We shall prove that $g_{x+m}(x) = g_{x+1}(x)$ for all $m > 1$ by induction on m; by construction we know that $g_n(x) = g_{n+1}(x)$ for n as above. Therefore if $m = 2$ we know that $g_{x+2}(x) = g_{x+1}(x)$, yielding the first step of the inductive proof. If we know the result for m, we can obtain it for $m + 1$ by once again applying the identity $g_n(x) = g_{n+1}(x)$. This proves that G satisfies the required property for the graph of a function from $\mathbf{N}$ to B.

Finally, we need to prove uniqueness. Suppose that f' is an arbitrary function satisfying the given properties, and let f be constructed as in the previous paragraphs. We shall prove that the restrictions of f and f' to each subset $\{0, \dots, n-1\}$ are equal by induction on n. If $n = 1$ then uniqueness follows because the assumptions imply that the values of both f and f' at 0 are equal to b_0. Suppose now that the restrictions of f and f' to the subset $\{0, \dots, n-1\}$ are equal; to prove the inductive step, it will suffice to show that $f(n) = f'(n)$. But this follows from the equalities

$$f(n) = H\left(f \mid \{0, \dots, n-1\}\right) = H\left(f' \mid \{0, \dots, n-1\}\right) = f'(n),$$

where the first equation is true by construction, the second is true by the induction hypothesis, and the third is true by the assumption on f'.■

Typical recursive definitions

In practice, recursive definitions are usually stated in a less formal manner than indicated by the existence and uniqueness result. Probably the best way to illustrate this is to give simple examples as one would see it in a semi – formal mathematical discussion and to analyze it in terms of the formal statement of the Recursive Definition Theorem. We begin with one which arises in numerous contexts.

Solutions to difference equations. Suppose that we are given a sequence of objects (say numbers, vectors, matrices or functions) $a(n)$ in a set A which has a reasonable notion of addition. We would like to create a new sequence $b(n)$ such that for each n the difference between consecutive terms $b(n + 1) – b(n)$ is equal to $a(n)$. Such an equation is often called a *first order difference equation*, and in some respects the theory of solutions to difference equations resembles the theory of solutions to differential equations. In particular, solutions to first order equations generally exist if one properly

specifies an ***initial value*** **b(0)** for the sequence. It should be clear that we can uniquely define **b(n)** by the conditions given here, but we would also like to explain how this fits into the framework of the Recursive Definition Theorem. According to that result, for each **n** we need to define a suitable function $\mathbf{H : A^{\{0, \cdots, n\}} \to A}$, and one simple way of doing so is to take $\mathbf{H(g) = g(n) + a(n)}$. The conditions of the Recursive Definition Theorem then imply that one obtains a unique function **b(n)** satisfying the given conditions.■

Here is a more abstract type of example within set theory itself.

Proposition 4. *Let* **A** *be an infinite subset of the nonnegative integers* $\mathbf{N}$. *Then there is a strictly order – preserving* **1 – 1** *mapping* **f** *from* $\mathbf{N}$ *to* **A.**

***Proof*. (***)** Define the function **f** recursively as follows: Take **f(0)** to be the least element of **A.** Suppose that we have a **1 – 1** strictly order – preserving mapping **f** defined from the finite set **{0, … , n – 1}** to **A.** Since **A** is infinite it follows that the image **f[{0, … , n – 1}]** is a proper subset of **A,** so that its complement is nonempty and there is some element of **A** which is greater than every element in **f[{0, … , n – 1}]**. Take **f(n)** to be the least such element of **A.** We claim the latter recursively defines **f;** this will be discussed further in the next paragraph. To complete the recursive step in the argument, we need to show that the newly extended function on **{0, … , n}** is also strictly order – preserving. This follows because **f** is already known is strictly order – preserving on **{0, … , n – 1}** and **f(n) > f(j)** for all **j < n.**■

Finally, to end the argument we need to show that the globally defined function **f** is also strictly order – preserving; if **x < y,** then **x** and **y** belong to **{0, … , y }** and since the restriction of f to the latter is strictly order – preserving it follows that **f(x) < f(y)** as required.■

We now need to analyze the construction of **f** and see how it can be formalized to fulfill all the conditions in the Recursive Definition Theorem. The main thing that does not appear in our discussion is a complete means for defining an element of **A** given an arbitrary mapping from **{0, … , n – 1}** to **A.** In our recursive definition we assumed that the function defined on the finite piece of $\mathbf{N}$ was strictly increasing, and at each step we showed that the extended function was also strictly increasing. Strictly speaking we need to define an element of **A** even for partial functions that are not strictly increasing, but the precise nature of these definitions is unimportant because we shall never need the definitions for functions that are not strictly increasing. Formally one can define the function for such irrelevant sequences by some simple arbitrary device. For example, in our setting we can simply take the value for one of the "irrelevant" partial functions to be the unique least element of **A.** If there are ever circumstances in which it is not clear how to define a value for **"**irrelevant**"** partial functions, one standard way is to work inside the slightly larger set $\mathbf{A \cup \{A\}}$ (recall this properly contains **A**) and simply define the value of the irrelevant functions to be the extra element **A.**■

V . 3 : Finite sets

(Halmos, §§ 11 – 13; Lipschutz, §§ 1.8, 3.2)

Courses in discrete structures and combinatorics study questions about finite sets extensively. In this section we shall develop a few basic aspects of this topic that will be needed or useful later in the course.

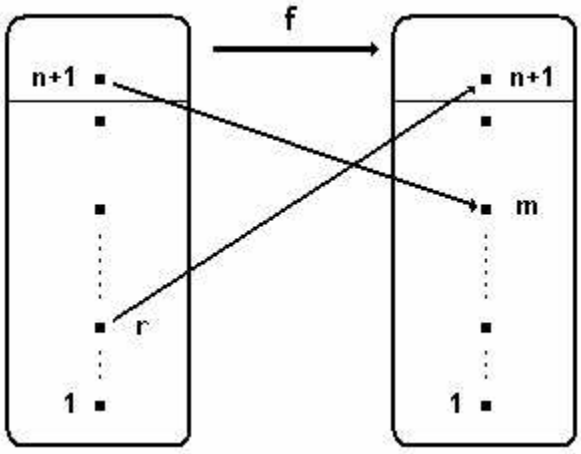
For our purposes a set **X** will be said to be ***finite*** if there is a positive integer n such that there is a **1 – 1** correspondence from **X** to **{1, … , n}.**

The pigeonhole principle

Experience indicates that if **X** is a finite set, then there is no **1 – 1** correspondence between **X** and a proper subset of itself. Our first objective is to give a rigorous proof of this basic fact.

Theorem 1. *Suppose that* **A** *is a finite set,* **B** *is a subset of* **A,** *and* **f : A → A** *is a* **1 – 1** *mapping with* **f[A] = B.** *Then* **B = A.**

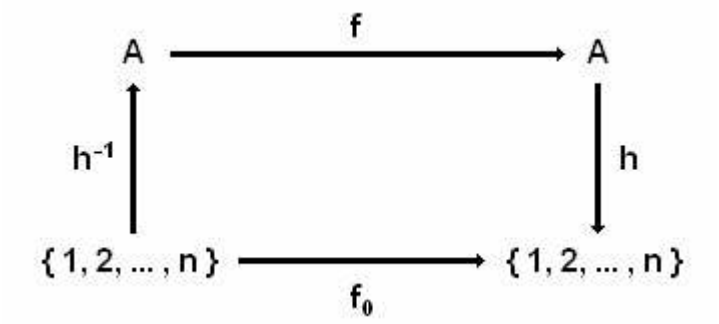
Proof. **(**)** We shall first consider the special case where **A = {1, 2, … , n}** and proceed by induction on **n.** If **n = 1** then the result is trivial. Suppose it is true for **n** and proceed to the case of **n + 1.** Call this set **A** and let **C** be the set of the first n elements. If **f[C]** is contained in **C** then by induction **f[C] = C** and we must then have **f (n + 1) = n + 1.** Suppose that **f[C]** is not contained in **C.** Since **f** is **1 – 1** it follows that **f(n + 1)** cannot be equal to **n + 1,** and it also follows that **f(r) = n + 1** for some **r < n + 1.** Suppose that **f(n + 1) = m < n + 1.** Define a new function **g : C → C** by setting **g(r) = m** and **g(k) = f(k)** otherwise.



CLAIM: **g** is a **1 – 1** mapping. Suppose that **g(i) = g(j) .** Since **f = g** for **x ≠ r** it follows that one of **i** and **j** must be equal to **r,** so say **j = r.** Then we have **g(i) = f(i)** and **g(r) = m = f(n + 1).** Since **i < n + 1** and **f** is **1 – 1** it follows that **g(i) ≠ g(r)** and consequently **g** is **1 – 1** mapping. By induction **g** is onto.

We shall use the preceding paragraph to prove that **f** is onto. If **y < n + 1** then **y = g(z)** for some **z ∈ C,** and since **g(z) = f(w)** for some **w,** it follows that the image of **f** contains all of **C.** Since we have shown that **n + 1 = f(r)** it follows that the image of **f** contains all of **A,** provided that **A = {1, 2, … , n}.**

To prove the general case, let **A** be a finite set with **n** elements, so that there is a **1 – 1** onto mapping h from **A** to **{1, 2, … , n}.** Given a **1 – 1** mapping **f : A → A** let $\mathbf{f_0}$ be the conjugate mapping from **{1, 2, … , n}** to itself defined by $\mathbf{f_0 = h\,f\,h^{-1}}$.



We claim that $\mathbf{f_0}$ is a **1 – 1** mapping. Suppose that $\mathbf{f_0(x) = f_0(y)}$; by definition of $\mathbf{f_0}$ we have $\mathbf{h\,f\,h^{-1}(x) = h\,f\,h^{-1}(y)}$. Since the mappings **h, f** and $\mathbf{h^{-1}}$ are all **1 – 1** we can successively use the injectivity of **h** to conclude that that $\mathbf{f\,h^{-1}(x) = f\,h^{-1}(y)}$, the injectivity of **f** to conclude that that $\mathbf{h^{-1}(x) = h^{-1}(y)}$, and the injectivity of $\mathbf{h^{-1}}$ to conclude that that **x = y.** Therefore $\mathbf{f_0}$ is **1 – 1,** and therefore the preceding argument shows that $\mathbf{f_0}$ is also onto.

To prove that **f** is onto, suppose that **z ∈ A,** and let **w = h(z).** By the special case established above, it follows that $\mathbf{w = f_0(v)}$ for some **v,** so that

$$\mathbf{z = h^{-1}(w) = h^{-1}(f_0(v)) = h^{-1}[h\,f\,h^{-1}(v)] = f\,h^{-1}(v)}$$

which implies that **f** must be onto**.**■

Counting elements of finite sets

If **X** is a finite set, there is a unique natural number n such that there is a **1 – 1** correspondence between **X** and **{1, … , n};** uniqueness follows from the previous discussion in this section**.** Following standard practice we say that **X** *has* **n** *elements* if this is the case, and we write **|X| = n.**

Our first result looks obvious, but we still need to prove it.

Proposition 2. ***If*** **B** *is a subset of* **A,** *then* **|B| ≤ |A|.**

Proof. We proceed by induction on **n = |A|.** If **n = 0** then the result is trivial because **A** is empty and hence **B** is also empty, so we have **|B| = 0 ≤ 0 = |A|.** Suppose the result is known for **|A| = k,** and consider the case where **|A| = k + 1.**

Let **f : {1, … , k + 1} → A** be a **1 – 1** correspondence, and let **B** be a subset of **A.** Let **C** be the subset of **A** obtained by removing **f(k + 1),** and let **D** denote the intersection of **B** and **C.** By construction **|C| = k** and **D** is a subset of **C**, and therefore by the induction hypotheses we have **|D| ≤ k.** There are now two cases depending upon whether or not **f(k + 1)** belongs to **B.** If so, then **D = B** and hence **|B| = |D| ≤ k < |A|.** If not, then **B = D ∪ {f(k + 1)}** and hence **|B| = |D| + 1 ≤ k + 1 = |A|.** This completes the proof of the inductive step**.**■

Corollary 3. ***If*** **B** *is a proper subset of* **A,** *then* **|B| < |A|.**

This follows immediately by combining the previous two results**.**■

The following basic formulas for counting elements of finite sets have important counterparts for infinite sets that will be discussed in Unit **V**.

Theorem 4. *Let* **A** *and* **B** *be sets with* **n** *and* **m** *elements respectively.*

1. *If* **A** *and* **B** *are disjoint, then* **|A ∪ B| = n + m.**
2. *For arbitrary finite sets* **A** *and* **B** *we have* **|A × B| = n · m.**
3. *If* **A** *and* **B** *are arbitrary finite sets and* $\mathbf{B}^{\mathbf{A}}$ *is the set of functions from* **A** *to* **B,** *then we have* $|\mathbf{B}^{\mathbf{A}}| = \mathbf{m}^{\mathbf{n}}$.

Proof. All of the proofs proceed by induction on **n = |A|.**

Verification of **(1):** If **n** = 0 then **A ∪ B = B** and therefore **m = |B| = |A ∪ B| = 0 + m.** Suppose the result is true for **n = k,** suppose also that **|A| = k + 1,** suppose we have a **1 – 1** correspondence between **A** and **{1, … , k + 1 },** let **C ⊂ A** correspond to **{1, … , k },** and let **z** be the unique element of **A** such that **A = C ∪ { z }.** By the induction hypothesis there is a **1 – 1** correspondence **g : {1, … , k + m } → C ∪ B.** Define a function **f : {1, … , k + m + 1 } → A ∪ B** such that **f = g** on the subset **{1, … , k + m }** and **f(k ++ m + 1) = z.**

We claim that **f** is **1 – 1** and onto. Suppose that **f(x) = f(y).** If neither **x** nor **y** is equal to **k + m + 1,** then **g(x) = f(x)** and **g(y) = f(y),** and since **g** is **1 – 1** it follows that **x = y.** Suppose now that, say, **x = k + m + 1.** Then **f(x) = z.** On the other hand, if **f(y) = z** then the only possibility is **k + m + 1,** and hence **x = y** in this case too**.** Therefore **f** is a **1 – 1** mapping. Suppose now that **w** belongs to **A ∪ B;** we need to show that **w** lies in the image of **f.** If **w** is not equal to **z** then we have **w = g(j)** for some **j < k + m + 1,** and thus we also have **w = f(j)** for the same choice of **j.** On the other hand, if **w = z**

then we have $z = f(k + m + 1)$. Therefore f is $1 - 1$ and onto, so this completes the proof of the inductive step.

Verification of **(2):** If $n = 0$ then $A \times B = \emptyset$ and therefore $0 = |\emptyset| = |\emptyset \times B| = 0 \cdot m$. Suppose once again the result is known to be true for $n = k$, suppose also that $|A| = k + 1$, suppose we have a $1 - 1$ correspondence between A and $\{1, \ldots, k + 1\}$, let $C \subset A$ correspond to $\{1, \ldots, k\}$, and let z be the unique element of A such that $A = C \cup \{z\}$. By the induction hypothesis we know there is a $1 - 1$ correspondence $g : \{1, \ldots, k \cdot m\} \rightarrow C \times B$. Let $h : \{1, \ldots, m\} \rightarrow B$ be a $1 - 1$ correspondence; Define $f : \{1, \ldots, k \cdot (m + 1)\} \rightarrow A \times B$ such that $f = g$ on $\{1, \ldots, k \cdot m\}$ and

$$f(k \cdot m + j) = (z, h(j))$$

for $j = 1, \ldots, m$.

We claim that f is $1 - 1$ and onto. If neither x nor y greater than $k \cdot m$, then $g(x) = f(x)$ and $g(y) = f(y)$, and since g is $1 - 1$ it follows that $x = y$. Suppose now that, say, we have $x > k \cdot m$. Then $f(x) = (z, b)$ for some b in B, and hence $f(y) = (z, b)$. By construction, the only way this can happen is if y is also greater than $k \cdot m$. Therefore we may write $x = k \cdot m + i$ and $y = k \cdot m + j$ for some integers i and j between 1 and m. Since $f(x) = f(y)$, it follows from the construction that $h(i) = h(j) = b$, and the latter in turn implies that $i = j$. Therefore we have $x = y$ and hence f is $1 - 1$. Suppose now that w belongs to $A \times B$; we need to show that w lies in the image of f. If the first coordinate of w is not equal to z then in fact we have $w = g(j)$ for some $j \leq k \cdot m$, and thus we also have $w = f(j)$ for the same choice of j. On the other hand, if the first coordinate of w is equal to z, then write $w = (z, b)$. By construction $b = h(j)$ for some j, and it then follows that $w = (z, b) = f(k \cdot m + j)$. Therefore f is $1 - 1$ and onto, so this completes the proof of the inductive step.

Verification of **(3): (***)** If $n = 0$ then there is a unique function from $A = \emptyset$ to B; namely, the function whose graph is the empty set. Therefore we have $|B^A| = |B^{\emptyset}| = 1 = m^0$. Suppose again the result is known to be true for $n = k$, suppose also $|A| = k + 1$, assume we have a $1 - 1$ correspondence between A and $\{1, \ldots, k + 1\}$, let $C \subset A$ correspond to $\{1, \ldots, k\}$, and let z be the element of A such that $A = C \cup \{z\}$. By the induction hypothesis we know there is a $1 - 1$ correspondence $g : \{1, \ldots, m^k\} \rightarrow B^C$.

By the result in the preceding part of the theorem, it will suffice to construct a $1 - 1$ correspondence between B^A and $B^C \times A$, for then one obtains the equations

$$|B^A| = |B^C \times A| = m^k \cdot m = m^{k+1}$$

which is what we need to prove in order to verify the inductive step. Suppose now that we are given a function $\mathbf{u : A \to B}$. Consider the mapping $\mathbf{\Omega: B^A \to B^C \times A}$ defined by $\mathbf{\Omega\,(u) = (u|C, u(z))}$; we claim that $\mathbf{\Omega}$ is **1 – 1** and onto**.**

Suppose first that $\mathbf{\Omega\,(u) = \Omega\,(v)}$. Then by construction we have $\mathbf{u|C = v|C}$ and $\mathbf{u(z) = v(z)}$. Combining these with $\mathbf{A = C \cup \{z\}}$, we see that $\mathbf{u(t) = v(t)}$ for all $\mathbf{t \in A}$, and therefore we must have $\mathbf{u = v}$. Therefore $\mathbf{\Omega}$ is **1 – 1**. Suppose now that we are given an arbitrary pair **(g, b).** Then there is a function **f** such that $\mathbf{f(t) = g(t)}$ for all $\mathbf{t \in C}$ and $\mathbf{f(z) = b}$; therefore $\mathbf{\Omega}$ is onto as required**.**■

The result in the third part of the theorem illustrates one important reason for using $\mathbf{B^A}$ to denote the set of all functions from **A** to **B.**

Boolean algebras of subsets

We shall prove a result relating the properties of finite sets to the Strong Principle of Finite Induction that was formulated in the preceding section.

Definition. Given a set **A,** let **P(A)** be the set of all subsets with the algebraic operations of union, intersection, and relative complementation. A ***Boolean subalgebra*** of **P(A)** is a subset $\mathbf{S \subset P(A)}$ such that **S** is contained in **P(A),** it contains **A** and the empty set, it is closed under taking finite unions and intersections, and it is also closed under taking relative complements.

The simplest examples of Boolean subalgebras are given by equivalence relations. Specifically, if ***R*** is an equivalence relation on **A** and **S** is the family of all subsets that are unions of ***R*** **–** equivalence classes, then it is a routine exercise to verify that **S** is a Boolean subalgebra of **P(A).** The result below shows that all Boolean subalgebras have this form if **A** is a finite set.

Proposition 5. *Let* **A** *be a set, and let* **S** *be a Boolean subalgebra of* **P(A).** *Then there is an equivalence relation such that the subsets of* **S** *are the unions of* ***R*** *– equivalence classes.*

Proof. **(**)** A subset $\mathbf{Y \in S}$ is said to be *atomic* for **S** if it is nonempty and there are no nonempty subsets $\mathbf{X \in S}$ that are properly contained in **Y.** We shall prove the proposition by verifying the following two assertions:

1. Every subset of **S** is a union of atomic subsets.
2. Two atomic subsets of **S** are either disjoint or identical.

By previous results, it will follow that the atomic subsets are the equivalence classes for some equivalence relation on **A.**

We shall prove the first statement by induction on **|A|.** If **A** has 0 or 1 element, then **S** must be equal to **P(A),** and for any finite set **A** a subset is atomic for **P(A)** if and only if it contains exactly one element. Suppose now that the result is true for all sets **B** such that $\mathbf{|B| < |A|}$. There are two cases depending upon whether **S** contains a nonempty

proper subset. If it does not, then **S** only consists of **A** and the empty set, and therefore **A** must be atomic. On the other hand, if **S** contains a nonempty proper subset **C,** then it also contains **A – C = D,** and **D** is also a nonempty proper subset. It follows that both **|C|** and **|D|** are strictly less than **|A|.**

Let **S|C** and **S|D** denote the set of all subsets in **S** that are contained in **C** and **D** respectively. We claim that these are Boolean subalgebras of **P(C)** and **P(D)** respectively; by our hypotheses we know that the empty set lies in both, that **C** and **D** are contained in **S|C** and **S|D** respectively, and that both of the latter are closed under finite unions or intersections (because the same is true for **S**). To show these families are closed under relative complementation, note that if **X** lies in **S|C** or then

$$C - X = C \cap A - X$$

shows that **C – X** also belongs to **S|C,** and similar considerations show that if **X** lies in **S|D** then **D – X** also lies in **S|D.** By the induction hypothesis it follows **C** and **D** are unions of atomic subsets, and therefore the same is true for $A = C \cup D$.

To complete the proof, we need to prove the second assertion given above; specifically, we need to prove that two atomic subsets are either disjoint or identical. But if **X** and **Y** are atomic subsets of **S,** then the Boolean subalgebra condition implies that $X \cap Y$ also belongs to **S.** Since it is contained in the minimal nonempty subsets **X** and **Y,** either the intersection is empty or else if it is nonempty then it must be equal to both **X** and **Y.**■

An abstract ***Boolean algebra*** is an algebraic system consisting of a set **A** together with three operations; namely, two binary operations $\cup$, $\cap$ and one unary operation (sending an element **x** to x') which have the formal properties of unions, intersections, and complements. Chapter 11 of Lipschutz contains further material on such structures, with emphasis on computational techniques. An entirely different perspective on Boolean algebras, which reflects their role in modern pure mathematics, is contained in the following reference (which is written at the graduate level):

> P. R. Halmos, ***Lectures on Boolean algebras*** (Originally published as Van Nostrand Math. Studies, No. 1). Springer – Verlag, New York, 1974. ISBN: 0 – 387 – 90094 – 2.

V . 4 : The real numbers

(Lipschutz, §§ 2.2 – 2.6, 7.7)

Following the approach of Section 1, we shall give an axiomatic description of the real numbers in terms of their basic properties. Many of these properties are also properties of the integers, but there are also some important new ones.

Basic rules for addition and multiplication. Formally, these are the conditions defining an abstract type of mathematical system known as a ***field.*** The first five of these are the previously introduced properties for a commutative ring with unit, and the final one reflects an important difference between the integers in the real numbers; in the latter one can divide by nonzero numbers, but usually this is not possible within the integers.

FIRST AXIOM GROUP FOR THE REAL NUMBERS. *The real numbers are a set* $\mathbf{R}$, *and they have binary operations* $A : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$, *which is normally expressed in the form* $A(u, v) = u + v$, *and* $M : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$, *which is normally expressed in the form* $M(u, v) = u\,v$ *or* $u \cdot v$ *or* $u \times v$, *such that the following algebraic conditions are satisfied:*

1. (Associative Laws). *For all* a, b, c *in* $\mathbf{R}$, $(a + b) + c = a + (b + c)$ *and* $(a\,b)\,c = a\,(b\,c)$.
2. (Commutative Laws). *For all* a, b *in* $\mathbf{R}$, $a + b = b + a$ *and* $a\,b = b\,a$.
3. (Distributive Law). *For all* a, b, c *in* $\mathbf{R}$, $a\,(b + c) = a\,b + a\,c$.
4. (Existence of 0 and 1). *There are distinct elements* $0, 1$ *in* $\mathbf{R}$ *such that for all* a *we have* $a + 0 = a$, $a \cdot 0 = 0$ *and* $a \cdot 1 = a$.
5. (Existence of negatives or additive inverses). *For each* a *in* $\mathbf{R}$ *there is an element* $-a$ *in* $\mathbf{R}$ *such that* $a + (-a) = 0$.
6. (Existence of reciprocals or multiplicative inverses). *For each* a *in* $\mathbf{R}$ *there is an element* a^{-1} *in* $\mathbf{R}$ *such that* $a \cdot a^{-1} = 1$.

Basic rules for ordering. These are the same as the ordering properties for the integers. When combined with the previous conditions, these yield a type of mathematical system known as an ***ordered field.***

SECOND AXIOM GROUP FOR THE REAL NUMBERS. *There is a linear ordering on* $\mathbf{R}$ *such that the following hold:*

1. *If* $a > 0$ *and* $b > 0$, *then* $a + b > 0$ *and* $a\,b > 0$.
2. *For all* a, b *in* $\mathbf{R}$, *we have* $a > b$ *if and only if* $a - b > 0$.

Basic rules for completeness of the ordering. The ordering on the real numbers satisfies an additional fundamental condition called the Dedekind completeness axiom

after R. Dedekind (1831 – 1916), who formulated this property. In order to state this axiom it is necessary to introduce some additional standard definitions.

Definitions. Let **(L, ≤)** be a linearly ordered set, and let **A** be a subset of **L.** An element **x ∈ L** is said to be an *upper bound* for **A** in **L** if for each **a ∈ A** we have **a ≤ x;** note that the definition contains no information on whether **x** belongs to **L.** An upper bound **x** is said to be a ***least** upper bound* (for **A** in **L**) if for **every** upper bound **y** for **A** we have **x ≤ y.**

Proposition 1. *If* **x** *and* **z** *are least upper bounds for a subset* **A** *as above, then* **x = z.**

Proof. Since **x** is a least upper bound and **z** is an upper bound, we have **x ≤ z.** Similarly, since **x** is a least upper bound and **z** is an upper bound, we have **z ≤ x.** Combining these, we conclude that **x = z.**■

If a set **A** has a least upper bound **x,** then we often write **x = L.U.B.(A)** or **x = sup(A).** The symbolism **sup** is an abbreviation for the quasi – Latin term for the least upper bound; namely, the *supremum***.**

There are dual notions for the reverse ordering on a linearly ordered set. Specifically, if **B** is a subset of **L** then a *lower bound* is a number **y** such that **y ≤ b** for all **b ∈ B;** note that the definition contains no information on whether **x** belongs to **L.** A ***greatest** lower bound* is a lower bound **y** such that **x ≤ y** for every lower bound **x.** It follows as above that if a greatest lower bound exist then it is unique. If a set **B** has a greatest lower bound **y,** then we often write **y = G.L.B.(B)** or **x = inf(B).** The symbolism **inf** is an abbreviation for the quasi – Latin term for the greatest lower bound; namely, the *infimum***.**

Notice that the least upper bound is a lower bound for the set of upper bounds and a greatest lower bound is an upper bound for the set of lower bounds.

DEDEKIND COMPLETENESS AXIOM FOR THE REAL NUMBERS. *If* **A** *is a nonempty subset of* **R** *which has an upper bound, then* **A** *has a least upper bound.*

Corollary 2. *If* **B** *is a nonempty subset of* **R** *which has a lower bound, then* **B** *has a greatest lower bound.*

The proof of this corollary depends upon the following elementary observation.

Lemma 3. If **x** *and* **y** *are distinct real numbers and* **x < y,** *then* **– y < – x.**

***Proof of Lemma* 3.** By the axioms we know that **y – x >** 0. However, the left hand side is equal to **– (x – y),** and therefore we have **– y < – x** as required**.**

Proof of Corollary 2. Let **A** be the set of all ***negatives*** of elements of **B.** Then the assumption that **B** has a lower bound implies that A has an upper bound, and hence by the Dedekind Completeness Axiom the set **A** has a least upper bound, say **u.** We claim that **– u** is a greatest lower bound for **B.** First of all, the lemma implies that since **u** is an

upper bound for **A** the element **– u** is a lower bound for **B.** Suppose now that **v** is an arbitrary lower bound for **B.** Then the lemma implies that **– v** is an upper bound for **A,** and therefore since **u** is a least upper bound it follows that **u ≤ – v.** Therefore the lemma implies that **v ≤ – u,** so that **– u** is a greatest lower bound for **B.**

Remarks. (1) If a set **A** does not have an upper bound, then this is often expressed symbolically as **sup(A) = + ∞.** Notice that in this context the symbol **"∞"** is not a number, but rather It is a short way to say that there is no number which is an upper bound for **A.** Similarly, if **B** has no lower bound, then **inf(B) = – ∞.**

(2) Two curious implications of the preceding notation are the "paradoxical" identities **sup(Ø) = – ∞** and **inf(Ø) = + ∞.** To see the first of these, notice that every **M ∈ R** is an upper bound for the empty set. This is because, given **M,** there is no **x ∈ Ø** such that **x ≥ M.** Thus, the set of upper bounds for **Ø** has no lower bound. To see the second, notice that every **M ∈ R** is a lower bound for the empty set. This is because, given **M,** there is no **x ∈ Ø** such that **x ≤ M.** Thus, the set of lower bounds for **Ø** has no upper bound. — In contrast to this result, if **A** is a nonempty subset of **L** then we always have **inf(A) ≤ sup(A)** if we agree that **– ∞** is less than every real number and **+ ∞** is greater than every real number (and of course **– ∞ < + ∞**)**.** In fact, if **x** is an arbitrary element of **A** then we have **inf(A) ≤ x ≤ sup(A).**

Clearly we want the real number system to contain the integers or a system equivalent to the integers. Here is one way of formulating this:

INTEGRAL COMPATIBILITY AXIOM. *There Is a* **1 – 1** *mapping* **J** *from the integers* $\mathbf{Z}$ *to the real numbers* $\mathbf{R}$ *with the following properties:*

1. **J** *maps the zero element of* $\mathbf{Z}$ *to the zero element of* $\mathbf{R}$**.**
2. **J** *maps the multiplicative unit of* $\mathbf{Z}$ *to the multiplicative unit of* $\mathbf{R}$**.**
3. *For all integers* **x** *and* **y,** *we have* **J(x + y) = J(x) + J(y) .**
4. *For all integers* **x** *and* **y,** *we have* **J(x y) = J(x) J(y) .**
5. *For all integers* **x** *and* **y,** *we have* **J(x) < J(y)** *if and only if* **x < y.**

Of course, the real numbers are also supposed to contain the ***rational numbers***, which are all numbers expressible as quotients of integers **a/b** where **b** is nonzero. Usually the rational numbers are denoted by **Q** (presumably for *quotients*). Note that the rational numbers clear satisfy all the properties of the real numbers aside from the Dedekind Completeness Property. Strictly speaking we cannot say formally that this property fails for the rational numbers, but if we grant that there should be a real number that is the square root of 2, then an argument going back to the ancient Greeks (possibly even to the Pythagoreans in the 6th century B. C. E.) implies that some real numbers, including the square root of 2, are not rational. Incidentally, the classical number π**,** denoting the ration of a circle's circumference to its diameter, is also irrational, but this was first established in relatively modern times by J. H. Lambert (1728 – 1777); it should be noted that the first use of the symbol π for the number was due to W. Jones (1675 – 1749) in 1706. As noted at the beginning of these notes, one of the important features of set theory is that it provided a mathematically sound way of describing such irrational numbers as well as their relation to the rationals, thus completing the answer to a question that first arose in ancient Greek mathematics.

Uniqueness of the real numbers

We have given a list of properties that the real number system is assumed to satisfy. In the next section we shall prove that any system satisfying these properties also has many other familiar properties we expect from real numbers. However, as in Section 1 (and the discussion at the beginning of this unit), we would like to say that if we are given two systems which satisfy our axioms for the real numbers, then they are the same for all mathematical purposes; in the terminology of Section 1, the mathematical way of saying this is that there is an ***isomorphism*** between the two systems. Here is the formal statement.

Theorem 4. *Suppose that* **X** *and* **Y** *are sets with notions of addition, multiplication, ordering and "integers" which satisfy all the conditions for the real number system. Then there exists a* ***unique*** 1 – 1 *correspondence from* **h** *from* **X** *to* **Y** *that is an* ***isomorphism*** *in the sense of Section* **1:** *For all elements* $\mathbf{u, v \in X}$ *we have* $\mathbf{h(u + v) = h(u) + h(v)}$, $\mathbf{h(u \cdot v) = h(u) \cdot h(v)}$, *and* $\mathbf{h(u) < h(v)}$ *if and only if* $\mathbf{u < v}$. *Furthermore, the map* **h** *sends the zero and unit of* **X** *to the zero and unit of* **Y**, *and accordingly it also sends the "integers" in* **X** *to the "integers" in* **Y** *(and similarly for the "rationals" in the appropriate systems)*.

By the "integers" in **X** and **Y** we mean the subsets described in the integral compatibility axiom, and the "rationals" denote the smallest subsets that are closed under addition, subtraction and multiplication and contain both the integers and the reciprocals of nonzero integers.

As before, the existence of an isomorphism has the following implication:

> *Every true reasonable mathematical statement about the addition, multiplication and linear ordering of* **X** *is also true about* **Y** *and conversely.*

A proof of Theorem 4 appears in Unit **VIII.** The proof itself is relatively straightforward and elementary but somewhat tedious; however, it is absolutely necessary to establish such a result if we want to talk about **THE** real number system.

V.5: Familiar properties of the real numbers

(Lipschutz, §§ 2.2, 4.5)

The crucial justification for the Dedekind approach to the real number system is that it yields all the known properties of the real numbers. In this section we shall consider a few important examples:

Density of the rationals. *If* **x** *and* **y** *are rational numbers such that* $x < y$, *then there is a rational number* **q** *such that* $x < q < y$.

Existence of positive n^{th} roots. *If* **x** *is a positive real number and* **n** *is a positive integer, then there is a unique positive real number* **y** *such that* $y^n = x$.

Base 10 and decimal expansions. *The axioms for real numbers developed above are adequate to prove all the familiar facts about base* 10 *and infinite decimal expansions.*

Any reasonable mathematical theory of the real numbers should yield all of these facts in a fairly straightforward fashion.

As we have already noted, it is possible to go much further and develop everything done in calculus courses (and beyond**!**) using the given axioms for the real number system. Deriving all these fundamental results in calculus from our axioms is beyond the scope of these notes and this course (it properly belongs to courses on functions of a real variable); one standard reference which contains all the details is the following classic text:

> W. Rudin, ***Principles of Mathematical Analysis*** (3rd Ed.), International Series in Pure and Applied Mathematics). McGraw-Hill, New York, 1976. ISBN: 0 – 07 – 054235 – X.

We shall refer to Rudin at various points in this section as needed.

Density of the rational numbers

Even though numbers like the square root of 2 are irrational, it is still possible to approximate them to any desired degree of accuracy by rational numbers. This fact was understood intuitively in most if not all ancient civilizations, and it was formalized and generalized by Eudoxus of Cnidus in the 4th century B. C. E. Subsequently, Euclid's ***Elements*** used one formulation of this principle as the basis for its theory of geometric proportions. The first step in proving this rigorously for our formulation of the real numbers is named after Archimedes, who used it extensively in his writings during the 3rd century B. C. E., but it had also been known to Eudoxus and other earlier Greek mathematicians.

Theorem 1. (Archimedean Law) *If* **a** *and* **b** *are positive real numbers, then there is a positive integer* **n** *such that* $na > b$.

By the well – ordering of the positive integers, there will be a ***(unique) minimal value*** of **n** for which this holds.

Proof. Assume the conclusion is false, so that for every positive integer **n** we have the inequality $na \leq b$. If **A** denotes the set of all products **na,** where **n** is a positive integer, it follows that **b** is an upper bound for A, and by the Dedekind Completeness Property the set **A** must have a least upper bound, which we shall call **c.** Since we have

$\mathbf{ma \leq c}$ for every positive integer **m,** if we set $\mathbf{m = n + 1}$ we see that $\mathbf{(n + 1)a \leq c}$ for every positive integer **n.** If we subtract **a** from both sides, we see that $\mathbf{na \leq c - a}$ for every positive integer **n.** But this implies that $\mathbf{c - a}$ is also an upper bound for **A,** and we had chosen **c** to be the least upper bound, so we have obtained a contradiction. The latter arises from our assumption that **b** was an upper bound for **A,** and therefore this must be false, which means that the conclusion of the theorem must be true**.**■

With this result at our disposal, we can prove the density of the rationals.

Theorem 2. *If* **a** *and* **b** *are positive real numbers such that* $\mathbf{a < b,}$ *then there is a rational number* **q** *such that* $\mathbf{a < q < b.}$

One can easily obtain the same result when **a** and **b** are not both positive from the theorem as follows. If **a** is negative and **b** is positive, then we may simply take $\mathbf{q = 0.}$ On the other hand if $\mathbf{a < b <} 0$ then we have $\mathbf{-a > -b > 0,}$ and therefore by the theorem there is a rational number **s** such that $\mathbf{-b < s < -a.}$ If we take $\mathbf{q = -s,}$ then it will follow that $\mathbf{a < q < b.}$

The proof of the theorem requires the following elementary facts.

Proposition 3. *If* **x** *is a positive real number, then its reciprocal* $\mathbf{x^{-1}}$ *is also positive.*

Proposition 4. *If* **x** *and* **y** *are positive real numbers such that* $\mathbf{x < y}$*, then their reciprocals satisfy the reverse inequalities* $\mathbf{x^{-1} > y^{-1}.}$

***Proof of Proposition* 3.** Suppose this is false, so that $\mathbf{x^{-1}}$ is negative. Then

$$\mathbf{-x^{-1} = (-1)x^{-1}}$$

Is positive, and therefore so is

$$\mathbf{-1 = x(-x^{-1}).}$$

Since the number –1 is not positive we have a contradiction, which arises from our assumption that the reciprocal of **x** was negative, and therefore it follows that the reciprocal of **x** must be positive as claimed**.**■

***Proof of Proposition* 4.** Suppose this is false, so that we have either $\mathbf{x^{-1} = y^{-1}}$ or else $\mathbf{x^{-1} < y^{-1}.}$ The first of these implies that

$$\mathbf{y = xx^{-1}y = xy^{-1}y = x}$$

which contradicts our assumption that $\mathbf{x < y.}$ To prove that $\mathbf{x^{-1} < y^{-1}}$ is impossible, note first that if positive real numbers satisfy $\mathbf{a < b}$ and $\mathbf{c < d}$ then

$$\mathbf{bd - ac = (bd - ad) + (ad - ac) = (b - a)d + a(d - c) > 0}$$

and hence $\mathbf{bd > ac.}$ Therefore $\mathbf{x < y}$ and $\mathbf{x^{-1} < y^{-1}}$ combine to imply that $\mathbf{xx^{-1}}$ is strictly less than $\mathbf{yy^{-1}.}$ However, each of the preceding two products is equal to 1 and thus we have a contradiction. Thus $\mathbf{x^{-1} < y^{-1}}$ is impossible, and the only remaining possibility is the one stated in the conclusion of the result**.**■

Proof of Theorem 2. By Proposition 3, if **a** is positive then so is its reciprocal, and thus the Archimedean law implies there is some positive integer **p** such that $\mathbf{p = p\,1 > a^{-1}}$**.** Taking reciprocals, we find that **0 < 1/p < a.** The Archimedean Law similarly implies the existence of some positive integer **r** such that **0 < 1/r < b – a.** If we take **m** to be the larger of **p** and **r,** then it will follow that both **0 < 1/m < a** and **0 < 1/m < b – a.** Applying the Archimedean Law one more time, we can find a ***first*** positive integer **n** such that **a < n/m.** If we also have **n/m < b,** then we may take **q = n / m** and the proof will be complete. To see that **n/m < b,** proceed as follows. Since **n** is the first positive integer such that **a < n/m,** it follows that **(n – 1)/m ≤ a,** and therefore we also have

$$\mathbf{n/m = (\,(n-1)/m\,) + (1/m) < a + (b-a) = b}$$

which is exactly what we needed**.**■

A statement and proof of the Condition of Eudoxus are given in the online document

http://math.ucr.edu/~res/math153/history03a.pdf

and the application of the condition to proportionality questions as in Euclid's ***Elements*** appears in the following related document:

http://math.ucr.edu/~res/math153/history03b.pdf

Existence of positive $\mathbf{n^{th}}$ *roots*

The main result is exactly what we would expect:

Theorem 5. *If* **r** *is a positive real number and* **n > 1** *is an integer***,** *then there is a unique positive real number* **y** *such that* $\mathbf{y^n = r.}$

The idea of the proof is simple. Given **r** and **n,** consider the set **A** of all positive real numbers **y** such that $\mathbf{y^n < r.}$ In order to prove the theorem, it will suffice to establish the following two points.

1. The set **A** has an upper bound (hence a least upper bound).
2. If **z** is the least upper bound of **A,** then $\mathbf{z^n = r.}$

Proof of the first step. There are two separate cases, depending upon whether **r ≤ 1** or **r > 1.** In the first case, if **z** belongs to **A** then we also have **y ≤ 1**, for **y > 1** implies that $\mathbf{z^n > 1.}$ Suppose now that **r > 1,** and let **n** be an integer such that **n > r.** We claim that **n** is an upper bound for **A;** as before, it will suffice to show that if **y > n** then **y** does not belong to **A.** This follows because **z > n** and **n > 1** imply that $\mathbf{z^n > n^n > n.}$

The proof of the second step of Theorem 5 will rely on the following standard algebraic fact.

Theorem 6. (Binomial Theorem). *Let* **x** *and* **y** *be real numbers, and let* **n** *be a positive integer. Then we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

where the numbers

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

are the usual binomial coefficients and **n!** *denotes the factorial of* **n,** *which is formally defined by* 0 **!** **=** 1 *and the usual description for* **n** **>** 0:

$$n! = \prod_{k=1}^{n} k$$

The proof of this result proceeds by induction on **n** and is based upon the standard triangular identities named after B. Pascal (1623 – 1662), which state that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

for non-negative integers **n** and **k** where **n** $\geq$ **k** and with the initial condition

$$\binom{n}{0} = \binom{n}{n} = 1.$$

In principle (at least), mathematicians in China and India had discovered the preceding identities centuries earlier, but we shall not elaborate on this point. Note that if we take **x = y =** 1, then the formula states that the corresponding sum of binomial coefficients is equal to $2^{\mathbf{n}}$**.** We shall use this fact at a few steps in the proof of Theorem 5. Some of these steps will be stated separately before we proof the second part of Theorem 5.

Proofs of the Binomial Theorem appear in many precalculus and discrete structures textbooks (*e.g*., see pages 327 – 328 of Rosen for an argument that is somewhat different from the inductive proof mentioned above), and therefore we shall not give a proof here.

Lemma 7. *If* **1 > t > 0** *then* $\mathbf{(1-t)^n > 1 - 2^n t}$**.**

Lemma 8. *If* **1 > y > 0** *and* **z > 1** *then* $\mathbf{(z+y)^n < z^n + 2^n z^n y}$**,** *and if* **1 > y > 0** *and* **z < 1** *then* $\mathbf{(z+y)^n < z^n + 2^n y}$**.**

Proof of Lemma 7. In the Binomial Theorem take **x = 1** and **y = – t .** Let **C(n, k)** denote the **(n, k)** binomial coefficient to avoid typesetting problems. For each **k >** 0**,** a

lower estimate for the k^{th} term of the expansion for $(1 - t)^n$ is given by $-C(n, k)t$. If we add these terms over all nonnegative values of **k** and use the fact that the sum of all the coefficients $C(n, k)$ is 2^n, we obtain the lower estimate in the statement of the lemma.■

***Proof of Lemma* 8.** In this case we take $z = x$. Once again let $k > 0$. Then an upper estimate for the k^{th} term of the expansion is given by $C(n, k)z^n y$ if $z > 1$, and by the expression $C(n, k)y$ if $z < 1$. Adding these terms over all nonnegative values of **k** and using the fact that the sum of all the coefficients $C(n, k)$ is 2^n, we obtain the desired upper estimates.■

We are now prepared to complete the proof of the result on the existence of n^{th} roots.

***Proof of the second part of Theorem* 5.** We again have separate cases where $r \leq 1$ or $r > 1$, and in each case we need to show that both $z^n < r$ and $z^n > r$ are impossible.

Before proceeding we make some elementary observations. If $r = 1$ then $z = 1$ and there is nothing to prove. We **CLAIM** that if $r < 1$ or $r > 1$ then **z** also satisfies $z < 1$ or $z > 1$ respectively. — If $r < 1$ then we claim there is a **v** such that $0 < v < 1$ and $v^n > r$. If this is true then **r** is an upper bound for **S** and therefore the least upper bound **z** must be strictly less than 1 (in fact, it must be strictly less than **v**). By the Lemma 7 we know that if $1 > t > 0$ then $(1 - t)^n > 1 - 2^n$ and therefore if we choose **v** such that $x = 1 - v$ satisfies $2^n x < 1 - r$ then v^n will be strictly greater than **r**. Finally, if $r > 1$ then $r^{-1} < 1$, and therefore it is possible to find some **w** such that $0 < w < 1$ and $w^n > r^{-1}$. If we set $v = w^{-1}$, we then obtain the inequalities $v > 1$ and $v^n < r$. But this means that $1 < v \leq z$.

Suppose now that $1 < r$ and $z^n < r$, where $z > 1$ by the preceding paragraph. If we have $w > z$ then $w^n \geq r$ because **z** is the least upper bound of all **x** such that $x^n < r$. Let $s = r - z^n$; it will suffice to find a number **v** such that v^n lies between z^n and **r**. If $1 > y > 0$ then Lemma 8 implies that $(z + y)^n$ is less than $z^n + 2^n z^n y$, and if we now choose **y** so that $2^n z^n y < s$, then $v = z + y$ will satisfy the desired condition $z^n < v^n < r$. Now suppose we have $1 > r$ and $z^n < r$, so that $z < 1$ by the preceding paragraph. Let **w** and **s** be as before. Then we still have $w^n \geq r$ and we would again like to find some **v** such that v^n lies between z^n and **r**. Taking **y** as before, we can now use Lemma 8 to conclude that $(z + y)^n < z^n + 2^n y$, and if we choose **y** so that $2^n y < s$ then $v = z + y$ satisfies the desired condition $z^n < v^n < r$. Observe that the main difference in the arguments for the two cases $1 < r$ and $1 > r$ is the estimate for $(z + y)^n$ given by the Binomial Theorem.

Suppose now that $z^n > r$. By the definition of a least upper bound, for every $h > 0$ there is some **w** such that $z - w > h$ and $w^n < r$. Hence if $x < z$ and $h = z - x$ then we can find a **w** such that $x < w < z$ and $w^n < r$. The latter in turn implies that $x^n < w^n < r$. Thus we have shown that if $x < z$ then $w^n < r$, while if $x > z$ then $w^n > z^n > r$. Once again it will suffice to find a number ***v*** such that v^n lies between z^n and **r**. Let $s = z^n - r$ and let $y > 0$ as before, but this time consider the quantity $(z - y)^n$. If $r > 1$ we then obtain the inequality

$$(z - y)^n > z^n - 2^n z^n y$$

while if $r > 1$ we obtain the inequality

$$(z + y)^n > z^n + 2^n y .$$

In each case if we choose **y** sufficiently small the right hand side will be strictly greater than **r**, which contradicts our previous observation that $x < z$ implies $w^n < r$. This completes the proof of Theorem 5.■

The next result is a simple consequence of Theorem 5 and the proof of Lemma **V.1.3,** but it provides an important relation between the algebraic and order structures on the real number system.

Corollary 8. *A real number* **x** *is nonnegative if and only if there is another real number* **y** *such that* $y^2 = x$.

Proof. The proof of Lemma **V.1.2** only depends upon algebraic and ordering properties that hold for both the integers and the real numbers**,** and thus it follows that Lemma **V.1.2** is also true for the real numbers; therefore for every real number **y** we see that the square y^2 is nonnegative. Conversely, by Theorem 5 we know that every nonnegative number is the square of some other real number.

Section 4.5 of Lipschutz discusses the use of Theorem 5 to define rational and irrational powers of a positive real number (in particular, see the subheading, "Exponential Functions," at the bottom of page 101).

Base 10 *and decimal expansions*

We shall only summarize the main points here, leaving the proofs to an Appendix for this section of the notes.

One of the most elementary facts about a positive real number **x** is that it can be written as the sum **[x] + (x)** of a nonnegative integer **[x]** and a nonnegative real number **(x)** that is strictly less than one, and this decomposition is unique. The integer **[x]** is often called the *greatest integer function* of **x** or the *integral part* of **x** or the *characteristic* of **x**, and the remaining number **(x)** is often called the *fractional part* or *mantissa* of **x**. The characteristic – mantissa terminology dates back to the original tables of base 10 logarithms published by H. Briggs (1561 – 1630); the literal meaning of the Latin root word ***mantisa*** is "makeweight," and it denotes something small that is placed onto a scale to bring the weight up to a desired value. We shall derive the decomposition of a nonnegative real number into a characteristic and mantissa from the axiomatic properties of the real numbers.

Theorem 9. *Let* **r** *be an arbitrary nonnegative real number. Then there is a unique decomposition of* **r** *as a sum* **n + s** *where* **n** *is a nonnegative integer and* $0 \leq s < 1$.

Here is the standard result on ***base*** **N** or **N – *adic*** expansions of positive integers. In the standard case when **N =** 10, this yields the standard way of writing a nonnegative integer in terms of the usual Hindu – Arabic numerals, while if **n =** 2 or 8 or 16 this yields the binary or octal or hexadecimal expansion respectively.

Theorem 10. *Let* $\mathbf{k}$ *be a positive integer, and let* $\mathbf{N > 1}$ *be another positive integer. Then there are unique integers* $\mathbf{a_j}$ *such that* $\mathbf{0 \le a_j \le N - 1}$ *and*

$$\mathbf{k = a_0 + a_1 \cdot N + \ldots + a_m \cdot N^m}$$

for a suitable nonnegative integer $\mathbf{m}$*.*

For both practical and theoretical reasons, a mathematically sound definition of the real numbers should yield the usual decimal expansions for base 10 as well as the corresponding expansions for other choices of the base $\mathbf{N}$. We shall verify this here and show that decimal expansions have several properties that are well – known from our everyday experience in working with decimals.

Although decimal expansions of real numbers are extremely useful for computational purposes, they are not particularly convenient for theoretical or conceptual purposes. For example, although every nonzero real number should have a reciprocal, describing this reciprocal completely and explicitly by infinite decimal expansions is awkward and generally unrealistic. Another difficulty is that decimal expansions are not necessarily unique; for example, the relation

$$\mathbf{1.0 = 0.99999999999999999999 \ldots}$$

reflects the classical geometric series formula

$$\mathbf{a/(1 - r) = a + ar + ar^2 + \ldots + ar^k + \ldots}$$

when $\mathbf{a = 9/10}$ and $\mathbf{r = 1/10.}$ A third issue is whether one gets an equivalent number system if one switches from base 10 arithmetic to some other base. It is natural to expect that the answer to this question is yes, but any attempt to establish this directly runs into all sorts of difficulties almost immediately. This is not purely a theoretical problem**;** the use of digital computers to carry out numerical computations implicitly assumes that one can work with real numbers equally well using infinite expansions with base 2 (or base 8 or 16 as in many computer codes, or even base 60 as in ancient Babylonian mathematics). One test of the usefulness of the abstract approach to real numbers is whether it yields such consequences.

The preceding discussion justifies the standard method for expressing the integral part of a positive real number. Of course, the next step is to justify the standard expression for the fractional part. A natural first step is to verify that the usual types of infinite decimal expansions always yield real numbers.

Theorem 11. (Decimal Expansion Theorem). *Every infinite series of real numbers having the form*

$$\mathbf{a_N \cdot 10^N + a_{N-1} \cdot 10^{N-1} + \ldots + a_0 + b_1 \cdot 10^{-1} + b_2 \cdot 10^{-2} + \ldots + b_k \cdot 10^{-k} + \ldots}$$

(*with* $\mathbf{0 \le a_i, b_j \le 9}$)

*is convergent***.** *Conversely, every positive real number is the sum of an infinite series of this type where the coefficients of the powers of* 10 *are integers satisfying the basic inequalities* $\mathbf{0 \le a_i, b_j \le 9.}$

This turns out to be a fairly direct consequence of standard results on convergence of infinite series whose terms are all nonnegative (see Rudin, Theorem 3.25, page 60, for a proof):

COMPARISON TEST. *Suppose that*

$$\sum a_n \quad \text{and} \quad \sum b_n$$

are two series whose terms are nonnegative and satisfy $a_n \leq b_n$ *for all* **n.** *If the second series converges, then the first one does also.*

Theorem 10 immediately yields the standard "scientific notation" for a positive real number:

Corollary 12. (Scientific Notation Representation). *Every positive real number has a unique expression of the form* $a \cdot 10^M$*, where* $1 \leq a < 10$ *and* **M** *is an integer.*

Decimal expansions of rational numbers

One basic test for the effectiveness of a mathematical theory is whether one can use it to shed light on patterns that run through many basic examples. The decimal expansions for rational numbers are an example of this type. If one computes the decimal expansions for some simple fractions, the results turn out to yield decimal expansions that are eventually repeating. Here are some examples:

1 / 3 = 0.33333333333333333333333333333333333 …
1 / 6 = 0.16666666666666666666666666666666666 …
1 / 7 = 0.14285714285714285714285714285714285 …
1 / 11 = 0.01010101010101010101010101010101010 …
1 / 12 = 0.08333333333333333333333333333333333 …
1 / 13 = 0.07692307692307692307692307692307692 …
1 / 17 = 0.05882352941176470588235294117647058 …
1 / 18 = 0.05555555555555555555555555555555555 …
1 / 19 = 0.05263157894736842105263157894736842 …
1 / 23 = 0.04347826086956521739130434782608695 …
1 / 27 = 0.03703703703703703703703703703703703 …
1 / 29 = 0.03448275862068965517241379310344827 5 …
1 / 31 = 0.03225806451612903225806451612903225 8 …
1 / 34 = 0.02941176470588235294117647058823529 4 …
1 / 37 = 0.02702702702702702702702702702702702 7 …

Motivated by such examples, it is natural to ask whether the decimal expansions for an arbitrary rational number must have the following special property:

Theorem 13. (Eventual Periodicity Property.) *Suppose that* **r** *is a rational number such that* $0 < r < 1$**,** *and let*

$$r = b_1 \cdot 10^{-1} + b_2 \cdot 10^{-2} + \ldots + b_k \cdot 10^{-k} + \ldots$$

be a decimal expansion. Then the sequence **{ b_k }** *is* ***eventually periodic*** **;** *i.e., there are positive integers* **M** *and* **Q** *such that* $b_k = b_{k+Q}$ *for all* $k > M$.

CONVERSELY, suppose that the statement in the claim holds for the decimal expansion of some number, and choose **m** and **Q** as above. Let **s** be given by the first **m – 1** terms in the decimal expansion of **y**, and let **t** be the sum of the next **Q** terms. It then follows that **y** is equal to $s + t\,(1 + 10^{-Q} + 10^{-2Q} + 10^{-3Q} + \ldots)$. Now **s, t** and the geometric series in parentheses are all rational numbers, and therefore it follows that **y** is also a rational number. Therefore we have the following result:

Theorem 14. *A real number between* **0** *and* **1** *has a decimal expansion that is eventually periodic if and only if it is a rational number.*

Similar results hold if the numerical base**10** is replaced by an arbitrary integer $N > 1$.

Uniqueness properties of decimal expansions

Finally, here is the standard criterion for two decimal expressions to be equal:

Theorem 15. *Suppose that we are given two decimal expansions that yield the same real number:*

$$a_N \cdot 10^N + a_{N-1} \cdot 10^{N-1} + \ldots + a_0 + b_1 \cdot 10^{-1} + b_2 \cdot 10^{-2} + \ldots + b_k \cdot 10^{-k} + \ldots =$$

$$c_N \cdot 10^N + c_{N-1} \cdot 10^{N-1} + \ldots + c_0 + d_1 \cdot 10^{-1} + d_2 \cdot 10^{-2} + \ldots + d_k \cdot 10^{-k} + \ldots$$

Then $a_j = c_j$ *for all* **j***, and one of the following is also true:*

1. *For each* **k** *we have* $b_k = d_k$.
2. *There is an* $L > 0$ *such that* $b_k = d_k$ *for every* $k < L$ *but* $b_{L+1} = d_L + 1$, *with* $b_k = 0$ *for* $k > L$ *and* $d_k = 9$ *for all* $k > L$.
3. *There is an* $L > 0$ *such that* $b_k = d_k$ *for every* $k < L$ *but* $d_{L+1} = b_L + 1$, *with* $d_k = 0$ *for* $k > L$ *and* $b_k = 9$ *for all* $k > L$ *(the opposite of the previous possibility).*

One can reformulate the preceding into a strict uniqueness result as follows:

Corollary 16. *Every positive real number has a unique decimal expansion of the form*

$$a_N \cdot 10^N + a_{N-1} \cdot 10^{N-1} + \ldots + a_0 + b_1 \cdot 10^{-1} + b_2 \cdot 10^{-2} + \ldots + b_k \cdot 10^{-k} + \ldots$$

such that b_k *is nonzero for infinitely many choices of* **k.**

EXAMPLE. We can use the preceding result to define real valued functions on an interval in terms of decimal expansions. In particular, if we express an arbitrary real number $x \in (0, 1]$ as an infinite decimal

$$x = 0.b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 b_9 \ldots$$

where infinitely many digits b_k are nonzero, then we may define a function **f** from **(**0**,** 1**]** to itself by the formula

$$f(x) = 0.b_1 0 b_2 0 b_3 0 b_4 0 b_5 0 b_6 0 b_7 0 b_8 0 b_9 0 \ldots$$

and if we extend this function by setting **f (0) = 0** then we obtain a strictly increasing function on the closed unit interval (verify that the function is strictly increasing**!**). Note that this function has a jump discontinuity at every finite decimal fraction.

Since every nondecreasing real valued function on a closed interval is Riemann integrable, we know that **f** can be integrated. It turns out that the value of this integral is a fairly simple rational number; finding the precise value is left as an exercise for the reader (this is a good illustration of the use of Riemann sums – a natural strategy is to partition the unit interval into pieces whose endpoints are finite decimal fractions with at most **n** nonzero terms and to see what happens to the Riemann sums as **n** increases).

V. 5. Appendix A : Proofs of results on number expansions

This appendix contains proofs of several results from Section 5:

Theorem V.5.9
Theorem V.5.10
Theorem V.5.11
Theorem V.5.12
Corollary V.5.13
Theorem V.5.14
Theorem V.5.15
Corollary V.5.16

We begin by proving that a positive real number can be written in an essentially unique manner as the sum of an integral part and a fractional part which lies between 0 and 1.

Theorem V.5.9. *Let* **r** *be an arbitrary nonnegative real number. Then there is a unique decomposition of* **r** *as a sum of the form* **n + s** *such that* **n** *is a nonnegative integer and* $0 \leq s < 1$**.**

Proof. By the Archimedean Law there is a nonnegative integer **m** such that **m > r,** and since the nonnegative integers are well – ordered there is a ***minimum*** such integer . Since **r** is nonnegative it follows that m_1 cannot be zero and hence must also be positive. Therefore $m_1 - 1$ is also nonnegative and by the minimality of the positive integer m_1 we must have $m_1 - 1 \leq r$. If we take $n = m_1 - 1$ and $s = r - n$ then we have $r = n + s$

where **n** and **s** have the desired properties. Suppose that we also have **r = q + v** where **q** is a nonnegative integer and $0 \leq q < 1$. By hypothesis we have

$$q \leq r < q + 1$$

and the right hand inequality implies $n + 1 \leq q + 1$, or equivalently $n \leq q$. The equation $r = n + s = q + v$ can therefore be rewritten in the form

$$0 \leq q - n = s - v$$

and since ***(i)*** $s - v \leq s < 1$ and ***(ii)*** $q - n$ is an integer, it follows that $n = q$ and $s = v$.■

Base **N** *expansions for natural numbers*

We shall use the long division property for natural numbers to derive the standard result on base **N** expansions of positive integers. In the standard case when $N = 10$, this yields the standard way of writing a nonnegative integer.

Theorem V.5.10. *Let* **k** *be a positive integer, and let* $N > 1$ *be another positive integer. Then there are unique integers* a_j *such that* $0 \leq a_j \leq N - 1$ *and*

$$k = a_0 + a_1 N + \ldots + a_m N^m$$

for a suitable nonnegative integer **m.**

In the course of proving this result it will be useful to know the following:

Lemma 1. *Suppose that integers* **N, k,** *and* a_j *are given as above. Then we have*

$$a_0 + a_1 N + \ldots + a_m N^m \leq N^{m+1}.$$

Proof of Lemma 1. Since $a_j \leq N - 1$ for each **j** we have

$$a_j N^j \leq (N - 1) N^j = N^{j+1} - N^j$$

and therefore we have the inequality

$$a_0 + a_1 N + \ldots + a_m N^m \leq N - 1 + (N^2 - N) + \ldots + (N^{m+1} - N^m) =$$
$$N^{m+1} - 1 < N^{m+1}.■$$

Proof of Theorem V.5.10. It is always possible to find an exponent **q** such that $2^q > k$, and since $k \geq 2$ it follows that we also have $N^q > 2^q > k$. Let $[\, S_m \,]$ be the statement of the statement that every positive integer less than N^{m+1} has a unique expression as above. If $m = 0$ then the result follows immediately from the long division theorem, for then $k = a_0$. Suppose now that $[\, S_{p-1} \,]$ is true and consider the statement $[\, S_p \,]$. If $k < N^{p+1}$ then we can use long division to write **k** uniquely in the form

$$k = k_0 + a_p N^p$$

where $a_p \geq 0$ and $0 \leq k_0 < N^p$. We claim that $a_p < N$. If this were false then we would have $k \geq a_p N^p \geq N N^p = N^{p+1}$ and this contradicts the assumption that $k < N^{p+1}$.

By induction we know that $\mathbf{k_0}$ has a unique expression as a sum

$$\mathbf{k_0 \;=\; a_0 \;+\; a_1 N \;+\; \ldots \;+\; a_{p-1} N^{p-1}}$$

for suitable $\mathbf{a_j}$**.** This proves existence. To prove uniqueness, suppose that we have

$$\mathbf{k \;=\; a_0 \;+\; a_1 N \;+\; \ldots \;+\; a_p N^p \;=\; b_0 \;+\; b_1 N \;+\; \ldots \;+\; b_p N^p .}$$

Denote all but the last terms of these sums by $\mathbf{A = a_0 + a_1 N + \ldots + a_{p-1} N^{p-1}}$ and $\mathbf{B = b_0 + b_1 N + \ldots + b_{p-1} N^{p-1}}$**.** Then we have $0 \le \mathbf{A, B} \le \mathbf{N^p} - 1$ by the lemma, and therefore by the uniqueness of the long division expansion of k it follows that $\mathbf{a_p = b_p}$ and $\mathbf{A = B}$**.** By the induction hypothesis the latter implies that $\mathbf{a_j = b_j}$ for all $\mathbf{j < p}$**.** Therefore we have also shown uniqueness**.**■

Decimal expansions for real numbers

As we have already noted, a mathematically sound definition of the real numbers should yield the usual decimal expansions for base 10 as well as the corresponding expansions for other choices of the base **N**. We shall verify this and show that decimal expansions have many properties that are more or less predictable on empirical grounds.

One such property is the well – known decimal equality $1.0 = 0.9999999\ldots$ so we begin by noting this reflects the geometric series formula

$$\mathbf{a / (}1 - \mathbf{r)} \;=\; \mathbf{a \;+\; a r \;+\; a r^2} + \ldots \;+\; \mathbf{a r^k} + \ldots$$

when $\mathbf{a} = 9 / 10$ and $\mathbf{r} = 1 / 10$. In fact, the geometric series plays a key role in proving that infinite decimal expansions always yield real numbers.

Theorem V.5.11. (Decimal Expansion Theorem.) *Every infinite series of real numbers having the form*

$$\mathbf{a_N 10^N + a_{N-1} 10^{N-1} + \ldots + a_0 + b_1 10^{-1} + b_2 10^{-2} + \ldots + b_k 10^{-k} + \ldots}$$

(*with* $0 \le \mathbf{a_i}, \mathbf{b_j} \le 9$)

*is convergent***.** *Conversely, every positive real number is the sum of an infinite series of this type where the coefficients of the powers of* 10 *are integers satisfying the basic inequalities* $0 \le \mathbf{a_i}, \mathbf{b_j} \le 9$**.**

As noted above, there are two ways of writing 1 as an infinite series of this type, so such a representation is not unique, but empirical evidence suggest that all ambiguities in decimal expansions arise from this example, and we shall verify this later.

PROOF OF THE DECIMAL EXPANSION THEOREM. The proof of this result splits naturally into two parts, one for each implication direction.

Formal infinite decimal expansions determine real numbers: If one can show this for positive decimal expansions, it will follow easily for negative ones as well, so we shall restrict attention to the positive case. Consider the formal expression given above:

$$\mathbf{(a_N 10^N + a_{N-1} 10^{N-1} + \ldots + a_0 + b_1 10^{-1} + b_2 10^{-2} + \ldots + b_k 10^{-k} + \ldots)}$$

For each integer $\mathbf{p} > 0$, define $\mathbf{s}_p$ to be the sum of all terms in this expression up to and including $\mathbf{b}_p 10^{-p}$ and let **S** be the set of all such numbers $\mathbf{s}_p$. Then the set **S** has an upper bound, and in fact we claim that 10^{N+1} is an upper bound for **S.** To see this, observe that $\mathbf{a}_N 10^N + \mathbf{a}_{N-1}10^{N-1} + \ldots + \mathbf{a}_0 \leq 10^{N+1} - 1$ by a previous lemma and

$$\mathbf{b}_1 10^{-1} + \mathbf{b}_2 10^{-2} + \ldots + \mathbf{b}_k 10^{-k} + \ldots \leq$$
$$9\,(10^{-1} + 10^{-2} + \ldots + 10^{-k} + \ldots) = 1$$

and the assertion about an upper bound follows immediately from this. The least upper bound **r** for **S** turns out to be the limit of the sequence of partial sums $\{\mathbf{s}_p\}$.

Real numbers determine infinite decimal expansions: Given (say) a positive real number ***r***, the basic idea is to find a sequence of finite decimal fractions $\{\mathbf{s}_p\}$ such that for every value of **p** the number $\mathbf{s_p}$ is expressible as a fraction whose denominator is given by 10^p and

$$\mathbf{s}_p \leq \mathbf{r} < \mathbf{s}_p + 10^{-p}.$$

More precisely, suppose that we already have $\mathbf{s}_p$ and we want to find the next term. By construction $10^p \mathbf{s}_p$ is a positive integer and $10^p \mathbf{s}_p \leq 10^p \mathbf{r} < 10^p \mathbf{s}_p + 1$, so that

$$10^{p+1}\mathbf{s}_p \leq 10^{p+1}\mathbf{r} < 10^{p+1}\mathbf{s}_p + 10.$$

Choose $\mathbf{b}_{p+1}$ to be the largest integer such that

$$\mathbf{b}_{p+1} \leq 10^{p+1}\mathbf{r} - 10^{p+1}\mathbf{s}_p.$$

The right hand side is positive so this means that $\mathbf{b_{p+1}} \geq 0$. On the other hand, the previous inequalities also show that $\mathbf{b}_{p+1} < 10$ and since $\mathbf{b}_{p+1}$ is an integer this implies $\mathbf{b}_{p+1} \leq 9$. If we now take $\mathbf{s}_{p+1} = 10\,\mathbf{s}_p + \mathbf{b}_{p+1}$ then it will follow that

$$\mathbf{s}_{p+1} \leq \mathbf{r} < \mathbf{s}_{p+1} + 10^{-(p+1)}.$$

To see that the sequence converges, note that it corresponds to the infinite series

$$\mathbf{s}_p + \sum_p (\mathbf{b}_{p+1} 10^{-p}),$$

which converges by a comparison with the modified geometric series $\mathbf{s}_p + \sum_p 10^{(1-p)}$.■

Corollary V.5.12. (Scientific Notation Representation). *Every positive real number has a unique expansion of the form* $\mathbf{a} \cdot 10^M$*, where* $1 \leq \mathbf{a} < 10$ *and* **M** *is an integer.*

Existence. If **x** has the decimal expansion

$$\mathbf{a}_N \cdot 10^N + \mathbf{a}_{N-1} \cdot 10^{N-1} + \ldots + \mathbf{a}_0 + \mathbf{b}_1 \cdot 10^{-1} + \mathbf{b}_2 \cdot 10^{-2} + \ldots + \mathbf{b}_k \cdot 10^{-k} + \ldots$$
$$(\textit{with } 0 \leq \mathbf{a}_i, \mathbf{b}_j \leq 9)$$

then $\mathbf{x} \cdot 10^{-N}$ lies in the interval **[1, 10)** by construction.

Uniqueness. Suppose that we can write **x** as $\mathbf{a} \cdot 10^M$ and $\mathbf{b} \cdot 10^N$. Then by the conditions on the coefficients, we know that $\mathbf{x} \in [10^M, 10^{M+1}) \cap [10^N, 10^{N+1})$. Since the half open intervals $[10^M, 10^{M+1})$ and $[10^N, 10^{N+1})$ are disjoint unless $\mathbf{M = N}$, it follows that the latter must hold. Therefore the equations $\mathbf{x} = \mathbf{a} \cdot 10^M = \mathbf{b} \cdot 10^N$ and $\mathbf{M = N}$ imply $\mathbf{a = b}$.■

Decimal expansions of rational numbers

In working with decimals one eventually notices that the decimal expansions for rational numbers have the following special property:

Theorem V.5.13. (Eventual Periodicity Property.) *Let* **r** *be a rational number such that* $0 < r < 1$*, and let*

$$r = b_1 10^{-1} + b_2 10^{-2} + \ldots + b_k 10^{-k} + \ldots$$

be a decimal expansion. Then the sequence $\{ b_k \}$ *is* ***eventually periodic*** *; i.e., there are positive integers* **M** *and* **Q** *such that* $b_k = b_{k+Q}$ *for all* $k > M$.

Proof. Let a / b be a rational number between 0 and 1**,** where **a** and **b** are integers satisfying $0 < a < b$. Define sequences of numbers r_n and x_n recursively, beginning with $r_0 = a$ and $x_0 = 0$. Given r_n and x_n express the product $10\, r_n$ by long division in the form $10\, r_n = b\, x_{n+1} + r_{n+1}$ where $x_{n+1} \geq 0$ and $0 \leq r_{n+1} < b$.

CLAIMS:

1. *Both of these numbers only depend upon* r_n .
2. *We have* $x_{n+1} < 10$.

The first part is immediate from the definition in terms of long division, and to see the second note that $x_{n+1} \geq 10$ would imply $10\, r_n \geq 10\, b$**,** which contradicts the fundamental remainder condition $r_n < b$.

Since r_n can only take integral values between 0 and $b - 1$, it follows that there are some numbers **Q** and m such that $r_m = r_{m+Q}$.

CLAIM: $r_k = r_{k+Q}$ *for all* $k \geq m$.

We already know this for $p = m$, so assume it is true for $p \leq k$. Now each term in the sequence r_n depends only on the previous term, and hence the relation $r_k = r_{k+Q}$ implies $r_{k+1} = r_{k+Q+1}$. Therefore the claim is true by finite induction.■

CONVERSELY, suppose that the statement in the claim holds for the decimal expansion of some number, and choose **m** and **Q** as above. Let **s** be given by the first $m - 1$ terms in the decimal expansion of **y**, and let **t** be the sum of the next **Q** terms. It then follows that **y** is equal to $s + t\,(1 + 10^{-Q} + 10^{-2Q} + 10^{-3Q} + \ldots)$. Now **s, t** and the geometric series in parentheses are all rational numbers, and therefore it follows that **y** is also a rational number. Therefore we have the following result:

Theorem V.5.14. *A real number between* 0 *and* 1 *has a decimal expansion that is eventually periodic if and only if it is a rational number*.■

In Section 5 we gave the following examples to illustrate the theorem:

1 / 3 = 0.33333333333333333333333333333333333 **...**
1 / 6 = 0.16666666666666666666666666666666666 **...**
1 / 7 = 0.14285714285714285714285714285714285 **...**
1 / 11 = 0.01010101010101010101010101010101010 **...**
1 / 12 = 0.08333333333333333333333333333333333 **...**
1 / 13 = 0.07692307692307692307692307692307692 **...**
1 / 17 = 0.05882352941176470588235294117647058 **...**
1 / 18 = 0.05555555555555555555555555555555555 **...**
1 / 19 = 0.05263157894736842105263157894736842 **...**
1 / 23 = 0.04347826086956521739130434782608695 **...**
1 / 27 = 0.03703703703703703703703703703703703 **...**
1 / 29 = 0.03448275862068965517241379310344827 **...**
1 / 31 = 0.03225806451612903225806451612903225 **...**
1 / 34 = 0.02941176470588235294117647058823529 **...**
1 / 37 = 0.02702702702702702702702702702702702 **...**

Note that the minimal period lengths in these examples are 1, 1, 6, 2, 1, 6, 16, 1, 18, 22, 3, 28, 15, 16 and 3. One is naturally led to the following question:

Given a fraction **a / b** *between* 0 *and* 1**,** *what determines the (minimal) period length* **Q**?

To illustrate the ideas, we shall restrict attention to the special case where **a / b = 1 / p,** where **p** is a prime not equal to 2 or 5 (the two prime divisors of 10). In this case the methods of abstract algebra yield the following result:

Theorem 2. *If* **p ≠** 2**,** 5 *is a prime, then the least period* **Q** *for the decimal expansion of* 1 **/ p** *is equal to the multiplicative order of* 10 *in the (finite cyclic) group of multiplicative units for the integers mod* **p.**■

We shall not verify this result here, but the proof is not difficult.

Corollary 3. *The least period* **Q** *divides* **p –** 1**.**

The corollary follows because the order of the group of units is equal to **p –** 1 and the order of an element in a finite group always divides the order of the group**.**■

One is now led to ask when the period is actually equal to this maximum possible value. Our examples show this is true for the primes 7, 19, 23 and 29 but not for the primes 11, 13, 31 or 37**.**

More generally, one can define a primitive root of unity in the integers mod **p** to be an integer **a** mod **p** such that **a** is not divisible by **p** and the multiplicative order of the class of **a** in the integers mod **p** is precisely **p –** 1**.** Since the group of units is cyclic, such primitive roots always exist, and one can use the concept of primitive root to rephrase the question about maximum periods for decimal expansions in the following terms:

For which primes **p** *is* 10 *a primitive root of unity mod* **p?**

A simple answer to this question does not seem to exist. In the 1920s E. Artin (1898 – 1962) stated the following conjecture:

Every integer **a > 1** *is a primitive root of unity mod* **p** *for infinitely many primes* **p.**

This means that 10 ***should*** be the primitive root for infinitely many primes **p**, and hence there should be infinitely many full – period primes. Quantitatively, the conjecture

amounts to showing that about 37% of all primes asymptotically have 10 as primitive root. The percentage is really an approximation to Artin's constant

$$C_{\text{Artin}} = \prod_{k=1}^{\infty} \left[1 - \frac{1}{p_k(p_k - 1)}\right] = 0.3739558136\ldots$$

where $\mathbf{p_k}$ denotes the $\mathbf{k^{th}}$ prime. Further information about this number and related topics appears in the following online reference:

http://mathworld.wolfram.com/ArtinsConstant.html

Uniqueness of decimal expansions

The criterion for two decimal expressions to be equal is well understood.

Theorem V.5.15. *Suppose that we are given two decimal expansions that yield the same real number:*

$$\mathbf{a_N 10^{N} + a_{N-1} 10^{N-1} + \ldots + a_0 + b_1 10^{-1} + b_2 10^{-2} + \ldots + b_k 10^{-k} + \ldots =}$$

$$\mathbf{c_N 10^{N} + c_{N-1} 10^{N-1} + \ldots + c_0 + d_1 10^{-1} + d_2 10^{-2} + \ldots + d_k 10^{-k} + \ldots}$$

Then $\mathbf{a_j = c_j}$ *for all* **j**, *and one of the following is also true:*

1. *For each* **k** *we have* $\mathbf{b_k = d_k}$.
2. *There is an* $\mathbf{L > 0}$ *such that* $\mathbf{b_k = d_k}$ *for every* $\mathbf{k < L}$ *but* $\mathbf{b_{L+1} = d_L + 1}$, *while* $\mathbf{b_k = 0}$ *for all* $\mathbf{k > L}$ *and* $\mathbf{d_k = 9}$ *for all* $\mathbf{k > L}$.
3. *There is an* $\mathbf{L > 0}$ *such that* $\mathbf{b_k = d_k}$ *for every* $\mathbf{k < L}$ *but* $\mathbf{d_{L+1} = b_L + 1}$, *while* $\mathbf{d_k = 0}$ *for all* $\mathbf{k > L}$ *and* $\mathbf{b_k = 9}$ *for all* $\mathbf{k > L}$ *(the opposite of the previous possibility).*

If **x** and **y** are given by the respective decimal expansions above, then $\mathbf{x = y}$ implies the greatest integer functions satisfy $\mathbf{[x] = [y]}$, which in turn implies that $\mathbf{a_j = c_j}$ for all **j**. Furthermore, we then also have $\mathbf{(x) = (y)}$ and accordingly the proof reduces to showing the result for numbers that are between 0 and 1.

The following special uniqueness result will be helpful at one point in the general proof.

Lemma 4. *For each positive integer* **k** *let* $\mathbf{t_k}$ *be an integer between* 0 *and* 9. *Then we have*

$$1 = \mathbf{t_1 10^{-1} + t_2 10^{-2} + \ldots + t_k 10^{-k} + \ldots}$$

if and only if $\mathbf{t_k = 9}$ *for all* **k**.

Proof. Let **t** be the summation on the right hand side. If $\mathbf{t_k = 9}$ for all **k** then $\mathbf{t = 1}$ by the geometric series formula. Conversely, if $\mathbf{t_m < 9}$ for a specific value of **m** then

$$t_1 10^{-1} + t_2 10^{-2} + \ldots + t_k 10^{-k} + \ldots < u_1 10^{-1} + u_2 10^{-2} + \ldots + u_k 10^{-k} + \ldots$$

where $u_k = 9$ for $k \neq m$ and $u_m \leq 8$. The latter implies that the right hand side is less than or equal to $1 - 10^{-m}$, which is strictly less than 1.■

Theorem 5. *If we are given two decimal expansions*

$$x = x_1 10^{-1} + x_2 10^{-2} + \ldots + x_k 10^{-k} + \ldots$$

$$y = y_1 10^{-1} + y_2 10^{-2} + \ldots + y_k 10^{-k} + \ldots$$

then $x = y$ *if and only if one of the following is true*:

1. *For all positive integers* k *we have* $x_k = y_k$.
2. *There is some positive integer* M *such that* [i] $x_k = y_k$ *for all* $k < M$, [ii] $x_M = y_M + 1$, [iii] $x_k = 0$ *for* $k > M$, *and* [iv] $y_k = 9$ *for* $k > M$.
3. *A corresponding statement holds in which the roles of* x_k *and* y_k *are interchanged: There is some positive integer* M *such that* [i] $x_k = y_k$ *for all* $k < M$, [ii] $y_M = x_M + 1$, [iii] $y_k = 0$ *for* $k > M$, *and* [iv] $x_k = 9$ *for* $k > M$.

Proof. Suppose that the first alternative does not happen, and let L be the first positive integer such that $x_L \neq y_L$. Without loss of generality, we may as well assume that the inequality is $x_L > x_L$ (if the inequality points in the opposite direction, then one can apply the same argument reversing the roles of x_k and x_k throughout). Let z be given by the first $L - 1$ terms of either x or y (these are equal).

CASE 1. Suppose that $x_L \geq y_L + 2$. Note that $y_L \leq 7$ is true in this case. We then have

$$y \leq z + 10^{-L} y_L + 9 \times 10^{-L} (10^{-1} + 10^{-2} + \ldots + 10^{-k} + \ldots) = z + 10^{-L} (y_L + 1) <$$

$$z + 10^{-L} (x_L) \leq z + 10^{-L} (x_L + x_{L+1} 10^{-1} + x_{L+2} 10^{-2} + \ldots + x_{L+k} 10^{-k} + \ldots) = x.$$

Therefore $x > y$ if we have $x_L \geq y_L + 2$.

CASE 2. Suppose that $x_L = y_L + 1$, and let $w = 10^{-L} y_L$, so that $x_L = w + 10^{-L}$. We may then write

$$x = z + (w + 10^{-L}) + 10^{-L} u \quad \text{and} \quad y = z + w + 10^{-L} v$$

where by construction u and v satisfy $0 \leq u, v \leq 1$. If $x = y$ then the displayed equations imply that $10^{-L} + 10^{-L} u = 10^{-L} v$. The only way such an equation can hold is if $u = 0$ and $v = 1$. The first of these implies that the decimal expansion coefficients for the sum

$$0 = u = x_{L+1} 10^{-1} + x_{L+2} 10^{-2} + \ldots + x_{L+k} 10^{-k} + \ldots$$

must satisfy $x_k = 0$ for all $k > L$, and by the lemma the second of these can only happen if the decimal expansion coefficients for the sum

$$1 = v = y_{L+1}10^{-1} + y_{L+2}10^{-2} + \ldots + y_{L+k}10^{-k} + \ldots$$

satisfy $y_k = 9$ for all $k > L$. Therefore the second alternative holds in Case 2.

Conversely, the standard geometric series argument shows that two numbers with decimal expansions given by the second or third alternatives must be equal. Of course, the two numbers are equal if the first alternative holds, so this completes the proof of the theorem.■

One can reformulate the preceding into a strict uniqueness result as follows:

Corollary V.5.16. *Every positive real number has a unique decimal expansion of the form*

$$a_N 10^N + a_{N-1}10^{N-1} + \ldots + a_0 + b_1 10^{-1} + b_2 10^{-2} + \ldots + b_k 10^{-k} + \ldots$$

such that b_k *is nonzero for infinitely many choices of* k.

This follows immediately from the preceding results on different ways of expressing the same real number in decimal form; there is more than one way of writing a number in decimal form if and only if it is an integer plus a finite decimal fraction, and in this case there is only one other way of doing so and all but finitely many digits of the alternate expansion are equal to 9.■