Technical Report

UCAM-CL-TR-469
ISSN 1476-2986

Number 469

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

A combination of geometry theorem proving and nonstandard analysis, with application to Newton's Principia

Jacques Désiré Fleuriot

August 1999

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted March 1999 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Clare College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986
DOI *https://doi.org/10.48456/tr-469*

Preface

Except where otherwise stated in the text, this dissertation is the result of my own work and is not the outcome of work done in collaboration.

This dissertation is not substantially the same as any I have submitted for a degree or diploma or any other qualification at any other university.

No part of my dissertation has already been, or is being currently submitted for any such degree, diploma or other qualification.

This dissertation does not exceed sixty thousand words, including tables, footnotes and bibliography.

Abstract

Sir Isaac Newton's *Philosophiæ Naturalis Principia Mathematica* (the *Principia*) was first published in 1687 and set much of the foundations that led to profound changes in modern science. Despite the influence of the work, the elegance of the geometrical techniques used by Newton is little known since the demonstrations of most of the theorems set out in it are usually done using calculus. Newton's reasoning also goes beyond the traditional boundaries of Euclidean geometry with the presence of both motion and infinitesimals.

This thesis describes the mechanization of Lemmas and Propositions from the *Principia* using formal tools developed in the generic theorem prover Isabelle. We discuss the formalization of a geometry theory based on existing methods from automated geometry theorem proving. The theory contains extra geometric notions, including definitions of the ellipse and its tangent, that enable us to deal with the motion of bodies and other physical aspects.

We introduce the formalization of a theory of filters and ultrafilters, and the purely definitional construction of the hyperreal numbers of Nonstandard Analysis (NSA). The hyperreals form a proper field extension of the reals that contains new types of numbers including infinitesimals and infinite numbers.

By combining notions from NSA and geometry theorem proving, we propose an "infinitesimal" geometry in which quantities can be infinitely small. This approach then reveals new properties of the geometry that only hold because infinitesimal elements are allowed. We also mechanize some analytic geometry and use it to verify the geometry theories of Isabelle.

We then report on the main application of this framework. We discuss the formalization of several results from the *Principia* and give a detailed case study of one of its most important propositions: the *Propositio Kepleriana*. An anomaly is revealed in Newton's reasoning through our rigorous mechanization.

Finally, we present the formalization of a portion of mathematical analysis using the nonstandard approach. We mechanize both standard and nonstandard definitions of familiar concepts, prove their equivalence, and use nonstandard arguments to provide intuitive yet rigorous proofs of many of their properties.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Larry Paulson for his guidance and constant encouragement throughout my research. His advice and suggestions, as well as his insight, have been invaluable.

I would also like to thank Tobias Nipkow for suggesting Nonstandard Analysis. Many thanks to my colleagues in the Automated Reasoning Group of the Computer Laboratory, who have provided a friendly environment to work in. In particular, I am grateful to James Margetson, Clemens Ballarin, and Florian Kammüller for the useful and inspiring discussions that we have had. My thanks also to Lewis, our librarian, for his help and to Margaret and Hanni who provided administrative assistance.

The research reported in this thesis was supported by a scholarship from the Cambridge Commonwealth Trust, which I gratefully acknowledge. I would also like to thank the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom for granting me an Overseas Research Students award. The Computer Laboratory, Data Connection, and Clare College also kindly provided funding for visits to conferences.

Many thanks to all my friends, both here and at home, for their encouragement. I am especially grateful to Jenny for her friendship and her help.

To Diya, for her unfaltering support, patience, and generosity at all times, a mere thank you is not enough.

Last but not least, my heartfelt thanks and deepest gratitude go to my parents and my brother for their love, support, and inspiration over the years. And it is to them that I dedicate this thesis with my love and affection.

Contents

List of Figures

Glossary

AC Axiom of Choice

CAS Computer Algebra System

GTP Geometry Theorem Proving

HOL Higher Order Logic

IST Internal Set Theory

NS Nonstandard

NSA Nonstandard Analysis

UFT Ultrafilter Theorem

WUF Weak Ultrafilter Theorem

ZF Zermelo-Fraenkel set theory

ZFC ZF + AC

Chapter 1

Introduction

It is often argued that the history of nineteenth-century mathematics is that of the replacement of geometry by algebra and analysis. Synthetic or coordinate-free geometric reasoning which had produced some of the greatest achievements of mankind — Euclid's *Elements* and Newton's *Philosophiæ Naturalis Principia Mathematica* (the *Principia*) to name but two — was finally shifted from its long-standing central position in mathematics to a marginal one.

The process, aimed at increasing rigour in mathematics, arguably began with the introduction of analytic geometry and culminated with the expulsion of the infinitesimal. It laid the foundations of modern mathematics where rigour became the central tenet, mostly at the cost of intuition.

Throughout the seventeenth century though, and in the *Principia* in particular, geometric reasoning in the tradition of the ancient Greek geometers prevailed over everything else. Even though Newton uses his own infinitesimal procedures, these are introduced and justified geometrically to keep the reasoning of the *Principia* essentially in the spirit of the Ancients. The *Principia*, in this sense, embodies concepts that one might not only view as devalued but also difficult to make rigorous within the conceptual bounds of modern mathematics.

In this work, we show that a formalization that respects much of Newton's original reasoning is possible by combining adequate synthetic geometric methods with rigorous notions of infinitesimals. We mechanize, within the theorem prover Isabelle, procedures of the *Principia* that have often been regarded as logically vague by investigating and applying concepts from both mechanical geometry theorem proving (GTP) and Nonstandard Analysis (NSA). This results in an enriched geometry with powerful and intuitive tools that have application to mechanical theorem proving in ordinary Euclidean geometry as well.

The formalization of infinitesimals using NSA also results in a framework that is important and powerful in its own right. One of its first and natural applications is also investigated in this work: the mechanization of the calculus. We apply the rich array of concepts from NSA to the treatment of mathematical analysis in an intuitive and often illuminating way.

As a further observation, the powerful intuitions that infinitesimals provide are often used in constructing proofs of theorems. However, one is not allowed to use them in the proofs themselves without formal justification. So, an overall result that this work achieves is a rigorous and powerful treatment of geometry and analysis that incorporates the intuitive notions often used as tools into the

actual proof.

1.1 A Brief History of the Infinitesimal

The infinitesimal throughout its history has been at the centre of many controversies. Its use was widespread in geometry and other mathematical fields prior to the nineteenth century until a return to mathematical rigour drove it out of mathematics once and for all, or so it seemed. The infinitesimal calculus of Newton and Leibniz was then reformulated by the methods of Cauchy and Weierstrass to meet the modern standards of rigour. Yet today, through the power and sophistication of mathematical logic, the infinitesimal has been revived and been made acceptable again. Abraham Robinson's Nonstandard Analysis [72] is viewed as a viable Calculus of Infinitesimals. Moreover, NSA is regarded by some as a vindication of the informal use of infinitesimals of eighteenth-century mathematics against the absolute rigour of the nineteenth century, adding a new development to the long lasting war between the finite and the infinite, the continuous and the discontinuous [26].

The first calculus textbook was written by the Marquis de L'Hospital in 1696. His enthusiasm for the use of infinitesimals is obvious since right at the outset it is stated as an axiom that two quantities differing by an infinitely small amount can be substituted for one another. Thus, two quantities are simultaneously considered to be equal to each other and not equal to each other. Moreover, a second supposition of de L'Hospital regards a curve as the "totality of an infinity of straight segments, each infinitely small". De L'Hospital axioms imply a belief in the existence of the infinitely small quantities since it was common in that time, in the tradition of Euclid and Archimedes, to view axioms as empirical facts. On the *Principia* itself, de L'Hospital observes that it is "almost wholly of this [infinitesimal] calculus" [81].

It should be noted that even though de L'Hospital was a student of Leibniz, the latter did not share his belief in the existence of the infinitesimal. Indeed Leibniz was rather critical of his student's enthusiasm and, though he was a proponent of the use of infinitesimal methods, he had more subtle ideas regarding the foundations of his calculus. He did not claim that infinitesimals really existed. Leibniz instead viewed infinitesimals and infinitely large numbers as 'ideal' or fictitious numbers that still obeyed the same laws of arithmetic as ordinary numbers of mathematics. However, just like de L'Hospital, he also stated that two quantities differing by an infinitesimal amount could be viewed as equal. The inconsistency of these two assumptions was clear right from the start.

The belief in the existence of the infinitesimals despite the obvious foundational flaws prevailed throughout the eighteenth century in most parts of Europe. Although after the nineteenth century infinitesimal methods were no longer allowed in any rigorous arguments, physicists and engineers never ceased relying on them. The powerful and intuitive tools that infinitesimals provided, and still provide, have made sure that they never lose their appeal to them.

The original formulation of calculus by either Leibniz or Newton left much to be desired. The vagueness and lack of solid foundations made it prone to attacks on technical, philosophical and even theological grounds. What were these infinitesimals that could be both equal to and different from zero? In-

deed Bishop Berkeley's famous *Analyst* [10] constitutes a brilliant and invective attack on both Newton's theory of fluxions and Leibniz Differential Calculus. Berkeley's criticism, seemingly addressed to an "infidel" mathematician (most probably the astronomer Edmund Halley), was logically valid, devastating and left unanswered. In this, Berkeley said of infinitesimals:

> They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?

Throughout the nineteenth century, as the foundations of modern analysis were being set, mathematicians made sure that it was free of contradictions by doing what the Greeks had done previously: they banned the use of infinitesimals. It is only recently, in such a long history, that infinitesimals have become acceptable again. Nonstandard Analysis shows, by slightly modifying the ideas of Leibniz and his followers, how a consistent theory can be obtained. NSA introduces a new equivalence relation, $\approx$, that relates two quantities that differ only by an infinitely small amount. Two such quantities are no longer claimed to be equal but equivalent in a well-defined sense and can be substituted for one another in some cases but not in others. This effectively solves the contradiction quoted above from Bishop Berkeley that required an infinitesimal to be both zero and yet not equal to zero. NSA effectively shows that the "infidel mathematician" is innocent [75].

1.2 The *Principia* and its Methods

The *Principia* is considered to be one of the greatest intellectual achievements in the history of exact science. It has, however, been influential for over three centuries rarely in the geometrical terms in which it was originally written but mostly in the analytico-algebraic form that was used very early to reproduce the work. We examine some of the original methods used in the *Principia* in the following sections.

1.2.1 Newton's Style and Reasoning

Newton's reasoning rests on both his own methods and on geometric facts that, though well known for his time (for example, propositions of Apollonius of Perga and of Archimedes), might not be easily accessible to modern readers. Moreover, the style of his proofs is notoriously convoluted due to the use of a repetitive, connected prose. Whiteside [81] notes the following:

> I do not deny that this hallowed ikon of scientific history is far from easy to read... we must suffer the crudities of the text as Newton resigned it to us when we seek to master the *Principia's* complex mathematical content.

According to Whiteside's analysis of the mathematical principles behind the *Principia*, Newton's work is written in a relatively standard late seventeenth century form, that is a mixture of synthetic and algebraic-geometric reasoning. Moreover, Newton also adds his own informal geometrical-limit (or *ultimate*) arguments to the reasoning procedures; these are essential reasoning techniques on which we shall expand in the sequel.

Newton suggested after the publication of the *Principia* that he first did the work using calculus and then wrote it up in synthetic form so that the mathematics would not be too unusual or innovative for his audience. These suggestions are now widely believed to have been circulated due to the priority dispute that raged at the time between Newton and Leibniz about the invention of the calculus. In fact, even Newton's preliminary work on the motion of bodies has the same sort of synthetic form as the *Principia*. Newton's style of reasoning is reminiscent to some extent of that of Isaac Barrow who had shown that all the infinitesimal and algebraic geometric results of the mid-seventeenth century could be reconstructed using more or less proper synthetic geometry.

1.2.2 From Prose to Mathematical Statements

Sometimes Newton's prose tends to obscure the aim of particular arguments or geometric constructions. One might also argue, however, that the inventiveness of Newton's use of geometry often requires dedication on the part of the reader.

However, as remarked by Lamport [56], the difficulty with prose-style proof is the inherent lack of information about its logical structure. This observation applies aptly to the *Principia* since the reader has the often difficult task of extracting a proof out of Newton's elaborate prose. Moreover, to formalize the *Principia*, there is the additional burden of unravelling Newton's reasoning from the standpoint of a modern reader. Indeed, the mathematical notation of the seventeenth century can be rather obscure and primitive when compared with that used by twentieth century mathematicians. This can lead to ambiguities and wrong interpretation if not examined and checked carefully. The noticeable effect is a rather long formalization process.

Due to the much greater wealth of notation and concepts in modern mathematics, variables can now be named and formulae structured. This produces succinct mathematical statements rather than long winded, possibly erroneous or ambiguous ones. Consider the enunciation of **Lemma 11**, as a typical example from Newton's *Principia* (see Figure 1.1):

> The evanescent subtense of the angle of contact, in all curves having a finite curvature at the point of contact, is ultimately in the duplicate ratio of the subtense of the conterminous arc.

In order to formalize this lemma, we first need to recast Newton's terminology using more familiar geometric notions. The accompanying figure, to which Newton makes no reference in his enunciation, becomes essential for this process:

subtense of the angle of contact:
BD is Newton's so-called *subtense of the angle of contact.* It is the perpendicular to the tangent AD meeting the curve at B.

subtense of the conterminous arc:
The *subtense* is chord AB and arc AB is its *conterminous arc.*

duplicate ratio: The duplicate ratio of a number x is x^2.

So essentially, in modern notation, Newton is saying in Lemma 11 that $BD \propto AB^2$ ultimately.

The next task in formalizing this theorem and making it amenable to mechanization requires making notions such as *evanescent* and *ultimately* rigorous. These are special concepts from Newton's reasoning that have been the object of the current research work and which will now be introduced in more detail.

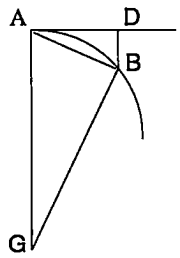


Figure 1.1: Diagram accompanying Newton's Lemma 11

In the various figures used by Newton, some elements must be considered as "very small": for example, we encounter lines that are infinitely or indefinitely small or arcs that may be nascent or evanescent. De Gandt [36] argues that there is a temporal infinitesimal that acts as the independent variable in terms of which other magnitudes are expressed. However, since time itself is often represented geometrically using certain procedures, the infinitesimal time or "particle of time" in Newton's own expression appears as distance or area.

1.2.3 The Infinitesimal Geometry of the *Principia*

On reading the enunciation of many of the lemmas of the *Principia*, one often comes across what Newton calls **ultimate** quantities or properties— for example, ultimate ratio (Lemmas 2,3,4 ...), ultimately vanishing angle (Lemma 6), and ultimately similar triangles (Lemma 8). Whenever Newton uses the term, he is referring to some "extreme" situation where, for example, one point might be about to coincide with another one thereby making the length of the line or arc between them vanishing, that is infinitesimal.

Furthermore, as points move along arcs or curves, deformations of the diagrams usually take place; other geometric quantities that, at first sight, might not appear directly involved can start changing and, as we reach the extreme situation, new ultimate geometric properties usually emerge. We need to be able to capture these properties and reason about them. The use of infinitesimals allows us to "freeze" the diagram when such extreme conditions are reached: we introduce, for example, the notion of the distance between two points being infinitesimal, that is, infinitely close to zero and yet not zero when they are about to coincide. With this done, we can then deduce new or ultimate properties about angles between lines, areas of triangles, similarity of triangles and so on. This is what distinguishes our geometry from ordinary Euclidean geometry.

The infinitesimal aspects of the geometry give it an intuitive nature that seems to agree with the notions of infinitesimals from Nonstandard Analysis. Unlike Newton's reasoning, for which there are no formal rules of writing and manipulation, the intuitive infinitesimals have a formal basis in Robinson's NSA. This enables us to master motion, which is part of Newton's geometry, and

consider the relations between geometric quantities when it really matters, that is at the point when the relations are ultimate.

1.3 On Nonstandard Analysis

The rise of Nonstandard Analysis in recent years has provided what is regarded as a good alternative to the classical treatment of mathematical analysis. It deals simply and elegantly with concepts and processes — essentially those associated with the notion of limits — that are at the heart of analysis. The nonstandard approach not only shortens many proofs, but also removes the huge gap that lies between an initial intuitive approach to analysis — found, say, in a first course in calculus — and a rigorous one offered in more advanced classical analysis.

Many people, who are interested in analysis primarily as a tool, are discouraged from using it when faced with the exacting, rather dry nature of the standard treatment. Viewed in this light, the nonstandard approach can make mathematical analysis available as a rigorous tool to a wider audience, while not forsaking intuition.

The general agreement about the power of NSA has already resulted in its application in several different fields: economics [71] and physics [2], for instance. However, its role in elementary mathematics education is still a matter of much discussion and controversy. The main drawback, it is often believed, lies in the need to make the logical foundations of NSA clear before it can be used safely and effectively. Such a task is not easy and opinions differ on how much mathematical background is needed. There has been progress in the exposition of NSA over the years: Robinson's original treatment required a rather advanced understanding of mathematical logic and model theory. Subsequent reformulations of the theory, using set theory for example, have made it more accessible to a wider and less specialized audience. The current work has benefited from these later developments. Once the foundations of NSA have been clearly understood though, its algebraic power and simplicity bear much fruit in teaching analysis.

Perhaps the best approach though lies in a combination of the standard and nonstandard treatment of analysis. The nonstandard treatment can then be used in situations where it brings a clear advantage to the classical treatment. Our work, in that regard, can be viewed as taking this middle way: standard and nonstandard definitions are provided for all concepts developed. The nonstandard methods then make use of the properties of standard mathematics to obtain those of the nonstandard extensions. Also, the nonstandard proofs are usually compared and contrasted with the corresponding standard ones to highlight any conceptual and mechanical benefits gained.

1.4 Objectives

One of the main objectives of this research is to study the geometric proofs of the *Principia* and investigate ways of mechanizing them using the rigorous framework provided by the computer proof assistant Isabelle [66]. From the outset, a major consideration has been to respect as much as possible New-

ton's reasoning. We want to demonstrate, with the help of (mathematical) logic and geometry theorem proving, that Newton's proof sketches are rigorous even though they might appear informal. We are especially interested in validating Newton's ultimate reasoning procedures which deal with vanishing or infinitesimal quantities.

To progress towards the above objective, a sound formalization of infinitesimals is necessary. Since Robinson's NSA makes the infinitely small respectable and rigorous, this makes it fit for mechanization. Our task therefore is to develop, through the use of definitions only, a theory which introduces the new types of numbers from NSA. Then by combining concepts from NSA and geometry, we expect to capture the reasoning that takes Newton's geometry beyond conventional Euclidean geometry.

Also, in our development of Euclidean and infinitesimal geometry, some basic analytic geometry is carried out. The approach is also purely definitional and is aimed at verifying the rules of geometry emanating from the methods formalized in Isabelle.

The last motivation for this work is to provide a mechanized treatment of analysis within a nonstandard framework. This aspect of our research arose as a result of the formal development of infinitesimals, and of the other classes of nonstandard numbers in Isabelle. Notions from NSA are used to give a simpler, algebraic formulation to many familiar concepts from standard calculus. We aim to show the numerous advantages that infinitesimals, and the nonstandard approach in general, bring to the mechanization of analysis in terms of shorter, simpler and more intuitive proofs.

1.5 Achieving our Goals

To achieve the above objectives, various theories had to be built in Isabelle. We now give an overview of the development by highlighting our contributions and describing the main components of this research. Figure 1.2 gives a diagrammatic overview and the relations between the various parts. Below, the numbers in brackets refer to the diagram.

Constructions up to the reals (1,2).
Since Isabelle only had the natural numbers and a simple integer theory (at the time), we constructed all the number systems up to the reals. These are the positive natural numbers, the positive rationals, the positive reals, and then the reals. All the important properties of each number system had to be proved before the next one could be constructed.

Mechanization of geometry (3).
To mechanize the geometric arguments of the *Principia*, a geometry theory has been developed in the theorem prover Isabelle. The main concern was to have methods powerful enough for us to derive the theorems that follow from Newton's diagrams. No geometry theory had been built in Isabelle previously; so, we had to investigate and formalize geometric methods that would fit nicely within Isabelle's framework while satisfying our aims.

Construction of hyperreals (4).
The theory of nonstandard reals extends that of the real numbers in Isabelle. The construction involved adding new theories about filters and

ultrafilters and also proving Zorn's Lemma in Isabelle HOL. The algebra of the operations on the nonstandard numbers was developed and new relations introduced. The hypernaturals were also constructed; these extend the natural numbers.

Combination of geometry and Nonstandard Analysis (5).
A theory combining geometry and infinitesimals has been developed. This contains proofs of infinitesimal geometry theorems. Nonstandard tools are developed that can also be applied to Euclidean geometry. An elementary theory of vectors is formalized which is used to verify the geometry axioms of Isabelle.

Mechanization of Principia (6).
The combination of geometry and nonstandard numbers have then been applied to Newton's *Principia*. Proofs of several important Lemmas and propositions of the *Principia* are formalized in the theory. An anomaly in Newton's demonstration of the Kepler Problem is revealed using the rigorous techniques.

Mechanization of Nonstandard Analysis (7).
Some elementary analysis has been investigated using the nonstandard approach. The hyperreals provide a wide array of numbers and notions that model familiar concepts from mathematical analysis.

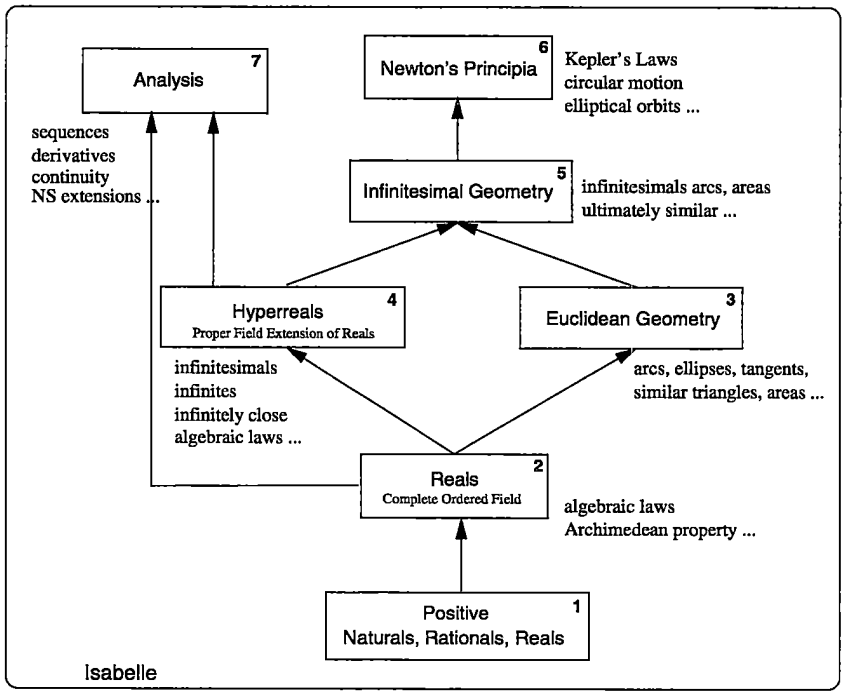


Figure 1.2: Overview of development

1.6 Organisation of Thesis

This thesis first provides a survey (Chapter 2) of the background of geometry theorem proving prior to the publication of Wu's deeply influential method. We then review and point out the relative merits of the algebraic and coordinate-free techniques that have been successfully applied to GTP since 1977. The choice of which geometric methods to formalize in Isabelle is discussed. The new notions with which the GTP methods are extended to deal with various (physical) concepts from the *Principia*, such as motion, are also introduced.

We then discuss the ultrapower construction of the hyperreals in Isabelle (Chapter 3). Using the reals and numbers beyond, the concept of an infinitesimal geometry is examined where quantities can be infinitely small and new geometric properties emerge (Chapter 4).

Hyperreal vectors and their associated algebra are then used to provide a definitional foundation to the geometry techniques axiomatized in Isabelle. The analytic geometry theory is used to capture the infinitesimal geometric notions as well. We then describe how the various tools — hyperreals, infinitesimal geometric relations etc. — are applied to the mechanization of some of the Lemmas and Propositions from Newton's *Principia* (Chapter 5). This shows how various notions from Newton's reasoning are formalized using a combination of NSA and geometry. The mechanization of one of the most important theorems of the *Principia* is examined as a detailed case study; the goal is to demonstrate clearly the interaction between geometry and the tools of nonstandard analysis. A flaw in Newton's reasoning emerges through the rigour of mechanization within the nonstandard framework.

As a further contribution, the hyperreals are applied to the formal development of some aspects of nonstandard real analysis or infinitesimal calculus (Chapter 6). This includes concepts from the theories of limits, series, continuous functions, and differentiation. Finally, we offer our conclusions and survey some possible further work (Chapter 7).

Chapter 2

Geometry Theorem Proving

This chapter first surveys the early development of mechanical theorem proving in geometry. GTP was initially viewed as an artificial intelligence problem that many believed would be easily tackled by machines. However, the initial optimism soon vanished as various difficulties associated with the domain emerged and no significant results could be proved. The developments over the last twenty years or so though have led to a revolution in the field. We survey several of the powerful methods proposed and analyse our own choice for formalization in Isabelle. Concepts that are important to deal with the geometry of the *Principia* are also introduced.

2.1 Historical Background

In 1899, Hilbert proposed five groups of axioms in his *Grundlagen der Geometrie* (Foundations of Geometry) [46]. In this classic work, Hilbert showed the consistency and independence of the sets of axioms and from them derived the various properties of plane (Euclidean) geometry. Hilbert's geometry consists of points, lines, and planes as primitives and of relations between these for angle congruence and incidence amongst others. The relationships between the primitives are completely determined by the axioms. Hilbert's insistence that no geometrical intuition was needed to prove any results — he suggested that the primitives could be replaced by chairs, tables, and beer mugs, as long as these satisfied the axioms — marked a clear departure from the geometry of Euclid. Ancient Greeks geometry was meant as an axiomatization of concepts that were intuitively obvious while Hilbert's abstracted geometry away from any concrete interpretation.

The role of Hilbert's *Grundlagen* in relation to mechanical GTP is quite an important one. Indeed, the formalization given by Hilbert made clear for the first time the possibility of mechanizing elementary geometry. This was realized by Poincaré who, with great prescience, argued the following in his review of the *Grundlagen* (1902) [70]:

> Thus Hilbert has, so to speak, tried to put the axioms in such a form that they could be applied by someone who did not understand their meaning because he had never seen a point, a straight line, or a plane. Reasoning should, according to him, be capable of

> being carried out according to purely mechanical rules, and for doing geometry it suffices to apply these rules to the axioms slavishly without knowing what they mean. In this way one could build up all of geometry, I will not say without understanding anything at all since one must grasp the logical sequence of the propositions, but at least without perceiving anything. One could give the axioms to a logic machine, for example the *logical piano* of Stanley Jevons, and one would see all of geometry emerge from it.

Hilbert's geometry, despite its great influence, has been criticised though for giving the same status to points, lines, and planes. It can be contrasted with Tarski's theoretical contribution published in 1926 which is an axiom system for Euclidean geometry. In Tarski's geometry, the universe contains only points with two primitive relations on them: betweenness and equidistance. A direct consequence was that much more work had to be done to prove results in the Tarski system.

As far as actual machine geometry is concerned, the first geometry theorem prover (the Geometry Machine) was developed by Gelernter in 1959; it was then extended and used to prove a number of theorems from high-school textbooks of the time [37]. Gelernter's Geometry Machine included specific heuristic knowledge about the geometry domain and had a backward chaining search strategy. The main heuristic built into the Machine was to use the diagram accompanying the statement of the geometry problem to reject false goal statements.

More work was also done using the same approach by Gilmore [38], Nevins [63] and Elcock [29] with additions such as forward chaining for example. The axiomatic or synthetic approaches used by Gelernter and the others above were, however, not very successful in proving or discovering any non-trivial theorems. The main problem, despite the numerous search strategies and heuristics that have been tried, is the high inefficiency; this is due to factors such as the huge search space of geometry rule applications. Koedinger argues that traditional geometry problem solving is hard [54] and outlines how the number of inferences that can be made rapidly increases at each layer in the proof (from seven at the beginning of the proof of a typical problem to over 100 000 at the third layer where a minimum of six layers are required). This makes it essential to add sophisticated search strategies and heuristic knowledge to GTP systems for them to have any chance of proving anything in the geometry domain. Koedinger further argues that the lack of success of these approaches lies in the fact that the underlying problem representation on which most of them are built has remained the same; namely one that has the formal geometry rules as operators and requires a search in the problem space for the rules that can be applied.

In 1969, Cerutti and Davis, rather ahead of their time, used symbolic manipulation in a system called FORMAC to prove theorems in elementary analytic geometry [15]. They used Descartes' method, that is an essentially algebraic approach that assigns coordinates to points, to prove Pappus' theorem. In the same paper, the authors also outlined how they obtained two new theorems by going through the output of the machine. Even though this algebraic approach was relatively successful, Wang thinks it unfortunate that the techniques described in it were not investigated any further to give a general GTP method [79]. It took several more years before the potential of the algebraic approach was finally recognized. This changed a rather stagnant area into one which has

achievements ranking, most probably, amongst the best in automated reasoning so far.

2.2 Algebraic Techniques

The algebraic techniques usually proceed through the introduction and use of some coordinate system. The geometry problem or statement is then translated into an algebraic form that can be dealt with using a number of powerful algorithms. We outline a few of the most successful algebraic techniques next.

2.2.1 Wu's Method

The work of Wu Wen-tsün laid the foundations for automatic theorem proving in geometry. Wu's seminal paper, first published in China in 1977, heralded a new era in GTP and radically improved the power of mechanical reasoning in geometry. The method has been successfully applied to Euclidean geometry and has been used to discover several non-trivial geometric theorems. Wu's technique, in contrast to Tarski's complicated decision procedure for elementary geometry, is a feasible method for mechanical GTP. To prove a particular geometry theorem, the method works as follows:

- First, it translates the geometry theorems into polynomial equations, polynomial inequations, and polynomial inequalities. This corresponds to the initial introduction of coordinates and symbolic transformation of the problem whereby geometric relations and entities are converted into corresponding algebraic ones. This approach can be traced back to Hilbert's *Foundations of Geometry*, where starting from an initial set of axioms, a number system and a coordinate system are introduced to provide a model for the axioms. The polynomial inequations mentioned above express the so-called degenerate cases that need to be ruled out (see Section 2.4.2).

- Next, the method decides algorithmically if the conclusion c of the geometric theorem follows from the relations that make up the hypothesis of the theorem under the non-degenerate conditions. This step involves triangulating the hypothesis polynomials $\{h_1, \ldots, h_n\}$ using pseudo-division to yield the characteristic set C of the h_i's and the pseudo-remainder r of c with respect to C. Then, if r is zero, the geometry statement is generally true. Otherwise, the set C is further decomposed by factoring and each component examined to decide the validity of the statement. As a final stage, the non-degenerate conditions can be analysed, if necessary and if possible [83, 79].

This simple algebraic procedure works well to prove a remarkable number of theorems. Moreover, it can even prove theorems that are under-specified and generate sufficient non-degeneracy conditions to make a particular geometry statement valid. Wu's method acts as a complete decision procedure for statements whose hypotheses and conclusion can be expressed by polynomial equations and has a rather deep underlying mathematical theory. We urge the interested reader to consult Wu's influential paper for a thorough exposition of this powerful mechanical procedure [83].

2.2.2 Gröbner Bases Method

Wu's work resulted in a rekindled interest in the field of GTP and led to the application of Buchberger's Gröbner bases method to automated GTP [51, 55]. As outlined by Wang in his survey paper [79], there are two main approaches to using Gröbner bases in geometry:

- The first approach, mainly due to Kutzler and Stifter [55], and Chou [21] involves computing the Gröbner bases, G, of the set of hypothesis polynomials and then the normal form h of the conclusion polynomial c modulo G. The polynomial c is in normal form modulo G if and only if it cannot be reduced to another polynomial using the elements of G. It has been shown that any polynomial can be reduced to normal form modulo some set of polynomials in finitely many steps. The theorem is then true if the normal form $h = 0$

- The other main approach is due to Kapur [51] and involves deciding whether a finite set of polynomials does not have a solution in an algebraically closed field. According to Kapur, using Hilbert's *Nullstellensatz*, this refutational approach is equivalent to checking whether 1 is in the ideal generated by the polynomials. Thus, as described by Wang [79], the method proceeds by computing a Gröbner basis G^* of $\{h_1, \ldots, h_n\} \cup \{zc - 1\}$ and then checking whether $1 \in G^*$.

There exist several other well-established algebraic methods for automated geometry theorem proving. We have, for example, those based on Cylindrical Algebraic Decomposition [24] developed by Arnon [4] and by Buchberger et al. [13] amongst others. These quantifier elimination methods can be applied to an array of decision problems that can be expressed in prenex form [79]. The method was developed as an improvement over the original complex, decision procedure proposed by Tarski for elementary geometry [77]. Wu's method has also been improved and even combined with Arnon's in an attempt to deal with theorems that involve inequalities [16].

2.2.3 On the Algebraic Methods

Computer algebra systems (CAS) seem ideally suited for dealing with the various GTP methods described above. Indeed, packages for Gröbner bases are routinely provided by most serious systems, for example. There has been work done by Fearnley-Sander [32], Kutzler et al. [55] and others in which geometric problems are translated into corresponding algebraic ones that are then solved by computer algebra methods.

Algebraic reasoning techniques despite their power are, however, not perfect and have several crucial limitations. They have problems, in general, dealing with geometric situations involving inequalities. Therefore, in geometric terms, this means that techniques such as Wu's and the Gröbner basis methods cannot handle adequately concepts involving *order relations* such as *betweenness* and *congruent angles*. Traditional synthetic (logical) geometric reasoning methods, however, have no problems with order relations; so an interesting approach might be using both of these approaches to tackle GTP. Matsuyama and Nitta [61] have done some work on integrating logical and algebraic reasoning. Their

system suffers from several problems though, especially on the side of the logical reasoner which is not strong enough to represent some of the geometric relations. Arnon [4] has also done some work on combining, rather than integrating, the two approaches but his work mainly uses the logical reasoner for expressing the geometric problem (through the use of predicates) and relies on the algebra component for actual geometric reasoning.

In the light of recent work done by Harrison [43], Ballarin and Paulson [7], and others in interfacing theorem provers such as Isabelle and HOL, with computer algebra systems, it might be interesting and fruitful to have the two components cooperate to solve geometry problems. The more powerful capabilities of theorem proving systems would improve logical reasoning, where needed, while the algebra system could then be used to deal (easily in many cases) with the aspects of the proof not involving order relations. The problems would be expressed logically in the theorem prover using primitive predicates which would also be responsible for transformation and interaction with the computer algebra system when needed.

2.3 Coordinate-Free Techniques

The main drawbacks of the algebraic techniques are the long and hard to read proofs they usually generate. This makes it difficult to explain the proofs geometrically and might be viewed as taking away the intuitive appeal usually associated with geometry reasoning. We next describe a few of the so-called coordinate-free techniques that have been used successfully in automated GTP. The techniques are generally not as powerful as those involving coordinates but have been quite successful in proving a large number of theorems in several provers. Moreover, the geometric interpretation that can be attached to these methods adds a lot to their appeal. We start with a description of Clifford algebra, which has been generating a lot of attention recently.

2.3.1 Clifford Algebra

A geometric algebra is one which can be used as a suitable grammar for representing basic geometrical relations. Clifford Algebra has been proposed as a unifying framework for various fields encompassing physics, mathematics and computer science since it admits a geometric interpretation [45]. In fact, Clifford (in 1878) originally suggested the name *Geometric algebra* for what he called a "grammar of space". Once the algebra is chosen, the next step is to construct a theory or calculus for representing complex geometrical relations and structures.

Directed Number Systems

The first step in developing a geometric calculus involves choosing and encoding the *basic* geometric concepts in symbolic form. The concepts of magnitude and direction are usually taken as basic, and the concept of vector as the basic kind of *directed number*.

A directed number is defined algebraically but interpreted geometrically. The number is defined implicitly by specifying rules for adding and multiplying vectors. The vectors are assumed to generate an associative algebra in which

the square of every vector is a scalar. Let $\mathcal{V}_n$ be an n-dimensional vector space over the real numbers. For any vectors u, v, w in $\mathcal{V}_n$, the *geometric product* is defined by the basic axioms:

$$\begin{aligned} &u(vw) = (uv)w \\ &u(v+w) = uv + uw \\ &(v+w)u = vu + wu \\ &uu = |u|^2 \quad \text{(contraction rule)} \end{aligned}$$

The contraction rule determines a measure of distance between vectors in $\mathcal{V}_n$ and the vector space can be regarded as an n-dimensional Euclidean space.

These simple rules defining vectors are the basic rules of *Clifford Algebra.* They determine the mathematical properties of vectors completely and generate a rich mathematical structure. The outstanding feature that differentiates Clifford Algebra from other algebraic systems is the geometric interpretation that can be attributed to multiplication. To make this apparent, the geometric product uv is decomposed into symmetric and antisymmetric parts defined by

$$\begin{aligned} u \cdot v &= \tfrac{1}{2}(uv + vu) \\ u \wedge v &= \tfrac{1}{2}(uv - vu) \end{aligned}$$

giving

$$uv = u \cdot v + u \wedge v$$

The symmetric product $u \cdot v$ is the conventional Euclidean dot or *inner product* on a vector space with its usual geometrical interpretation. The antisymmetric product $u \wedge v$ is called the *outer product* and is neither a scalar nor vector. The outer product $u \wedge v$ is called a *bivector* and geometrically can be interpreted as a directed area, in the same way that a vector represents a directed line. The two products combined together yield the geometric product uv and determine its geometric interpretation:

$$\begin{aligned} uv = -vu &\iff u \cdot v = 0 \\ uv = vu &\iff u \wedge v = 0 \end{aligned}$$

This means that the geometric product uv of nonzero vectors u and v anticommutes if and only if u and v are *orthogonal,* and commutes if and only if the vectors are *collinear.* The geometric product thus completely describes the relative directions and magnitudes of the vectors u and v since, for example, the degree of commutativity of the product uv is a measure of the direction of u relative to v. In fact, the quantity $u^{-1}v = uv/uu$ is a measure of the relative direction and magnitude of u and v. Another geometric interpretation can regard $u^{-1}v$ as an operator which transforms u into v, as expressed algebraically by the equation $u(u^{-1}v) = v$. We thus have a complete and powerful geometric interpretation for the product of two vectors. This can be extended to the product of many vectors.

Clifford Algebra and GTP

Clifford algebraic techniques have been studied extensively in recent times and used for automated theorem proving in several geometries. Li and Cheng, for example, have done some notable work in applying a combination of Clifford

algebra and Wu's method to differential geometry [59]. The work is able to prove theorems in the local theory of space curves. Also worth mentioning is the work by Wang who derived sets of rewrite rules for simplification of Clifford algebraic expressions [80]. Wang's approach was also applied to problems in computer vision. The keen interest in applying Clifford algebra to GTP results from its expressiveness: it allows numerous concepts to be expressed in various geometries. Furthermore, it provides an efficient approach to automatic GTP while remaining more or less meaningful geometrically when compared to the algebraic methods involving coordinates.

Despite the advantages of Clifford Algebra, it does not seem, at first sight, directly relevant to our work since it does not match closely the geometric concepts and infinitesimal nature of Newton's proofs. However, the simple and intuitive algebra is of intrinsic interest since its formalization would open up application in many areas. In Chapter 4, we outline a related but simpler and more familiar theory based on vectors. This is formalized in Isabelle mainly for verifying its geometric theory.

In the next sections, we introduce more "traditional" methods that are based on *geometric invariants* [19, 20] and high level geometry lemmas about these invariants. These are powerful methods developed originally in China (by J. Z. Zhang) for geometry education.

2.3.2 The Area Method

A particular property is ideal as an invariant if it ensures that the proofs generated are short. Also, the methods should be powerful enough to prove many properties without adding auxiliary points or lines. The other important aspect is to achieve diagram independence for the proofs, that is, the same proof can be applied to several diagrams.

In this method, there are basic lemmas about geometric properties called signed areas. Other rules are obtained by combining several of the basic ones to cover frequently-used cases and simplify the search process.

The line segment from point A to point B is represented by $A --- B$, its length by $\texttt{len}\ (A --- B)$, and the *signed* area $\mathtt{S}_\triangle ABC$ of a triangle is the usual notion of area with its sign depending on how the vertices are ordered[1]. The usual approach consists in having $\mathtt{S}_\triangle ABC$ as positive if A–B–C is in anti-clockwise direction and negative otherwise. Familiar geometric properties such as collinearity, `coll`, and parallelism, $\|$, can then be defined as follows

$$\begin{aligned} \texttt{coll}\, a\, b\, c &\equiv (\mathtt{S}_\triangle\, a\, b\, c = 0) \\ a --- b \,\|\, c --- d &\equiv (\mathtt{S}_\triangle\, a\, b\, c = \mathtt{S}_\triangle\, a\, b\, d) \end{aligned}$$

New points can be introduced and the *signed area of a quadrilateral* $\mathtt{S}_\square\, a\, b\, c\, d$ defined in terms of signed areas of triangles:

$$\mathtt{S}_\triangle\, a\, b\, c = \mathtt{S}_\triangle\, a\, b\, d + \mathtt{S}_\triangle\, a\, d\, c + \mathtt{S}_\triangle\, d\, b\, c \tag{2.1}$$

$$\mathtt{S}_\square\, a\, b\, c\, d \equiv \mathtt{S}_\triangle\, a\, b\, c + \mathtt{S}_\triangle\, a\, c\, d \tag{2.2}$$

[1] Isabelle's notation `s_delta` will also be used to denote the signed area.

Properties about $\mathtt{S}_\square$ can now be derived by changing the order of the points and sign of the geometric quantity. We have, for example:

$$\begin{aligned} \mathtt{S}_\square\, a\,b\,c\,d &= \mathtt{S}_\square\, b\,c\,d\,a \\ \mathtt{S}_\square\, a\,b\,c\,d &\equiv \mathtt{S}_\triangle\, a\,b\,d + \mathtt{S}_\triangle\, b\,c\,d \qquad (2.3) \\ -\mathtt{S}_\square\, a\,b\,c\,d &\equiv \mathtt{S}_\triangle\, b\,a\,c + \mathtt{S}_\triangle\, a\,d\,c \qquad (2.4) \end{aligned}$$

An example

The power and elegance of the area method can be seen at work in the proof of the relatively hard Pascal's theorem[2].

EXAMPLE 2.3.1.. *Let A, B, and C be three points on one line and let P, Q, and R be three points on another line. If* $C - Q \parallel B - P$ *and* $R - B \parallel Q - A$ *then* $C - R \parallel A - P$ (Figure 2.1).

Below we give a rather detailed overview of how this example can be stated and proved in Isabelle. Each step shows the currents subgoal (after the $\Longrightarrow$) to be proved by applying one or more tactics. The premises, especially if they have been modified, are also shown in some cases.

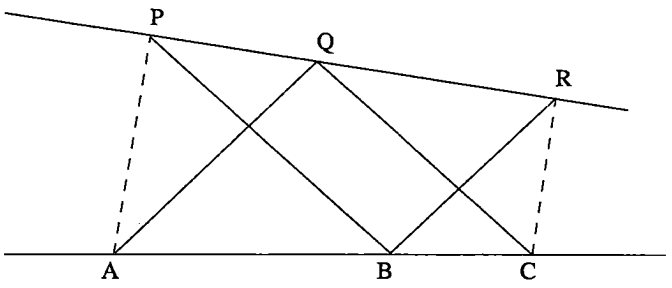


Figure 2.1: Pascal's Theorem

Main Goal
$[|\mathtt{coll}\ A\,B\,C; \mathtt{coll}\ P\,Q\,R; C - Q \parallel B - P; R - B \parallel Q - A|]$
$\Longrightarrow C - R \parallel A - P$

Our aim is to show that the area of $\triangle CAP = \triangle RAP$ which follows from the definition of parallel lines. To assist the simplification later on, we substitute by $\mathtt{S}_\triangle RAP = -\mathtt{S}_\triangle RPA$:

$[|\mathtt{coll}\ A\,B\,C; \mathtt{coll}\ P\,Q\,R; C - Q \parallel B - P; R - B \parallel Q - A|]$
$\Longrightarrow \mathtt{s\_delta}\ C\ A\ P = -\mathtt{s\_delta}\ R\ P\ A$

Next, the theorem *para_sum_area*, shown below, is used to express one of the signed areas in the conclusion of the subgoal as the sum of the areas of two triangles.

para_sum_area
$[|\mathtt{coll}\ A\,B\,C; C - Q \parallel B - P|]$
$\Longrightarrow \mathtt{s\_delta}\ C\ A\ P = \mathtt{s\_delta}\ Q\ B\ P + \mathtt{s\_delta}\ B\ A\ P$

[2]This theorem is sometimes known as Pappus' theorem (Artin [5]).

This theorem is based on property 2.1 of section 2.3.2 and on the definition of parallel lines. The following goal with its associated premises now results:

$$\begin{array}{l}[|\texttt{s\_delta}\ C\ A\ P = \texttt{s\_delta}\ Q\ B\ P + \texttt{s\_delta}\ B\ A\ P;\\ \quad \texttt{s\_delta}\ R\ P\ A = \texttt{s\_delta}\ B\ Q\ A + \texttt{s\_delta}\ Q\ P\ A|]\\ \Longrightarrow \texttt{s\_delta}\ C\ A\ P = -\texttt{s\_delta}\ R\ P\ A\end{array}$$

The next steps are trivial and follow from the theorems we proved about areas of quadrilaterals. The subgoals are routinely proved by Isabelle's simplifier, thereby proving Pascal's Theorem. We give a rather more detailed Isabelle proof to show the area method at work. Replacing with the decompositions of $\texttt{s\_delta}\ C\ A\ P$ and $\texttt{s\_delta}\ R\ P\ A$, we simplify the goal to:

$$\Longrightarrow \texttt{s\_delta}\ Q\ B\ P + \texttt{s\_delta}\ B\ A\ P = -(\texttt{s\_delta}\ B\ Q\ A + \texttt{s\_delta}\ Q\ P\ A)$$

By applying theorems (2.3) and (2.4), we can reduce the sums of signed triangular areas in the goal to one involving the signed areas of quadrilaterals. This new goal is then trivially true.

$$\Longrightarrow \texttt{s\_quad}\ Q\ B\ A\ P = - - \texttt{s\_quad}\ Q\ B\ A\ P$$

2.3.3 The Full-Angle Method

The concept of the angle and its associated properties provide powerful tools that have been used traditionally in geometry theorem proving. In their work on producing automated readable proofs, Chou et al. [20] also propose a method based on the concept of *full-angles* that can be used to deal with classes of theorems that pose problems to the area method. A full-angle $\langle u, v\rangle$ is the angle from line u to line v measured anti-clockwise. We note that u and v are lines rather than rays; this has the major advantage of simplifying proofs by eliminating case-splits in certain cases.

The full-angle is then used to express other familiar geometric properties and augment the reasoning capabilities of the geometry theory. For example, Chou et al. express that two lines are perpendicular, $\perp$, as follows,

$$u \perp v \equiv \langle u, v\rangle = \langle 1\rangle$$

where $\langle 1\rangle$ is a constant in their theory satisfying $\langle 1\rangle + \langle 1\rangle = 0$. In addition, there are other rules that enable new lines to be introduced such as:

$$p -\!\!- x \parallel u -\!\!- v \Longrightarrow \langle a -\!\!- b, p -\!\!- x\rangle = \langle a -\!\!- b, u -\!\!- v\rangle$$

In their paper [20], Chou et al. propose fourteen basic rules about the properties of full-angles and some seven extra rules obtained (without proofs) by combining the basic ones. In our approach, described in Section 2.4.1, we define the notion of full-angle between lines and derive some of the basic and combined rules needed for our development.

A Simple Example: Euclid I.29

Euclid's proposition 29 of Book I [30] can be easily proved using the full-angle method. The proposition states that if $A -\!\!- B \parallel C -\!\!- D$ and the transversal

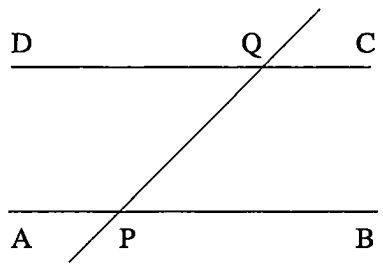


Figure 2.2: Euclid's Proposition I.29

$P — Q$ intersects $A — B$ and $C — D$, then $\langle A — B, P — Q\rangle = \langle C — D, P — Q\rangle$ (Figure 2.2).

This theorem admits a straightforward proof by using the rules about full-angles given in Section 2.3.3, since the angle between two parallel lines is zero.
Proof:

$$
\begin{aligned}
A -- B \parallel C -- D \implies & \langle A -- B, C -- D\rangle =_a 0 \\
& \langle A -- B, P -- Q\rangle + \langle P -- Q, C -- D\rangle =_a 0 \\
& \langle A -- B, P -- Q\rangle =_a -\langle P -- Q, C -- D\rangle \\
& \langle A -- B, P -- Q\rangle =_a \langle C -- D, P -- Q\rangle
\end{aligned}
$$

This demonstration also shows how an angle can be split to introduce a line into the goal. This typically happens when doing a forward proof construction in Isabelle. On the other hand, in a backward proof, this corresponds to eliminating a line from the current goal and reflects the way in which the automatic method actually proceeds. By working in the interactive environment of Isabelle, there is the added flexibility of being able to work in the forward direction, which is not possible in the fully automatic provers. Another advantage is that variable instantiation can be delayed in Isabelle, so one can bind a newly introduced line when sufficient information becomes available.

2.3.4 Which Method?

The choice of which particular geometric method to develop in Isabelle was an interesting one. On the one hand, there are the highly successful algebraic methods, as we have outlined above, and on the other hand, the relatively less successful synthetic techniques. The efficiency of the algebraic GTP techniques makes the former ideally suited for automatic theorem provers but in our case they do have several limitations.

A first possible candidate for formalization in Isabelle was Hilbert's geometry, especially considering Poincaré's comment on the classic work. In fact, we did formalize the first two groups of axioms proposed by Hilbert as a simple case study on geometric proofs in Isabelle. The first group of seven axioms establishes the connection between the fundamental concepts that exist in Hilbert's universe, namely points, lines, and planes. The two theorems stated by Hilbert (without proof) about lines and planes intersections were verified. The second group of five axioms defines the concept of a line segment AB as the set of points lying between A and B. These axioms effectively axiomatise the idea of

betweenness which was implicitly assumed by Euclid. With the second set of axioms, we were able to prove several of Hilbert's results including the important theorem that any simple polygon divides the plane into two disjoint regions, an interior and an exterior, and that the line joining any point in the interior with any point in the exterior must have a point in common with the polygon.

There was scope for formalizing the remaining three groups of axioms in Isabelle but our partial mechanization indicated that a rather large amount of work would be involved to get to the results needed for our proofs of Newton's *Principia.* Our main aim was not the formalization of geometry in Isabelle just for the sake of mechanical theorem proving in Euclidean geometry: there are already many successful automatic provers dealing with this, as outlined above. We were interested in formalizing concepts from powerful, established methods that would fit well with the type of reasoning present in the *Principia.*

Thus, we chose the full angle method since it deals extremely well with theorems about circles (and angles). Moreover, its blend of algebra and synthetic deduction provides geometrically intuitive and high-level steps that enable proofs to remain short. The area method has similar advantages. Also, Newton himself uses area properties and ratios of segments in many cases when proving his Lemmas and Propositions. As a result, rules from the area method were also formalized to produce Isabelle's geometry theory.

2.4 Formalizing Geometry in Isabelle

As just mentioned, the rules and definitions that we use constitute a mixture of the algebraic and synthetic approach. The algebra is, however, much simpler than that arising from the introduction of coordinates (c.f. Section 2.2) and most of the work can be viewed as being done in the spirit of logical deduction rather than pure algebraic reasoning. The area axioms in Isabelle are related to the ones proposed by Chou, Gao, and Zhang in their work on Euclidean geometry using the area method [17]. In this work, Chou et al. present a strict set of axioms that enables their GTP technique to work for affine geometries over any fields. This suits our purpose well as we certainly want a minimal set of axioms. In fact, by introducing definitions, we manage to derive some of the axioms of Chou et al. as theorems.

2.4.1 Defining the Theories

The Axioms of the Area Method

We present below the axioms that are used for GTP in Isabelle. Two types of geometric concepts, namely, points and lines are introduced. If a point x is on a line l ($\mathtt{on}\ (x, l)$), or equivalently, l goes through $\{x\}$, then they are said to be *incident.* This is the only basic geometric relation:

$$\mathtt{incident}\ A\ l \equiv \forall x \in A.\ \mathtt{on}\ (x, l)$$

The properties that are required are:

1. $a \neq b \Longrightarrow \mathtt{incident}\ \{a, b\}\ (a \mathtt{---} b)$
2. $\mathtt{len}\ (a \mathtt{---} a) = 0$

3. $a \neq b \Longrightarrow \exists p.\ \texttt{incident}\ \{p\}\ (a --- b) \wedge$
   $\texttt{len}\ (a --- p) + \texttt{len}\ (p --- b) = \texttt{len}\ (a --- b)$

4. Three points a, b, and c determine a signed area $\texttt{s\_delta}\ a\ b\ c \in \mathbb{R}^*$ satisfying $\texttt{s\_delta}\ a\ b\ c = \texttt{s\_delta}\ c\ a\ b = \texttt{s\_delta}\ b\ c\ a = -\texttt{s\_delta}\ b\ c\ a = -\texttt{s\_delta}\ b\ c\ a = -\texttt{s\_delta}\ c\ b\ a = -\texttt{s\_delta}\ a\ c\ b$

5. There exists at least three non-collinear points i.e. $\exists a\ b\ c.\texttt{s\_delta}\ a\ b\ c \neq 0$

6. A new point d can be introduced or eliminated using the following rule: $\texttt{s\_delta}\ a\ b\ c = \texttt{s\_delta}\ a\ b\ d + \texttt{s\_delta}\ a\ d\ c + \texttt{s\_delta}\ d\ b\ c$

7. Lengths of segments are given in terms of areas using the following rule: $[|\texttt{incident}\ \{a,b,c,d\}\ (L);\ \texttt{len}\ (a --- b) = \alpha \cdot \texttt{len}\ (c --- d)|]$
   $\Longrightarrow \texttt{s\_delta}\ p\ a\ b = \alpha \cdot \texttt{s\_delta}\ p\ c\ d$

The definition for parallelism given in Section 2.3.2 is also present in the theory. After assuming the basic rules as axioms, we formally verified that the various theorems, as well as the combined rules built in as high level lemmas, in the area method hold [19]. We also proved a number of theorems about the sign of $\texttt{S}_\square\ a\,b\,c\,d$ that depend on the ordering of the vertices, for example $\texttt{S}_\square\ a\,b\,c\,d = -\texttt{S}_\square\ a\,d\,c\,b$. With these few basic rules, a surprisingly large number of Euclidean geometry theorems can be proved — some of which, like Pascal's theorem in Section 2.3.2, are relatively advanced.

The Full-angle Method

As mentioned, our aim is not to improve Chou's approaches to GTP, since they are essentially algorithmic and designed to perform automatic proofs. We have, however, provided a definition for the equality between full-angles. In Isabelle, we define the relation of *angular* equality as follows:

$$x =_a y \equiv \exists n \in \mathbb{N}.\ |x - y| = n\pi$$

where π is introduced as a constant in the theory.

We can then easily prove that $\pi =_a 0$ and $\frac{3\pi}{2} =_a \frac{\pi}{2}$. Moreover, this enables us to combine the area and full-angles methods when carrying out our proofs and deduce, for example, that $\langle a --- b, b --- c\rangle =_a 0 \Longleftrightarrow \texttt{S}_\triangle\ a\,b\,c = 0$. The problems, such as $\pi = 0$, that would arise if the ordinary equality for angles was used are avoided.

The relation $=_a$ is an equivalence relation that is also used to express the properties that we might want. For example the idea of two lines being perpendicular becomes

$$a --- b \perp c --- d \equiv \langle a --- b, c --- d\rangle =_a \frac{\pi}{2}$$

The two other main properties of full-angles deal with their sign and how they can be split or joined. The same rule therefore either introduces a new line x or eliminates it from the full angles depending on the direction in which it is used.

- $\langle u, v\rangle =_a -\langle v, u\rangle$
- $\langle u, v\rangle =_a \langle u, x\rangle + \langle x, v\rangle$

2.4.2 Formulating Degenerate Conditions

When dealing with geometry proofs, we often take for granted conditions that need to be stated explicitly for machine proofs: for example, two points making up a line should not coincide. The machine proofs are valid only if these conditions are met. These subsidiary requirements are known as *non-degenerate* conditions and are required in our case mostly to prevent the denominators of fractions from becoming zero in the various algebraic statements.

It is often easy to omit non-degenerate conditions especially when the user formulates a theorem to be proved using a diagram as a guide. We tend to draw on paper "well-formed" diagrams that enable us to picture the property we are trying to prove. However, for the automatic theorem prover, if the necessary conditions are not available then it might fail to find a proof. For example, in a parallelogram $ABCD$, the two diagonals AC and BD bisect each other *if* A, B, C, *and* D *are not collinear*. The case where AB and CD are on the same line is a degeneracy.

In practice, it is usually unreasonable to have the user of an automatic geometry theorem prover specify the many non-degenerate conditions of complicated theorems. Thus, one approach, is to have the automatic prover generate these conditions and present them to the user. There are various techniques that have been developed to find conditions that make particular geometry conjectures false; for example, in the algebraic cases, the solutions of the hypothesis equations that are not solutions of the equations in the conclusions can be sought. These extra solutions usually represent conditions that falsify the theorem and are then returned as desired side conditions to the user.

In our case, however, since we are dealing with an interactive theorem prover, it often becomes obvious rather early in the proof if any non-degenerate conditions have been missed out. Indeed, a mismatch is usually noticed between the assumptions of the goal in hand with those of the theorems that we have available as rules to apply. When this occurs, the statement of the problem needs to be refined and the extra (non-degenerate) conditions added. From our own experience, we feel that this often provides us with a deeper understanding of many theorems that we are trying to prove. We become more aware of the amount of information that resides in a diagram and that is often assumed implicitly when trying to prove a theorem on paper. In this respect, an interactive approach to GTP might be more rewarding as an educational tool since one learns how to refine the statement of a geometry problem. More importantly, many subtleties are discovered that might otherwise go unnoticed or be taken for granted if returned as an automatic side-condition, for example. A challenging experiment would be in designing a graphical front end to this or some other geometry theory formalized in Isabelle as has been done for some automatic provers such as *Euclid* [19].

It is also worth noting that often theorems can be true even in some degenerate cases. For example, the sum of the three interior angles of a triangle is π even if the three vertices are collinear. This sometimes makes it hard for automatic provers to rule out degenerate conditions that genuinely falsify a theorem. As will be seen in Chapter 5, an important aspect of our approach is that it enables us to deduce how various geometric quantities behave when we reach conditions that existing GTP techniques would consider degenerate since they are infinitesimal. We thus provide powerful notions that extend the abilities

of the geometric methods and enable them to probe into situations that they cannot handle otherwise. Furthermore, we observe that geometric theorems and lemmas that hold at the infinitesimal level do not necessarily hold in general and our approach provides tools to prove them.

2.4.3 The Geometry of Motion

The *Principia* is mainly concerned with the mathematics of the motion of bodies such as planets. Thus, our geometry theory needs to provide definitions and rules that are required for such proofs [34]. These include length of arcs, length of chords and area of sectors. Since Newton deals with circular motion, and the paths of planets around the sun are elliptical, definitions for the circle, the ellipse and tangents to these figures are also provided. We review some of these definitions next.

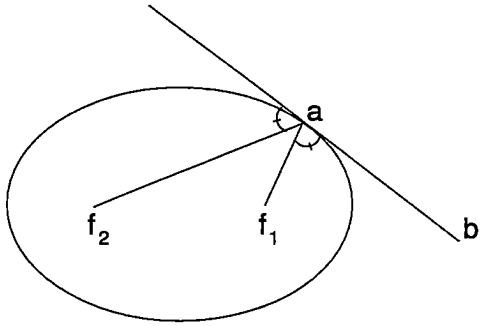


Figure 2.3: The ellipse and its tangent

The Ellipse and the Circle

The definition of the ellipse uses the familiar string and tack construction in which the distance from one focus of the ellipse to a point on the ellipse back to the other focus is constant [40]. The definition in Isabelle specifies the corresponding set of points:

$$\texttt{ellipse}\, f_1\, f_2\, r \equiv \{p.\, |\texttt{len}(f_1 \mathrel{--} p)| + |\texttt{len}(f_2 \mathrel{--} p)| = r\}$$

The ellipse is especially important since one of the major tasks of the *Principia* lies in providing the mathematical analysis that explains and confirms Kepler's guess that planets travelled in ellipses round the sun [81]. In our work, the circle is viewed as a special case of the ellipse where the foci coincide:

$$\texttt{circle}\, x\, r \equiv \texttt{ellipse}\, x\, x\, (2 \cdot r)$$

We show that the circle has another, equivalent definition by proving the following theorem:

$$\texttt{circle}\, x\, r \Longleftrightarrow \{p.\, |\texttt{len}\ (x \mathrel{--} p)| = r\}$$

The Circular Arc

The arc is an important tool in Newton's reasoning procedures. When analysing motion at a particular point on an ellipse or circle, it is the (infinitesimal) arc at that point that is usually considered. Through the use of suitable devices, the circular arc is used in most cases as an adequate approximation of any arc of finite curvature. Thus, the following definitions are provided for the length and curvilinear area (i.e. area of sector) of an arc ab with centre of curvature at x (see Figure 2.4):

$$\texttt{arc\_len}\ x\ a\ b \equiv |\texttt{len}\ (x --- a)| \cdot \langle a --- x, x --- b\rangle$$
$$\texttt{arc\_area}\ x\ a\ b \equiv 1/2 \cdot \texttt{len}\ (x --- a)^2 \cdot \langle a --- x, x --- b\rangle$$

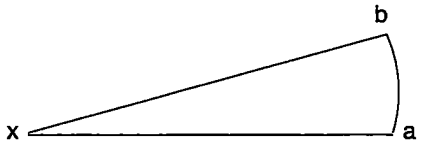


Figure 2.4: The circular arc

The Tangent

The tangent to the ellipse (and the circle) at a point is an important concept since it gives the direction of motion at that particular point. This is the direction in which the body would move in the absence of the centripetal force. The deviation from the tangent can thus be viewed as a measure of the force acting on a body as we shall see in Section 5.2.5. We use the following definition for the tangent to the ellipse:

$$\begin{aligned}\texttt{is\_e\_tangent}\ (a --- b)\ f_1\ f_2\ E \ \equiv\ & (\texttt{is\_ellipse}\ f_1\ f_2\ E \ \wedge\ a \in E \ \wedge \\ & \langle f_1 --- a, a --- b\rangle =_a \langle b --- a, a --- f_2\rangle)\end{aligned}$$

The definition of the tangent to an ellipse relies on a nice property of the curve (which also could provide an alternative definition): light emitted from one focus, say f_1, will reflect at some point p on the ellipse to the other focus f_2 (Figure 2.3). Thus, light reflects from the curve in exactly the same way as it would from the tangent line at p. Since the law of reflection means that the angle of incidence is the same as the angle of reflection, the above definition follows. Given a point on an ellipse, we are also interested in finding the set of points that belong to the tangent at that point. We add the following definition to deal with this situation:

$$\texttt{e\_tangent}\ x\ f_1\ f_2\ E \equiv \{p.\ \texttt{is\_e\_tangent}\ (x --- p)\ f_1\ f_2\ E\}$$

Definitions relating to the tangent to the circle are also made straightforwardly in Isabelle. This is also an important notion since circular motion is a significant aspect of the *Principia*.

$$\begin{aligned}&\texttt{is\_c\_tangent}\ (a --- b)\ x\ C \equiv \\ &\qquad (\texttt{is\_circle}\ x\ C \wedge a \in C \wedge x --- a \perp a --- b)\end{aligned}$$

$$\texttt{c\_tangent}\ a\ x\ C \equiv \{p.\ \texttt{is\_c\_tangent}\ (a --- p)\ x\ C\}$$

Various theorems can be proved about ellipses, circles, and their tangents. Examples of useful and interesting results are (see Figure 2.5):

- $\texttt{is\_c\_tangent}\,(a --- b)\,x\,C \iff \texttt{is\_e\_tangent}\,(a --- b)\,x\,x\,C$
- $[|\texttt{is\_c\_tangent}\,(a --- d)\,x\,C;\ b \in C|] \Longrightarrow 2 \cdot \langle b --- a, a --- d\rangle =_a \langle b --- x, x --- a\rangle$
- $[|\texttt{is\_c\_tangent}\,(a --- d)\,x\,C;\ \{b,c\} \subseteq C;\ \texttt{coll}\,c\,b\,d; \neg\texttt{coll}\,a\,c\,b; b \neq d|] \Longrightarrow \texttt{SIM}\ a\,b\,d\,c\,a\,d$
- $[|\texttt{is\_c\_tangent}\,(a --- d)\,x\,C;\ \{b,c\} \subseteq C; b \neq c|] \Longrightarrow \langle b --- c, c --- a\rangle =_a \langle b --- a, a --- d\rangle$
- $[|\texttt{is\_c\_tangent}\,(a --- d)\,x\,C;\ \{b,c\} \subseteq C; b \neq c|] \Longrightarrow \texttt{len}\ (d --- a)^2 = \texttt{len}\ (d --- b) \cdot \texttt{len}\ (b --- c)$

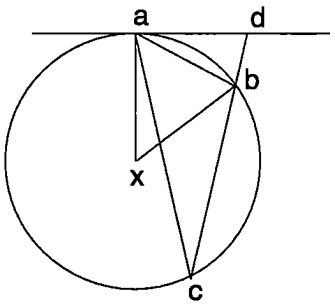


Figure 2.5: Geometric constructions for various circle theorems

More Properties of the Ellipse

We define a few other notions associated with the ellipse that are required in the development of the geometry theory. These are all geometric notions that were well known by Newton and on whose properties he sometimes relied implicitly. The centre of the ellipse is defined as the point collinear with the foci and halfway between them:

$$\texttt{is\_centre\_ellipse}\,c\,f_1\,f_2\,E \equiv (\texttt{is\_ellipse}\,f_1\,f_2\,E \wedge \texttt{coll}\,f_1\,f_2\,c \wedge |\texttt{len}\,(f_1 --- c)| = |\texttt{len}\,(c --- f_2)|)$$

If a chord $p --- g$ goes through the centre of the ellipse, then it is called a diameter of the ellipse. Assuming it meets the ellipse at two points p and g, the following definition is used:

$$\texttt{is\_diameter\_ellipse}\,(p --- g)\,f_1\,f_2\,E \equiv (p \in E \wedge g \in E \wedge (\forall c.\ \texttt{is\_centre\_ellipse}\,c\,f_1\,f_2\,E \longrightarrow \texttt{coll}\,p\,c\,g))$$

Consider Figure 2.6: take the diameter $p --- g$ of an ellipse E, the diameter $u --- v$ parallel to the tangent at p (or to the one at g) is known as the conjugate

diameter [3]:

$$\begin{aligned}&\texttt{is\_conj\_diameter\_ellipse}\ (u --- v)\ (p --- g)\ f_1\ f_2\ E \equiv \\ &\quad (\texttt{is\_diameter\_ellipse}\ (p --- g)\ f_1\ f_2\ E \wedge \\ &\quad\ ((\forall t \in \texttt{e\_tangent}\ p\ f_1\ f_2\ E.\ u --- v \parallel p --- t) \vee \\ &\quad\ \ (\forall t \in \texttt{e\_tangent}\ g\ f_1\ f_2\ E.\ u --- v \parallel g --- t)))\end{aligned}$$

From this definition, it follows that the conjugate of the conjugate of a diameter is the original diameter:

$$\begin{aligned}&[|\texttt{is\_conj\_diameter\_ellipse}\ d_1\ d_2\ f_1\ f_2\ E; \\ &\ \texttt{is\_conj\_diameter\_ellipse}\ d_2\ d_3\ f_1\ f_2\ E|] \Longrightarrow d_1 = d_3\end{aligned}$$

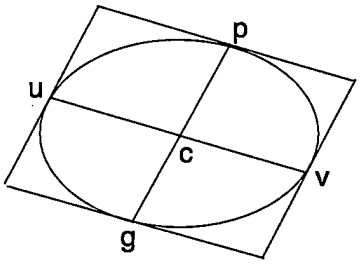


Figure 2.6: Conjugate diameters of the ellipse

A number of other properties relating to the ellipse are proved such as the one stating that *all parallelograms described around a given ellipse have the same area* (Figure 2.7).

This relationship appears (in slightly different wording) as **Lemma 12** of the *Principia* where it is employed in the solution of the famous *Propositio Keplerian*a or Kepler problem. Newton refers us to the "writers on the conics sections" for a proof of the lemma. This is demonstrated in Book 7, Proposition 31 in the *Conics* of Apollonius of Perga [3]. Unlike Newton, we have to prove this result explicitly in Isabelle to make it available to other proofs. The formalization is rather involved and proceeds by a series of construction to show that the area of parallelograms *cgov* is equal to that of parallelogram *catd* and hence that areas of circumscribing parallelograms *rxoy* and *ltzk* are equal.

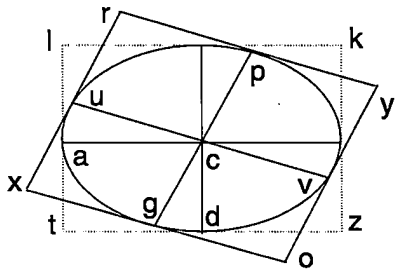


Figure 2.7: The areas of the two bounding parallelograms are the same.

2.4.4 Other Geometric Properties

The other main geometric notions include those of similar and congruent triangles. We look at the second notion more closely. Intuitively, two triangles are congruent if one figure can be moved without changing its size or shape, so as to coincide with the other. There are two possible approaches for the mathematical treatment of congruence. One way is to state enough postulates to describe its essential properties, and then prove the theorems that follow. However, congruence can also be treated definitionally in terms of lengths (distance) and angles.

Two triangles $\triangle abc$ and $\triangle a'b'c'$ are congruent if they have equal angles at a and a', at b and b', and at c and c' and every pair of corresponding sides have equal lengths. In fact in Isabelle, congruence is defined by building on the notion of similar (SIM) triangles. Thus, these are defined by

$$\begin{aligned}\texttt{SIM}\ a\ b\ c\ a'\ b'\ c' \equiv\ & \langle b -\!\!- a, a -\!\!- c\rangle =_a \langle b' -\!\!- a', a' -\!\!- c'\rangle \ \wedge \\ & \langle a -\!\!- c, c -\!\!- b\rangle =_a \langle a' -\!\!- c', c' -\!\!- b'\rangle \ \wedge \\ & \langle c -\!\!- b, b -\!\!- a\rangle =_a \langle c' -\!\!- b', b' -\!\!- a'\rangle\end{aligned}$$

and

$$\begin{aligned}\texttt{CONG}\ a\ b\ c\ a'\ b'\ c' \equiv\ & \texttt{SIM}\ a\ b\ c\ a'\ b'\ c' \wedge \\ & |\texttt{len}\ (a -\!\!- b)| = |\texttt{len}\ (a' -\!\!- b')| \wedge \\ & |\texttt{len}\ (a -\!\!- c)| = |\texttt{len}\ (a' -\!\!- c')| \wedge \\ & |\texttt{len}\ (b -\!\!- c)| = |\texttt{len}\ (b -\!\!- c')|\end{aligned}$$

An additional postulate is needed to prove theorems about congruence. This is known as the *SAS* (Side Angle Side) axiom:

> **SAS.** If two sides and the included angle of the first triangle are congruent to the corresponding parts of the second triangle, then the two triangles are congruent.

In Isabelle, this is stated as the following rule:

$$\begin{aligned}&[|\neg\texttt{coll}\ a\,b\,c; \neg\texttt{coll}\ a'\,b'\,c'; \langle c -\!\!- b, b -\!\!- a\rangle =_a \langle c' -\!\!- b', b' -\!\!- a'\rangle \\ &\quad |\texttt{len}\ (a -\!\!- b)| = |\texttt{len}\ (a' -\!\!- b')|; |\texttt{len}\ (b -\!\!- c)| = |\texttt{len}\ (b' -\!\!- c')|; \\ &|] \Longrightarrow \texttt{CONG}\ a\ b\ c\ a'\ b'\ c'\end{aligned}$$

Using the definition of congruence and the SAS postulate, the basic congruence theorems are proved. We list some of these:

- If two sides of a triangle are equal, then the angles opposite them are equal. This is an easy consequence of the SAS postulate. In Isabelle, we have:

$$\begin{aligned}&[|\neg\texttt{coll}\ a\,b\,c;\ |\texttt{len}\ (a -\!\!- b)| = |\texttt{len}\ (a -\!\!- c)|| \\ &\qquad \Longrightarrow \langle c -\!\!- b, b -\!\!- a\rangle =_a \langle a -\!\!- c, c -\!\!- b\rangle\end{aligned}$$

- The *ASA* theorem. If two angles and the included side of the first triangle are equal to the corresponding parts of the second, then the two triangles are congruent.

$$\begin{aligned}&[|\neg\texttt{coll}\ a\,b\,c; \neg\texttt{coll}\ a'\,b'\,c'; \langle c -\!\!- b, b -\!\!- a\rangle =_a \langle c' -\!\!- b', b' -\!\!- a'\rangle; \\ &\quad \langle a -\!\!- c, c -\!\!- b\rangle =_a \langle a' -\!\!- c', c' -\!\!- b'\rangle; |\texttt{len}\ (b -\!\!- c)| = |\texttt{len}\ (b' -\!\!- c')| \\ &|] \Longrightarrow \texttt{CONG}\ a\ b\ c\ a'\ b'\ c'\end{aligned}$$

- The *SSS* theorem. Given two triangles, if all three pairs of corresponding sides are equal, then the triangles are congruent.

$$[|\neg \texttt{coll}\ a\,b\,c; \neg \texttt{coll}\ a'\,b'\,c'; |\texttt{len}\ (a --- b)| = |\texttt{len}\ (a' --- b')|; \\ |\texttt{len}\ (a --- c)| = |\texttt{len}\ (a' --- c')|; |\texttt{len}\ (b --- c)| = |\texttt{len}\ (b' --- c')| \\ |] \Longrightarrow \texttt{CONG}\ a\ b\ c\ a'\ b'\ c'$$

The proof of the SSS theorem is more involved than the other congruence theorems due to case splits. In addition to these theorems, we prove that triangle congruence is an equivalence relation. Also, analogous properties for similarity of triangles (e.g. SAS and SSS) are proved.

2.5 Concluding Remarks

This chapter has outlined the theories developed for Euclidean geometry in Isabelle. These are based on well-known methods in automated GTP that are designed to produce short and human-readable proofs. The work has involved adding concepts such as similar and congruent triangles since they are needed for formalizing Newton's proofs. Such notions have traditionally been used in geometry, though Chou et al. note that they have limitations when dealing with automated GTP [19]. However, our proofs are not affected since we are not concerned with completely automatic proofs. To deal with some of the main types of motion analysed by Newton, definitions of ellipses, circles, tangents, and arcs amongst others have also been added to the theory. Many properties resulting from constructions based on these geometric objects have been proved. The proofs in Isabelle tend to be short since the geometry theory is powerful. We have also verified many of the results of the signed area and full-angles methods.

As far as related work is concerned, we know of few examples of geometry theory developed in interactive theorem provers. A theory that can deal with collinearity and betweenness was developed in IMPS [31] but to our knowledge can only prove a limited number of theorems. The possibility of developing such a powerful geometric framework in Isabelle owes much to the versatility of the system; its theory mechanism enabled us to formalize the main concepts behind the traditional methods of Chou et al., while its various proof procedures and simplification routines helped preserve many of their important properties.

As mentioned previously, using Isabelle's theory, many theorems of Euclidean geometry can be proved. However, this is not sufficient to prove results from the *Principia* that involve Newton's ultimate reasoning. For that more tools are needed that can capture the notion of infinitesimal. In the next chapter, we describe the formalization of the infinitesimal in Isabelle using the ultrapower construction of Nonstandard Analysis. Several other classes of numbers are also constructed that have application not only in geometry and the formalization of Newton's reasoning but also in other areas.

Chapter 3

Constructing the Hyperreals

In the early 1960's, Abraham Robinson finally provided a rigorous foundation for the use of infinitesimals in analysis by developing the new concept of **Nonstandard Analysis** (NSA). The idea was to introduce a new number system known as the *hyperreals* which contains the real numbers but also infinitesimals and infinite numbers. The notions of infinitesimals and other nonstandard numbers introduce many subtleties into the theory that need to be dealt with.

In this chapter, we describe the constructions of Robinson's hyperreals in Isabelle. Our approach is purely definitional to ensure that infinitesimals and other nonstandard numbers have a sound foundation in the system. To reach our goal has required constructing the various number systems leading to the reals and then going one step further to define the hyperreals by working on sequences of reals. The hyperreals have considerable intrinsic interest since they exhibit many new properties. Moreover, as a tool, they are of great value to the formalization of analysis — an aspect that will be described as we report on the mechanization of nonstandard real analysis.

We start by giving a description of Isabelle and of the HOL object logic in which this work was carried out.

3.1 Isabelle/HOL

Isabelle [66] is a generic theorem prover, written in ML, into which the user can encode their own object-level logics. Examples of such object logics are higher order logic (HOL), Zermelo-Fraenkel set theory (ZF), and first order logic (FOL). Terms from the object logics are represented and manipulated in Isabelle's intuitionistic higher order meta-logic, which supports polymorphic typing.

3.1.1 Theories in Isabelle

Isabelle's theories provide a hierarchical organization for the syntax, declarations and axioms of a mathematical development and are developed using theory definition files [66]. A typical theory file will organize the definitions of types and

functions. It may also contain the primitive axioms that are asserted (without proofs) by the user. A particular theory will usually collect (in a separate file) the proven named theorems and make them available to all its children theories.

The meta-level connectives are implication ($\Longrightarrow$), universal quantifier and equality. In Figure 3.1, we give the description of some of the notations used in Isabelle/HOL. Throughout the presentation, we will be using mostly conventional mathematical notations when describing our development. However, there are cases where we might use the ASCII notations actually used to express terms and rules in Isabelle as explicit examples.

An inference rule with n premises or antecedents has the following form in Isabelle:

$$[|\phi_1; \ldots; \phi_n|] \Longrightarrow \psi$$

This abbreviates the nested implication $\phi_1 \Longrightarrow (\ldots \phi_n \Longrightarrow \psi \ldots)$. Such a rule can also be viewed as the proof state with subgoals $\phi_1, \ldots, \phi_n$ and main goal ψ [66]. Alternatively, this can be viewed as meaning "if $\phi_1 \wedge \ldots \wedge \phi_n$ then ψ".

3.1.2 Proof Construction

Rules can be combined in various ways to derive new ones using higher order resolution; this process is known as proof construction and can proceed in both backward and forward directions:

- In backward fashion, the user supplies a goal and reduces it to simpler subgoals by applying existing rules until they are solved. A goal is solved when it becomes the instance of some previously proved theorem.
- In forward proofs, the antecedents or assumptions of a rule can be resolved with other rules to derive new assumptions. This process can be carried on until either the conclusion is the instance of some assumption or the goal is an instance of a theorem.

3.1.3 Higher Order Logic in Isabelle

One of Isabelle's logics is HOL, a higher order logic that supports polymorphism and type constructors. Isabelle/HOL is based on Gordon's HOL theorem prover [41] which itself originates from Church's paper [23]. Isabelle/HOL is well developed and widely used. It has a wide library of theories defined in it including the natural numbers, set theory, well-founded recursion, inductive definitions, and equivalence relations. Isabelle/HOL has been applied to reasoning in many fields including the verification of security protocols [67] and verifying the type system of the Java programming language [64].

Though Isabelle is mainly used interactively as a proof assistant, it also provides substantial support for automation. It has a generic simplification package, which is set up for many of the logics including HOL. Isabelle's simplifier performs conditional and unconditional rewritings and makes use of context information [66]. The user is free to add new rules to the simplification set (the *simpset*) either permanently or temporarily. Isabelle also provides a number of generic automatic tactics that can execute proof procedures in the various logics. The automatic tactics provided by Isabelle's *classical reasoner* include a fast tableau prover called `Blast_tac` coded directly in ML and `Auto_tac` which

syntax	*description*
`&`	$\wedge$, and
`~`	$\neg$, not
`==>`	$\Longrightarrow$, implication (meta level)
`-->`	$\longrightarrow$, implication (object level)
`=`	$\equiv$, if and only if
`! or ALL`	$\forall$, for all
`? or EX`	$\exists$, exists
`@`	ϵ, Hilbert choice
`%`	λ, lambda abstraction
`- A`	$\overline{A}$, set complement
`Union c`	$\bigcup c$, union over sets of sets

Figure 3.1: ASCII notation for HOL

attempts to prove all subgoals by a combination of simplification and classical reasoning. Other powerful theorem proving tactics include those which, unlike `Blast_tac`, construct proofs directly in Isabelle: for example, `Fast_tac` implements a depth-first search automatic tactic.

The HOL methodology

Isabelle/HOL has been chosen as the logic in which to carry out our proofs. One of the main reasons is that it provides strong typing and therefore ensures that only type correct terms are accepted. Moreover, the HOL methodology is to admit only conservative extensions to a theory. This means defining and deriving the required mathematical notions rather than postulating them. The definitional approach of HOL requires that assertions are proved about some model instead of being postulated. Such a rigorous definitional extension guarantees consistency, which cannot be ensured when axioms are introduced. As pointed out by Harrison [43], such an approach provides a simple logical "basis that can be seen to be correct once and for all". With regards to the foundations of infinitesimals, the definitional approach is certainly advisable when one considers the numerous inconsistent axiomatizations that have been proposed in the past [26]. Of course, care still needs to be exercised, as a wrong definition will almost certainly yield the wrong properties.

3.2 Properties of an Infinitesimal Calculus

We first look at some of the requirements for a set of infinitesimals that could be useful for the development of an infinitesimal calculus. Keisler [52] and Vesley [78], for example, discuss the various properties that need to hold for developing a calculus for the infinitesimals. Let the set `Infinitesimal` denote the set of infinitesimals where an infinitesimal can, for the time being, be viewed intuitively as a number smaller in magnitude than all positive reals.

We would like the following properties:

1. 0 is an `Infinitesimal`
2. there is a nonzero infinitesimal

3. `Infinitesimal` is a ring

It might seem reasonable also to want the following:

4. `Infinitesimal` is a subring of the real numbers $\mathbb{R}$

5. `Infinitesimal` is an ideal in $\mathbb{R}$:

$$\forall r \in \mathbb{R} \forall x \in \texttt{Infinitesimal}.\ rx \in \texttt{Infinitesimal}$$

6. also we expect `Infinitesimal` to be non-Archimedean:

$$\exists x \in \texttt{Infinitesimal}.\ \forall n.\ nx < 1$$

The above, (1)–(6), look sufficient for a simple theory of infinitesimals but unfortunately such a theory would be inconsistent. Furthermore, as Vesley [78] notes, if $\mathbb{R}$ is the set of classical reals, then *any nontrivial ideal in $\mathbb{R}$ is equal to $\mathbb{R}$*. Thus, if `Infinitesimal` satisfies (2), (4), (5) then `Infinitesimal` $= \mathbb{R}$. This problem is tackled in NSA by dispensing with property (4). Instead, using the axioms of classical set theory, a set $\mathbb{R}^*$ of hyperreals is obtained with properties that include `Infinitesimal` $\subseteq \mathbb{R}^*$, $\mathbb{R} \subseteq \mathbb{R}^*$, (1)–(3), (6), but *not* `Infinitesimal` $\subseteq \mathbb{R}$ and therefore not (4). As a result, (5) now requires `Infinitesimal` to be an ideal in the set of finite members of $\mathbb{R}^*$. This set includes the reals and the infinitesimals amongst other numbers.

Though an axiomatic approach seems the easiest way to get quickly to the infinitesimals, there is always the possibility that the set of axioms might lead to an inconsistency, as we saw above. We would rather have a development of infinitesimals that is guaranteed to be sound — especially with regards to the stormy history of infinitesimals.

3.3 Internal Set Theory

There is, in the literature, an axiomatic version of NSA introduced by Nelson and based on ZF set theory with the Axiom of Choice (ZFC)[62]. Nelson's approach is known as Internal Set Theory (IST) and adds three additional axioms to those of ZFC. We have not developed Nelson's theory, even though ZF is one of the object-logics of Isabelle, because there are aspects of the additional axioms that seem hard to formalize in Isabelle.

We give a brief description of IST which formalizes a portion of Robinson's NSA. The language of IST adds the new undefined unary predicate standard to the usual undefined binary $\in$ relation of ZFC. A formula of IST that does not involve the new predicate, that is a formula of ZFC, is called *internal*; otherwise, a formula is called *external*. As mentioned by Nelson, both of these concepts are metamathematical notions — they are properties of formulae and do not apply to sets themselves.

New axioms are added to those of ZFC by IST to govern the behaviour of standard. The following abbreviations are first introduced:

$$\forall^{st} x\ \phi \text{ for } \forall x[\mathsf{standard}(x) \Longrightarrow \phi]$$
$$\exists^{st} x\ \phi \text{ for } \exists x[\mathsf{standard}(x) \wedge \phi]$$

Similarly, with finite(x) meaning that set x is finite, these further abbreviations become possible:

$$\forall^{st\ fin} x\ \phi \text{ for } \forall x[\mathsf{standard}(x) \wedge \mathsf{finite}(x) \Longrightarrow \phi]$$
$$\exists^{st\ fin} x\ \phi \text{ for } \exists x[\mathsf{standard}(x) \wedge \mathsf{finite}(x) \wedge \phi]$$

The axiom schemas can now be given for IST:

(T) Transfer

Let A be an internal formula whose only free variables are $x, t_1, \ldots, t_n$. Then the *transfer principle* is:

$$\forall^{st} t_1 \ldots \forall^{st} t_n[\forall^{st} x A \Longleftrightarrow \forall x A]$$

It is convenient to interpret the $t_1, \cdots, t_n$ as parameters; we are mainly concerned with x. The transfer principle asserts, to quote Nelson [62], that "if we have an internal formula A, and all the parameters have standard values, and if we know that A holds for all standard x, then it holds for all x".

(I) Idealization

The *idealization principle* enables us to prove the existence of nonstandard objects. Let A be an internal formula. Then the idealization principle is:

$$\forall^{st\ fin} z \exists x \forall y \in z A(x, y) \Longleftrightarrow \exists x_0 \forall^{st} y A(x_0, y)$$

To illustrate how this schema ensures the existence of nonstandard objects, let our internal formula be $A(x, y) \equiv x \in \mathbb{N} \wedge (y \in \mathbb{N} \Longrightarrow y \leq x)$. We choose x to be $\mathsf{Max}\, z \cap \mathbb{N}$ if $z \cap \mathbb{N} \neq \emptyset$, otherwise we choose $x = 0$. Thus, the idealization principle says that there exists a natural number x_0 greater (or equal) than all *standard* natural numbers; therefore, it ensures the existence of (nonstandard) infinitely large natural numbers.

(S) Standardization

The final assumption that is asserted about the predicate standard is known as the *standardization principle*. Let A be any formula, external or internal, not containing y. Then the standardization principle is:

$$\forall^{st} x \exists^{st} y \forall^{st} z[z \in y \Longleftrightarrow z \in x \wedge A]$$

It follows from (S) that for any formula A, whether it contains the predicate standard or not, for all standard set x, there exists a unique standard $y \subseteq x$ whose standard elements z are precisely those of x and such that A holds. The set y is usually denoted by ${}^s\{z \in x.\, A\}$. A point worth noting is that if A involves the predicate standard or something based on the latter such as the concept of being infinitely small, then A is no longer a formula of ZFC and so the axiom of separation cannot be applied to it.

The (S) axiom above thus points out an important subtlety of IST which should not be overlooked: since no axiom of ZFC refers to the new predicate standard, these cannot be applied to any external formula (i.e. one involving

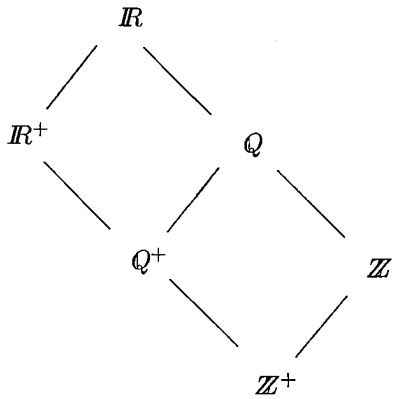
standard directly or indirectly). We cannot, for instance, apply the separation axiom to any external formula: it is therefore illegal to form the set $\{x \in \mathbb{R}. x \in \texttt{Infinitesimal}\}$ in IST.

We also observe that, since the addition of axiom (I) to ZFC set theory enables us to show the existence of infinitely large natural numbers without introducing any contradiction, the set $\mathbb{N}$ does not consist solely of the usual natural numbers. When expressed formally, this is sometimes known as Reeb's thesis.

Nelson also proves that IST is a conservative extension of ZFC. This means that if A is an internal formula provable in IST, then A has a proof in the standard set theory, that is A has a proof in ZFC. The rather different method that was chosen for NSA in Isabelle is examined next. Whereas Nelson's IST can be viewed as an intentional approach, this is an explicit extensional one.

3.4 Constructions Leading to the Reals

There are various classical methods in existence in the literature on the construction of the various number systems. The usual approach is to arrange them in a lattice respecting the inclusions between the sets. Let $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ be the sets of integers, rationals, and reals respectively, and $\mathbb{Z}^+$, $\mathbb{Q}^+$, $\mathbb{R}^+$ be their positive elements. Note that $\mathbb{Z}^+$ is the set of elements of type `pnat`.



As can be seen in the figure, there are several ways to reach $\mathbb{R}$ from $\mathbb{Z}^+$. These various paths, however, differ greatly in the technical details of the constructions along them. Conway [25] suggests that there is a best way through the lattice to the reals that avoids, as much as possible, **case splits**. These are tedious and unnecessary complications that are often treated superficially in textbooks giving "natural" constructions of number systems. Conway proposes the following general methods that we implement in Isabelle.

To add negative numbers, that is to proceed, for example, from $\mathbb{R}^+$ to $\mathbb{R}$, the signed number, $x \in \mathbb{R}^+$, is represented as an ordered pair of unsigned numbers (a, b) meaning $a - b$ and the equivalence relation

$$(a, b) \sim (c, d) \iff a + d = b + c \tag{3.1}$$

is used. This is better than the obvious approach of adding zero and $-x$ for each $x \in \mathbb{X}^+$ which leads to too much case-splitting.

Similarly, one can go from $\mathbb{Z}$ to $\mathbb{Q}$ or from $\mathbb{Z}^+$ to $\mathbb{Q}^+$ by taking ordered pairs (a, b) meaning a/b and the equivalence relation

$$(a, b) \sim (c, d) \iff a \cdot d = b \cdot c \tag{3.2}$$

To proceed from $\mathbb{Q}$ to $\mathbb{R}$ or from $\mathbb{Q}^+$ to $\mathbb{R}^+$, the method of **Dedekind cuts** is used. There are several other methods available such as Cauchy sequences and positional expansions [43]. The best path, as suggested by Conway, is $\mathbb{Z}^+ \longrightarrow \mathbb{Q}^+ \longrightarrow \mathbb{R}^+ \longrightarrow \mathbb{R}$.

3.4.1 Equivalence Relations in Isabelle/HOL

We use Isabelle's `Equiv` theory, which defines equivalence relations in higher-order set theory, to define the new type of positive rationals. First, we recall the definitions of equivalence relations, set quotients and equivalence classes:

DEFINITION 3.4.1. *A relation $\sim$ is said to be an equivalence relation if and only if it is reflexive $(x \sim x)$, symmetric $(x \sim y \Longrightarrow y \sim x)$, and transitive $(x \sim y \wedge y \sim z \Longrightarrow x \sim z)$.*

DEFINITION 3.4.2. *Given an equivalence relation $\sim$ on a set S, then the* quotient *of S with respect to $\sim$ is the set of all* equivalence classes, *and is defined by $S/\sim \equiv \{[x] \mid x \in S\}$ where $[x] \equiv \{y \in S \mid x \sim y\}$.*

The set of all equivalence classes $S/\sim$ is called the *quotient set of S by $\sim$* and a member of an equivalence class is often referred to as a *representative* of the class.

3.4.2 Example: Constructing $\mathbb{Q}^+$ from $\mathbb{Z}^+$

In this section, we illustrate, by means of an example, how a new type can be introduced in Isabelle as the quotient set of some equivalence relation. We also show how primitive functions are defined on the new type using **abstraction** and **representation** functions. Other operations derived from the primitive functions are also introduced.

The theory `PRAT`, shown in Figure 3.2 and developed on our way to the reals (and beyond), defines the type `prat` of positive rational numbers and its associated operations. The new type is defined on pairs of elements of `pnat` which denotes the positive natural numbers introduced as an explicit type in Isabelle.

Under the `constdefs` keyword, we declare and define the equivalence relation (3.2) specified at the beginning of Section 3.4 above that enables us to proceed from $\mathbb{Z}^+$ to $\mathbb{Q}^+$ in the lattice:

$$\texttt{pratrel} \equiv \{p.\, \exists a\, b\, c\, d.\, p = ((a, b), (c, d)) \wedge a \cdot d = b \cdot c\}$$

Using `typedef`, we declare the new type `prat`:

$$\texttt{prat} \equiv \{x.\, \texttt{True}\}/\texttt{pratrel} \quad (\texttt{Equiv.quotient\_def})$$

The representing set of elements is defined as the set of equivalence classes of fractions, that is the set of equivalence classes consisting of ordered pairs of

```
PRAT = PNAT + Equiv +

constdefs
 (* equivalence relation *)
 pratrel :: "((pnat * pnat) * (pnat * pnat)) set"
 "pratrel ≡ {p. ∃ a b c d. p = ((a,b),(c,d)) ∧ ad = bc}"

typedef
 prat =
    "{x::(pnat*pnat). True}/pratrel" (Equiv.quotient_def)

instance
 prat :: {ord, plus, times}

constdefs
 prat_of_pnat :: pnat ⇒ prat       ("$#_" [80] 80)
 "prat_of_pnat m ≡ Abs_prat(pratrel^^{(m,Abs_pnat 1)})"

 qinv    :: prat ⇒ prat
 "qinv Q  ≡ Abs_prat(⋃(x,y)∈Rep_prat(Q). pratrel^^{(y,x)})"

defs
 prat_add_def
 "P + Q ≡ Abs_prat(⋃p∈Rep_prat(P). ⋃q∈Rep_prat(Q).
   split(λa b. split(λc d. pratrel^^{(ad + bc, bd)}) q) p)"

    ...

 prat_less_def
 "P < (Q::prat) ≡ ∃T. P + T = Q"

end
```

Figure 3.2: Isabelle/HOL Theory for Rationals using Equivalence Classes

positive natural numbers. The theorem `quotient_def` (from the theory `Equiv`) acts as witness to prove the non-emptiness of the new type and is given in brackets next to the new type. Non-emptiness needs to be proved to ensure that the quantifier rules of HOL are sound [68], otherwise the new type is rejected.

Once a new type has been introduced successfully, Isabelle provides coercion functions — the abstraction and representation functions — that enable us to define basic operations on the new type. Thus, in this particular example, the functions

$$\texttt{Abs\_prat} :: (\texttt{pnat} * \texttt{pnat})\ \texttt{set} \Rightarrow \texttt{prat}$$
$$\texttt{Rep\_prat} :: \texttt{prat} \Rightarrow (\texttt{pnat} * \texttt{pnat})\ \texttt{set}$$

are added to the theory such that `prat` is isomorphic to $\{x.\,\texttt{True}\}/\texttt{pratrel}$ by `Rep_hyprat` and its inverse `Abs_prat`. Using these functions and other operations from Isabelle's `Set` and `Equiv` theories, we are now ready to define operations on the positive rationals. For example, the inverse function `qinv`, which swaps the elements of the ordered pairs (x, y) representing x/y around to give y/x, is constructed in Isabelle by:

$$\texttt{qinv}\ Q \equiv \texttt{Abs\_prat}\ (\bigcup(x,y) \in \texttt{Rep\_prat}\ (Q).\texttt{pratrel``}\{(y,x)\})$$

where

- $\bigcup x \in A.\, B[x] \equiv \{y.\, \exists x \in A.\, y \in B\}$ (union of family of sets).
- $r``s \equiv \{y.\, \exists x \in s.\, (x,y) \in r\}$ (image of set s under relation r).

Once the primitive operations such as addition and multiplication are defined, we can use them to derive other operations such as the ordering relation:

$$P < Q \equiv \exists T.\, P + T = Q$$

We then show that the operations on the new type respect the various field properties and that we have indeed defined the densely ordered (but not Dedekind-complete) field of the positive rationals.

3.4.3 A Few Important Theorems

In this section, some of the more important theorems that we proved during our constructions leading up to the reals are given. We are especially concerned with those that will be needed to establish properties of hyperreals and nonstandard real analysis later on.

1. **Completeness of the reals.** The *Supremum Property*, which states that every nonempty set of reals X that has an upper-bound has a least upper bound is proved:

$$\forall X.\ (\exists x.\, x \in X) \wedge (\exists U.\, \forall x \in X.\, x \leq U)$$
$$\Longrightarrow \exists u.\, (\forall x \in X.\, x \leq u) \wedge \forall u'.\, (\forall x \in X.\, x \leq u') \Longrightarrow u \leq u'$$

2. **The Archimedean property for the reals.** This simple result has far-reaching implications since it rules out the existence of infinitely small quantities or infinitesimals in $\mathbb{R}$. Any such infinitesimal in $\mathbb{R}$ would mean that its reciprocal is an upper bound of $\mathbb{N}$ in $\mathbb{R}$ thereby contradicting the Archimedean property.

$$\forall x.\, \exists n.\, x < n$$

Various mechanizations of standard analysis (see for example Harrison's work in HOL [42, 43]) have developed theories of limits, derivatives, continuity of functions and so on, taking as their foundations the real numbers. Our work, however, will now go one step further and show how the reals can be used to build a richer number system.

3.5 Filters and Ultrafilters

In this section, the preliminaries necessary to our construction are presented. The definitions and theorems that we need and their formalization in the set theory of Isabelle/HOL are reviewed. Our aim is to establish an equivalence relation on the set of all infinite sequences of reals and use the system of equivalence classes as a model for $\mathbb{R}^*$. We start with the concept of a *filter*.

DEFINITION 3.5.1. *Let S be any non-empty set. A filter $\mathcal{F}$ over S is a collection of subsets of S such that*

F1) $S \in \mathcal{F} \wedge \emptyset \notin \mathcal{F}$

F2) $X \in \mathcal{F} \wedge Y \in \mathcal{F} \Longrightarrow X \cap Y \in \mathcal{F}$

F3) $X \in \mathcal{F} \wedge X \subseteq Y \subseteq S \Longrightarrow Y \in \mathcal{F}$

Every filter is a *nonempty* collection of subsets since $S \in \mathcal{F}$, and filters are closed under finite intersection and supersets. There are numerous examples of filters including the *trivial filter* $\{S\}$ and, if S is infinite, the *Fréchet* or *cofinite* filter $\{X.\,\texttt{finite}\ (S - X)\}$. In Isabelle, we develop a theory `Filter` and formalize the notions described above as follows:

$$\begin{aligned}\texttt{isFilter}\ F\ S \equiv\ & (F \subseteq \texttt{Pow}\,S \wedge S \in F \wedge \\ & \forall X \in F.\, \forall Y \in F.\, X \cap Y \in F \wedge \\ & \forall X\ Y.\, X \in F \wedge Y \subseteq S \longrightarrow Y \in F)\end{aligned}$$

$$\texttt{Filters}\ S \equiv \{X.\ \texttt{isFilter}\ X\ S\}$$

We note in the above definitions the occurrence of `Filters` S which is defined to be the set of all filters over S. We adopt this general approach of defining sets of the various structures that are dealt with for clarity; this is possible since in Isabelle/HOL's set theory the type α *set* is isomorphic to the type $\alpha \Rightarrow$ *bool* [66].

Let us mention some of the terminology often encountered when filters and related concepts are used. A set $X \subseteq S$ is sometimes said to be **large** [74] or **quasi-big** [44] if $X \in \mathcal{F}$. Other terms used include *residual* or *generic* when dealing with directed sets or Baire category theory. Moreover, and of relevance to our development, a condition P on points $x \in S$ is said to be satisfied **almost everywhere** (a.e.) or **almost always**, or is $\mathcal{F}$-**true** or **almost true**, if the set $\{x \in S.\, P \text{ is satisfied at } x\}$ is a member of $\mathcal{F}$.

A refinement of the concept of a filter is now introduced by defining the notion of an *ultrafilter* over the nonempty set S:

DEFINITION 3.5.2. *An ultrafilter $\mathcal{U}$ over S is a filter over S such that*

U1) $\mathcal{U} \subseteq \mathcal{F} \wedge \mathcal{F} \in \texttt{Filters}\ S \Longrightarrow \mathcal{U} = \mathcal{F}$

An ultrafilter is thus a **maximal** filter, that is a filter that cannot be enlarged. An ultrafilter (and hence a filter) is said to be *free* if and only if it does not contain any finite sets. A filter which is not free is said to be *fixed*. We are mainly interested in free ultrafilters. The definitions used in Isabelle's `Filter` theory follow:

$$\texttt{Ultrafilters}\ S \equiv \{X.\ X \in \texttt{Filters}\ S \wedge \forall G \in \texttt{Filters}\ S.\ X \subseteq G \longrightarrow X = G\}$$

$$\texttt{FreeUltrafilters}\ S \equiv \{X.\ X \in \texttt{Ultrafilters}\ S \wedge \forall x \in X.\ \neg\texttt{finite}\ x\}$$

We proceed to prove various properties of filters, ultrafilters and so on from these definitions. These include a theorem about ultrafilters that states that $\mathcal{U}$ is an ultrafilter on S if and only if for any subset A of S, either A belongs to $\mathcal{U}$ or else its complement $S - A$ belongs to $\mathcal{U}$, but not both:

$$\mathcal{U} \in \texttt{Ultrafilters}\ S \Longleftrightarrow (\mathcal{U} \in \texttt{Filters}\ (S) \wedge \forall A \in \texttt{Pow}\, S.\ A \in \mathcal{U} \vee S - A \in \mathcal{U})$$

The content of this theorem is critically important to our development and an outline of its proof in Isabelle is given below:

Proof: Suppose that $\mathcal{U}$ is a filter such that for every $A \subseteq S$ either $A \in \mathcal{U}$ or $S - A \in \mathcal{U}$. Let G be a *superfilter* of $\mathcal{U}$ i.e. a filter such that $\mathcal{U} \subseteq G$ and suppose that $B \in G$ and $B \notin \mathcal{U}$. But then, from our initial assumption, it follows that $S - B \in \mathcal{U} \subseteq G$, and so $\emptyset = B \cap (S - B) \in G$ which contradicts property (F1) for a filter. Hence there is no proper filter containing G, and so $\mathcal{U}$ is an ultrafilter.

Conversely, suppose that $\mathcal{U}$ is an ultrafilter and $A \notin \mathcal{U}$. Define a set $G \equiv \{X \subseteq S. \exists J \in \mathcal{U}.\ A \cap J \subseteq X\}$. Then $\mathcal{U} \subseteq G$ and $\mathcal{U} \neq G$ since $A \in G$, and so G cannot be a filter since by assumption $\mathcal{U}$ is maximal. But G is not empty, and if $B, C \in G$ and $B \subseteq D$ then $B \cap C \in G$ and $D \in G$ (verifying conditions (F2) and (F3) for G to be a filter). Since $S \in G$, G can fail to be a filter only if $\emptyset \in G$. That is, we have $A \cap J = \emptyset$ for some $J \in \mathcal{U}$ for which we must then have $J \subseteq (S - A)$. It follows that $S - A \in \mathcal{U}$. □

From this result, it can be seen that the Fréchet filter on an infinite set $\mathcal{S}$ is not an ultrafilter though it follows that it is free. What is needed to progress any further in the development is to show **the existence of a free ultrafilter on any infinite set** — this result is a corollary of the important Ultrafilter Theorem [48, 74]. Using the result above, we can see that for an ultrafilter $\mathcal{U}$ to be free, every cofinite subset of $\mathcal{S}$, and hence the Fréchet filter, has to be contained in $\mathcal{U}$. This result will be useful to us in Section 3.5.2 but first, we give an overview of our proof of *Zorn's Lemma* and how we appeal to it to guarantee the existence of an ultrafilter. We then extend this result and show that the ultrafilter can be free as well.

3.5.1 Zorn's Lemma

The existence of free ultrafilters is not obvious at first sight. To show that the ultrafilter theorem holds and to carry out our construction, we need Zorn's

lemma. This is an equivalent form of the axiom of choice (AC) and first needs to be proved in Isabelle/HOL.

> **Zorn's Lemma.** Let S be a nonempty set of sets such that each chain $c \subseteq S$ has an *upper bound* in S. Then S has a *maximal* element y, i.e. a set $y \in S$ such that no member of S properly contains y.

The statement of Zorn's Lemma involves the idea of a partially ordered set and related concepts. We present briefly various mathematical concepts and theorems about them needed in Isabelle/HOL to express Zorn's Lemma.

Paulson has already proved Zorn's Lemma in Isabelle's Zermelo-Fraenkel set theory (Isabelle/ZF) [69] by mechanizing a paper by Abrial and Laffitte [1]. Reporting on the mechanization, Paulson remarks that the formal language used by Abrial and Laffitte is close to higher order logic and thus should be useful to Isabelle/HOL amongst other proof assistants. In our current work, we have adapted the mechanization of Zorn's Lemma developed in Isabelle/ZF to Isabelle/HOL. Below, we briefly mention how our formalization in Isabelle/HOL compares with the one in Isabelle/ZF.

The definitions used by Abrial and Laffitte require the **choice** operator since starting from AC, they prove Hausdorff's Maximal Principle and then derive Zorn's Lemma. Unlike its ZF counterpart, Isabelle/HOL provides such an operator, the so-called Hilbert **description** operator, ε. Thus, the formulation of the various theorems in Isabelle/HOL is somewhat simpler than that given by Paulson for ZF. The latter requires that the existence of the choice function be stated explicitly as a temporary additional assumption [69].

We also use Isabelle's inductive package to define a set that is totally ordered by set inclusion. In general, the construction of the inductive set relies on defining a suitable successor function which, in our case, is defined using the choice or description operator:

$$\begin{aligned} \mathtt{succ}\ S\ c \equiv\ & \mathtt{if}\ (c \notin \mathtt{chain}\ S \vee c \in \mathtt{maxchain}\ S) \\ & \mathtt{then}\ c\ \mathtt{else}\ (\varepsilon c'.\ c' \in \mathtt{super}\ S\ c) \end{aligned}$$

Our other definitions of set of chains, super chains and maximal chains are similar to those in Isabelle/ZF. Note that the definitions suppose that the set S has some *partial ordering* defined on it which is denoted by $\leq$:

$$\begin{aligned} &\mathtt{chain}\ S \equiv \{F.\ F \subseteq S \wedge (\forall x \in F.\ \forall y \in F.\ x \leq y \vee y \leq x)\} \\ &\mathtt{super}\ S\ c \equiv \{d.\ d \in \mathtt{chain}\ S \wedge c \subset d\} \\ &\mathtt{maxchain}\ S \equiv \{c.\ c \in \mathtt{chain}\ S \wedge \mathtt{super}\ S\ c = \emptyset\} \end{aligned}$$

We tried to simplify these definitions at first by removing references to the inductive set S, since it is actually used by Abrial and Laffitte to provide typing in their version of ZF. Thus, S as a parameter seems redundant when working in Isabelle's typed higher order logic. However, relying on the type made some of our proofs about ultrafilters unnecessarily complicated and prompted us to refer explicitly to the underlying set in definitions and hence in our proof of Zorn's Lemma. In outline, with these definitions, we prove the theorem of Hausdorff: every partially-ordered set contains a maximal chain. So, taking the subset relation as the partial ordering on S, we have

$$\exists c.\ c \in \mathtt{maxchain}\ S$$

and then consider an upper bound u of such a maximal chain c — guaranteed to exist according to the premise of Zorn's Lemma — which we prove to be a maximal element. Expressed formally in Isabelle, the following theorem is established:

$$\begin{aligned}\forall c \in \texttt{chain}\ S.\ \exists u \in S.\ \forall x \in c.\ x \subseteq u \\ \Longrightarrow \exists y \in S.\ \forall x \in S.\ y \subseteq x \longrightarrow y = x\end{aligned}$$

3.5.2 The Ultrafilter Theorem

The Ultrafilter Theorem (UFT) is a complicated but important principle that lies midway between AC and the Axiom of Choice for Finite Sets [74]. Moreover, UFT like the Axiom of Choice has many important equivalent forms. Schechter presents and discusses twenty five of these occurring in many areas of mathematics [74] and points to the many more equivalents occurring in the literature. The version that we are interested in is

> **(UFT) Ultrafilter Theorem (Cartan).** If $\mathcal{F}$ is a filter on a set S then there is an ultrafilter $\mathcal{U}$ on $\mathcal{S}$ with $\mathcal{F} \subseteq \mathcal{U}$.

This result can be proved using Zorn's Lemma. In fact, we are really interested in proving a corollary of UFT about the existence of free ultrafilters:[1]

> **Corollary.** On every infinite set there exists a free ultrafilter

Expressed in Isabelle, we want to prove

$$\neg\texttt{finite}\ S \Longrightarrow \exists u.\ u \in \texttt{FreeUltrafilters}\ S$$

To do so, we define in the theory `Filter` the set, `SuperFrechet` S, of all filters on S that contain the Fréchet filter (i.e. the set of superfilters of the Fréchet filter):

$$\begin{aligned}&\texttt{Frechet}\ S \equiv \{A.\ \texttt{finite}\ S - A\} \\ &\texttt{SuperFrechet}\ S \equiv \{G.\ G \in \texttt{Filters}\ S \wedge \texttt{Frechet}\ S \subseteq G\}\end{aligned}$$

Our proof consists first in showing that `SuperFrechet` S contains a maximal element, that is, an ultrafilter on the (infinite) set S, and then in showing that this maximal element does not contain any finite sets. Stated formally in Isabelle, the following goal needs to be established:

$$\begin{aligned}\neg\texttt{finite}\ S \Longrightarrow\ &\exists U \in \texttt{SuperFrechet}\ S. \\ &\forall G \in \texttt{SuperFrechet}\ S.\ U \subseteq G \longrightarrow U = G \wedge \\ &\forall x \in U.\ \neg\texttt{finite}\ x\end{aligned}$$

Existence of Ultrafilter

We split the main goal above in two parts and outline in this section how the existence of the ultrafilter is proved. Formally, we need to prove

$$\begin{aligned}\neg\texttt{finite}\ S \Longrightarrow\ &\exists U \in \texttt{SuperFrechet}\ S. \\ &\forall G \in \texttt{SuperFrechet}\ S.\ U \subseteq G \longrightarrow U = G\end{aligned}$$

[1] Some authors like Hoskins [48] and Keisler [52] state the corollary (or even one of its special cases) as the actual Ultrafilter Theorem.

Applying Zorn's Lemma (as an introduction rule in Isabelle) and with some simplification, this reduces the above to the following new subgoal:

$$[|\neg \texttt{finite}\ S;\ c \in \texttt{chain}\ (\texttt{SuperFrechet}\ S)|] \\ \Longrightarrow \exists u \in \texttt{SuperFrechet}\ S.\ \forall x \in c.\ x \subseteq u$$

Thus, we now have to show that each chain of SuperFrechet S has an upper bound in SuperFrechet S. Since the empty set is also a chain, we need to consider the two possibilities for the chain c,

1) $c = \emptyset$: We simply use the fact that Frechet $S \in$ Filters S and hence that Frechet $S \in$ SuperFrechet S to prove the theorem for this case.[2]

2) $c \neq \emptyset$: This case is trickier. The proof consists in choosing the union of the nonempty chain c, $\bigcup c$, as the upper bound we are looking for. It is trivially true that $x \subseteq \bigcup c$ for all $x \in c$. To bring the proof to conclusion, it just remains to show that SuperFrechet S is closed under the union of nonempty chains:

$$[|c \neq \emptyset;\ \neg \texttt{finite}\ S; c \in \texttt{chain}\ (\texttt{SuperFrechet}\ S)|] \\ \Longrightarrow \bigcup c \in \texttt{SuperFrechet}\ S$$

The proof requires showing that $\bigcup c$ is a filter. Property (F1) for a filter is proved easily using Isabelle's classical reasoner. In outline, to prove (F2), we choose $x_0 \in \bigcup c$, and $x_1 \in \bigcup c$. Then $x_0 \in G_0$ and $x_1 \in G_1$ for some filters G_1 and G_2 in the chain c. Since c is a chain we have that $G_1 \subseteq G2$ or $G_2 \subseteq G_1$. If $G_1 \subseteq G_2$ then $x_0, x_1 \in G_2$ and so, by (F1), $x_0 \cap x_1 \in G_2 \subseteq \bigcup c$; the case $G_2 \subseteq G_1$ is proved in a similar way. Finally, we prove that Property (F3) also holds from the properties of chains and unions. We shall omit the details for this last step since they are easily deduced.

Freeness Property

The second part of the main goal consists in proving that the ultrafilter does not contain any finite set. Making use of the statement proved in the previous part, this reduces to solving the following subgoal (i.e. deriving a contradiction) in Isabelle:

$$[|U \in \texttt{SuperFrechet}\ S; x \in U; \texttt{finite}\ x|] \Longrightarrow \texttt{False}$$

To prove this, we first deduce that $(S - x) \in U$ since finite $(S - (S - x))$ and Frechet $S \subseteq U$. Hence, since U is closed under set intersection, it follows that $\emptyset = x \cap (S - x) \in U$ which is a contradiction of Property (F1) of the filter. Thus U is free.

This concludes our proof of the existence of a free ultrafilter on any infinite set. This important theorem will be used in the next section to define the hyperreals by considering a special case known as the **Weak Ultrafilter Theorem**.

[2] We have noticed that many proofs given in the literature neglect to consider the case where c is the empty chain.

We have described so far the mathematical foundations set up in Isabelle to enable the new types of numbers going beyond the traditional number systems to be defined. After carrying out constructions up to the reals, proving Zorn's Lemma in Isabelle and developing a theory of filters, we are now ready to apply the so-called **ultrapower** construction to get the **hyperreals**.

3.6 Ultrapower Construction of the Hyperreals

Our aim is to construct a linearly ordered field $\mathbb{R}^*$ that contains an isomorphic copy of the reals $\mathbb{R}$ extended with other elements. This new, strictly larger field is known as a *nonstandard* or hyperreal number system and obeys the same laws as the reals.

As several authors have pointed out [75, 49], the construction of the hyperreals is reminiscent of the construction of the reals from the rationals using equivalence classes induced by Cauchy sequences. In this case, however, we use a free ultrafilter to partition the set of all sequences of real numbers into equivalence classes. The set of these equivalence classes, that is the quotient set, is used to define the new type `hypreal` denoting the hyperreal numbers.

3.6.1 Choosing a Free Ultrafilter

To start the construction, a free ultrafilter $U_{\mathbb{N}}$ is chosen on the set of natural numbers $\mathbb{N}$. There exists one according to the Weak Ultrafilter Theorem:

> **(WUF) Weak Ultrafilter Theorem.** There exists a free ultrafilter on $\mathbb{N}$.

As can be seen, this is a special case of the UFT corollary from the last section. In fact, we have the implications AC $\Rightarrow$ UFT $\Rightarrow$ WUF, which are not reversible. Thus, UFT is strictly weaker than AC, and WUF is weaker still. To prove WUF, we show that the set of naturals is not finite by an inductive proof and then discharge the premise of the UFT corollary.

This ultrafilter need not be explicitly defined: it does not matter which ultrafilter on $\mathbb{N}$ is used. The set of all free ultrafilters on $\mathbb{N}$ determines a set of isomorphic fields from which we can choose any member to be the set of hyperreal numbers. Thus, in our formalization, we use Hilbert's ϵ-operator to define $U_{\mathbb{N}}$:

$$U_{\mathbb{N}} \equiv (\varepsilon U.\, U \in \texttt{FreeUltrafilters}\ (\texttt{UNIV} :: \texttt{nat set}))$$

Here (`UNIV` :: `nat set`) denotes the set $\{n::\texttt{nat}.\,\texttt{True}\}$ i.e. the set $\mathbb{N}$. Higher order logic provides a typed set theory in which the universal set exists.

Once we have defined $U_{\mathbb{N}}$, its properties that will be used in the proofs involving the hyperreals are established. We give here a list of the theorems that we proved, many of which follow from the definitions given in the previous sections:

1) (`UNIV` :: `nat set`) $\in U_{\mathbb{N}}$

2) $\emptyset \notin U_{\mathbb{N}}$

```
HYPREAL = REAL + FILTER +

constdefs
 U_IN :: "nat set set"
 "U_IN ≡ (@U. u ∈ FreeUltrafilters (UNIV::nat set))"

 (* equivalence relation *)
 hyprel "((nat ⇒ real) * (nat ⇒ real)) set"
 "hyprel ≡ {p. ∃ r s. p = (r,s) ∧ {n. r n = s n}∈U_IN}"

typedef
 hypreal ≡ "{x::(nat ⇒ real). True}/hyprel"     (Equiv.quotient_def)

instance
 hypreal :: {ord, plus, times}

defs

 hypreal_zero_def  "0hr ≡ Abs_hypreal(hyprel^^{λn::nat. 0r})"
 hypreal_one_def   "1hr ≡ Abs_hypreal(hyprel^^{λn::nat. 1r})"

constdefs
 hypreal_minus :: hypreal ⇒ hypreal
 "- P          ≡ Abs_hypreal(⋃X∈Rep_hypreal(P).
                    hyprel^^{λn::nat. - (X n)})"

 (* embedding for the reals *)
 hypreal_of_real  :: real ⇒ hypreal
 "hypreal_of_real r ≡ Abs_hypreal(hyprel^^{λn::nat. r})"

 hrinv          :: hypreal ⇒ hypreal
 "hrinv P       ≡ Abs_hypreal(⋃X∈ Rep_hypreal(P).
   hyprel^^{λn. if X n = 0r then 0r else rinv (X n)})"

defs
 hypreal_add_def
 "P + Q  ≡ Abs_hypreal(⋃X∈Rep_hypreal(P).
    ⋃Y∈Rep_hypreal(Q). hyprel^^{λn::nat. X n + Y n})"

      ...

 hypreal_less_def
 "P < (Q::hypreal)  ≡ ∃X Y. X∈Rep_hypreal(P) ∧
       Y∈Rep_hypreal(Q) ∧ {n::nat. X n < Y n} ∈U_IN"
```

Figure 3.3: Isabelle/HOL Theory for Hyperreals

3) $[|X \in U_{\mathbb{N}}; Y \in U_{\mathbb{N}}|] \Longrightarrow X \cap Y \in U_{\mathbb{N}}$

4) $[|X \in U_{\mathbb{N}}; X \subseteq Y|] \Longrightarrow Y \in U_{\mathbb{N}}$

5) $X \in U_{\mathbb{N}} \Longrightarrow \neg\, \texttt{finite}\, X$

6) $X \in U_{\mathbb{N}} \iff -X \notin U_{\mathbb{N}}$

7) $\{n.\, P(n)\} \in U_{\mathbb{N}} \Longrightarrow \exists n.\, P(n)$

8) $X \cup Y \in U_{\mathbb{N}} \Longrightarrow X \in U_{\mathbb{N}} \vee Y \in U_{\mathbb{N}}$

3.6.2 Equality

Using $U_{\mathbb{N}}$, the hyperreals are constructed by considering the set of all sequences of real numbers indexed by $\mathbb{N}$ and defining the following equivalence relation on this set.

DEFINITION 3.6.1. *Given two sequences of real numbers* $\langle r_n \rangle$ *and* $\langle s_n \rangle$,

$$\langle r_n \rangle \sim_{U_{\mathbb{N}}} \langle s_n \rangle \iff \{n \in \mathbb{N} \mid r_n = s_n\} \in U_{\mathbb{N}}$$

The sequences $\langle r_n \rangle$ and $\langle s_n \rangle$ are sometimes said to be equal *almost everywhere* (a.e.). This terminology is used to mean that the entries of a sequence determine some set in the ultrafilter $U_{\mathbb{N}}$.

Figure 3.3 shows Isabelle's theory HYPREAL in which the new type hypreal is introduced using the definition above. The relation hyprel denotes $\sim_{U_{\mathbb{N}}}$ in the theory:

$$\texttt{hyprel} \equiv \{p.\, \exists r\ s.\, p = (r, s) \wedge \{n.\, r(n) = s(n)\} \in U_{\mathbb{N}}\}$$

The first property that we prove is that hyprel is an equivalence relation.

PROPOSITION 3.6.1. *The relation* $\sim_{U_{\mathbb{N}}}$ *is an equivalence relation.*

Proof: Let $\langle a_n \rangle$,$\langle b_n \rangle$, $\langle c_n \rangle$ be sequences of real numbers.
reflexivity: Since $\mathbb{N} \in U_{\mathbb{N}}$, we have $\langle a_n \rangle \sim_{U_{\mathbb{N}}} \langle a_n \rangle$ and thus $\sim_{U_{\mathbb{N}}}$ is reflexive.

symmetry: if $\langle a_n \rangle \sim_{U_{\mathbb{N}}} \langle b_n \rangle$ then, by symmetry of equality, $\langle b_n \rangle \sim_{U_{\mathbb{N}}} \langle a_n \rangle$ implying that $\sim_{U_{\mathbb{N}}}$ is symmetric.

transitivity: Now, given $\langle a_n \rangle \sim_{U_{\mathbb{N}}} \langle b_n \rangle$ and $\langle b_n \rangle \sim_{U_{\mathbb{N}}} \langle c_n \rangle$, let $A = \{n \in \mathbb{N} \mid a_n = b_n\}$ and $B = \{n \in \mathbb{N} \mid b_n = c_n\}$, and $C = \{n \in \mathbb{N} \mid a_n = c_n\}$ then $A \cap B \subseteq C$. Since $A, B \in U_{\mathbb{N}}$, it follows that $A \cap B \in U_{\mathbb{N}}$ since $U_{\mathbb{N}}$ is $\cap$-closed, and hence $C \in U_{\mathbb{N}}$ since $U_{\mathbb{N}}$ is also $\subseteq$-closed. Therefore, $\langle a_n \rangle \sim_{U_{\mathbb{N}}} \langle c_n \rangle$. □

3.6.3 Defining Operations on the Hyperreals

Arithmetic operations on the new type, that is on the equivalence classes, are usually defined in terms of the pointwise operations on the sequences. Let $[\langle X_n \rangle]$ denote the equivalence class containing $\langle X_n \rangle$. Addition, for example, is defined by

$$[\langle X_n \rangle] + [\langle Y_n \rangle] \equiv [\langle X_n + Y_n \rangle] \tag{3.3}$$

In Isabelle, however, using the abstraction and representation functions, we define addition on hyperreals P and Q as follows,

$$\begin{aligned} P + Q \equiv\ & \texttt{Abs\_hypreal}\,(\bigcup X \in \texttt{Rep\_hypreal}(P). \\ & \bigcup Y \in \texttt{Rep\_hypreal}(Q).\, \texttt{hyprel``}\{\lambda n.\, X\,n + Y\,n\}) \end{aligned}$$

Then we prove equation (3.3) above as a theorem. It can then be supplied to the simplifier for use in many of the proofs. In Isabelle, equation (3.3) takes the following form:

$$\begin{aligned} &\texttt{Abs\_hypreal}\,(\texttt{hyprel``}\{\lambda n.\ Xn\}) + \texttt{Abs\_hypreal}\,(\texttt{hyprel``}\{\lambda n.\ Yn\}) \\ &\qquad = \texttt{Abs\_hypreal}\,(\texttt{hyprel``}\{\lambda n.\ Xn + Yn\}) \end{aligned} \tag{3.4}$$

Properties such as commutativity and associativity follow straightforwardly from the corresponding properties of the reals. We can similarly prove $\texttt{0hr} + P = P$ when 0hr is defined as shown in Figure 3.3. Multiplication is defined in a similar way to addition. Associativity, commutativity, and distributivity of multiplication are all directly inherited from the reals and easily proved.

3.6.4 Ordering

The ordering relation on the hyperreals is defined as follows:

$$\begin{aligned} P < Q \equiv\ & \exists X \in \texttt{Rep\_hypreal}\ P. \\ & \exists Y \in \texttt{Rep\_hypreal}\ Q.\, \{n.\, X\,n < Y\,n\} \in U_{I\!N} \end{aligned}$$

We prove the following simplification theorem expressing the order relation in terms of equivalence classes of sequences of real numbers. A hyperreal $[\langle X_n \rangle]$ is less than a hyperreal $[\langle Y_n \rangle]$ if and only if X_n is less than Y_n *almost everywhere*:

$$\begin{aligned} &\texttt{Abs\_hypreal}\,(\texttt{hyprel``}\{X\,n\}) < \texttt{Abs\_hypreal}\,(\texttt{hyprel``}\{Y\,n\}) \\ &\qquad \Longleftrightarrow \{n.\, X\,n < Y\,n\} \in U_{I\!N} \end{aligned}$$

Also, the system of hyperreal numbers generated by the free ultrafilter is a totally ordered field. To show this, we first prove that the ordering relation is total. This proof is relatively simple and follows from the fact that given any two hyperreals $[\langle x_n \rangle]$ and $[\langle y_n \rangle]$, either they are equal leading to

$$\{n \in I\!N \mid x_n = y_n\} \in U_{I\!N}$$

or else, by the Complement property of the ultrafilter as given in Section 3.6.1, we have that

$$\{n \in I\!N \mid x_n \neq y_n\} \in U_{I\!N}$$

In the second case, since the reals are totally ordered, we have to consider the sets $\{n \in I\!N \mid x_n < y_n\}$ and $\{n \in I\!N \mid y_n < x_n\}$. We know that only one of these can belong to the free ultrafilter $U_{I\!N}$ (since otherwise, closure of $U_{I\!N}$ under intersection would entail that $\emptyset \in U_{I\!N}$ which contradicts property (F1) of the filter).

3.6.5 Multiplicative Inverse

To show that $\mathbb{R}^*$ is a field, we need only prove that each non-zero element $[\langle X_n \rangle] \in \mathbb{R}^*$ has a multiplicative inverse. For any non-zero element, we have

$$\{n \in \mathbb{N} \mid X_n = 0\} \notin U_{\mathbb{N}}$$

and therefore, once more by the Complement property of $U_{\mathbb{N}}$,

$$\{n \in \mathbb{N} \mid X_n \neq 0\} \in U_{\mathbb{N}}$$

Therefore, define $Y_n = 1/X_n$ for each value of n for which $X_n \neq 0$ and set $Y_n = 0$ otherwise. Then the set $\{n \in \mathbb{N}. X_n \cdot Y_n = 1\} \in U_{\mathbb{N}}$, so that $[\langle X_n \rangle] \cdot [\langle Y_n \rangle] = [\langle 1 \rangle]$. This motivates the following definition, in Isabelle, for the inverse function `hrinv`:

$$\begin{aligned} \texttt{hrinv}\ P \equiv\ & \texttt{Abs\_hypreal}\ (\bigcup X \in \texttt{Rep\_hypreal}(P). \\ & \texttt{hyprel``}\{\lambda n.\ \texttt{if}\ X\,n = \texttt{0r}\ \texttt{then}\ \texttt{0r}\ \texttt{else}\ \texttt{rinv}\ (X\,n)\}) \end{aligned}$$

It is easily proved that for all non-zero x, $\texttt{hrinv}\ x \cdot x = \texttt{1hr}$ as required. A few points worth mentioning are that `hrinv` x stands for the more conventional notation x^{-1} when x is an hyperreal; the inverse function for the reals is itself denoted by `rinv`, while `0r` and `1r` are defined as the zero and one respectively of the real field. Once again, for simplification purposes, we prove the useful theorem about inverse involving the equivalence classes of real sequences:

$$\begin{aligned} &\texttt{hrinv}\ (\texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{X\,n\})) \Longleftrightarrow \\ &\texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{\texttt{if}\ X\,n = \texttt{0r}\ \texttt{then}\ \texttt{0r}\ \texttt{else}\ \texttt{rinv}\ (X\,n)\}) \end{aligned}$$

We have shown in the above that $\mathbb{R}^*$ is a totally ordered field. The next important step is to show that $\mathbb{R}^*$ contains a proper subfield that is isomorphic to the reals $\mathbb{R}$.

3.7 Structure of the Hyperreal Number Line

In this section, we continue our investigation by introducing and defining the various elements that make up the new totally ordered field which we show to be a proper extension of the reals. We also define a number of concepts that follow from this classification of the elements of $\mathbb{R}^*$.

3.7.1 Embedding the Reals

Since our free ultrafilter has been fixed, we have effectively restricted our attention to one particular totally ordered field $\mathbb{R}^*$, though as we mentioned previously, there are infinitely many distinct but isomorphic number systems. We now embed the reals in our hyperreals by defining a map `hypreal_of_real` :: `real` $\Rightarrow$ `hypreal` in Isabelle. This embedding is defined by

$$\texttt{hypreal\_of\_real}\ r = [\langle r, r, r, \ldots \rangle]$$

and expressed in Isabelle as

$$\texttt{hypreal\_of\_real}\ r \equiv \texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{\lambda n{::}\texttt{nat}.\ r\})$$

In what follows, any embedded real r will be denoted by $\widetilde{r}$ unless the embedding function `hypreal_of_real` is used explicitly. Thus, the additive identity element `0hr` and the multiplicative identity element `1hr` of the hyperreals are the explicit images of the real numbers zero (`0r`) and one (`1r`) respectively under the embedding. To show that `hypreal_of_real` maps $\mathbb{R}$ to a proper subfield of $\mathbb{R}^*$, we first define the following hyperreal number:

$$\omega \equiv \texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{\lambda n{::}\texttt{nat}.\ \texttt{real\_of\_nat}\ n\})$$

where `real_of_nat` :: `nat` ⇒ `real` maps its natural argument n to the real $n+1$. For clarity, we omit the details of the various intermediate embeddings (`nat` ⇒ `pnat`, `pnat` ⇒ `prat`, `prat` ⇒ `preal`, etc.) required for defining `real_of_nat`, though we do need to prove their various properties (e.g. they are injective and order preserving) explicitly in Isabelle. This sort of detail is not usually mentioned in textbooks where it is assumed that one can define a map in one step.

We can now exhibit a member of $\mathbb{R}^*$ that is not equal to any real number, since there is no r such that $\widetilde{r} = \omega$. This is because the set on which $\langle r, r, r, \ldots\rangle$ and $\langle 1, 2, 3, \ldots\rangle$ coincide can consist of at most one element. Hence, by the definition of ultrafilter $U_{\mathbb{N}}$, the two sequences cannot be equivalent since no finite set can belong to $U_{\mathbb{N}}$. In fact, as we shall see shortly, $\widetilde{r} < \omega$ for any real number r, that is, ω is a so-called **infinite** number. Similarly, $\epsilon = \omega^{-1} = [\langle 1, \frac{1}{2}, \frac{1}{3}, \ldots\rangle]$ is an **infinitesimal**.

We will call all members of $\mathbb{R}^*$ that are images of the reals, the **standard** elements of $\mathbb{R}^*$. We then define the set of standard reals `SReal` in the theory `NSA` as follows,

$$\texttt{SReal} \equiv \texttt{range}\ (\texttt{hypreal\_of\_real})$$

where

$$\texttt{range}\ f = \{y.\ \exists x.\ y = f\ x\}$$

We can now view `SReal` as the real numbers embedded in $\mathbb{R}^*$, that is as a sub-ordered field if we agree to identify each real number r with the corresponding standard element $\widetilde{r}$ of $\mathbb{R}^*$. We then have that $\mathbb{R}^*$ is an extension or enlargement of $\mathbb{R}$. We shall come across the general concept of set extensions once more in Section 6.1.

3.7.2 Nonstandard Numbers

We have exhibited in the previous section a hyperreal, ω, that does not belong to `SReal`. There are infinitely many of these so-called **nonstandard** hyperreal numbers. They can be classified into various sets that include, for example, infinitesimals and the infinite numbers. We start this section with a preamble where the absolute value function for the hyperreals is introduced. This function is needed in order to define the various types of numbers found in our theory. Moreover, it also shows some of the characteristics that we will encounter later on when dealing with functions in a nonstandard setting. We present the important theorems proved in Isabelle as we proceed in our exposition.

The definitions of infinitesimal, finite, and infinite numbers use the absolute value function. This function, which we also defined on the reals, needs to be extended to the hyperreal numbers. The definition that we use is analogous to

that used for the reals. Using the if-then-else construct of Isabelle HOL, we have

$$\texttt{hrabs}\ x \equiv \texttt{if}\ \texttt{0hr} \leq x\ \texttt{then}\ x\ \texttt{else}\ -x$$

In fact, an alternative definition exists in which the (real) absolute value function is simply applied pointwise to an equivalence class representative in $\mathbb{R}^*$. In Isabelle, with `rabs` denoting the absolute value function for the reals, this takes the form of the following theorem:

$$\begin{aligned}&\texttt{hrabs}\ (\texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{X\})) = \\ &\qquad \texttt{Abs\_hypreal}\ (\texttt{hyprel``}\{\lambda n.\ \texttt{rabs}\ (X\ n)\})\end{aligned}$$

This result, taken in conjunction with the definitions of the operations such as addition, multiplication, and reciprocal hints at a general technique in which functions can be defined on the hyperreals through *extensions* of the analogous ones defined on the reals using our free ultrafilter $U_{\mathbb{N}}$. We shall be examining this notion of extension in Section 6.1.

The intuitive notion of an infinitesimal number can now be formally defined. Sets of finite and infinite numbers are also introduced formally.

DEFINITION 3.7.1. *An element x of $\mathbb{R}^*$ is said to be an **infinitesimal** if and only if for every positive standard real number r we have $|x| < r$. It is **finite** if and only if for some standard real number r we have $|x| < r$; and **infinite** if and only if for every standard real number r we have $r < |x|$.*

In the literature, the definition will often just say that an infinitesimal is less in magnitude than any positive (standard) *real* number. Here, since we have different types, it becomes explicit that such a definition is actually referring to the standard copy in $\mathbb{R}^*$. This leads to the following definition in Isabelle for the set of `Infinitesimal`

$$\begin{aligned}&\texttt{Infinitesimal} :: \texttt{hypreal set} \\ &\texttt{Infinitesimal} \equiv \{x.\ \forall r \in \texttt{SReal}.\ \texttt{0hr} < r \longrightarrow \texttt{hrabs}\ x < r\}\end{aligned}$$

This definition can be considered as a high level one. Indeed, it is possible to define the set of infinitesimals by going down to the level of our free ultrafilter $U_{\mathbb{N}}$ itself. We thus prove the next theorem, which turns out to be useful when supplied to Isabelle's simplifier in cases where one wants to deal with properties of real sequences rather than notions of infinitesimals.

$$\begin{aligned}(x \in \texttt{Infinitesimal}) \Longleftrightarrow\ &(\exists X \in \texttt{Rep\_hypreal}\ x.\ \forall u.\ \texttt{0r} < u \\ &\longrightarrow \{n.\ \texttt{rabs}\ (X\ n) < u\} \in U_{\mathbb{N}})\end{aligned}$$

We give below the definitions for the sets `Finite` and `Infinite` of finite and infinite numbers respectively, as declared in Isabelle, and the equivalent theorems derived in terms of the free ultrafilter:

$$\texttt{Finite} :: \texttt{hypreal set}$$
$$\texttt{Finite} \equiv \{x.\, \exists r \in \texttt{SReal}.\, \texttt{hrabs}\ x < r\}$$

$$(x \in \texttt{Finite}) \Longleftrightarrow (\exists X \in \texttt{Rep\_hypreal}\ x.\ \exists u.\, \{n.\, \texttt{rabs}\ (X\, n) < u\} \in U_{\mathbb{N}})$$

$$\texttt{Infinite} :: \texttt{hypreal set}$$
$$\texttt{Infinite} \equiv \{x.\, \forall r \in \texttt{SReal}.\, r < \texttt{hrabs}\ x\}$$

$$(x \in \texttt{Infinite}) \Longleftrightarrow (\exists X \in \texttt{Rep\_hypreal}\ x.\ \forall u.\, \{n.\, u < \texttt{rabs}\ (X\, n)\} \in U_{\mathbb{N}})$$

In fact, we can view the various low-level theorems as lemmas that enable us to translate properties involving the hyperreals into those depending on the ultrafilter. This is useful in our mechanization when we deal with real functions and their extensions.

An important point, highlighted through the definition of infinite and infinitesimal numbers, and already mentioned in Section 3.2, is that the set of hyperreal numbers is non-Archimedean. This is because not every bounded subset of $\mathbb{R}^*$ is guaranteed to have a least upper bound or greatest lower bound. For example, the set of infinite numbers is bounded below by any finite number but has no greatest lower bound. In Section 6.1, we consider sets of hyperreals that do have least upper bounds and use their special properties.

3.7.3 On Infinitesimals, Finite and Infinite Numbers

We have proved various properties of infinitesimals, finite and infinite numbers. A few of the theorems are listed below:

- The set `Finite` of finite elements is a **subring** of $\mathbb{R}^*$ i.e. sums, differences, and products of finite elements are finite.
- The set `Infinitesimal` of infinitesimals is also a subring of $\mathbb{R}^*$.
- The set `Infinitesimal` is an ideal in `Finite` i.e. the product of an infinitesimal and a finite number is infinitesimal.
- x is infinite if and only if `hrinv` x is infinitesimal for all non-zero x.

The hyperreal number ω defined in Section 3.7.1 is a member of `Infinite`: for any given real number x, for all sufficiently large values of n, we have $x < n$. The infinitesimal number ϵ defined by the equivalence class containing the sequence $\langle 1/n \rangle$ is a member of `Infinitesimal` since for any given x, for all sufficiently large value of n, we have $0 < 1/n < x$. We have also proved that ω is the multiplicative inverse of ϵ, since

$$\begin{aligned}\omega \cdot \epsilon &= [\langle 1, 2, 3, \ldots \rangle] \cdot [\langle 1, 1/2, 1/3, \ldots \rangle] \\ &= [\langle 1 \cdot 1, 2 \cdot 1/2, 3 \cdot 1/3, \ldots \rangle] \\ &= [\langle 1, 1, 1, \ldots \rangle] \\ &= \texttt{1hr}.\end{aligned}$$

We next introduce an important equivalence relation that will be extremely useful to our mechanization.

DEFINITION 3.7.2. *Two hyperreal numbers x and y are said to be* ***infinitely close****, $x \approx y$, if and only if their difference $x - y$ is infinitesimal.*

It is easily proved that x is an infinitesimal if and only if $x \approx 0$. To show that $\approx$ is an equivalence relation is trivial. In addition, we prove the following theorems (amongst others):

1) $[|a \approx b; c \approx d|] \Longrightarrow a + c \approx b + d$

2) $[|a \approx b; c \approx d|] \Longrightarrow a - c \approx b - d$

3) $(a + b \approx a + c) \Longleftrightarrow b \approx c$

4) $[|a \approx b; c \in \texttt{Finite}|] \Longrightarrow a \cdot c \approx b \cdot c$

5) $[|a \approx b; c \approx d; b \in \texttt{Finite}; c \in \texttt{Finite}|] \Longrightarrow a \cdot c \approx b \cdot d$

6) $[|a \in \texttt{Finite}; a \approx b|] \Longrightarrow b \in \texttt{Finite}$

7) $[|a \in \texttt{SReal}; a \neq \texttt{0hr}|] \Longrightarrow (a \cdot x \approx a \cdot y) = (x \approx y)$

8) $[|x \in \texttt{SReal}; y \in \texttt{SReal}|] \Longrightarrow (x \approx y) = (x = y)$

9) $[|x \approx y; y \in \texttt{Finite} - \texttt{Infinitesimal}|] \Longrightarrow \texttt{hrinv}\ x \approx \texttt{hrinv}\ y$

10) $x \approx y \Longrightarrow \texttt{hrabs}\ x \approx \texttt{hrabs}\ y$

We continue in the next section with another basic fact about the structure of $\mathbb{R}^*$, which defines a function from the set of finite numbers onto the reals.

3.7.4 The Standard Part Theorem

The standard part of a finite nonstandard number is defined to be the unique real infinitely close to it. We use Hilbert's choice operator, ϵ, to express this in Isabelle:

$$\texttt{st}\ x \equiv (\varepsilon r.\ r \in \texttt{SReal} \wedge r \approx x)$$

We now prove the existence and uniqueness of the standard part. Existence needs to be demonstrated in any case whenever Hilbert's operator is used.

PROPOSITION 3.7.1. *Let x be a finite hyperreal number. Then, there exists a unique standard real number r such that $r \approx x$.*

Proof: Let $A = \{y \in \mathbb{R} \mid y \leq x\}$. Since x is finite, A is nonempty and is bounded above. Let r be the least upper bound of A. For any real $\epsilon > 0$, $r - \epsilon \in A$ and $r + \epsilon \notin A$ and thus $r - \epsilon \leq x < r + \epsilon$. So $|r - x| \leq \epsilon$ from which it follows that $r \approx x$.

To show uniqueness, suppose that there exists a real number s such that $s \approx x$. Then, since $\approx$ is transitive, $s \approx r$ and so $r - s \approx 0$. But $r - s$ is real, so $r - s = 0$ and $r = s$. □

The proof given above glosses over many of the details that need to be satisfied for mechanization. Indeed, the completeness of the reals, and hence of the embedded reals, is needed in the form of the *supremum property*, which ensures that any nonempty set of reals that is bounded above has a least upper

bound. We first proved the property for the positive real numbers (`preal`) and then extended it to the reals (`real`). Now, since we are dealing with the hyperreals and identifying the reals with the proper subfield of $\mathbb{R}^*$ which is isomorphic to $\mathbb{R}$, we have to transfer this theorem explicitly to the isomorphic copy of $\mathbb{R}$, namely `SReal`.

Once the existence of the standard part has been proved, we prove various of their properties: for any $x, y \in$ `Finite`, we have,

- $x \approx y \Longleftrightarrow \mathtt{st}\ x = \mathtt{st}\ y$
- $x \approx \mathtt{st}\ x$
- $x \in \mathtt{SReal} \Longrightarrow \mathtt{st}\ x = x$
- $\mathtt{st}\ (x + y) = \mathtt{st}\ x + \mathtt{st}\ y$
- $\mathtt{st}\ (x \cdot y) = \mathtt{st}\ x \cdot \mathtt{st}\ y$
- if $\mathtt{st}\ y \neq \mathtt{0hr}$ then $\mathtt{st}\ (x \cdot \mathtt{hrinv}\ (y)) = \mathtt{st}\ x \cdot \mathtt{hrinv}\ (\mathtt{st}\ y)$
- if $x \leq y$ then $\mathtt{st}\ x \leq \mathtt{st}\ y$
- $\mathtt{st}\ (\mathtt{st}\ x) = \mathtt{st}\ x$
- $\mathtt{st}\ (\mathtt{hrabs}\ x) = \mathtt{hrabs}\ (\mathtt{st}\ x)$

From some of these theorems, we can see that the map preserves algebraic structure. The standard part function can be defined in other ways for an ultrapower. For example, it corresponds to the order homomorphism of `Finite` with kernel `Infinitesimal` onto $\mathbb{R}$ [76]. The standard part is an important concept that can be used when formulating the nonstandard definition for the limit of a sequence of reals (see Section 6.3.1) and also when defining the *slope* of a real function at a real point, as we shall see in Section 6.6.

3.8 The Hypernatural Numbers

We can construct a set of numbers $\mathbb{N}^*$ that contains both finite elements, identifiable with the ordinary natural numbers themselves, and infinite numbers greater than all natural numbers. This discrete set is known as the **hypernaturals**. We are interested in this type of numbers as they will be needed in the nonstandard formalization of real sequences and series (see Section 6.3).

The construction of the hypernaturals in Isabelle is analogous to that of the hyperreals: we use the same free ultrafilter $U_{\mathbb{N}}$ but replace sequence of reals by sequences of natural numbers. Thus, $\mathbb{N}^*$ is now characterized explicitly as the set of equivalence classes $[\langle m_n \rangle]$ determined by sequences m_n of natural numbers. The new equivalence relation on sequences is denoted by `hypnatrel` in Isabelle. In what follows, we make some observations on the construction and properties that apply to members of $\mathbb{N}^*$. These are interesting in their own right but also in view of the applications to mechanization of analysis using nonstandard methods.

We define an embedding function that identifies each natural number m with the hypernatural number determined by the constant sequence $\langle m, m, ..., m\rangle$. In Isabelle, we define the function `hypnat_of_nat :: nat ⇒ hypnat`:

$$\texttt{hypnat\_of\_nat}\ m \equiv \texttt{Abs\_hypnat}\ (\texttt{hypnatrel}\hat{}\hat{}\{\lambda n{::}\texttt{nat}.\, m\})$$

Using the map `hypnat_of_nat`, we easily define the set `SHNat` of *standard* natural numbers embedded in $\mathbb{N}^*$:

$$\texttt{SHNat} \equiv \texttt{range}\ (\texttt{hypnat\_of\_nat})$$

In what follows, a natural number n embedded in the hypernaturals will also be denoted by $\overline{n}$ in some cases.

3.8.1 Infinite Hypernaturals

We define a hypernatural Ω denoting $[\langle n\rangle] = [\langle 0, 1, 2, ...\rangle]$ by

$$\Omega \equiv \texttt{Abs\_hypnat}\ (\texttt{hypnatrel}\hat{}\hat{}\{\lambda n{::}\texttt{nat}.\, n\})$$

We prove that for any embedded natural number $n \in$ `SHNat`, $\Omega \neq n$ meaning that $\mathbb{N}^*$ properly includes $\mathbb{N}$. This motivates the following definition for the set of non-standard hypernaturals:

$$\texttt{HNatInfinite} \equiv -\ \texttt{SHNat}$$

To establish that the only non-standard hypernaturals are the infinite ones, we prove the following equivalence theorem:

$$\texttt{HNatInfinite} \Longleftrightarrow \{N. \forall n \in \texttt{SHNat}.\, n < N\}$$

Thus, $\mathbb{N}^*$ consists of the finite standard copies of the ordinary natural numbers and of the infinite hypernatural numbers only.

3.8.2 Properties of the Hypernaturals

Some of the properties proved for the hypernatural numbers are these:

1) $\mathbb{N}^*$ is a discrete subset of $\mathbb{R}^*$

2) $\mathbb{N}^*$ is closed under addition and multiplication.

3) Every infinite number has an immediate predecessor which is also infinite.

The first property can be proved either by defining directly an embedding function from the hypernaturals to the hyperreals or by taking the nonstandard extension of the set of natural numbers (embedded in the reals).

An important observation, following from the third theorem above, is that the non-empty set of infinite hypernatural numbers, `HNatInfinite`, does not have a **least** element. Thus, the well-ordering property of the natural numbers does not extend to the hypernaturals. This observation shows that, though most properties of the natural numbers are transferred to the hypernaturals, there are important exceptions. It will be seen in our subsequent exposition that properties such as the one above and the Archimedean property extend only to *special* subsets of the hypernaturals and hyperreals respectively.

3.9 An Alternative Construction for the Reals

The construction of the reals using Dedekind cuts is well established in the literature. However, the method has many critics due to the problems that arise with negative cuts, case splitting and so on. We avoided these by defining only the positive reals.

Nonstandard Analysis can be used to provide an alternative construction for the reals. The idea is to apply the same techniques that we have used to enlarge the reals into the hyperreals to the rationals, $\mathbb{Q}$, in order to get the *hyperrationals*, $\mathbb{Q}^*$. The sets of infinitesimal rationals and finite rational numbers are then defined in a similar way to the sets `Infinitesimal` and `Finite` respectively. The set of infinitesimal rationals is a maximal ideal and the reals can then be seen as the quotient ring of the finite elements of $\mathbb{Q}^*$ modulo the maximal ideal of the infinitesimal rationals. This quotient set defines an Archimedean field isomorphic to the reals.

It should be relatively straightforward to carry out this construction in Isabelle, given that we already have the necessary framework to extend the rationals. Our experience with extending the reals and the naturals shows that much of the code can be reused. Most of the proofs (e.g. that $\mathbb{Q}^*$ is an ordered field) will not only follow directly from the rationals but will also have exactly the same steps as those for the hyperreals. This possible construction of the reals in Isabelle is left as a potentially instructive exercise — especially when compared to Dedekind cuts or other more conventional constructions.

3.10 Related Work

The reals were first constructed in Automath in 1977 by Jutting [50] who translated Landau's famous monograph on the foundations of analysis [57]. More recently, Harrison has constructed the reals and formalized a substantial amount of analysis in HOL [43]. The work of Harrison has influenced some of our decisions during mechanization, especially when formalizing analysis, where we have benefited from the observations made by him on notations, for example. We shall be coming back to these aspects when describing our formalization of analysis using the hyperreals. As far as our own constructions up to the reals are concerned, we have followed mostly the presentation given by Gleason [39], since it matches the sequence of constructions that Conway advocates [25].

3.11 Concluding Remarks

As far as we are aware, there has not been any previously published construction of the hyperreals using a mechanical theorem prover. This chapter has described the construction process resulting in a proper field extension of the reals. Various classes of numbers, including the notorious infinitesimals, have been introduced and their properties formalized. The $\approx$ (infinitely close) relation has been introduced, which is crucial to the formalization of both Newton's *Principia* and of nonstandard real analysis. The framework has been shown to be flexible by allowing the hypernaturals, and their associated properties, to be formalized with minimal effort.

To reach the hyperreals has involved all the constructions up to the reals (which we have not described in much detail) and proving the various properties of each number system introduced; it also involved working in Isabelle/HOL set theory to formalize Zorn's Lemma and the theory of filters and ultrafilters. As might be expected, a number of interesting remarks emerge from this development. We outline some of these next.

The formalization of filters is an important contribution since these are useful concepts with numerous applications in set theory, logic, algebra etc. They can also be used to study the various notions of convergence: they yield essentially the same results as (convergence) nets [74]. Nets provide a natural generalization of sequences and are commonly used in analysis. In fact, nets are also useful to the mechanization of analysis, as was shown by Harrison in HOL [43]. Thus, Isabelle's theory of filters could be used for a general theory of convergence.

Since this work formalizes the Ultrafilter Theorem, the ultrapower construction becomes available for the development of other nonstandard number systems. For instance, the hyperintegers could be introduced. In particular, it becomes possible to construct the *hyperhyperreal* numbers from the hyperreals. These numbers are introduced by Henle and Kleinberg [44], for example, and are shown to contain, in addition to the field $\mathbb{R}^*$, numbers even smaller than the infinitesimals. The new hyperhyperreal field can be used, with benefits, for analysis over the hyperreals. On the other hand, ultrapowers also have other independent uses: they are important concepts in the study of Banach spaces, for instance.

In Chapter 6, the development of some real analysis in Isabelle, using nonstandard techniques, is covered. This formalization is compared with the work carried out by Harrison in HOL. The advantages that the more algebraic and often more intuitive nonstandard formulation of familiar concepts has over the standard approach are pointed out. Of course, the work still proceeds strictly through definitions.

Before this, we examine in the next chapter some of the notions that arise with the introduction of infinitesimals in geometry. In particular, we mention non-Archimedean geometry and some of the new concepts are defined using the infinitely close relation. Using the hyperreals, we then build a theory of elementary vector geometry and use it to verify the geometric methods of Isabelle.

Chapter 4

Infinitesimal and Analytic Geometry

This chapter covers some of the concepts and properties that arise when infinitesimal notions are introduced in the geometry theory. The hyperreal space is shown to have a practical and rich application in geometry. A few theorems and proofs are examined. Also as an important part of this work, some algebraic geometry is developed using hyperreal vectors. This is a definitional approach used to formalize notions from the traditional GTP methods and to verify their basic axioms. We start with a brief review of Non-Archimedean geometry.

4.1 Non-Archimedean Geometry

The *Axiom of Archimedes* or *Axiom of Continuity* from Hilbert's *Foundations of Geometry* [46] may be stated as follows:

> Let A, B, C, and D be four distinct points. Then on the ray AB there is a finite set of distinct points, A_1, A_2,..., A_n such that each segment A_iA_{i+1} is congruent to the segment CD and such that B is between A and A_n.

This means that given any line segment of length l and any measure m, there exists an integer n such that n units of measure yield a line segment greater than the given line segment i.e. $l < n \cdot m$. Geometrically speaking, this means that the length of a line has no limit, which is a tacit assumption of Euclid. This axiom of Hilbert can therefore be viewed as stating that the points on the line are in one-to-one correspondence with the real numbers $\mathbb{R}$.

After introducing the various groups of axioms, Hilbert proceeds to show their consistency and mutual independence. This is done by interpreting every geometric concept arithmetically and making sure that all the axioms are satisfied in the interpretation. For example, a point is identified with the ordered pair of real numbers (a,b) and a line with the ratio $(u{:}v{:}w)$ in which u and v are both non-zero. A point lies on a line if $ua + vb + w = 0$. Properties such as convergence are interpreted algebraically by means of the expressions for translation and rotation of analytic geometry. Thus, a model is constructed for the

axioms of geometry and any contradiction deduced from these would mean that the axioms of arithmetic are inconsistent.

The possibility of a non-Archimedean geometry is exposed when proving the mutual independence of Hilbert's sets of axioms. Indeed, it is possible to construct a model that satisfies all the various axioms except the Axiom of Archimedes. In such a geometry, our measure m can be laid off successively upon our line segment of length l an arbitrary number of times without ever reaching the end point of the line. This geometry might be seen, intuitively, as one in which infinitesimal notions are allowed. Of course, the most famous example of an axiom being denied in geometry is that of the parallel axiom, which leads to non-Euclidean geometry.

It is worth noting that one of the first to attempt a systematic investigation of non-Archimedean geometry was the Italian mathematician Veronese in his *Fundamenti di Geometria.* As observed by Fisher [33], his work was often unacknowledged by contemporary mathematicians such as Hilbert and Poincaré and only recently have historians given its influence due recognition. Veronese's poor and tortuous exposition has been blamed to some extent for this.

In his review of Hilbert's Foundations of Geometry, Henri Poincaré makes the following important observation about non-Archimedean geometry [70]:

> ...the coordinates of a point would be measured not by ordinary numbers but by non-Archimedean numbers, while the usual operations of the straight lines and the plane would hold, as well as the analytic expressions for angles and lengths. It is clear that in this space all the axioms would remain true except that of Archimedes.

And moreover, he notes

> On every straight line new points would be interpolated between ordinary points.

This matches our approach in which we effectively replace the real number line with a hyperreal one. Poincaré also gives a geometric example where an ordinary line is compared with a non-Archimedean one:

> If, for example, D_0 is an ordinary straight line, and D_1 the corresponding non-Archimedean straight line; if P is any ordinary point of D_0, and if this point divides D_0 into two half rays S and S' (I add, for precision, that I consider P as not belonging to either S or S') then there will be on D_1 an infinity of new points as well between P and S as between P and S'. Then there will be on D_1 an infinity of new points which will lie to the right of all the ordinary points of D_0. In short, our ordinary space is only a part of the non-Archimedean space.

This geometrical representation means that points can be infinitely close to each other on line D_1. Indeed, the first infinity of new points mentioned by Poincaré corresponds to those infinitely close to P. And then, we also have the new points on D_1 that lie beyond those of D_0. These points to the right of all of the points of D_0 thus correspond to the infinite hyperreals. These observations motivate the establishment of a one-to-one correspondence f between the hyperreals and a line L instead of the usual correspondence with the reals. A *coordinate system*, f, for L is then such that each point P on it has a unique hyperreal coordinate given by $x = f(P)$.

4.2 New Definitions and Relations

Since we have explicitly defined the notion of equality between angles, we also need to define the idea of two angles being infinitely close to one another. We use the infinitely close relation to do so:

$$a_1 \approx_a a_2 \equiv \exists n \in I\!N.\ |a_1 - a_2| \approx n\pi$$

This is an equivalence relation. We prove in Isabelle the following property, which could provide an alternative definition for $\approx_a$:

$$a_1 \approx_a a_2 \Longleftrightarrow \exists \epsilon \in \texttt{Infinitesimal}.\ a_1 =_a a_2 + \epsilon$$

We also have the theorem $a_1 \approx a_2 \Longrightarrow a_1 \approx_a a_2$. We now introduce a novel property that can be expressed using the concepts that we have developed so far in our theory— that of two triangles being *ultimately similar*. Recall that two triangles $\triangle abc$ and $\triangle a'b'c'$ are similar if they have equal angles at a and a', at b and b', and at c and c'. The definition of ultimately similar triangles follows:

$$\begin{aligned}\texttt{USIM}\ a\ b\ c\ a'\ b'\ c' \equiv\ & \langle b -\!\!- a, a -\!\!- c\rangle \approx_a \langle b' -\!\!- a', a' -\!\!- c'\rangle \ \wedge \\ & \langle a -\!\!- c, c -\!\!- b\rangle \approx_a \langle a' -\!\!- c', c' -\!\!- b'\rangle \ \wedge \\ & \langle c -\!\!- b, b -\!\!- a\rangle \approx_a \langle c' -\!\!- b', b' -\!\!- a'\rangle\end{aligned}$$

This property allows the treatment of triangles that are being deformed and tending towards similarity as points move in Newton's dynamic geometry. Elimination and introduction rules are developed to deal with the USIM property. We only need to know that two corresponding angles are infinitely close to deduce that two triangles are ultimately similar:

$$\begin{aligned}[|&\langle b -\!\!- a, a -\!\!- c\rangle \approx_a \langle b' -\!\!- a', a' -\!\!- c'\rangle; \\ &\langle a -\!\!- c, c -\!\!- b\rangle \approx_a \langle a' -\!\!- c', c' -\!\!- b'\rangle|] \Longrightarrow \texttt{USIM}\ a\,b\,c\,a'\,b'\,c'\end{aligned}$$

It follows also, trivially, that $\texttt{SIM}\,\texttt{a}\,\texttt{b}\,\texttt{c}\,\texttt{a'}\,\texttt{b'}\,\texttt{c'} \Longrightarrow \texttt{USIM}\,\texttt{a}\,\texttt{b}\,\texttt{c}\,\texttt{a'}\,\texttt{b'}\,\texttt{c'}$. Similarly, we also define the geometric relation of *ultimate congruence* UCONG. Areas, angles and lengths can be made infinitesimal as needed when carrying out the proofs leading to theorems such as these (Figure 4.1):

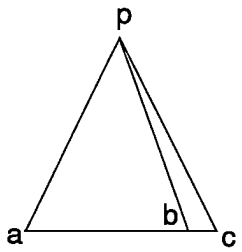


Figure 4.1: Infinitely close areas

- $[|\texttt{coll}\ a\,b\,c; \texttt{s\_delta}\ p\ b\ c \approx \texttt{0hr}|] \Longrightarrow \texttt{s\_delta}\ p\ a\ c \approx \texttt{s\_delta}\ p\ a\ b$
- $[|\texttt{coll}\ a\ b\,c; \texttt{len}\ (b -\!\!- c) \approx \texttt{0hr}|] \Longrightarrow \texttt{s\_delta}\ p\ a\ c \approx \texttt{s\_delta}\ p\ a\ b$

- $[|\texttt{coll}\, a\, b\, c; \langle b --- p, p --- c\rangle \approx_a \texttt{Ohr}|] \Longrightarrow \texttt{USIM}\, p\, a\, b\, p\, a\, c$

The above theorems formalize intuitive properties that result when parts of $\triangle abc$ are allowed to become infinitesimal. Such new relations do not hold in pure Euclidean geometry and therefore cannot be derived from the diagram alone.

4.3 Infinitesimal Geometry Proofs

We now give an example that illustrates the use of infinitesimal arguments in geometry. This is an argument attributed to Nicholas of Cusa who lived in the fifteenth century and which we quote from the book by Davis and Hersh [26]

> We wish to find the relation between the area of a circle and its circumference. For simplicity we suppose that the radius of the circle is 1. Now, the circle can be thought of as composed of infinitely many straight-line segments, all equal to each other and infinitely short. The circle is then the sum of infinitesimal triangles, all of which have altitude 1. For a triangle the area is half the base times the altitude. Therefore the sum of the areas of the triangles is half the sum of the bases. But the sum of the areas of the triangles is the area of the circle, and the sum of the bases of the triangles is its circumference. Therefore the area of the circle of radius 1 is equal to one half of its circumference.

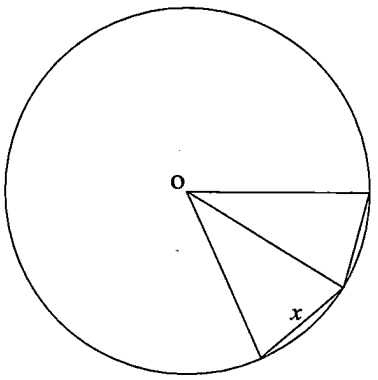


Figure 4.2: Infinitesimal triangles inscribed in a circle

The conclusion of this proof is true though the proof itself without a proper theory of infinitesimals was prone to attacks. As Davis and Hersh mention, the notion of a triangle with infinitely small base was viewed as elusive: either the base, hence the area of the triangle was zero or else it was greater than zero. In the first case, the resulting sum of the areas would also be zero, while in the second an infinitely large sum would result since infinitely many terms were being added together. In neither case, could the circle of finite circumference be obtained as a sum of infinitely many identical pieces. We show that this "proof" by Nicholas of Cusa is correct once we have infinitesimals and infinite numbers available. Indeed, one can prove that the area of the circle is the *standard part*

of the sum of infinitely many infinitesimals. We assume that the area of the circle and its circumference are both real, non-zero quantities i.e.

$$\text{area of circle} \in \mathbb{R} - \{0\} \text{ and circumference of circle} \in \mathbb{R} - \{0\}$$

Let the base of each triangle be of length $x \in \texttt{Infinitesimal}$ and let there be $N \in \texttt{HNatInfinite}$ of them. Now, following Nicholas of Cusa's argument we have

$$\text{area of circle} \approx N \cdot (1/2 \cdot x) \tag{4.1}$$

$$\text{circumference of circle} \approx N \cdot x \tag{4.2}$$

Now we have as a theorem that any real number infinitely close to another number is the standard part of that number (stated as a theorem of Isabelle):[1]

$$[|x \in \texttt{SReal}; x \approx y|] \Longrightarrow x = \texttt{st}\ y \tag{4.3}$$

Therefore by (4.3), (4.1) and (4.2) become

$$\text{area of circle} = \texttt{st}\ (N \cdot (1/2 \cdot x)) \tag{4.4}$$

$$\text{circumference of circle} = \texttt{st}\ (N \cdot x) \tag{4.5}$$

We want to divide (4.4) by (4.5) to get the required result. However, first we need to simplify (4.4). We know that $N \cdot (1/2 \cdot x) = 1/2 \cdot (N \cdot x) \in \texttt{Finite}$ since it is infinitely close to area of circle $\in \mathbb{R}$. Thus, we can use the following theorem to deduce that $N \cdot x \in \texttt{Finite}$:

$$[|a \in \texttt{SReal}; a \cdot y \in \texttt{Finite}|] \Longrightarrow y \in \texttt{Finite}$$

We can now simplify the standard part of the product (4.4) using the following theorems about products of standard parts:

$$[|a \in \texttt{SReal};\ y \in \texttt{Finite}|] \Longrightarrow \texttt{st}\ (a \cdot y) = \texttt{st}\ (a) \cdot \texttt{st}\ y$$

$$a \in \texttt{SReal} \Longrightarrow \texttt{st}\ (a) = a$$

From which, we get the desired conclusion

$$\frac{\text{area of circle}}{\text{circumference of circle}} = \frac{1/2 \cdot \texttt{st}\ (N \cdot x)}{\texttt{st}\ (N \cdot x)} = \frac{1}{2}$$

Our proof is coherent because

- infinitesimals are not necessarily zero and hence we can have a sum greater than zero.
- also, $\mathbb{R}^*$ is a non-Archimedean field, and the sum of finitely many infinitesimals is always an infinitesimal, that is

$$\forall x \in \texttt{Finite}.\, \forall y \in \texttt{Infinitesimal}.\, x \cdot y \in \texttt{Infinitesimal}$$

[1] In stating the various theorems, for clarity, we do not show the embedding functions explicitly; however, the reader should bear in mind that we are working in the hyperreals and so when referring to a real number, for example, we are actually working with its embedded copy.

where Finite contains arbitrarily large *real* numbers since $\mathbb{R} \subseteq$ Finite. Thus, the infinite number of line-segments, N, needs to be a nonstandard natural number. That is, $N \in \mathbb{N}^* - \mathbb{N}$ (an infinite hypernatural), for the sum of areas of the triangles to be infinitely close to the area of the circle which is assumed to be a real quantity (i.e. non-infinitesimal).

4.3.1 Infinitesimal Notions in Euclid's *Elements*

We mentioned that our geometry differs from Euclidean geometry in its infinitesimal aspects. However, while investigating Euclid's *Elements* [30], we did come across a proposition which immediately struck us as involving infinitesimal notions.

Book III of the *Elements* contains definitions relating to the geometry of circles and discusses properties of chords, tangents, inscribed angles etc. The theorem we are interested in is *Proposition 16*, which we quote:

> The straight line drawn at right angles to the diameter of a circle from its extremity will fall outside the circle, and into the space between the straight line and the circumference another straight line cannot be interposed; further the angle of the semi-circle is greater, and the remaining angle less than any acute rectilinear angle.

Euclid's Proposition thus deals with the space between the tangent TA and the arc ACE (Figure 4.3). He notes firstly that no line can be drawn in that space and going through A that falls entirely outside the circle. Secondly, he considers the angle between the tangent TA and the arc ACE, known as a *horn* angle. Our investigation revealed that the exact magnitude of horn angles was the subject of much controversy amongst the ancient Greek geometers and indeed for many centuries afterwards. Kline notes, in his survey of ancient and modern mathematics, that although Proposition 16 states that the angle is smaller in magnitude than any acute angles between straight lines, it does not mention that the angle is of zero magnitude [53].

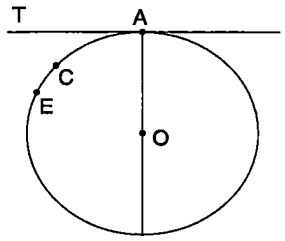


Figure 4.3: The horn angle from Proposition 16 of Euclid's *Elements*

The horn angle posed numerous problems to mathematicians since one would intuitively expect the size of the angle to increase as circles of smaller and smaller diameters pass through A and the tangent to TA; however, according to the preceding proposition, this is not possible. On the other hand, if two horn angles are of zero magnitude (i.e. equal) they should be superposable; but this is not so. This led to suggestions that horn angles were not actual angles.

Formulated in terms of the geometric notions in our framework, horn angles can be viewed as infinitesimal angles since they are smaller in magnitude than

any finite angles. This also explains why, without infinitesimals, it is impossible to notice any increase in the size of the horn angle when circles of decreasing diameters are drawn. Viewed in this light, horn angles probably represent one of the first areas of mathematics where notions of infinitesimals appeared. Rather more recently, there has been a revival of interest in these angles and they can now be encountered in fields such as conformal mappings and indirectly in non-Archimedean analysis.

4.3.2 Useful Infinitesimal Geometric Theorems

As we worked with the infinitesimal geometry, we investigated the effects of allowing infinitely small elements. We wanted to prove results that we felt intuitively should hold and also discover other more subtle ones. As expected, many new properties emerged from the formalization of Newton's ultimate reasoning. We give below a few of the interesting theorems that we proved; more are presented when we go through proofs of Newton's Lemmas and Propositions in the next chapter.

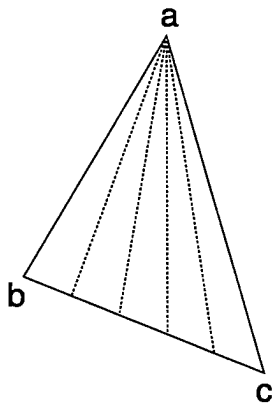


Figure 4.4: A "shrinking" triangle

Consider a triangle bac with sides of finite lengths as the angle $\langle b — a, a — c\rangle$ becomes infinitesimal but the altitude does not change. The following theorems follow:

$$[|\neg \texttt{coll}\ a\,b\,c; \langle b — a, a — c\rangle \approx_a \texttt{0hr}; \texttt{len}\ (a — b) \in \texttt{Finite};\ \texttt{len}\ (a — c) \in \texttt{Finite}|] \Longrightarrow \texttt{len}\ (b — c) \approx \texttt{0hr}$$

$$[|\neg \texttt{coll}\ a\,b\,c; \langle b — a, a — c\rangle \approx_a \texttt{0hr}; \texttt{len}\ (a — b) \in \texttt{Finite};\ \texttt{len}\ (a — c) \in \texttt{Finite}|] \Longrightarrow \texttt{s\_delta}\ a\ b\ c \approx \texttt{0hr}$$

$$[|\neg \texttt{coll}\ a\,b\,c; \langle b — a, a — c\rangle \approx_a \texttt{0hr}; \texttt{len}\ (a — b) \in \texttt{Finite};\ \texttt{len}\ (a — c) \in \texttt{Finite}|] \Longrightarrow \texttt{len}\ (a — b) \approx \texttt{len}\ (a — c)$$

As can be seen from Figure 4.4, we should expect these theorems to hold when point b, moving towards point c, becomes infinitely close to it. These results become useful when dealing with more complex geometric constructions that arise when modelling motion. Consider, for example, the case in which a point c is moving along a circle towards a point a as in Figure 4.5.

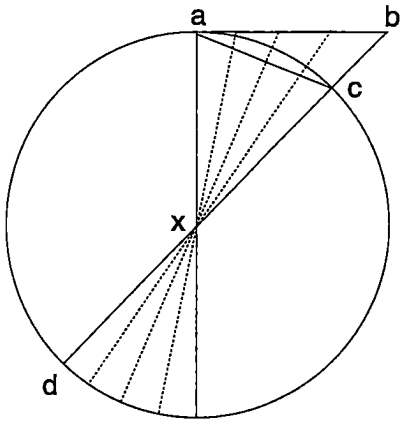


Figure 4.5: When point c is infinitely close to a, bc is infinitesimal

The *ultimate* situation, where we have infinitesimal quantities involved in the geometry, is of interest to us. This is when c is infinitely close to a. When this occurs, we prove that the angle between the chord ac and the tangent ab is infinitesimal, that is $\langle b --- a, a --- c\rangle \approx_a \texttt{0hr}$ and so it follows that length bc is also infinitesimal in this situation. From these, we can deduce that the length of segment bd is infinitely close to that of segment cd. It is obvious that this theorem does not hold in general. Moreover, this shows how various simpler, intuitive theorems combine to derive new ones when several geometric concepts are involved — with some of the latter changing (e.g. $\triangle bac$) while others remain fixed (the circle i.e. the path of motion). The proof of some of these theorems will be described in more detail in the next chapter since they occur in the *Principia*.

4.4 Verifying the Axioms of Geometry

The geometric theories of Isabelle are built around the basic rules given by Chou et al. for their area and full-angle methods. In their main papers about these methods [19, 20], the authors simply assert these rules as facts. They do provide, however, a definition for the tangent of the full-angle in terms of the signed area and the Pythagorean difference when discussing a complete method for GTP based on their notion of angles.

As already mentioned, many of the combined rules that are used as high level lemmas by Chou et al. have been verified in Isabelle. One of our goals, after a successful initial investigation of the *Principia*, was to justify the axioms used in Isabelle's geometry theory. It is our experience that in most automated geometry theorem provers using axiomatic methods, the basic facts or rules seem to be chosen on the basis of intuition. Moreover, Chou et al. do not usually justify the choice of their basic rules for their traditional methods; they use their common experience to decide on which extra lemmas to add to the prover. Such an approach is open to criticism from a formal point of view, but it should be mentioned the traditional concern in mechanical GTP has mainly been about the power and performance, rather than rigour.

Apart from using an interactive (hence slower) approach to GTP, the current work also differs from the traditional approach since it is done within the framework of Isabelle/HOL. This ensures that all the rules are rigorously proved and applied. To achieve a fully strict treatment of geometry in Isabelle/HOL involves verifying that the set of basic axioms proposed for the area method is consistent. It is worth noting that the reluctance to accept the axiomatic approach to geometry, despite the influence of Euclid's *Elements* for over two millenia, is not new. One can quote this anecdotal, yet important, observation from Russell's autobiography [73]:

> I had been told that Euclid proved things, and was much disappointed that he started with axioms. At first I refused to accept them unless my brother could offer me some reason for doing so, but he said: 'if you don't accept them we cannot go on', and as I wished to go on, I reluctantly admitted them *pro tem.* The doubt as to the premises of mathematics which I felt at that moment remained with me, and determined the course of my subsequent work.

One way to verify the axioms, is to show, in the spirit of Hilbert's *Grundlagen*, that there is a number system (say a field such as the hyperreals) associated with the geometry and reducing consistency of Isabelle's geometric theory to that of hyperreal arithmetic. This is achieved when working within the context of Isabelle/HOL, by developing a geometry theory according to the HOL-methodology i.e. strictly through definitions that capture the notions (points, lines, signed areas, etc.) that are being dealt with and then prove that the various properties follow. Of course, getting the **right** definitions is once more crucial since otherwise, we are likely to end up with the **wrong** properties.

To carry out this task, the hyperreal theories of Isabelle are extended with the notions of hyperreal vectors. In essence, this is an algebraic approach which develops geometric objects and relations between these objects in the Cartesian product $\mathbb{R}^{*n}$ of the field of hyperreals, where $n = 3$. We developed a theory of vectors in three dimensions, although we were only interested in plane properties since this has an algebra rich enough to capture the various notions we want to deal with. Thus, it also has more scope for future use. The hyperreals are chosen rather than the reals since we can then express infinitesimal geometric notions as well. The definitions that are used in the theories are given next. One theory introduces the algebraic operations on vectors while the other deals with the development of simple analytic geometry.

4.4.1 Euclidean Vector Space

In general, the simplest definition for a real vector in n dimensions is as an n-tuple of real numbers, $(r_1, \ldots, r_n)$. However, a more geometric definition can be provided that suits our purpose well.

DEFINITION 4.4.1. *Given two points $P = (x_1, y_1, z_1)$ and $Q = (x_2, y_2, z_2)$ in $\mathbb{R}^{*3}$, the vector $Q - P$ is called the directed line segment from P to Q. The components of the directed line segment are the terms in the 3-tuple $(x_2 - x_1, y_2 - y_1, z_2 - z_1)$.*

In this definition, we implicitly assume that the origin is given by the hyperreal coordinates (0hr, 0hr, 0hr) and hence that a particular point is specified by the

vector whose components correspond to its Cartesian coordinates. In Isabelle, we formulate a theory of three-dimensional vectors by first introducing vectors as a new type corresponding to a triple of hyperreal numbers:

$$\texttt{hypvec} \equiv \texttt{UNIV} :: (\texttt{hypreal} * (\texttt{hypreal} * \texttt{hypreal}))\ \texttt{set}$$

We can then define the various operations on the new type. For example, the *inner product* or *dot product* of two vectors P and Q is defined, using tuples as patterns in abstractions [68], by:[2]

$$\begin{aligned} P \cdot Q \equiv\ & (\lambda((x_1, y_1, z_1), (x_2, y_2, z_2)).\\ & x_1x_2 + y_1y_2 + z_1z_2)\\ & (\texttt{Rep\_hypvec}\ P, \texttt{Rep\_hypvec}\ Q) \end{aligned}$$

This definition is slightly more complicated than the usual textbook one since it uses an explicit λ-abstraction and the representation function. Just as in the other cases where we used the coercion functions (see Section 3.4.2, for example), we prove theorems that capture the more familiar definitions and which can then be fed to Isabelle's simplifier. So for the dot product, we have the expected:

$$\texttt{Abs\_hypvec}\ (x_1, y_1, z_1) \cdot \texttt{Abs\_hypvec}\ (x_2, y_2, z_2) = x_1x_2 + y_1y_2 + z_1z_2$$

Similarly, we also define other important operations, such as *cross product* and *scalar* multiplication ($\cdot_s$). For clarity, we give their definitions as the simplification theorems proved in Isabelle rather than the actual definitions in terms of Rep_hypvec and λ-abstractions. The Isabelle definitions unfortunately tend to be slightly cluttered and become somewhat hard to read, especially in the case of the cross product. So for cross and scalar products we prove the following rules:

$$\begin{aligned} &\texttt{Abs\_hypvec}\ (x_1, y_1, z_1) \times \texttt{Abs\_hypvec}\ (x_2, y_2, z_2) =\\ &\quad \texttt{Abs\_hypvec}\ (y_1z_2 - z_1y_2, z_1x_2 - x_1z_2, x_1y_2 - y_1x_2) \end{aligned}$$

$$a \cdot_s \texttt{Abs\_hypvec}\ (x, y, z) = \texttt{Abs\_hypvec}\ (ax, ay, az)$$

For any two vectors P and Q, the cross product can be viewed as defining the vector area of a parallelogram, with the vectors as two of the sides of the parallelogram and $P \times Q$ perpendicular to the plane containing P and Q (see Figure 4.6). With this nice geometric interpretation in mind, the next step involves proving various properties of the cross product that will enable us to capture the notion of signed area when verifying the axioms of the area method. The following theorem, which shows that the cross product is not commutative, is thus proved:

$$P \times Q = (-Q) \times P$$

Geometrically, this means a change in the direction of the vector while its magnitude remains unaffected. The negative of a vector P, for its part, is defined by negating its various components. In Isabelle:

$$-P \equiv (\lambda(x_1, x_2, x_3).\ \texttt{Abs\_hypvec}\ (-x_1, -x_2, -x_3))(\texttt{Rep\_hypvec}\ P)$$

In the next section, the definition of signed area of a triangle follows directly from the geometric interpretation and algebraic behaviour associated with the cross product.

[2] In what follows, the multiplication sign ($\cdot$) between hyperreal variables is omitted whenever no ambiguity is likely to result.

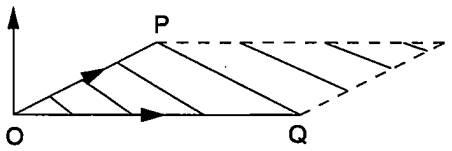


Figure 4.6: Geometric representation of cross product

Various other algebraic properties of the operations introduced so far are proved in Isabelle. A few straightforward ones that are useful to the development are as follows:

- $P \cdot Q = Q \cdot P$
- $P \cdot (Q + R) = P \cdot Q + P \cdot R$
- $(a \cdot_s P) \cdot (b \cdot_s Q) = ab \cdot_s (P \cdot Q)$

- $P \times (Q + R) = P \times Q + P \times R$
- $-(P \times Q) = (-P) \times Q$
- $-(P \times Q) = P \times (-Q)$
- $-(P \times Q) = Q \times P$
- $(a \cdot_s P) \times (b \cdot_s Q) = ab \cdot_s (P \times Q)$
- $P \times P = \texttt{0hv}$

- $P \cdot (P \times R) = \texttt{0hr}$
- $P \times (Q \times R) = (P \cdot R) \cdot_s Q - (P \cdot Q) \cdot_s R$
- $(a \cdot_s P + b \cdot_s Q) \times R = a \cdot_s (P \times R) + b \cdot_s (Q \times R)$

In these theorems, `0hv` denotes the *zero vector* and is defined in Isabelle as

$$\texttt{0hv} = \texttt{Abs\_hypvec}\ (\texttt{0hr}, \texttt{0hr}, \texttt{0hr})$$

Another important concept that has not yet been introduced is that of the *length* or *norm* of a vector. For a vector P, this is usually denoted by $|P|$ and defined by taking the square root of the dot product $P \cdot P$. In Isabelle,

$$\texttt{hypveclen}\ P = \texttt{hsqrt}\ (P \cdot P)$$

The square root operation over the hyperreals, denoted by `hsqrt` in Isabelle, is defined as the nonstandard extension of the square root operation (`sqrt`) over the reals. Details of these concepts will be given in Chapter 6. It is sufficient for the time being to regard taking the square root of a hyperreal as a well-defined operation with the usual properties.

After proving some further results of vector algebra, we developed a simple geometry theory based on the geometric interpretation of vectors and their operations. Our main motivation, as outlined earlier, was to verify the basic rules of the area method. However, having formalized hyperreal vectors rather than real vectors, we also performed some direct investigation of the infinitesimal geometric notions expressed using vectors. In the next sections, the definitions and results of the vector geometry development are outlined.

4.4.2 Using Vectors in Euclidean Geometry

Chou, Gao, and Zhang have also used vector calculations in automated geometry theorem proving [18]. They assert a set of basic rules about the operations that can be carried out on vectors. Theorems are then derived using these basic axioms of the theory. The algorithm used by Chou et al. is relatively simple: given a construction sequence for a geometric configuration, the points (i.e. vector variables) are eliminated one at a time from the vector expression standing for the conclusion, until only independent vector variables are left. The conclusion that results is then tested to see if it is identically zero.

In contrast to the above approach, we proceed by means of definitions only and having introduced hyperreal vectors and defined the operations on them, there is enough algebraic power for the theories to express geometric concepts: orthogonality and parallelism, signed areas, congruence of angles, infinitesimal geometric notions and much more. Moreover, we proceed mostly through simplification and substitution steps that are be applied to both the conclusion and premises of the current goal. That is, the proof steps in Isabelle are not limited to point elimination only.

Using the theory, we can attain our aim and derive the basic axioms of Chou et al. about signed areas in our theory. We first introduce as basic geometric objects the notions of points and lines by defining the following types in Isabelle:

$$\texttt{pt} \equiv \texttt{UNIV} :: \texttt{hypvec set}$$
$$\texttt{line} \equiv \texttt{UNIV} :: (\texttt{pt} * \texttt{pt})\ \texttt{set}$$

From these definitions, a point is therefore specified by a position vector and a (directed) line given by a pair of vectors representing its end-points. These definitions give the theory a separate, nicer geometric interpretation in which geometric objects (points and lines) are dealt with rather than vectors of hyperreal numbers. The abstraction and representation functions of Isabelle enable us to deal with the underlying vector theory to prove basic properties of parallelism, perpendicularity, collinearity etc. Once this is done, we can hope to work at a higher abstract level which deals with geometric relations and interact rather minimally with the underlying vector constructions. This is similar in spirit with our construction of numbers, say the reals by Dedekind cuts, where initially for each operation we have to prove cut properties but as more theorems are proved, we deal less and less with the actual cuts and more with the algebra of the reals.

However, in the subsequent exposition we shall regard position vectors and points as being interchangeable when giving the definitions and describing properties proved. This abuse of notation is simply to make the definitions more readable on paper since it avoids the use of the coercion functions. We will show the definitions or theorems as actually formulated if the need ever arises. Therefore, for each geometric condition, we have the corresponding vector definition:

1) That C is on line AB:

$$\texttt{incident}\ C\ (A --- B) \equiv (C - A) \times (B - A) = \texttt{0hv}$$

2) That AB is parallel to CD:

$$A --- B \parallel C --- D \equiv (B - A) \times (D - C) = \texttt{0hv}$$

3) That AB is perpendicular to CD:

$$A -\!\!- B \perp C -\!\!- D \equiv (B - A) \cdot (D - C) = \mathtt{0hr}$$

4) The length of a line AB:

$$\mathtt{len}\ (A -\!\!- B) \equiv \mathtt{hypveclen}\ (B - A)$$

5) The signed vector area of triangle ABC:

$$\mathtt{s\_delta}\ A\ B\ C \equiv 1/2 \cdot_s (A - B) \times (C - B)$$

6) That the $\triangle ABC$ and $\triangle XYZ$ are similar, with vertices in the same direction:

$$\begin{aligned}\mathtt{SIM}\ A\,B\,C\,X\,Y\,Z \equiv\ & \neg\mathtt{coll}\ A\,B\,C \wedge \neg\mathtt{coll}\ X\,Y\,Z \wedge \\ & (\mathtt{hypveclen}\ (C - A)/\mathtt{hypveclen}\ (B - A) = \\ & \mathtt{hypveclen}\ (Z - X)/\mathtt{hypveclen}\ (Y - X))\end{aligned}$$

7) That the angles $\langle A -\!\!- B, B -\!\!- C\rangle$ and $\langle X -\!\!- Y, Y -\!\!- Z\rangle$ are congruent. For this, we define the cosine and sine functions:

$$\mathtt{Cos}\ \langle A -\!\!- B, B -\!\!- C\rangle \equiv \mathtt{unitvec}\ (A - B) \cdot \mathtt{unitvec}\ (C - B)$$

where `unitvec` is the unit vector which is defined in the next section, and

$$\begin{aligned}\mathtt{Sin}\ \langle A\!-\!\!- B, B\!-\!\!- C\rangle \equiv\ & \mathtt{hypveclen}\ ((A - B) \times (C - B)) \cdot \\ & \mathtt{hrinv}\ (\mathtt{hypveclen}\ (A - B) \cdot \mathtt{hypveclen}\ (C - B))\end{aligned}$$

Then, since we want directed angles and distinguish between an angle and its supplement, we have the following definition for congruence of angles:

$$\begin{aligned}&\langle A -\!\!- B, B -\!\!- C\rangle =_a \langle X -\!\!- Y, Y -\!\!- Z\rangle \equiv \\ &(\mathtt{Cos}\ \langle A -\!\!- B, B -\!\!- C\rangle = \mathtt{Cos}\ \langle X -\!\!- Y, Y -\!\!- Z\rangle \wedge \\ &\ \mathtt{Sin}\ \langle A -\!\!- B, B -\!\!- C\rangle = \mathtt{Sin}\ \langle X -\!\!- Y, Y -\!\!- Z\rangle)\end{aligned}$$

With these definitions set up, we prove that the basic properties of signed areas actually hold and justify the statements of geometric relations that were made in terms of them. The rules about the sign of the area depending on the ordering of the vertices of the triangle (Property 4 of Section 2.4.1) are all proved without any problems since our definition makes them direct consequences of the algebraic properties of the cross product. Consider, for example:

$$\begin{aligned}-\mathtt{s\_delta}\ c\ b\ a &= -1/2 \cdot_s (c - b) \times (a - b) \\ &= -1/2 \cdot_s (-(a - b)) \times (c - b) \\ &= --1/2 \cdot_s (a - b) \times (c - b) \\ &= \mathtt{s\_delta}\ a\ b\ c\end{aligned}$$

This and similar rules are proved with the help of Isabelle's automatic tactic and added to the simplifier. The definition of parallelism in terms of signed areas is

also easily verified and the following theorem defines incidence (or collinearity) in terms of signed area:

$$\texttt{incident}\ a\ (b --- c) \Longleftrightarrow (\texttt{s\_delta}\ a\ b\ c = \texttt{0hv})$$

We also extend the definition of incidence to that of a set of points incident on a line, thereby enabling us to verify axiom 7 of Section 2.4.1. We can deal with the ratios of oriented lines by proving theorems such as these:

- $A --- B \parallel C --- D\ (C \neq D)$:

$$\frac{\texttt{len}\ (A --- B)}{\texttt{len}\ (C --- D)} = \frac{(B - A) \cdot (D - C)}{(D - C) \cdot (D - C)}$$

- if R is the foot of the perpendicular from point A to line PQ $(P \neq Q)$:

$$\frac{\texttt{len}\ (P --- R)}{\texttt{len}\ (P --- Q)} = \frac{(A - P) \cdot (Q - P)}{\texttt{len}\ (P --- Q)^2}$$

- if two non-parallel lines intersect at a point R:

$$\begin{aligned}&\texttt{len}\ (P --- R) \cdot_s (Q - P) \times (V - U) = \\ &\texttt{len}\ (P --- Q) \cdot_s (U - P) \times (V - U)\end{aligned}$$

Some of the results above are high level lemmas stated by Chou et al. as being used in their automated GTP method based on vectors [18]. We verify all of them in Isabelle and store them as lemmas that become valuable when proving complicated geometry theorems.

4.4.3 Using Vectors in Infinitesimal Geometry

We start by extending some of the definitions used for the hyperreals to their vectors.

DEFINITION 4.4.2. *A hyperreal vector P is said to be infinitesimal, finite, or infinite if its length $|P|$ is infinitesimal, finite or infinite respectively. Moreover, P is infinitely close to Q if and only if $B - A$ is infinitesimal.*

With this definition formalized in Isabelle, the following useful and equivalent theorems about infinitely close vectors are proved:

$$\begin{aligned}P \approx_v Q \Longleftrightarrow\ &(\lambda((x_1, y_1, z_1), (x_2, y_2, z_2)).\\ &\quad x_1 \approx x_2 \wedge y_1 \approx y_2 \wedge z_1 \approx z_2)\\ &\quad (\texttt{Rep\_hypvec}\ P, \texttt{Rep\_hypvec}\ Q)\end{aligned}$$

$$\begin{aligned}&\texttt{Abs\_hypvec}\ (x_1, y_1, z_1) \approx_v \texttt{Abs\_hypvec}\ (x_2, y_2, z_2)\\ &\quad \Longleftrightarrow x_1 \approx x_2 \wedge y_1 \approx y_2 \wedge z_1 \approx z_2\end{aligned}$$

In other words, two hyperreal vectors are infinitely close if and only if their components in corresponding positions are infinitely close to one another. The second theorem is the most useful one as it can be added to the simplifier. It is also simpler to work with than the actual definition since most of the infinitely

close properties of vectors are then inherited directly from those of the hyperreals and can therefore be proved automatically using Isabelle's `auto_tac`. The two theorems just given are formalized without much difficulty though they involve dealing with properties of nonstandard extensions of functions. We also prove the following theorems:

1) P is infinitesimal if and only if all its components are infinitesimal.

2) P is finite if and only if all its components are finite.

3) P is infinite if and only if at least one of its components is infinite.

Just as the concept of two lines being parallel was introduced, using hyperreal vectors the weaker notion of two lines being *almost parallel* is defined:

$$A -\!\!- B \parallel_a C -\!\!- D \equiv (B - A) \times (D - C) \approx \texttt{0hv}$$

The notion of the unit vector is also introduced and defined for a vector P by $1/|P| \cdot_s P$. In Isabelle,

$$\texttt{unitvec}\ P = \texttt{hrinv}\ (\texttt{hypveclen}\ P) \cdot_s P$$

Any vector P can be classified according to the behaviour of its length $|P|$ and unit vector. Moreover, the following geometric results follow (with $A \neq B$ and $C \neq D$):

$$\begin{aligned} A -\!\!- B \parallel_a C -\!\!- D \iff\ & \texttt{unitvec}\ (B - A) \approx_v \texttt{unitvec}\ (D - C) \vee \\ & \texttt{unitvec}\ (B - A) \approx_v -\texttt{unitvec}\ (D - C) \end{aligned}$$

Interestingly, other "almost relations" seem possible in principle and might be worth investigating: for example, an ellipse with infinitely close foci is *almost* a circle.

More relevant to this work, various useful infinitesimal geometric theorems are proved now using infinitesimal vector geometry. These include the ones shown in Section 4.3.2, and others such as this one:

$$\begin{aligned} & [|\texttt{incident}\ a\ (b -\!\!- c); \texttt{s\_delta}\ p\ c\ b \approx_v \texttt{0hv}|] \\ & \Longrightarrow \texttt{s\_delta}\ p\ a\ c \approx_v \texttt{s\_delta}\ p\ a\ b \end{aligned}$$

The proofs tend to require theorems about the infinitely close relation as proved especially for hyperreal vectors. For example, to prove the theorem above, one needs a cancellation law:

$$[|a \in \texttt{SReal}; a \neq \texttt{0hr}|] \Longrightarrow (a \cdot_s w \approx_v a \cdot_s z) = (w \approx_v z)$$

as well as various others involving associativity and commutativity of vector addition to perform AC-rewriting.

4.5 Concluding Remarks

In this chapter, we have proposed the notion of an infinitesimal geometry. We have related aspects of non-Archimedean geometry to our geometry and introduced new concepts based on the introduction of hyperreals and their relations.

Various theorems have been proved that have no direct counterparts in Euclidean geometry since the latter only deals with real numbers.

Vector algebra offers an attractive approach to mechanical geometry theorem proving. We have already mentioned the active research going on using the related field of Clifford algebra, which is generally regarded as being more expressive. In our case, since we are doing interactive rather than automatic theorem proving, vectors provide a simple and adequate approach to analytic geometry. Also, as was shown by Dieudonné, inner (dot) and cross products of vectors are sufficient to develop elementary geometry [28].

We have shown that hyperreal vectors obey the usual algebraic rules for vectors since they form an inner product space over the field $\mathbb{R}^*$. By using the extended vectors instead of real vectors, it is possible to describe, in addition to ordinary geometric concepts, the novel notions of infinitesimal geometry presented at the beginning of this chapter.

The analytic geometry development was carried out to provide a definitional foundation in which to verify the basic rules of the geometric methods as postulated by Chou et al. These could be seen to be intuitively correct though no formal proofs had been provided. In so doing, we have ensured that the geometric theory respects as much as possible the HOL methodology.

Chapter 5

Mechanizing Newton's *Principia*

Results obtained through the combination of geometric and NSA techniques can now be presented. The methods have been used to investigate the infinitesimal geometry. Some of the results confirm what one intuitively might expect to hold when elements are allowed to be infinitesimal. In what follows, the formalization of various notions found in Newton's prose is examined. Important theorems about motion along arcs, circular paths and elliptical orbits are mechanized.

5.1 Formalizing Newton's Properties

We first describe how we formalize some of the expressions used by Newton in his various proofs. The technical meaning of some of the *ultimate* properties are represented in a consistent way as follows:

Ultimately vanishing or *evanescent* quantity
: We represent this as being **infinitely close** to zero i.e. x is vanishing means $x \in \texttt{Infinitesimal}$ (or equivalently $x \approx 0$). If x is an angle then we can have $x \approx_a 0$. No claims are made that ultimately vanishing quantities are eventually *equal* to zero. This is also used to model Newton's *nascent* quantities.

Finite but not vanishing
: $x \in \texttt{Finite} - \texttt{Infinitesimal}$. Sometimes this property needs to be stated explicitly, for example when taking the ratios of two quantities.

Ultimately equal quantities
: x is ultimately equal to y means $x \approx y$. Here again, we use the infinitely close relation because the quantities, though their difference may become infinitely small, do not necessarily ever become equal. The infinitely close relation gives us the freedom of making the quantities arbitrarily close without ever claiming that they are equal.

Ultimately increased without limit
: Depending on whether the quantity is continuous or discrete, we have $x \in \texttt{Infinite}$ or $x \in \texttt{HNatInfinite}$ respectively.

Ultimately similar triangles
`USIM a b c a' b' c'`

Ultimately congruent triangles
`UCONG a b c a' b' c'`

5.2 Mechanized Propositions and Lemmas

Some of the results obtained with the help of notions discussed in Chapter 4 can now be presented. We analyse several significant properties mechanized using our combination of NSA and geometry.

5.2.1 Newton's First Lemma

Newton's **Lemmas** are derivations that are set up at the beginning of the *Principia*, and to which Newton appeals afterwards when needed in his proofs. We examine the proofs of various of these theorems. In some cases, we compare our proofs with Newton's. The relatively long enunciation of these lemmas shows some of the difficulties encountered when dealing with Newton's mathematical prose. Newton's first Lemma enables us to introduce our infinitely close relation in the geometry.

Lemma 1. *Quantities, as well as ratios of quantities, which in any finite time you please constantly tend towards equality, and before the end of that time approach nearer to each other than by any given difference you please, become ultimately equal.*

This lemma is formalized, with x and y as Newton's quantities and D as the difference, by the following theorem of Isabelle:

$$\forall D \in \texttt{SReal}.\ |x - y| < D \Longrightarrow x \approx y$$

The proof is trivial since it follows directly from our definition of the *infinitely close* relation and of infinitesimals. In fact, by having as a condition that the difference D is greater than zero, we have an equivalence relation. Newton's proof is short and proceeds by *reductio ad absurdum*:

> If you deny it, let them become ultimately unequal, and let their ultimate difference be D. Therefore, they cannot approach nearer to each other than by the given difference D, contrary to the hypothesis.

One point worth noting is that time is absent from our formalization, even though Newton's quantities can be viewed as depending continuously on it. This is because we are only interested in a specific time: when the situation is ultimate. This is the time when the new and useful properties emerge since quantities are infinitesimal. As can be seen, Newton's proof itself only considers the ultimate quantities, which provides support for our approach.

We also note that if the quantities themselves, and not just the difference between them, are vanishing then greater care is necessary. Newton explains in the scholium to the lemmas

> ... by the ultimate ratio of evanescent quantities is to be understood the ratio of quantities not before they vanish, nor afterwards, but with which they vanish.

We can allow ratios of evanescent quantities by insisting that such evanescent quantities can be infinitesimal but not necessarily zero i.e. $x \approx 0$ but not necessarily $x \neq 0$. Thus, the existence of the infinitely close relation enables the denominator of the ratio to be smaller than any other real quantity and still be a well-defined quantity. However, care needs to be exercised when manipulating such ratios and the rules of NSA make explicit the conditions that need to be satisfied before one can multiply and divide them, for example. We shall describe later how rigorous concepts from NSA enable us to deal formally with ratios of vanishing or infinitesimal quantities. Moreover, we shall also examine the geometric tools developed by Newton to deal with his ratios of evanescent quantities soundly.

5.2.2 Motion along an Arc of Finite Curvature

Lemma 6. *If any arc ACB, given in position, is subtended by its chord AB, and in any point A, in the middle of the continued curvature, is touched by a right line AD, produced both ways; then if the points A and B approach one another and meet, I say, the angle BAD, contained between the chord and the tangent, will be diminished in infinitum, and ultimately will vanish.*

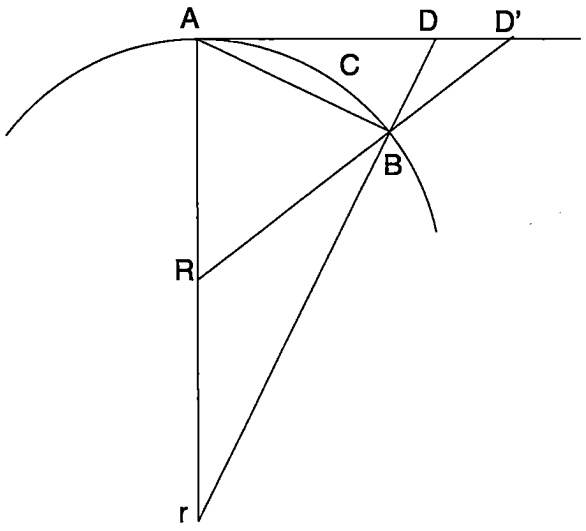


Figure 5.1: Figure based on Newton's diagram for Lemma 6

In Figure 5.1, let R be the centre of curvature and let line AD touch the arc at A. We follow Newton's assumption, given in the scholium to the lemmas, that the curvature is finite. We assume that our (vanishing) circular arc has the same curvature as the general curve at point A and hence the same tangent. The circle of contact at A is extended to its antipodal point at r. The latter is not used but shown for completeness. We prove that $\langle B --- A, A --- D \rangle \approx_a 0$ when B is infinitely close to A.

We give below an overview of our reasoning and theorems proved to reach the conclusion. We show that the angle subtended by the arc becomes infinitesimal as B approaches A and that the angle between the chord and the tangent is always half that angle:

$$[|\texttt{arc\_len}\ R\ A\ B \approx \texttt{0hr}; \texttt{len}\ (A --- R) \in \texttt{Finite} - \texttt{Infinitesimal}|]$$
$$\Longrightarrow \langle B --- R, R --- A\rangle \approx_a \texttt{0hr}$$

$$[|\texttt{is\_c\_tangent}\ (A --- D)\ R\ \mathit{Circle}; B \in \mathit{Circle}|]$$
$$\Longrightarrow \langle B --- A, A --- D\rangle =_a \langle B --- R, R --- A\rangle/2$$

We use the theorem from NSA that `Infinitesimal` is an ideal in `Finite` and the results above to prove that the angle between the chord and the tangent also becomes infinitesimal.

$$[|\langle B --- R, R --- A\rangle \in \texttt{Infinitesimal}; 1/2 \in \texttt{Finite}|]$$
$$\Longrightarrow \langle B --- R, R --- A\rangle \cdot 1/2 \in \texttt{Infinitesimal}$$

$$[|\texttt{is\_c\_tangent}\ (A --- D)\ R\ \mathit{Circle}; B \in \mathit{Circle}; \texttt{arc\_len}\ R\ A\ B \approx \texttt{0hr};$$
$$\texttt{len}\ (A --- R) \in \texttt{Finite} - \texttt{Infinitesimal}|] \Longrightarrow \langle B --- A, A --- D\rangle \approx_a \texttt{0hr}$$

We next use our concept of *ultimately similar* triangles to prove part of Newton's **Lemma 8**. According to this Lemma, the ultimate form of evanescent $\triangle$ABR and $\triangle$AD'R is that of similitude in Figure 5.1.

It is clear from the diagram that $\triangle$ABR and $\triangle$AD'R are not similar in ordinary Euclidean Geometry. Infinitesimals notions reveal that we are tending towards similarity of these triangles when the point B is about to meet point A. This property cannot be deduced from just the *static* diagram. Understanding the dynamics of Newton's geometry requires the use of imagination to incorporate motion and see what is happening to the relations between various parts of the diagram as points are moving. This task is not always trivial. The relation USIM $A\ B\ R\ A\ D'\ R$ can be illustrated by considering the relation between the various parts of the diagram as point B moves towards point A (Figure 5.2).

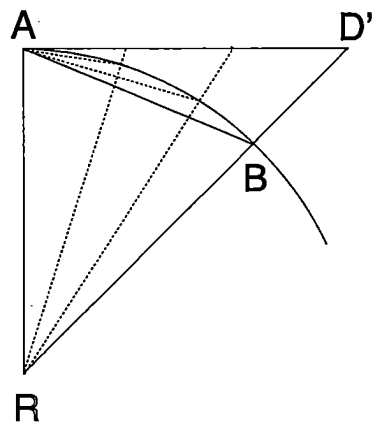


Figure 5.2: Ultimately similar triangles

We have already proved in **Lemma 6** that the angle between the chord and the tangent is infinitesimal i.e. $\langle B -\!\!- A, A -\!\!- D' \rangle \approx_a 0$. From this result, we can deduce that $\langle R -\!\!- A, A -\!\!- D' \rangle$ and $\langle R -\!\!- A, A -\!\!- B \rangle$ are infinitely close:

$$\langle B -\!\!- A, A -\!\!- D' \rangle \approx_a \texttt{0hr} \Longrightarrow \langle R -\!\!- A, A -\!\!- D' \rangle \approx_a \langle R -\!\!- A, A -\!\!- B \rangle$$

Finally, since $\triangle$ABR and $\triangle$AD'R have two corresponding angles that are infinitely close (they have one common angle in fact), we can show that they are ultimately similar (and even ultimately congruent):

$$\begin{array}{l} \langle B -\!\!- R, R -\!\!- A \rangle =_a \langle B -\!\!- R, R -\!\!- A \rangle \\ \Longrightarrow \langle B -\!\!- R, R -\!\!- A \rangle \approx_a \langle B -\!\!- R, R -\!\!- A \rangle \end{array}$$

$$\begin{array}{l} [|\langle B -\!\!- R, R -\!\!- A \rangle \approx_a \langle B -\!\!- R, R -\!\!- A \rangle; \\ \quad \langle R -\!\!- A, A -\!\!- D' \rangle \approx_a \langle R -\!\!- A; A -\!\!- B \rangle |] \Longrightarrow \texttt{USIM}\ A\,B\,R\,A\,D'\,R \end{array}$$

In the various proofs just described, the use of a circular arc for the arc of finite curvature is justified because it is possible to construct a circle at the point A that represents the best approximation to the curvature there (see section 5.2.4 for more details on circular approximations). We can also list a few more properties that are proved about motion along an arc (see Figure 5.1):

- $\triangle BDA$ and $\triangle ABr$ are **similar** and hence $\texttt{len}\ (A -\!\!- B)^2 = \texttt{len}\ (A -\!\!- r) \cdot \texttt{len}\ (D -\!\!- B)$. The latter result is stated and used but not proved by Newton in **Lemma 11**.
- $\texttt{len}\ (A -\!\!- B)$, $\texttt{arc\_len}\ R\ A\ B$, and $\texttt{len}\ (A -\!\!- D')$ are infinitely close and, in fact, their **ultimate ratio** is infinitely close to 1. This is **Lemma 7**.

We now further apply our infinitesimal techniques to various kinds of motions that are studied in the *Principia*. Infinitesimals and Newton's Lemmas are required to deal with the ultimate situation and enable various kinds of approximations where a particular figure can be replaced by another one in whole or in parts.

5.2.3 Circular Motion

Motion along a circle is the simplest type of conic motion and was originally thought to be the type of orbits in which planets moved. We examine next the various procedures used to investigate circular motion and to derive the physical laws governing it.

The Polygonal Approximation and Kepler's Law of Equal Areas

In this technique, the circle or circular path is approximated by a circumscribed polygon or polygonal path of n sides. The number of sides, n, is *ultimately* increased without limit, that is $n \in \texttt{HNatInfinite}$ in the ultimate situation. As the number of sides increases, the polygonal path approaches the circular one and, when the number of sides is infinite, Newton no longer distinguishes between the polygon and the actual circle. This approach is, of course, reminiscent of the one used to prove the relation between the area and the circumference of the circle that we formalized in Section 4.3. Newton's approximation is justified

by assuming that an impulsive force acts intermittently to deviate the motion of the body from a rectilinear path as shown in Figure 5.3. The impulsive force is known as the centripetal force which acts towards the centre of the circle.

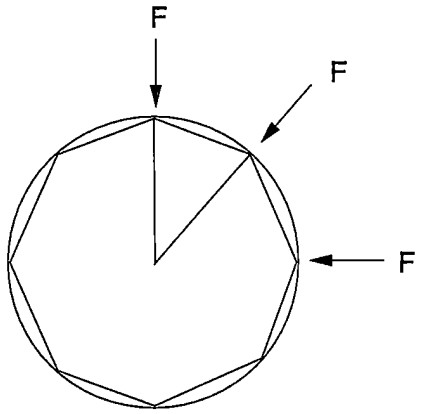


Figure 5.3: A circular path approximated by a polygon (octagon) with an impulsive force F acting at each intersection point

Using the polygonal approximation, Newton sets up the first important theorem of the *Principia*, namely Kepler's second law[1] or Kepler's Law of Equal Areas. This was published in 1609 and was often regarded until Newton's *Principia* as the least important of Kepler's Laws. It is established by Newton as the first mathematical Proposition of the *Principia*.

In Newton's diagram (Figure 5.4), the polygons $ABCDEF$ are used to approximate the continuous motion of a planet in its orbit. The motion between any two points such as A and B of the path is not influenced by any force though there are impulsive forces, all directed towards the fixed centre S, that act at $A, B, C, \ldots$. Newton proved that if the time interval between successive impulses is fixed then all the triangular areas $SAB, SBC, \ldots$, are equal, that is equal areas are described in equal times. The demonstration of this law makes no assumption about how this force varies with distance from the centre of force S; its only restriction is that it be directed toward S. Newton reduces the discontinuous motion along the straight edges $AB, BC, \ldots$, to continuous motion along a smooth orbital path by using an infinitesimal process that lets the size of the triangles become infinitely small.

We follow Newton's argument and prove that the area of SAB is equal to that of SBC using our geometric tools. We quote from the exposition of Proposition 1 in the *Principia*:

> Let time be divided into equal parts, and in the first part of the time let the body, by its inherent force, describe the straight line AB. In the second part of the time, the same body, if nothing were to impede it, would pass on by means of a straight line to c (by Law 1), describing the line Bc equal to AB, with the result that, radii AS, BS, cS being drawn to the centre, the areas ASB, BSc would come out equal.

[1] This was actually Kepler's first observation originally.

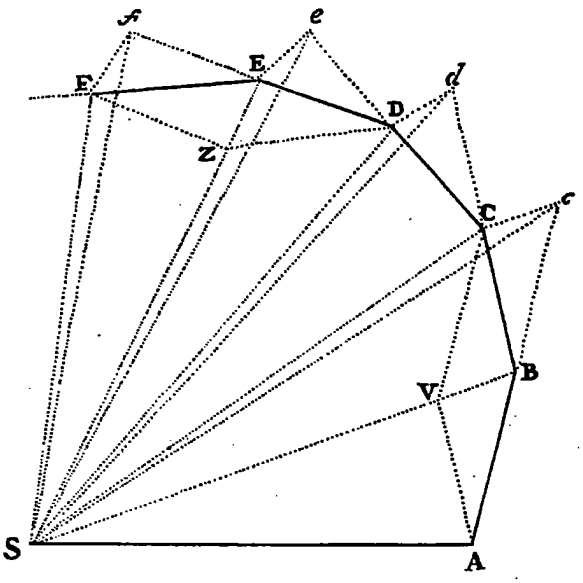


Figure 5.4: Original diagram from the *Principia* showing a body moving under the influence of a series of impulsive centripetal forces

We first observe that the area of SAB equals that of SBc because the triangles have equal bases (since the times are equal and no force has acted to change the velocity) and the same height:

$$[|\texttt{coll}\ A\ B\ c; \texttt{len}\ (A --- B) = \texttt{len}\ (B --- c)|] \\ \Longrightarrow \texttt{s\_delta}\ S\ A\ B = \texttt{s\_delta}\ S\ B\ c$$

The impulsive centripetal force at B makes the body depart from motion in a straight line and Newton makes the following construction (using the Parallelogram Law of Forces):

> Let cC be drawn parallel to BS, meeting BC at C; and, the second part of the time being completed, the body (by Corollary I of the laws) will be located at C, in the same plane as the triangle ASB ... Connect SC, and because of the parallels SB, Cc, triangle SBC will be equal [in area] to triangle SBc, and therefore to triangle SAB.

This leads to the following lemma, which is also easily proved in Isabelle since it follows from the definition of parallel lines:

$$[|S --- B \parallel c --- C|] \Longrightarrow \texttt{s\_delta}\ S\ B\ c = \texttt{s\_delta}\ S\ B\ C$$

The proof that the areas are equal follows. In fact, this first part of Kepler's Law of Equal Areas is proved automatically in one step by Isabelle thanks to the presence of powerful proof tactics.

The next step in Newton's proof is to decrease the breadth of the triangles to be infinitesimally small. By formalizing Newton's Lemma 3 and its corollaries, we can substitute the straight edge by a curved line:

$$\langle A --- S, S --- B\rangle \approx \texttt{0hr} \Longrightarrow \texttt{len}\ (A --- B) \approx \texttt{arc\_len}\ S\ A\ B$$

And furthermore using the same lemma, the area of the infinitesimal triangle SAB is infinitely close to the area of the arc and can be substituted:

$$\langle A -\!\!- S, S -\!\!- B \rangle \approx \texttt{0hr} \Longrightarrow \texttt{s\_delta}\ S\ A\ B \approx \texttt{arc\_area}\ S\ A\ B$$

As the triangles become infinitesimal, the perimeter of the path becomes infinitely close to a curvilinear one and the force can be viewed as acting continuously since the times between the impulses are infinitesimal. We note here the geometrical representation of time since making the triangles infinitesimal effectively makes the time intervals also infinitely close to zero. The result that the area described is proportional to time still holds for the evanescent triangles and hence also holds for the infinitely close curvilinear areas.

The Parabolic Approximation

The parabolic approximation to circular motion also introduces infinitesimals in the analysis but in a different way. This technique, developed after the polygonal approximation, is usually used right at the beginning or right at the end of motion of a body. It derives its name from Galileo's demonstration that the combination of uniform rectilinear motion and uniform accelerated rectilinear motion (at right angle to one another) gives rise to a parabolic path [12]. Thus, the portion of the circle infinitely close to some point is approximated by the corresponding portion of a parabola since the centripetal force acting perpendicularly disturbs the rectilinear motion of the body along the tangent. This approximation enables Newton to use Galileo's results about motion to relate the force F with the deviation denoted by BC in Figure 5.5.

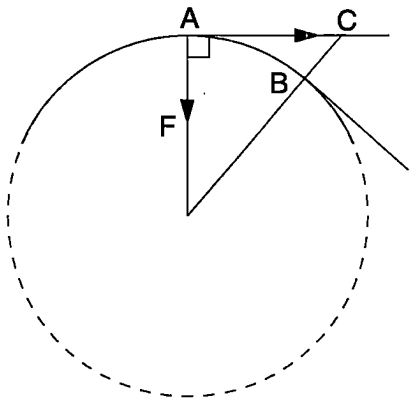


Figure 5.5: The parabolic approximation

5.2.4 Elliptical Motion

The analysis of elliptical motion is of utmost importance since planets were observed to move in elliptical orbits around the sun. Newton's analysis of the force required for elliptical motion relies on the circle of curvature.

The Circular Approximation using Osculating Circles

A circle is used to approximate the ellipse at a point (and in its infinitesimal neighbourhood). For example, in Figure 5.6, we have a circle at A and one at B approximating the curvature at these two points.

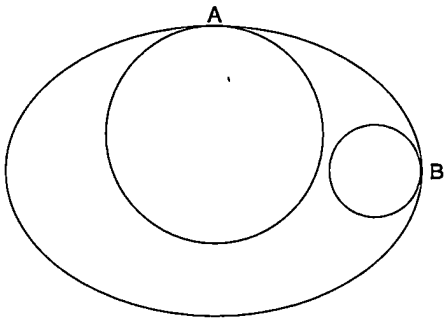


Figure 5.6: Osculating circles

This circle, sometimes known as the *osculating circle*[2], has the same first and second derivative as the curve at the given point A. Thus, the osculating circle has the same curvature and tangent at A as the general curve and therefore both have the same infinitesimal arc at A. Brackenridge gives more details on the technique [11, 12] and notes that the circle of curvature is small where the ellipse curves more rapidly, for example at B, and larger where it curves more slowly, such as at point A. This approximation also enables us to use a circular arc to derive results about any arcs with finite curvatures as we already observed. We can then make use of the definitions for areas of sectors, length of arcs etc., for example (c.f. Section 5.2.3).

5.2.5 Geometric Representation for the Force

With the help of Kepler's Law of Equal Areas and the circular approximation, we can now derive a completely **geometric** representation for the force acting on the orbiting body. Consider Figure 5.7, in which a point P is moving along an arc of finite curvature under the influence of a centripetal force acting towards S. Let Q be a point infinitely close to P, that is the length of the arc from P to Q is infinitesimal. QR, parallel to SP, represents the displacement from the rectilinear motion (along the tangent) due to the force acting on P. Line segment QT is the perpendicular dropped to SP. From Newton's **Lemma 10, Corollary 3** (one of Galileo's results about parabolic motion in *Two New Sciences* [35]), we have that the displacement "**in the very beginning of motion**" is proportional to the force and to the inverse square of the time. Hence, formalized in our framework, we have (for some real proportionality constant k_1) that

$$\text{force} \approx k_1 \cdot \frac{\texttt{len(Q --- R)}}{\texttt{Time}^2} \tag{5.1}$$

Since the distance between P and Q is infinitesimal, the angle $\langle P \text{---} S, S \text{---} Q \rangle$ is infinitely small, and hence the area of the sector SPQ ($\texttt{S}_{\texttt{arc}}\,\texttt{S P Q}$) is infinitely

[2]From the Latin *osculare* meaning to kiss– the term was first used by Leibniz.

close to that of the triangle SPQ:

$$\begin{aligned}\langle \mathtt{P} -\!\!- \mathtt{S}, \mathtt{S} -\!\!- \mathtt{Q}\rangle \approx_{\mathtt{a}} 0 &\implies \mathtt{S_{arc}\,S\,P\,Q} \approx \mathtt{S_{\triangle}\,S\,P\,Q} \\ &\implies \mathtt{S_{arc}\,S\,P\,Q} \approx 1/2 \cdot \mathtt{len}(\mathtt{Q} -\!\!- \mathtt{T}) \cdot \mathtt{len}(\mathtt{S} -\!\!- \mathtt{P}) \quad (5.2)\end{aligned}$$

From Kepler's Law of Equal Areas (Section 5.2.3), we have that the area of sector SPQ is proportional to the time taken (`Time`) to describe it. Thus, we can replace `Time` by $\mathtt{S_{arc}SPQ}$ in (5.1) and, hence, using (5.1) and (5.2), we have the following geometric representation for the force (for some new proportionality constant k):

$$\text{force} \approx k \cdot \frac{\mathtt{len}(\mathtt{Q} -\!\!- \mathtt{R})}{\mathtt{len}(\mathtt{Q} -\!\!- \mathtt{T})^2 \cdot \mathtt{len}(\mathtt{S} -\!\!- \mathtt{P})^2} \qquad (5.3)$$

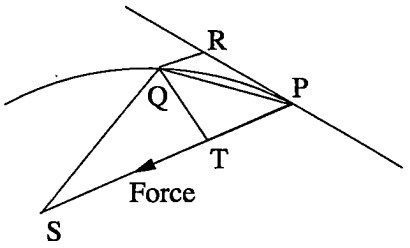


Figure 5.7: Representing the centripetal force geometrically

The geometric expression given above is a general theorem: it holds for any body moving under the influence of a centripetal force. In Section 5.4, the result is applied to the study of elliptical motion; this results in a new, more specific geometric expression for the force.

5.3 Ratios of Infinitesimals

We know that `Infinitesimal` is not an ideal in $\mathbb{R}$. It follows that the ratio of two infinitesimals (equivalently, the product of an infinitesimal and an infinite number) is not necessarily an infinitesimal. In other words, we cannot be sure whether the ratio of two infinitesimals is infinitesimal, finite or infinite. For any given positive infinitesimal ϵ, we have $\epsilon \approx \epsilon^2$, but see that

$$\begin{aligned}\epsilon^2 \cdot 1/\epsilon &\in \mathtt{Infinitesimal} \\ \epsilon \cdot 1/\epsilon &\in \mathtt{Finite} \\ \epsilon \cdot 1/\epsilon^2 &\in \mathtt{Infinite}\end{aligned}$$

Therefore, whenever an infinitesimal quantity is divided by another in one of Newton's ultimate situation, the type of number obtained for the ratio will need to be known to be of any use in subsequent deductions. Failure to establish the nature of a ratio of vanishing quantities within our framework might prevent a theorem from being applicable (since one or more of its premises cannot be discharged). This is due to the rigour with which infinitesimals, and relations based on them, are treated in NSA to prevent unsound steps.

Finite Geometric Witnesses

How does Newton deal with ratios of vanishing quantities? Whenever he is manipulating the ratio of infinitely small quantities, he usually makes sure that this can be expressed in terms of some finite (geometric) quantity in the proof. Thus, the ratio of infinitesimals is shown to be infinitely close or even equal to some finite quantity. We illustrate this important aspect by means of a simple example.

Let $\triangle abc$ be an "infinitesimal" triangle, i.e. its sides are all of infinitesimal lengths. Based on this information only, it is not possible to determine the nature of the ratio of any two of its sides since it could be finite, infinitesimal or infinite. If we have a second triangle $\triangle a'b'c'$ with sides of finite but not vanishing lengths, say all the sides have real lengths, and given furthermore that $\triangle abc$ and $\triangle a'b'c'$ are similar (or even ultimately similar), then it becomes possible to deduce the following (see Figure 5.8):

$$\begin{aligned}\texttt{SIM}\ a\ b\ c\ a'\ b'\ c' &\Longrightarrow \texttt{len}\ (a --- b)/\texttt{len}\ (a' --- b') = \texttt{len}\ (a --- c)/\texttt{len}\ (a' --- c') \\ &\Longrightarrow \texttt{len}\ (a --- b)/\texttt{len}\ (a --- c) = \texttt{len}\ (a' --- b')/\texttt{len}\ (a' --- c') \\ &\Longrightarrow \texttt{len}\ (a --- b)/\texttt{len}\ (a --- c) \in \texttt{SReal}\end{aligned}$$

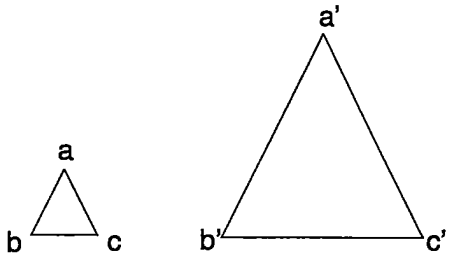


Figure 5.8: Geometric witness: similar "infinitesimal" and "real" triangles

Thus, because the two triangles are similar, we have been able to deduce that the ratio of infinitesimal is a finite (real) quantity. We are now free to manipulate the ratio and use it safely and soundly within our framework. The triangle $a'b'c'$ is a simple example of a *finite geometric witness*. Setting up or identifying such witnesses is a crucial step of Newton's analysis. This enables him to reason about infinitesimals and their ratios in ultimate situations by relating them to macroscopic features of the geometric diagram. In some cases, exhibiting a witness can be a rather complicated task that involves proving a large number of intermediate theorems. This is the case for the most famous result of Newton's *Principia*: the **Propositio Kepleriana** or **Kepler Problem**.

5.4 Case Study: Propositio Kepleriana

This is **Proposition 11** of Book 1 of the *Principia*. This Proposition is important for both mathematical and historical reasons, as it lays the foundations for Kepler's first law of Gravitation. It provides the mathematical analysis that could explain and confirm Kepler's guess that planets travelled in ellipses round the sun [81].

The proof of this proposition will be shown in detail as it gives a good overview of the mixture of geometry, algebra and limit reasoning that is so characteristic of Newton's *Principia*. It also gives an idea of the depth and amount of mathematical expertise involved in Newton's proof. The proof that Newton describes, though relatively short on paper, becomes a major demonstration when expanded and reproduced using Isabelle. The elegance of many of the constructions, which could be glossed over, is revealed through the detailed analysis.

We give formal justifications for the steps made by Newton in ultimate situations through our rigorous and logical use of infinitesimals. Infinitesimal reasoning is notorious for leading to contradictions. However, since nonstandard analysis is generally believed to be consistent, it ensures that our mechanization is rigorous. We will give the enunciation of the Proposition and the proof (sketch) provided by Newton. We will then expand on the sketch and provide detailed proofs of the steps that are made by Newton. This will require the use of the rules from the geometric and NSA theories developed in Isabelle. Moreover, an anomaly is revealed in Newton's reasoning through our rigorous formalization.

5.4.1 Proposition 11 and Newton's Proof

Proposition 11 is in fact stated as a problem by Newton at the start of Section 3 of the *Principia*. This section deals with "*the motion of bodies in eccentric conic section*". Particular orbits and laws governing forces that are relevant to the universe are investigated. The mathematical tools are developed for later use in Book III of the *Principia*, when natural phenomena of our world are investigated. Our task consists in expressing Newton's result as a goal which is then proved. Figure 5.9 shows Newton's original diagram used for this Proposition.

Proposition 11 *If a body revolves in an ellipse; it is required to find the law of the centripetal force tending to the focus of the ellipse.*

Newton's Solution: *Let S be the focus of the ellipse. Draw SP cutting the diameter DK of the ellipse in E, and the ordinate Qv in x; and complete the parallelogram $QxPR$. It is evident that EP is equal to the greater semiaxis AC: for drawing HI from the other focus H of the ellipse parallel to EC, because CS, CH are equal, ES, EI will be also equal; so that EP is the half-sum of PS, PI, that is (because of the parallels HI, PR, and the equal angles IPR, HPZ), of PS, PH, which taken together are equal to the whole axis $2AC$. Draw QT perpendicular to SP, and putting L for the principal latus rectum of the ellipse (or for $\frac{2BC^2}{AC}$), we shall have*

$$L \cdot QR : L \cdot Pv = QR : Pv = PE : PC = AC : PC,$$
$$\text{also, } L \cdot Pv : Gv \cdot Pv = L : Gv, \text{ and, } Gv \cdot Pv : Qv^2 = PC^2 : CD^2$$

By Corollary 2, Lemma 7, when the points P and Q coincide, $Qv^2 = Qx^2$, and Qx^2 or $Qv^2 : QT^2 = EP^2 : PF^2 = CA^2 : PF^2$, and (by Lemma 12) $= CD^2 : CB^2$. Multiplying together corresponding terms of the four proportions, and by simplifying, we shall have

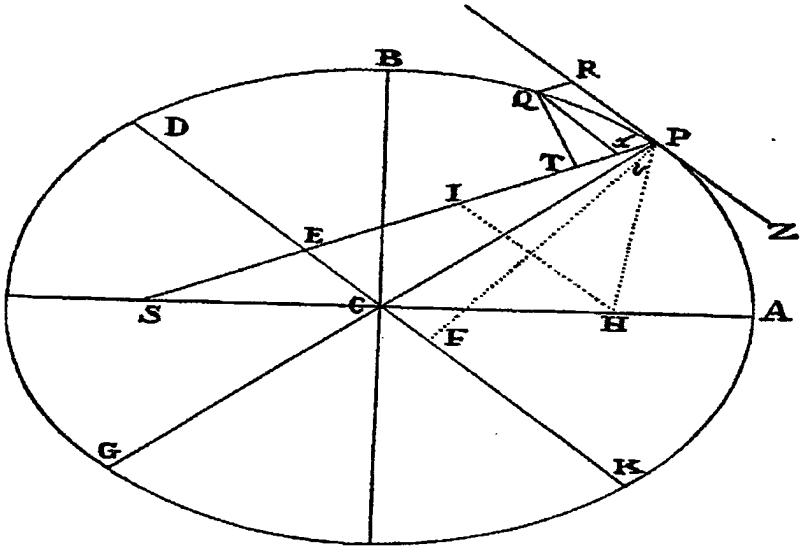


Figure 5.9: Newton's original diagram for Proposition 11

$$L \cdot QR : QT^2 = AC \cdot L \cdot PC^2 \cdot CD^2 : PC \cdot Gv \cdot CD^2 \cdot CB^2 = 2PC : Gv,$$

since $AC \cdot L = 2BC^2$. *But the points* Q *and* P *coinciding,* $2PC$ *and* Gv *are equal. And therefore the quantities* $L \cdot QR$ *and* QT^2, *proportional to these, will also be equal. Let those equals be multiplied by* $\frac{SP^2}{QR}$, *and* $L \cdot SP^2$ *will become equal to* $\frac{SP^2 \cdot QT^2}{QR}$. *And therefore (by Corollary 1 and 5, Proposition 6) the centripetal force is inversely as* $L \cdot SP^2$, *that is, inversely as the square of the distance* SP. *Q.E.I.*

Newton's derivation concludes that the centripetal force, for a body moving in an ellipse, is inversely proportional to the square of the distance.

Our proof proceeds in several steps. We set up various relationships that we will need for the conclusion. This involves proving Newton's intermediate results.

5.4.2 Expanding Newton's Proof

Newton's argument for Proposition 11 is complex and represents a major mechanization task. In what follows, we highlight the main results that were proved and, in some cases, details of the properties that needed to be set up first. We mention the constraints that needed to be satisfied within our framework before the various ratios that were proved could be combined. Our mechanization was broken down into several steps that roughly followed from Newton's original proof. The analysis provided by Brackenridge [12] and Densmore [27] clarified several aspects of the proof and its formalization. The main results that are set up are as follows (see Figure 5.9):

- $\texttt{len}(\texttt{E} - \texttt{P}) = \texttt{len}(\texttt{A} - \texttt{C})$

- $\texttt{len(A - C)/len(P - C)} = \texttt{L} \cdot \texttt{len(Q - R)/L} \cdot \texttt{len(P - v)}$
- $\texttt{L} \cdot \texttt{len(P - v)/(len(G - v)} \cdot \texttt{len(P - v))} = \texttt{L/len(G - v)}$
- $\texttt{len(G - v)} \cdot \texttt{len(P - v)/len(Q - v)}^2 = \texttt{len(P - C)}^2\texttt{/len(C - D)}^2$
- $\texttt{len(Q - v)}^2\texttt{/len(Q - T)}^2 \approx \texttt{len(C - D)}^2\texttt{/len(C - B)}^2$

Step 1: Proving $\texttt{len(E - P)} = \texttt{len(A - C)}$

This results shows that the length of EP is independent of P and Newton's proof uses several properties of the ellipse. A rather detailed overview of this particular proof is provided as it gives an idea of the amount of work involved in mechanizing Newton's geometric reasoning. Moreover, the reader can then compare Newton's proof style and prose with our own proof and see the GTP methods we have formalized in action.

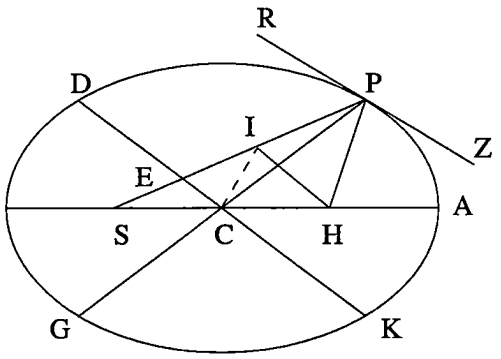


Figure 5.10: Construction for Step 1 of Proposition 11

In Figure 5.10, the following holds

- C is the centre of the ellipse with S and H as the foci
- P is a point on the curve
- RZ is the tangent at P
- the conjugate diameter $D - K \parallel P - Z$
- $P - S$ intersects $D - K$ at E
- $H - I \parallel E - C$ and $H - I$ intersects $P - S$ at I

Since $\texttt{H - I} \parallel \texttt{E - C}$, the following theorem holds,

$$\texttt{H - I} \parallel \texttt{E - C} \implies \texttt{S}_{\texttt{delta}}\,\texttt{C E I} = \texttt{S}_{\texttt{delta}}\,\texttt{C E H} \tag{5.4}$$

But the foci are collinear with and (by Apollonius III.45 [3]) equidistant from the centre of the ellipse; so the following can be derived using the signed-area method,

$$\begin{aligned} \texttt{coll S C H} &\implies \texttt{len(S - C)} \cdot \texttt{S}_{\texttt{delta}}\,\texttt{C E H} = \texttt{len(C - H)} \cdot \texttt{S}_{\texttt{delta}}\,\texttt{C S E} \\ &\implies \texttt{S}_{\texttt{delta}}\,\texttt{C E H} = \texttt{S}_{\texttt{delta}}\,\texttt{C S E} \end{aligned} \tag{5.5}$$

Also, points S, E and I are collinear and therefore combining with (5.4) and (5.5) above, we verify Newton's **"ES,EI will also be equal"**

$$\begin{aligned} \texttt{coll SEI} &\implies \texttt{len(S-E)} \cdot \texttt{S}_{\texttt{delta}}\,\texttt{CEI} = \texttt{len(E-I)} \cdot \texttt{S}_{\texttt{delta}}\,\texttt{CSE} \\ &\implies \texttt{len(S-E)} = \texttt{len(E-I)} \end{aligned} \tag{5.6}$$

Next, the following derivations can be made, with the help of the last result proving Newton's **"EP is the half-sum of PS, PI"**

$$\begin{aligned} \texttt{coll EIP} &\implies \texttt{len(E-P)} = \texttt{len(E-I)} + \texttt{len(I-P)} \\ &\implies \texttt{len(E-P)} = \texttt{len(S-E)} + \texttt{len(I-P)} \\ &\implies 2 \cdot \texttt{len(E-P)} = \texttt{len(E-P)} + \texttt{len(S-E)} + \texttt{len(I-P)} \\ &\implies 2 \cdot \texttt{len(E-P)} = \texttt{len(S-P)} + \texttt{len(I-P)} \\ &\implies \texttt{len(E-P)} = \frac{\texttt{len(S-P)} + \texttt{len(I-P)}}{2} \end{aligned} \tag{5.7}$$

Note the use of the following theorem in the derivation above

$$\texttt{coll SEP} \implies \texttt{len(S-E)} + \texttt{len(E-P)} = \texttt{len(S-P)}$$

Next, Newton argues that in fact (5.7) can be written as

$$\texttt{len(E-P)} = \frac{\texttt{len(S-P)} + \texttt{len(H-P)}}{2} \tag{5.8}$$

So, a proof of $\texttt{len(I-P)} = \texttt{len(H-P)}$ is needed to progress further. This will follow if it can be shown that $\triangle$PHI is an isosceles, that is

$$\langle \texttt{P-H}, \texttt{H-I} \rangle = \langle \texttt{H-I}, \texttt{I-P} \rangle \tag{5.9}$$

To prove (5.9), both $\texttt{H-I} \parallel \texttt{P-Z}$ and $\texttt{H-I} \parallel \texttt{P-R}$ are derived first using

$$[|\texttt{H-I} \parallel \texttt{E-C}; \texttt{E-C} \parallel \texttt{P-Z}|] \implies \texttt{H-I} \parallel \texttt{P-Z} \tag{5.10}$$

$$[|\texttt{H-I} \parallel \texttt{P-Z}; \texttt{coll PZR}|] \implies \texttt{H-I} \parallel \texttt{P-R} \tag{5.11}$$

From (5.10), (5.11), and the proof of Euclid I.29 given in Section 2.3.3

$$\texttt{H-I} \parallel \texttt{P-Z} \implies \langle \texttt{P-H}, \texttt{H-I} \rangle = \langle \texttt{H-P}, \texttt{P-Z} \rangle \tag{5.12}$$

$$\begin{aligned} \texttt{H-I} \parallel \texttt{P-R} &\implies \langle \texttt{H-I}, \texttt{I-P} \rangle = \langle \texttt{R-P}, \texttt{P-I} \rangle \\ &\implies \langle \texttt{H-I}, \texttt{I-P} \rangle = \langle \texttt{R-P}, \texttt{P-S} \rangle \end{aligned} \tag{5.13}$$

From the definition of the tangent to an ellipse and the collinearity of P, I, and S (also recall that full-angles are angles between *lines* rather than rays and are measured anti-clockwise),

$$\begin{aligned} \texttt{is\_e\_tangent (P-Z) S H Ellipse} &\implies \langle \texttt{H-P}, \texttt{P-Z} \rangle = \langle \texttt{R-P}, \texttt{P-I} \rangle \\ &\implies \langle \texttt{H-P}, \texttt{P-Z} \rangle = \langle \texttt{R-P}, \texttt{P-S} \rangle \end{aligned} \tag{5.14}$$

From (5.12), (5.13) and (5.14), the following is deduced as required

$$\langle \texttt{P-H}, \texttt{H-I} \rangle = \langle \texttt{H-I}, \texttt{I-P} \rangle$$

Thus, we have $\texttt{len}(\texttt{I} - \texttt{P}) = \texttt{len}(\texttt{H} - \texttt{P})$ (Euclid I.6 [30]), and hence (5.8) is proved, that is, Newton's assertion that **"[EP is the half sum of] PS, PH"**. It then follows from the definition of an ellipse that the sum of $\texttt{len}(\texttt{S} - \texttt{P})$ and $\texttt{len}(\texttt{P} - \texttt{H})$ is equal to the length of the major axis, that is,

$$\texttt{P} \in \texttt{Ellipse} \Longrightarrow \texttt{len}(\texttt{S} - \texttt{P}) + \texttt{len}(\texttt{P} - \texttt{H}) = 2 \cdot \texttt{len}(\texttt{A} - \texttt{C}) \tag{5.15}$$

From (5.15) and (5.8), we can finally derive the property that Newton states as being evident: **"EP is equal to the greater semiaxis AC"**

$$\texttt{len}(\texttt{E} - \texttt{P}) \;=\; \texttt{len}(\texttt{A} - \texttt{C}) \tag{5.16}$$

The first step has shown Newton's geometric reasoning in action. For the next steps, as the various ratios are derived, the detailed derivations of the geometric theorems are not always shown. We will concentrate on the setting up of the proportions and how everything is put together to get the final result. We will state Newton's Lemmas when they are used and theorems about infinitesimals that we use.

Step 2: Showing $\frac{L \cdot QR}{L \cdot Pv} = \frac{QR}{Pv} = \frac{PE}{PC} = \frac{AC}{PC}$

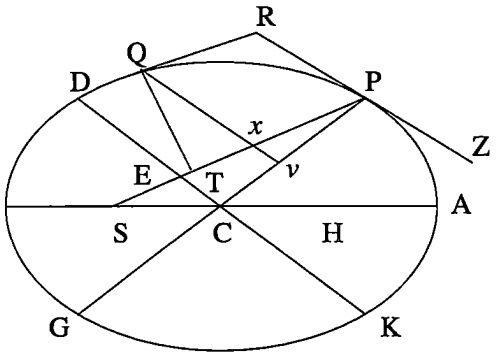


Figure 5.11: Construction for Steps 2—4 of Proposition 11

In Figure 5.11, in addition to properties already mentioned, the following relations hold

- $QT \perp SP$
- $QxPR$ is a parallelogram
- Q, x, and v are collinear
- Q is infinitely close to P

The (finite and non-zero) constant L, present in the ratio above, is known as the *latus rectum* of the ellipse at A (see Figure 5.9) and is defined as

$$L \equiv 2 \cdot \texttt{len}(\texttt{B} - \texttt{C})^2 / \texttt{len}(\texttt{A} - \texttt{C})$$

It is easily proved that $\texttt{v} - \texttt{x} \parallel \texttt{C} - \texttt{E}$ and so the theorem below follows

$$\texttt{v} - \texttt{x} \parallel \texttt{C} - \texttt{E} \implies \langle \texttt{P} - \texttt{v}, \texttt{v} - \texttt{x} \rangle = \langle \texttt{P} - \texttt{C}, \texttt{C} - \texttt{E} \rangle \tag{5.17}$$

From (5.17) and the fact that $\triangle$Pvx and $\triangle$PCE share P as a common vertex, it follows that they are *similar*. Also, since QxPR is a parallelogram, we have $\texttt{len(Q}-\texttt{R)} = \texttt{len(P}-\texttt{x)}$. Thus, the following derivations can be made:

$$\texttt{SIM PVxPCE} \Longrightarrow \frac{\texttt{len(P}-\texttt{E)}}{\texttt{len(P}-\texttt{C)}} = \frac{\texttt{len(P}-\texttt{x)}}{\texttt{len(P}-\texttt{v)}} = \frac{\texttt{len(Q}-\texttt{R)}}{\texttt{len(P}-\texttt{v)}} = \frac{\texttt{len(A}-\texttt{C)}}{\texttt{len(P}-\texttt{C)}} \tag{5.18}$$

One of the substitutions used in (5.18) follows from (5.16) proved in the previous step. The equations above verify Newton's ratios.

Step 3: Showing $\frac{L \cdot Pv}{Gv \cdot Pv} = \frac{L}{Gv}$

The proof of the ratio

$$\frac{\texttt{L} \cdot \texttt{len(P}-\texttt{v)}}{\texttt{len(G}-\texttt{v)} \cdot \texttt{len(P}-\texttt{v)}} = \frac{\texttt{L}}{\texttt{len(G}-\texttt{v)}} \tag{5.19}$$

is trivial and we will not expand on it.

Step 4: Showing $\frac{Gv \cdot Pv}{Qv^2} = \frac{PC^2}{PD^2}$

By Apollonius I.21 [3], *if the lines DC and Qv are dropped ordinatewise to the diameter PG, the squares on them DC^2 and Qv^2 will be to each other as the areas contained by the straight lines cut off GC, CP, and Gv, vP on diameter PG.* Algebraically, we proved the following property of the ellipse,

$$\begin{aligned} \frac{\texttt{len(D}-\texttt{C)}^2}{\texttt{len(Q}-\texttt{v)}^2} &= \frac{\texttt{len(G}-\texttt{C)} \cdot \texttt{len(P}-\texttt{C)}}{\texttt{len(G}-\texttt{v)} \cdot \texttt{len(P}-\texttt{v)}} \\ &= \frac{\texttt{len(P}-\texttt{C)}^2}{\texttt{len(G}-\texttt{v)} \cdot \texttt{len(P}-\texttt{v)}} \end{aligned}$$

Rearranging the terms, we get the required ratio,

$$\frac{\texttt{len(G}-\texttt{v)} \cdot \texttt{len(P}-\texttt{v)}}{\texttt{len(Q}-\texttt{v)}^2} = \frac{\texttt{len(P}-\texttt{C)}^2}{\texttt{len(D}-\texttt{C)}^2} \tag{5.20}$$

Step 5: Showing $\frac{Qv^2}{QT^2} \approx \frac{CD^2}{CB^2}$ and intermediate ratios

In Figure 5.12, we have the additional property,

- $PF \perp DK$

Again, it can be easily proved that $\texttt{Qx} \parallel \texttt{EF}$. The next theorem then follows from Euclid I.29 as given in Section 2.3.3

$$\begin{aligned} \texttt{Q}-\texttt{x} \parallel \texttt{E}-\texttt{F} &\Longrightarrow \langle \texttt{Q}-\texttt{x}, \texttt{x}-\texttt{E} \rangle = \langle \texttt{F}-\texttt{E}, \texttt{E}-\texttt{x} \rangle \\ &\Longrightarrow \langle \texttt{Q}-\texttt{x}, \texttt{x}-\texttt{T} \rangle = \langle \texttt{F}-\texttt{E}, \texttt{E}-\texttt{P} \rangle \end{aligned} \tag{5.21}$$

Since $\langle \texttt{P}-\texttt{F}, \texttt{F}-\texttt{E}, \rangle = \langle \texttt{x}-\texttt{T}, \texttt{T}-\texttt{Q} \rangle = \pi/2$ and (5.21), it follows that $\triangle$PEF and $\triangle$QxT are similar. The next theorems (using (5.16) where needed) then hold and verify Newton's intermediate results for the current step.

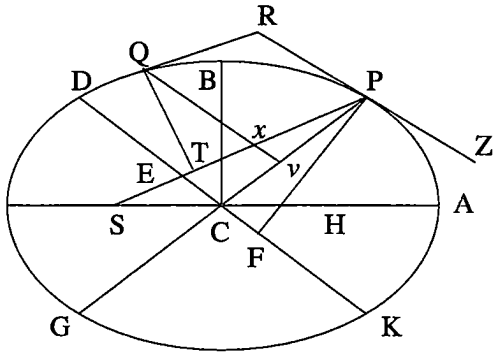


Figure 5.12: Construction for Step 5 of Proposition 11

$$\texttt{SIM PEFQxT} \Longrightarrow \frac{\texttt{len(Q - x)}^2}{\texttt{len(Q - T)}^2} = \frac{\texttt{len(P - E)}^2}{\texttt{len(P - F)}^2} = \frac{\texttt{len(C - A)}^2}{\texttt{len(P - F)}^2} \tag{5.22}$$

Newton's **Lemma 12** (See Figure 2.7) is now needed for the next result. According to the Lemma, the parallelogram circumscribed about DK and PG is equal to the parallelogram circumscribed about the major and minor axes of the ellipse. Thence, we have the following theorem

$$\texttt{len(C - A)} \cdot \texttt{len(C - B)} = \texttt{len(C - D)} \cdot \texttt{len(P - F)} \tag{5.23}$$

Rearranging (5.23) gives $\texttt{len(C - A)}/\texttt{len(P - F)} = \texttt{len(C - D)}/\texttt{len(C - B)}$ and substituting in (5.22) leads to

$$\frac{\texttt{len(Q - x)}^2}{\texttt{len(Q - T)}^2} = \frac{\texttt{len(C - D)}^2}{\texttt{len(C - B)}^2} \tag{5.24}$$

By Newton's **Lemma 7, Corollary 2**, when the distance between Q and P becomes infinitesimal as they coincide, we have the following result:

$$\frac{\texttt{len(Q - v)}}{\texttt{len(Q - x)}} \approx 1 \tag{5.25}$$

Now, to reach the final result for this step, we need to substitute $\texttt{len(Q - v)}$ for $\texttt{len(Q - x)}$ in (5.24). However, we cannot simply carry out the substitution even though the quantities are infinitely close. Indeed, one has to be careful when multiplying the quantities on both sides of the $\approx$ relation because they might no longer be infinitely close after the multiplication (c.f. Section 5.3). Consider, the non-zero infinitesimal ϵ,

$$\epsilon \approx \epsilon^2 \text{ but } \epsilon \cdot 1/\epsilon \not\approx \epsilon^2 \cdot 1/\epsilon$$

It is possible, however, to multiply two infinitely close quantities by any finite quantity; the results are still infinitely close. This is a consequence of the following theorem, proved in Isabelle's NSA theory:

$$[|x \approx y; u \in \texttt{Finite}|] \implies x \cdot u \approx y \cdot u \tag{5.26}$$

Now, assuming that $\texttt{len(C - D)}$ and $\texttt{len(C - B)}$ are both finite but not infinitesimal (for example, $\texttt{len(C - D)}, \texttt{len(C - B)} \in \mathbb{R}$), then $\texttt{len(C - D)/len(C - B)}$ is `Finite`. Hence, the ratio of *infinitesimals* $\texttt{len(Q - x)/len(Q - T)}$ is `Finite`. Therefore, from (5.25), (5.26) and using (5.24) the following theorem is derived

$$\frac{\texttt{len(Q - v)}^2}{\texttt{len(Q - T)}^2} \approx \frac{\texttt{len(C - D)}^2}{\texttt{len(C - B)}^2} \tag{5.27}$$

This gives the result that we wanted for the fifth step of the proof of Proposition 11. We are now ready to put all the various results together in the next and final step. This will then conclude the formal proof of the Proposition.

Step 6: Putting the ratios together Combining (5.20) and (5.27), with the help of theorem (5.26) and some algebra, yields:

$$\begin{aligned}[|\frac{\texttt{len(Q - v)}^2}{\texttt{len(Q - T)}^2} \approx \frac{\texttt{len(C - D)}^2}{\texttt{len(C - B)}^2};\ & \frac{\texttt{len(G - v)} \cdot \texttt{len(P - v)}}{\texttt{len(Q - v)}^2} \in \texttt{Finite}|] \\ \implies & \frac{\texttt{len(G - v)} \cdot \texttt{len(P - v)}}{\texttt{len(Q - T)}^2} \approx \frac{\texttt{len(P - C)}^2}{\texttt{len(C - B)}^2}\end{aligned} \tag{5.28}$$

This is combined with (5.19) to derive the next relation between ratios. The reader can check that both sides of the $\approx$ relation are multiplied by finite quantities ensuring the results are infinitely close:

$$\frac{\texttt{L} \cdot \texttt{len(P - v)}}{\texttt{len(Q - T)}^2} \approx \frac{\texttt{len(P - C)}^2 \cdot \texttt{L}}{\texttt{len(C - B)}^2 \cdot \texttt{len(G - v)}} \tag{5.29}$$

The next task is to combine the last result (5.29) with (5.18) to yield the following ratio which is equivalent to Newton's $L \cdot QR : QT^2 = AC \cdot L \cdot PC^2 \cdot CD^2 : PC \cdot Gv \cdot CD^2 \cdot CB^2$:

$$\frac{\texttt{L} \cdot \texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \approx \frac{\texttt{len(P - C)} \cdot \texttt{L} \cdot \texttt{len(A - C)}}{\texttt{len(C - B)}^2 \cdot \texttt{len(G - v)}} \tag{5.30}$$

But, we know that $\texttt{L} = 2 \cdot \texttt{len(B - C)}^2/\texttt{len(A - C)}$, so (5.30) can be further simplified to give Newton's other ratio "$L \cdot QR : QT^2 = 2PC : Gv$"

$$\frac{\texttt{L} \cdot \texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \approx \frac{2 \cdot \texttt{len(P - C)}}{\texttt{len(G - v)}} \tag{5.31}$$

Once these ratios have been derived, Newton says **"But the points Q and P coinciding, $2PC$ and Gv are equal. And therefore the quantities $L \cdot QR$ and QT^2, proportional to these are also equal."**

We formalize this by showing that $\texttt{len(P - v)} \approx 0$ as the distance between Q and P becomes infinitesimal; thus, it follows that $2 \cdot \texttt{len(P - C)/len(G - v)} \approx 1$ and so, using (5.31) and the transitivity of $\approx$, we have the result

$$\frac{\texttt{L} \cdot \texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \approx 1 \tag{5.32}$$

The final step in Newton's derivation is **"Let those equals be multiplied by $\frac{SP^2}{QR}$ and $L \cdot SP^2$ will become equal to $\frac{SP^2 \cdot QT^2}{QR}$"**. This final ratio gives the geometric representation for the force, as we showed in Section 5.2.5, and hence enables Newton to deduce immediately that the centripetal force obeys an inverse square law.

We would like to derive Newton's result in the *same way*, but remark that

$$[|\texttt{len(S - P)} \in \texttt{Finite} - \texttt{Infinitesimal};\ \texttt{len(Q - R)} \in \texttt{Infinitesimal}|] \Longrightarrow \frac{\texttt{len(S - P)}^2}{\texttt{len(Q - R)}} \in \texttt{Infinite} \tag{5.33}$$

as Q and P become coincident. So, there seems to be a problem with simply multiplying (5.32) by Newton's ratio SP^2/QR: we cannot ensure that the results are infinitely close. Our formal framework *forbids* the multiplication that Newton performs.

Therefore, we need to find an *alternative* way of arriving at the same result. Recall from Section 5.2.5, that we have proved the following geometric representation for the centripetal force:

$$\mathit{force} \approx k \cdot \frac{\texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \cdot \frac{1}{\texttt{len(S - P)}^2} \tag{5.34}$$

Now from (5.32), we can deduce that since $\texttt{L} \in \texttt{Finite} - \texttt{Infinitesimal}$, the following theorems hold

$$\frac{\texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \in \texttt{Finite} - \texttt{Infinitesimal} \tag{5.35}$$

$$\frac{\texttt{len(Q - T)}^2}{\texttt{len(Q - R)}} \approx \texttt{L} \tag{5.36}$$

$$\frac{\texttt{len(Q - T)}^2}{\texttt{len(Q - R)}} \in \texttt{Finite} \tag{5.37}$$

Since (5.35) holds and $1/\texttt{len(S - P)}^2 \in \texttt{Finite}$, it follows that $\mathit{force} \in \texttt{Finite}$ and so we can now use the following theorem about the product of *finite*, infinitely close quantities

$$|]a \approx b; c \approx d; a \in \texttt{Finite}; c \in \texttt{Finite}|] \Longrightarrow a \cdot c \approx b \cdot d$$

with (5.34), (5.36), and (5.37) to yield

$$\begin{aligned} \mathit{force} \cdot \texttt{L} &\approx k \cdot \frac{\texttt{len(Q - R)}}{\texttt{len(Q - T)}^2} \cdot \frac{1}{\texttt{len(S - P)}^2} \cdot \frac{\texttt{len(Q - T)}^2}{\texttt{len(Q - R)}} \\ &\approx k \cdot \frac{1}{\texttt{len(S - P)}^2} \end{aligned} \tag{5.38}$$

Note that we also used the symmetry of $\approx$ in the derivation above. Finally from (5.38), we get to the celebrated result. Since L is finite (real) and constant for

a given ellipse,

$$\begin{aligned} \mathit{force} &\approx \frac{k}{\mathtt{L}} \cdot \frac{1}{\mathtt{len}(\mathtt{S}-\mathtt{P})^2} \\ \mathit{force} &\propto_{\mathit{ultimate}} \frac{1}{\mathtt{len}(\mathtt{S}-\mathtt{P})^2} \end{aligned} \tag{5.39}$$

5.4.3 Conclusions

We have described in detail the machine proof of Proposition 11 of the *Principia* and shown how the theories developed in Isabelle can be used to derive Newton's geometric representations for physical concepts. We have used a combination of geometry and NSA rules to confirm, through a study of one of the most important Propositions of the *Principia*, that Newton's geometric and ultimate procedures can be cast within the rigour of our formal framework. The discovery of a step in Newton's reasoning that could not be justified formally– in contrast with other ones where Newton explicitly sets up finite witnesses– is noteworthy. The alternative derivation presented in this work is original, as far as we know. It shows how to use our rules to deduce the same result soundly.

Once again, the mechanization of results from the *Principia* has been an interesting and challenging exercise. Newton's original reasoning, though complex and often hard to follow, displays the impressive deductive power of geometry. The addition of infinitesimal notions results in a richer, more powerful geometry in which new properties can emerge in ultimate situations. These tools can be used to model physical phenomena, and also to provide rigorous proofs in geometry that make use of infinitesimal arguments.

Chapter 6

Nonstandard Real Analysis

Classical or standard analysis is mostly concerned with the study of the real numbers and with the properties of functions defined on them. We shall now describe the use of the hyperreals as valuable tools for mathematical analysis. Through the existence of infinitesimals, finite, and infinite numbers, NSA provides us with a rich structure which we use to formalize alternative treatments of topics in classical analysis. Such treatments are not only valuable for the additional light that they cast on the processes of analysis, but also for the simplification they bring to many concepts and arguments. As will be seen, the mechanization of analysis can benefit directly from this simplification, since difficult instantiation steps in proofs are simply eliminated in many cases. We start by showing how functions defined over the reals and naturals can be systematically extended to the hyperreals and hypernaturals, respectively. These notions are crucial to nonstandard real analysis. We then proceed to develop some elementary analysis that will make use of the new classes of numbers, the infinitely close relation, and other notions induced on them.

6.1 Extending a Relation to the Hyperreals

There are systematic methods through which functions defined on the reals are extended to the hyperreals. This process of extending a relation from $\mathbb{R}$ to $\mathbb{R}^*$ is known as the $*$-transform.

6.1.1 Internal Sets and Nonstandard Extensions

Many properties of the reals, suitably reinterpreted, can be transferred to the hyperreal number system. For example, we have seen that $\mathbb{R}^*$, like $\mathbb{R}$, is a totally ordered field. Also, just as $\mathbb{R}$ contains the natural numbers $\mathbb{N}$ as a discrete subset with its own characteristic properties, $\mathbb{R}^*$ contains the hypernaturals $\mathbb{N}^*$ as a corresponding discrete subset with analogous properties. Moreover, subsets $\mathbb{Z}^*$ (the hyperintegers) and $\mathbb{Q}^*$ of $\mathbb{R}^*$ exhibit relations to $\mathbb{N}^*$ similar to those that $\mathbb{Z}$ and $\mathbb{Q}$ bear to $\mathbb{N}$ in $\mathbb{R}$.

However, there are properties of $\mathbb{R}$ that do not transfer to $\mathbb{R}^*$. This is the case for the fundamental supremum property of the reals stated in Section 3.4.3. It is easy to see that this upper bound property does not necessarily hold by

considering, for example, the set $\mathbb{R}$ itself, which we regard as embedded into the hyperreals (i.e. the set SReal). This is a non-empty set which is bounded above (by any of the infinite numbers in $\mathbb{R}^*$) but does not have a least upper bound in $\mathbb{R}^*$.

Proof: Suppose that r is the least upper bound of $\mathbb{R}$. Then, it follows that r is infinite since it is an upper bound. But as $r \in$ Infinite, it follows that $r - 1 \in$ Infinite, so $r - 1$ is a smaller upper bound which is a contradiction. □

We now introduce an important refinement which classifies subsets of $\mathbb{R}^*$ into two types: **internal** and **external** subsets. With this done, we shall be able to prove the following statement, for example, about the supremum property for the hyperreals:

> Every non-empty *internal* subset of $\mathbb{R}^*$ which has an upper bound in $\mathbb{R}^*$ has a least upper bound in $\mathbb{R}^*$

DEFINITION 6.1.1. *Let A_n, $n \in \mathbb{N}$, be any sequence of sets of real numbers. This sequence determines a certain set A of hyperreals according to the rule: The hyperreal number $x = [\langle X_n \rangle]$ is a member of set $A \subseteq \mathbb{R}^*$ if and only if the set $\{n \in \mathbb{N}. X_n \in A_n\}$ belongs to the ultrafilter $U_{\mathbb{N}}$.*

This definition is analogous to the one we used to define hyperreals in terms of sequence of reals. The sequences of sets of real numbers can then be used to define the so-called internal sets of hyperreals. In Isabelle, we have the following declaration and definition for an internal set:

$$*\mathtt{sn}* :: (\mathtt{nat} \Rightarrow \mathtt{real\,set}) \Rightarrow \mathtt{hypreal\,set}$$
$$*\mathtt{sn}* A \equiv \{x. \forall X \in \mathtt{Rep\_hypreal}(x). \{n. X(n) \in A(n)\} \in U_{\mathbb{N}}\}$$

We are particularly interested in the special case when the sequence is constant, that is $A_n = A$ for all (or almost all) n. The internal set determined by such a sequence is called the **nonstandard extension** of A and, since this is the actual property that will be used more often in the course of our mechanization, it is defined explicitly:

$$*\mathtt{s}* :: \mathtt{real\,set} \Rightarrow \mathtt{hypreal\,set}$$
$$*\mathtt{s}* A \equiv \{x. \forall X \in \mathtt{Rep\_hypreal}(x). \{n. X(n) \in A\} \in U_{\mathbb{N}}\}$$

Thus, it follows that $*\mathtt{s}* A = *\mathtt{sn}* (\lambda n. A)$. In the literature, the nonstandard extension of a set A is usually denoted by A^*. We shall make use of this conventional mathematical notation as well. However, the actual Isabelle/HOL notation ($*\mathtt{s}* A$) will also be used in many cases, especially to show how a particular concept is expressed in the theorem prover.

It can be shown that any non-empty, internal subset of $\mathbb{R}^*$ has the supremum property though the proof will not be given here. In fact, for any subset of S of $\mathbb{R}^*$ that fails to have a least upper bound one can infer that it is not internal. Any subset of hyperreals that is not internal is called **external set.**

The process of extending a set of real numbers to a set of hyperreals has shown an example of the $*$-transform at work. In general, this transformation procedure can be applied to any n-ary relation on the reals, extending it to an n-ary relation on the hyperreals. This is done using the rule that P holds on

an n-tuple in $(\mathbb{R}^*)^n$ if the index set where P holds on the representative real n-tuple sequence is in the chosen free ultrafilter. More instances of $*$-transforms will be met when nonstandard extensions of functions are introduced.

6.1.2 Properties of Extended Sets

Various properties of nonstandard extensions of sets of real numbers can now be derived. The first result proved (in one step using Isabelle's automatic tactic) is that $\mathbb{R}^*$ is the nonstandard extension of $\mathbb{R}$. The nonstandard extensions of sets of reals will, in general, be different from the original set. The exception occurs for finite sets since then the extension function simply degenerates to the embedding function. This is confirmed by the following theorem, where " denotes the image operator for relations:

$$\texttt{finite}\ A \Longrightarrow *\texttt{s}*\ A = \texttt{hypreal\_of\_real}\ ``A$$

If the set A is infinite, however, then we prove that A^* contains elements that are not standard copies of the members of A. The nonstandard extension provides us with a new set that is an enlargement of A. Thus, the enlargement of $\mathbb{R}$ yields a new set that contains infinitesimals and infinite elements that have no counterparts in the real number system. A number of useful results involving boolean operations on nonstandard extensions of sets are proved:

1) $\texttt{hypreal\_of\_real}\ ``A \subseteq *\texttt{s}*\ A$

2) $A \subseteq B \Longrightarrow (*\texttt{s}*\ A) \subseteq (*\texttt{s}*\ B)$

3) $(*\texttt{s}*\ \emptyset) = \emptyset$

4) $*\texttt{s}*\ (A \cup B) = (*\texttt{s}*\ A) \cup (*\texttt{s}*\ B)$

5) $*\texttt{s}*\ (A \cap B) = (*\texttt{s}*\ A) \cap (*\texttt{s}*\ B)$

6) $*\texttt{s}*\ (-A) = -(*\texttt{s}*\ A)$

7) $\forall n.\ X(n) \notin M \Longrightarrow \texttt{Abs\_hypreal}\ (\texttt{hyprel}\hat{}\hat{}\{X\}) \notin (*\texttt{s}*\ M)$

The proofs of these various theorems all follow from the basic and derived properties of the free ultrafilter (see Section 3.6.1). For example, property 3 follows quite straightforwardly from the fact that no filter contains the empty set. Property 5 is a direct consequence of the fact that filters are closed under the $\cap$ and $\subseteq$ operations. Proving properties 4 and 6 needs the fact that for any subset A of $\mathbb{N}$, either A or $\texttt{Compl}\ A$ belongs to the ultrafilter. The proofs are all straightforwardly carried through with the help of Isabelle's automatic tactic. The tactic can, in general, prove many of the theorems about sets operations automatically given the right rules in the simplification set.

6.1.3 Internal Functions and Nonstandard Extensions

Given a *standard* function which takes real arguments, we want to be able to define an analogous one that will also take *nonstandard* arguments. This leads to the notions of internal functions and to nonstandard extensions. These concepts are crucial as they will enable the formulation of familiar concepts in analysis

using nonstandard definitions. Also, they give a systematic way of extending any function over the reals to one over the hyperreals. We give the definition for the case dealing with function of one real variable [48]:

DEFINITION 6.1.2. *Let $\langle F_n \rangle$ be any sequence of standard functions from $\mathbb{R}$ to $\mathbb{R}$. This sequence determines an internal function $f \equiv [\langle F_n \rangle]$ from $\mathbb{R}^*$ to $\mathbb{R}^*$ according to the rule $x = [\langle X_n \rangle] \in \mathbb{R}^*$ maps into $y = [\langle Y_n \rangle] = f(x) \in \mathbb{R}^*$ if and only if $\{n \in \mathbb{N}. Y_n = F_n(X_n)\} \in U_{\mathbb{N}}$.*

Expressed in Isabelle, we have this rather more concise definition for the internal function:

```
*fn* :: (nat ⇒ (real ⇒ real)) ⇒ hypreal ⇒ hypreal
*fn* F x ≡ Abs_hypreal (⋃X ∈ Rep_hypreal(x). hyprel^^{λn. (Fn)(Xn)})
```

Thus, according to this definition, with F and x defined as above, the value of the internal function (`*fn*` F) at x is given by

$$(\texttt{*fn*}\ F)\, x = [\langle F_1(X_1), F_2(X_2), \cdots, F_n(X_n), \cdots \rangle]$$

Of interest, here as well, is the *nonstandard* extension of a standard function F as an important special type of internal functions. The nonstandard extension is obtained by having a constant sequence of functions i.e. one for which $F_n = F$ for (almost) all n. Once again, we define the special case explicitly:

```
*f* :: (real ⇒ real) ⇒ hypreal ⇒ hypreal
*f* F x ≡ Abs_hypreal (⋃X ∈ Rep_hypreal(x). hyprel^^{λn. F(Xn)})
```

We will denote the nonstandard extension of a given function either by f^* or by the equivalent Isabelle notation (`*f*` f). Referring back to the construction of the hyperreals in Isabelle, the definitions given for the field operations on them can all be viewed as nonstandard extensions of the analogous operations on the reals (e.g. addition on the hyperreals is actually $+^*$). We also note that our definition for nonstandard extension corresponds to Keisler's **Function Axiom** which states that "for each real function f of n variables there is a corresponding function f^* of n variables, called the natural extension of f" [52].

6.1.4 Properties of Extended Functions

We prove, as we did for set extensions, a number of useful properties about nonstandard extensions of functions. One of the first and most useful simplification theorem shows that the nonstandard extension of a function, f^*, is equivalent to applying f entrywise to an equivalence class representative in $\mathbb{R}^*$:

```
(*f* f) (Abs_hypreal (hyprel^^{λn. Xn})) =
        (Abs_hypreal (hyprel^^{λn. f(Xn)}))
```

This enables us to prove theorems about nonstandard functions by using the properties of the corresponding standard real function, the reals, and of the free ultrafilter. We then prove various theorems about preservation of rules across the $*$-transformation and other properties. Some of these Isabelle theorems are listed next. Most of the proofs are mechanized in two steps or fewer with the help

of Isabelle's automatic tactic **auto_tac**; the latter is supplied with simplification rules such as the theorem above and others about addition, multiplication and other operations (Equation 3.4, for example)[1]:

1) $(\texttt{*f*}\ (\lambda y.\ f\,y + g\,y))\,x = (\texttt{*f*}\ f)\,x + (\texttt{*f*}\ g)\,x$

2) $(\texttt{*f*}\ (\lambda y.\ f\,y \cdot g\,y))\,x = (\texttt{*f*}\ f)\,x \cdot (\texttt{*f*}\ g)\,x$

3) $(\texttt{*f*}\ (f \circ g)) = (\texttt{*f*}\ f) \circ (\texttt{*f*}\ g)$

4) $(\texttt{*f*}\ \lambda y.\ k)\,x = \widetilde{k}$

5) $(\texttt{*f*}\ (\lambda y.\ -f\,y))\,x = -\,(\texttt{*f*}\ f)\,x$

6) $(\texttt{*f*}\ (\lambda y.\ y))\,x = x$

7) $(\texttt{*f*}\ f)\,(\widetilde{a}) = \widetilde{f(a)}$

8) $(\texttt{*f*}\ (\lambda h.\ f(y+h)))\,x = (\texttt{*f*}\ f)\,(\widetilde{y} + x)$

9) $(\texttt{*f*}\ (\lambda h.\ f(g(y+h))))\,x = (\texttt{*f*}\ (f \circ g))\,(\widetilde{y} + x)$

10) $\texttt{*f*}\ \texttt{rabs} = \texttt{hrabs}$

11) $x \neq \texttt{0hr} \Longrightarrow (\texttt{*f*}\ \texttt{rinv})\,x = \texttt{hrinv}\ x$

12) $(\texttt{*f*}\ f)\,x \in \texttt{*s*}\ A \Longrightarrow x \in \texttt{*s*}\ \{y.\ f\,y \in A\}$

Theorem 7 is important as it tells us that the extended function has the same solutions as its standard counterpart for all (embedded) real arguments. Theorems 8 and 9 are proved because of their importance in the nonstandard definition of derivatives. From Theorem 11, we prove specific cases such as $(\texttt{*f*}\ \texttt{rinv})\,\epsilon = \texttt{hrinv}\ \epsilon$. Theorem 12 is a general lemma needed for proofs in elementary real topology. One might try and picture these various theorems mentally to get a better, more intuitive feel for the properties. If we combine $*$-transforms of both sets and functions, we can derive further theorems such as

- $\texttt{*s*}\ (\texttt{range}\ f) = \texttt{range}\ (\texttt{*f*}\ f)$
- $\texttt{*s*}\ \{x.\ \texttt{rabs}\ (f\,x - y) < r\} = \{x.\ \texttt{hrabs}\ ((\texttt{*f*}\ f)\,x - \widetilde{y}) < \widetilde{r}\}$

We note that any real constant is mapped to its embedded counterpart in the transform, as expected, while the functions are replaced by their nonstandard extensions.

The importance of internal sets and functions cannot be overstated. Lindstrøm calls them the "nice" subsets and functions of nonstandard analysis [60], and draws an analogy to topology where, for example, the nice sets and functions are the open sets and continuous functions. Nice concepts are those that we are interested in whenever a new mathematical structure is introduced. In NSA, they are important because they enable hyperreal sets and functions to inherit properties from their standard counterparts in a natural way. Moreover, they also enable us to express familiar concepts for our new mathematical structure that may be only partially inherited (such as the supremum property which only applies to internal subsets of $\mathbb{R}^*$). They introduce new subtleties

[1] Recall from Section 3.7.1 that $\widetilde{r}$ stands for the image of real number r in $\mathbb{R}^*$.

into the mathematics that it is essential to grasp in order to use the richer NS concepts adequately and to benefit. The strict typing of Isabelle/HOL makes the new concepts clearer to understand and definitions ensure that their use is rigorous. We will later introduce some further extensions that enable us to deal with functions from $\mathbb{N}$ to $\mathbb{R}$, for example.

6.2 Towards an Intuitive Calculus

Consider the real function $f(x) = x^2$. This extends naturally to a function f^* over $\mathbb{R}^*$. Now, if a is finite and ϵ is infinitesimal then $f^*(a+\epsilon) = (a+\epsilon)^2 = a^2 + \epsilon(2a+\epsilon) \approx a^2 = f^*(a)$ since the set `Infinitesimal` is an ideal in `Finite`. Thus, an infinitesimal change in the argument x only produces an infinitesimal change in f. This is, intuitively, the behaviour expected from a continuous function such as $f(x)$ above: broadly speaking, one does not expect any sudden gap or jumps in the graph that represents the behaviour of the function. As pointed out by Keisler [52] and others [75], students who are just beginning to study calculus often find it difficult to cope with formulas involving quantifiers. The traditional epsilon-delta ('$\epsilon-\delta$') approach, for example, is a sudden leap from the intuitive calculus of school to the rigour and formality of real analysis. One of the advantages of introducing the hyperreals is the simplification that this brings to the statement of many properties such as limits and continuity. For example, the $\epsilon-\delta$ condition for a function f to be continuous at a:

$$\forall \epsilon.\ (0 < \epsilon \longrightarrow \exists \delta.\ (0 < \delta \wedge \forall x.\ (0 < |x-a| < \delta \longrightarrow |f(x) - f(a)| < \epsilon)))$$

is equivalent to the simpler formula

$$\forall x.\ x \approx \widetilde{a} \longrightarrow f^*(x) \approx \widetilde{f(a)}$$

The approach, through the formal use of infinitesimals and relations such as $\approx$, retains much of the intuition that was present in school mathematics. The nonstandard treatment has been expounded in textbooks by Keisler [52], Henle and Kleinberg [44], and more recently by Hoskins [48], for example. Keisler's text has even been used successfully as an introductory textbook in calculus courses. There is much to be gained from carrying out proofs using a nonstandard formulation in general, and as this work shows next, even mechanization of analysis becomes simpler and shorter due to the more algebraic nature of nonstandard analysis.

In applying nonstandard analysis to the formalization, we first introduce the standard and nonstandard formulations for the basic definitions in the theory. In the next step we prove that the standard and nonstandard definitions are equivalent. The nonstandard equivalents are then applied, whenever appropriate, to produce (often shorter) proofs of standard results. In the next sections, we will illustrate these points by mechanizing basic notions from the theories of limits for real sequences and series, elementary topology on the reals, limits and continuity of functions, and differentiability. We introduce and prove in Isabelle propositions stating that the standard and nonstandard definitions for the various concepts are equivalent.

6.3 Real Sequences and Series

A real sequence $\langle a_n \rangle$ is viewed as a standard function, a, mapping the natural numbers into the reals i.e. $a : \mathbb{N} \to \mathbb{R}$. The notation $a(n)$ is also used to denote a typical term a_n of the sequence.

The function a has a nonstandard extension a^* which maps the hypernaturals into the hyperreals. The $*$-transform of a is thus the function $a^* : \mathbb{N}^* \to \mathbb{R}^*$ where $a^*([\langle X_n \rangle]) = [\langle a(X_n) \rangle]$ for any $[\langle X_n \rangle] \in \mathbb{N}^*$. We therefore define this in a similar fashion to the extension `*f*` for real functions. In Isabelle, the nonstandard extension of a is given by (`*fNat*` a) and defined as

```
*fNat* :: (nat ⇒ real) ⇒ hypnat ⇒ hypreal
*fNat* a N ≡ Abs_hypreal (⋃ X ∈ Rep_hypnat(N). hyprel^^{λn. a(Xn)})
```

As can be seen, the nonstandard extension results in a sequence of hyperreals, meaning that we are actually dealing with a sequence of sequences of real numbers, and this sequence is indexed not by the natural numbers but by the hypernaturals. For this reason, the extended sequence is also known as a **hypersequence**.

Similar theorems to those presented in Section 6.1.4 about `*f*` are proved together with some new ones such as[2]

- (`*fNat*` $(\lambda n.\ a(\texttt{Suc}\ n)))\ N = (\texttt{*fNat*}\ a)\ (N + 1)$

Of particular importance is the theorem (`*fNat*` $a)(\overline{n}) = \widetilde{a(n)}$ which shows that the hypersequence agrees with the original sequence on $\mathbb{N}$; that is for any $n \in \mathbb{N}$, a_n^* is simply the image of a_n in the hyperreals. We recall that we are actually referring to the embedded copy of $\mathbb{N}$ in the hypernaturals and to the embedded copy of $\mathbb{R}$ in the hyperreals. This explains the appearance of the mapping functions `hypnat_of_nat` (in $\overline{n}$) and `hypreal_of_real` (in $\widetilde{a(n)}$), from sections 3.8 and 3.7.1 respectively, in the formulation of the theorem.

6.3.1 On Limits

The hyperreals are now used to define the concept of limit. A few observations about the notation need to be made first. The symbol ∞ is usually used in the real number system to denote that which is potentially arbitrarily large. The expression $\lim_{n\to\infty} a_n$ thus denotes the limiting value of a as n becomes an arbitrarily large natural number. In $\mathbb{R}^*$, the symbol ∞ can be viewed as having a similar meaning but this time, by arbitrarily large, one effectively means a number larger than any *finite* number in $\mathbb{R}^*$. So the expression $\lim_{n\to\infty} a_n^*$ denotes the value a^* approaches as n becomes an arbitrarily large hypernatural number or, more formally, the value infinitely close to a_n^* for any **infinite** hypernatural number n. This motivates the nonstandard definition for sequential limit that is given below.

With regards to the formalization in Isabelle, we decided to follow an approach similar to that used by Harrison in HOL [43] and formulate both a relational and functional form for sequential limits. We declare and define an

[2]From now on, we shall assume, for clarity, that 0 and 1 are overloaded over all the various types of numbers and refrain from using `0r`, `0hr`, `1hr`, etc.

infix 'tends to' relation '----->' and use it to express statements such as a_n tends to l by a_n ----> l. The standard definition used in Isabelle is:

$$X \texttt{ ----> } l \equiv \forall r.\, (0 < r \longrightarrow (\exists N.\, \forall n.\, N \leq n \longrightarrow \texttt{rabs}\ (Xn - L) < r))$$

Our formalization, however, also has a second version of the predicate denoted by '----NS>'; this second notion of convergence is defined using nonstandard concepts and expressed by the following simpler statement not involving any existential quantifiers:

$$X \texttt{ ----NS> } l \equiv (\forall N \in \texttt{HNatInfinite}.\, (\texttt{*fNat*}\ X)N \approx \widetilde{l})$$

The first task is to prove the equivalence of the two definitions. Before coming to this, we briefly make some remarks about the functional form of sequential limit. We declare a constant `lim` and use it to denote the statement $\lim_{n\to\infty} a_n$ by `lim` a (equivalent by η-expansion to `lim` $(\lambda n.\, a_n)$). A nonstandard version of the function is also introduced that is denoted by `nslim`. The following definitions are made, using Hilbert's ϵ-operator to denote the unique limit (if it exists):

- `lim` $a \equiv \varepsilon l.\, a$ ----> ℓ
- `nslim` $a \equiv \varepsilon l.\, a$ ----NS> ℓ

The relational form is effectively used to prove properties about limits rather than the functional form since, as mentioned by Harrison, the latter is less powerful. This is because all functions in HOL and, of course, Isabelle/HOL are total. The interested reader should consult Harrison's PhD thesis for an extended discussion on binders, relational versus functional forms of mathematical statements, and other related issues arising from HOL's lack of partial functions [43]. These points are equally relevant to the aspects of analysis that we have formalized in Isabelle. One last point that is worth noting is that, for a convergent sequence, the following theorem is proved which could be used as an alternative definition for `nslim` (and hence `lim`):

$$\texttt{nslim}\ a = \texttt{st}\ ((\texttt{*fNat*}\ a)\ \Omega)$$

where Ω denotes the infinite hypernatural $[\langle n \rangle]$ defined in Section 3.8. This is an interesting characterization of limit that arises due to the nonstandard framework.

We will now outline the steps needed to prove the equivalence of the standard and nonstandard definitions for limits.

6.3.2 Equivalence of Standard and NS Definitions

Proving the equivalence of the standard and nonstandard (NS) formulations of a particular property is important as it will guarantee that a standard theorem proved using nonstandard methods is true. The proof that the NS definition implies the standard definition is usually the trickier part. We need to go down to the level of the ultrafilter and use the theorems that recast properties such as belonging to the set `Infinitesimal` in terms of membership of $U_{\mathcal{N}}$ (as presented in Section 3.7.2).

THEOREM 6.1. *A sequence $a : \mathbb{N} \to \mathbb{R}$ converges to the real number l as its limit if and only if for each infinitely large hypernatural number $\eta = [\langle m_n \rangle] \in {}^*\mathbb{N} - \mathbb{N}$ we have that a^*_η is infinitely close to l. In symbols, a ----> $l \iff a$ ----NS> l.*

Proof:

1) a ----> $l \Longrightarrow a$ ----NS> l. Assume that the sequence $\langle a_n \rangle$ converges to l. Let $0 < r$ be given and let $\eta = [\langle m_n \rangle]$ be any given infinite hypernatural number. a ----> l implies that there exists a natural number N such that $|a_n - l| < r$ for all $N \leq n$. Now since η is an infinite hypernatural with representative sequence $\langle m_n \rangle$, we know, from Section 3.8, that $N \leq m_n$ for almost all the m_n i.e. $\{n.\ N \leq m_n\} \in U_{\mathbb{N}}$. But, we can also prove that

$$\{n.\ N \leq m_n\} \subseteq \{n.\ |a_{m_n} - l| < r\}$$

from which it immediately follows that $\{n.\ |a_{m_n} - l| < r\} \in U_{\mathbb{N}}$. Thus, given any positive real number r, we have that $|a_{m_n} - l| < r$ for almost all the a_{m_n}. From this it follows (again from Section 3.8) that $a^*_\eta - \widetilde{l}$ is infinitesimal i.e. $a^*_\eta \approx \widetilde{l}$. □

2) a ----NS> $l \Longrightarrow a$ ----> l. Suppose that $\langle a_n \rangle$ does not converge to l. Then, there is some standard real $r > 0$ and a function $f : \mathbb{N} \to \mathbb{N}$ satisfying $n \leq f(n)$ and $r \leq |a_{f(n)} - l|$ for all $n \in \mathbb{N}$. Now, writing $f(n) \equiv f_n$, the sequence $\langle f_n \rangle$ defines a hypernatural number η, which we prove to be infinite. We have $\{n.\ r \leq |a_{f_n} - l|\} \in U_{\mathbb{N}}$ since it coincides with $\mathbb{N}$. Thus, it follows that $a^*_\eta - \widetilde{l}$ is not infinitesimal. □

6.3.3 Remarks on the Proof

There are several points that need to be made about the mechanical proof of the theorem above. As we mentioned already, the first part of the proof was relatively easy to mechanize given that we had already proved various theorems expressing each class of hyperreal numbers in terms of the free ultrafilter: The second part needed several lemmas since it is more complicated. It involves, for example, a Skolemization step that textbook proofs often fail to mention explicitly. This requires the use of the Axiom of Choice, which here can be proved using Hilbert's description operator. It enables the existential quantifier to be pulled across the universal quantifier:

$$\forall x.\ \exists y.\ Q\,x\,y \Longrightarrow \exists f.\ \forall x.\ Q\,x\,(f\,x)$$

This theorem allows us to introduce a function from $\mathbb{N}$ to $\mathbb{N}$— effectively a sequence of natural numbers— that can then be used to define an infinite hypernatural number. The following lemma is thus proved on the way to the main result:

$$\forall n.\ n \leq f\,n \Longrightarrow \texttt{Abs\_hypnat}\ (\texttt{hypnatrel\^\^}\{f\}) \in \texttt{HNatInfinite}$$

Another important observation is that the structure of the proof follows a general pattern that will occur again when we mechanize the equivalence proofs for other properties. Indeed, the need to use AC when proving that a paticular NS

definition implies the standard one is a typical situation. Mechanical theorem proving benefits from re-use of code and of important theorems.

The fact that there is a general pattern in the proofs is not a coincidence and can be related to one of the central features of nonstandard analysis known as the **Transfer Principle**. This provides a context in which true statements about $\mathbb{R}$ are transformed into statements about $\mathbb{R}^*$ through a general procedure. Within a typed logic, this procedure would involve lifting results from the type `real` to the type `hypreal`, from `nat` to `hypnat` or viewed more generally, from any particular type to its extended counterpart.

In the subsequent survey of the development of NSA in Isabelle, we shall state the standard and nonstandard formulations of various concepts but often omit explicit details of the equivalence proof unless they differ considerably in nature from the proof just given. We shall, however, mention any interesting lemmas that were needed, as well as any particular difficulties encountered.

6.3.4 Properties of Sequential Limits

With the nonstandard formulation, the proofs of basic properties of sequences all become trivial. Indeed, their mechanization mostly involves simple algebraic manipulations that can be handled automatically by Isabelle's simplifier. We prove the following theorems:

1) $[|X$ `----NS>` $a; Y$ `----NS>` $b|] \Longrightarrow (\lambda n.\, X\,n + Y\,n)$ `----NS>` $a + b$

2) $[|X$ `----NS>` $a; Y$ `----NS>` $b|] \Longrightarrow (\lambda n.\, X\,n \cdot Y\,n)$ `----NS>` $a \cdot b$

3) X `----NS>` $a \Longrightarrow \lambda n.\, -X\,n$ `----NS>` $-a$

4) $[|X$ `----NS>`$a; a \neq 0|] \Longrightarrow (\lambda n.$ `rinv` $Xn)$ `----NS>` `rinv` (a)

5) $[|X$ `----NS>` $a; X$ `----NS>` $b|] \Longrightarrow a = b$

For the proof of (1) above, for example, we have that $X_n^* \approx \widetilde{a}$ and $Y_n^* \approx \widetilde{b}$, and hence $X_n^* + Y_n^* \approx \widetilde{a} + \widetilde{b}$ for any infinite hypernatural n by Theorem 1 of Section 3.7.3. The proof is done in one step using Isabelle's automatic tactic. The other theorems are all proved as simply, the only slight exception being (4). This requires a bit more work and the following simple lemma shown here with a mixture of conventional and Isabelle notation:

- $X^*(N) \neq 0 \Longrightarrow (\lambda m.$ `rinv` $(X\,m))^*(N) =$ `hrinv` $(X^*(N))$

This result effectively performs the $*$-transform over both the inverse function and the sequence function since `hrinv` $=$ `rinv`*, as mentioned in the previous chapter. Once these various basic properties are proved, we can deal with the important concept of Cauchy sequences and their associated theorems.

6.3.5 Sequences

In this section, we examine some of the important properties of sequences formalized in Isabelle. We first examine the concept of a bounded sequence.

Boundedness and Monotonicity

We define the standard and nonstandard notions of a bounded sequence as follows:

$$\texttt{Bseq}\ X \equiv \exists K.\ (\texttt{0r} < K \wedge \forall n.\ \texttt{rabs}\ (X\,n) \leq K)$$
$$\texttt{NSBseq}\ X \equiv \forall N \in \texttt{HNatInfinite}.\ (\texttt{*fNat*}\ X)\,N \in \texttt{Finite}$$

The equivalence of the standard and nonstandard definitions for boundedness is first proved, thereby making two characterizations of the concept available for use in our proofs. The NS definition, NSBseq, makes it immediately obvious that boundedness is a necessary condition for convergence i.e. we have the following theorem:

$$\texttt{NSconvergent}\ X \Longrightarrow \texttt{NSBseq}\ X$$

where

$$\texttt{NSconvergent}\ X \equiv (\exists l.\ X\ \texttt{----NS>}\ l)$$

This reduces, in Isabelle, to proving the following (simple) goal:

$$\exists l.\ \forall N \in \texttt{HNatInfinite}.\ (\texttt{*fNat*}\ X)\,N \approx \widetilde{l}$$
$$\Longrightarrow \forall N \in \texttt{HNatInfinite}.\ (\texttt{*fNat*}\ X)\,N \in \texttt{Finite}$$

Proof: Suppose that $\langle X_n \rangle$ converges to some $\alpha \in \mathbb{R}$ then $X_n^* \approx \widetilde{\alpha}$ for every infinite hypernatural n and must therefore be finite by the following lemma:

$$[|x \in \texttt{Finite}; x \approx y|] \Longrightarrow y \in \texttt{Finite}$$

The theorem is proved in one step by Isabelle's `blast_tac`. We also prove that boundedness is a sufficient condition for convergence provided a given sequence is **monotone**:

$$[|\texttt{Bseq}\ X; \texttt{monoseq}\ X|] \Longrightarrow \texttt{convergent}\ X$$

where the monotonicity of a sequence X is defined by

$$\texttt{monoseq}\ X \equiv\ ((\forall (m{::}\,\texttt{nat})\,n.\ m \leq n \longrightarrow Xm \leq Xn)\ \vee$$
$$(\forall m\,n.\ m \leq n \longrightarrow Xn \leq Xm))$$

The proof of the above theorem proceeds through a mixture of both standard and nonstandard arguments: for some of the lemmas needed, it is easier to prove a standard version rather than a nonstandard one. This is the case for the following result, for example:

$$\forall n.\ m \leq n \longrightarrow X\,n = X\,m \Longrightarrow \exists l.\ X\ \texttt{---->}\ l$$

The standard proof is trivial since the variables are easy to instantiate by a routine examination of the goal. Isabelle's automatic tactic then proves the theorem without difficulty. A nonstandard proof, however, would require proving a more demanding theorem:

$$\forall n.\ m \leq n \longrightarrow X\,n = X\,m \Longrightarrow \exists l.\ \forall N \in \texttt{HNatInfinite}.\ (\texttt{*fNat*}\ X)N \approx \widetilde{l}$$

This is one of the few cases where we have noticed that a nonstandard proof seems to be more complicated that its standard counterpart. The main difficulty here lies in finding the right instantiation for the existential variable. This happens to be easier for the standard theorem in this particular case.

Cauchy Sequences

The following statements are equivalent

1) **Convergence.** The sequence $\langle a_n \rangle$ converges i.e. $\exists l.\ a_n$ ----> l.

2) **Hyperreal Cauchy Condition.** For all infinite hypernatural numbers N and M, $a_N^* \approx a_M^*$.

3) **Real Cauchy Condition.** For all $0 < \epsilon$ there is an integer M such that for all $m, n \geq M$, $|a_m - a_n| < \epsilon$.

The standard proof that a sequence is Cauchy if and only if it is convergent can be obtained from most traditional textbooks on analysis. Harrison [43], for example, uses the proof from Burkill and Burkill [14] in HOL. Although, the mechanization is reported as being a direct formalization in HOL, Harrison's proof is rather complicated and long. This is partly due to difficulties inherent in finding the right instantiations for variables in ϵ–δ proofs, especially since HOL does not allow unknown variables whose instantiations can be delayed. Owing to this problem, Harrison suggests that Isabelle might provide a more natural environment for ϵ–δ proofs since it allows unknown variables to propagate and be instantiated later in the proof. Although this in itself seems a reasonable argument, we actually go one step further by using nonstandard arguments: our formalization avoids the need for ϵ–δ arguments altogether.

To prove the Cauchy criterion for convergence, Burkill and Burkill, and hence Harrison in HOL, define the extra notion of a subsequence. They then prove that every sequence has a monotonic subsequence. Although the main theorem is not difficult to reach once this result and a few other lemmas have been set up, one might feel that the need for various auxiliary notions (Harrison also needs to define a 'reindexing' function in his formalization, for example) diverts attention from what is actually being proved. The need to introduce and use the properties of subsequences is not immediately obvious to anyone trying to prove the theorem (without the help of a textbook, for example).

Our formalization avoids the notion of a subsequence and goes for a direct and more intuitive proof. First we prove the equivalence of the real (standard) and hyperreal (nonstandard) Cauchy conditions. This proceeds in a similar way to that of Theorem 6.1. The formalization is, in fact, more straightforward in this particular case and we shall not go into the details. With this equivalence set up, the proof of the main result is simple and direct within the nonstandard framework.

THEOREM 6.2. *The sequence $\langle X_n \rangle$ converges if and only if it is a Cauchy sequence.*

Proof: If $\langle X_n \rangle$ converges to l then $X_n^* \approx \widetilde{l} \approx X_m^*$ for all infinite n and m by the NS definition of convergence; so $\langle X_n \rangle$ is a Cauchy sequence by the NS definition of Cauchy criterion.

Conversely, if $\langle X_n \rangle$ is a Cauchy sequence then $\langle X_n \rangle$ is bounded and so X_n^* is finite for all infinite n. Therefore, using the Standard Part Theorem, there exists a standard (embedded) real number l infinitely close to X_Ω^* where Ω is our infinite hypernatural number from Section 3.8. Thus, we have that $X_n^* \approx X_\Omega^* \approx l$ for all infinite n (NS Cauchy criterion), and so $\langle X_n \rangle$ converges to l (NS formulation for convergence). □

This formalization is simple and short as was pointed out previously. One lemma, also needed by Harrison in HOL, requires proving that every Cauchy sequence is bounded. We actually use the nonstandard version of this theorem involving the hyperreal formulations of both the Cauchy and boundedness properties.

As a historical note, it is interesting to observe that though infinitesimals do not appear in the standard definition of Cauchy convergence, Cauchy used them as a tool in his *Cours d'analyse* (1821) [58]. Indeed, Cauchy explicitly states the following as an alternative version of convergence: "in other words, it is necessary and sufficient that, for infinitely large values of the number n, the sums $s_n, s_{n+1}, s_{n+2}, \ldots$ differ from the limit s, and consequently among themselves, by infinitely small quantities". Reinterpreted, within the context of nonstandard real analysis, this corresponds exactly to the hyperreal Cauchy condition. Laugwitz (further) mentions that Euler was the first one, much earlier in 1735, to state that $s_n - s_m$ be infinitesimal for infinitely large m, n was a necessary and sufficient condition for convergence [58]. Such use of infinitesimals, especially by the rigorous Cauchy, gives yet another indication of their power as a tool in analysis throughout centuries.

An important general result proved in our theory of sequences concerns the existence of the n-th root of any positive real numbers[3]. For any positive real number a and natural number n, there exists a unique positive real number r such that $r^n = a$. The proof proceeds by considering the set $S = \{x \in \mathbb{R}.\, x^n \leq a \wedge 0 < x\}$ and showing that S is non-empty and bounded above. The number r is then shown to be the supremum of S. With this theorem formalized, we can, for example, define the square root function and its nonstandard extension in Isabelle.

Sequences and Hyperreals

There is, as expected, a close relationship between the various properties of sequences and hyperreal numbers. Indeed since the development of the hyperreals has been based on the use of sequences of real numbers, we can prove the following theorems:

- If $\langle a_n \rangle$ is bounded then $[\langle a_n \rangle]$ is finite; expressed as a theorem of Isabelle we have,

$$\texttt{NSBseq}\ X \Longrightarrow \texttt{Abs\_hypreal}\ (\texttt{hyprel``\{}X\texttt{\}}) \in \texttt{Finite}$$

- If $\langle a_n \rangle$ converges to zero then $[\langle a_n \rangle]$ is an infinitesimal.
- If $\langle a_n \rangle$ is an unbounded sequence then $[\langle a_n \rangle]$ is an infinite hyperreal.

6.3.6 Series

In standard analysis an infinite series is the limit of a sequence of finite sums. Despite the notation

$$\sum_{i=0}^{\infty} a_i$$

[3]The proof mechanized in Isabelle is from J. L. Orr's *Webnotes* found on the Web at `http://www.math.unl.edu/~webnotes/contents/chapters.htm`

one does not try in classical analysis to interpret it literally as an infinite number of additions. Instead, one considers the sums of finitely many of the terms of the series, and examines the behaviour of such sums as an increasingly large, but still finite, number of terms are allowed. Using our framework, however, it is possible to use the nonstandard criterion for sequential convergence to define *literally* infinite sums.

Infinite Sums and Infinite Series

Given a real sequence (f_n), we define the standard notion of finite sum $(\sum_{i=m}^{n-1} f_i)$ using Isabelle's **recdef** package, which implements well founded recursion:[4]

$$\begin{array}{ll} \texttt{consts} & \texttt{sumr} :: (\texttt{nat} * \texttt{nat} * (\texttt{nat} \Rightarrow \texttt{real})) \Rightarrow \texttt{real} \\ \texttt{recdef} & \texttt{sumr measure } (\lambda(m,n,f).\, n) \\ & \texttt{sumr } (m, 0, f) = 0 \\ & \texttt{sumr } (m, \texttt{Suc } n, f) = \texttt{ if n < m then } 0 \\ & \qquad\qquad\qquad\qquad \texttt{else sumr } (m,n,f) + f(n) \end{array}$$

The first line declares `sumr` to be a constant. The well-founded relation is `measure` $(\lambda(m,n,f).\, n)$ which is given to show that the argument of `sumr` is "smaller" at each recursive call, and hence that it terminates [68]. The operator `measure` is part of a suite of operators recognized and used by Isabelle/HOL to automatically prove that the constructed relation is well-founded. The function "$\lambda(m,n,f).\, n$" is called a **measure function** which specifies that the recursion terminates because n decreases.

The `recdef` definition also provides an induction rule specialized for `sumr` which enables direct proofs of the theorem. The expected theorems about finite sum are then all derived. We shall not list them here but instead describe how the canonical nonstandard extension of `sumr` is defined.

Consider a sequence of finite sums: this constitutes a mapping from $\mathbb{N}$ to $\mathbb{R}$ which has a unique nonstandard extension defined, for any infinite hypernatural numbers $M = [\langle X_n \rangle]$ and $N = [\langle Y_n \rangle]$, as

$$\sum_{i=M}^{N} {}^{*} a_i \equiv \left[\left\langle \sum_{i=X_1}^{Y_1} a_i, \sum_{i=X_2}^{Y_2} a_i, \sum_{i=X_3}^{Y_3} a_i, \ldots \right\rangle \right] \tag{6.1}$$

This enables one to talk of the sum being taken to N terms (M can be set to 0), where N is any hypernatural number. The value of such an *infinite* sum is a hyperreal number which depends on the number of terms taken. The formalization of the nonstandard extension in (6.1) is given in Isabelle by

$$\begin{array}{l} \texttt{sumhr} :: (\texttt{hypnat} * \texttt{hypnat} * (\texttt{nat} \Rightarrow \texttt{real})) \Rightarrow \texttt{hypreal} \\ \texttt{sumhr } p \equiv (\lambda(M,N,f). \\ \qquad \texttt{Abs\_hypreal}(\bigcup X \in \texttt{Rep\_hypnat } M. \\ \qquad\qquad \bigcup Y \in \texttt{Rep\_hypnat } N. \\ \qquad\qquad \texttt{hyprel``}\{\lambda n.\, \texttt{sumr } ((Xn),(Yn),f)\}))\, p \end{array}$$

As is usual in such cases, the corresponding simplification theorem is proved; it

[4] This function could also be defined using Isabelle's primitive recursion (`primrec`) package.

can be added to Isabelle's simplifier when needed:

$$\begin{aligned}&\texttt{sumhr}\ (\texttt{Abs\_hypnat}\ (\texttt{hypnatrel\^\^}\{\lambda n.\ X\,n\}),\\ &\qquad \texttt{Abs\_hypnat}\ (\texttt{hypnatrel\^\^}\{\lambda n.\ Y\,n\}), f)\\ &\qquad = \texttt{Abs\_hypreal}\ (\texttt{hyprel\^\^}\{\lambda n.\ \texttt{sumr}\ (X\,n, Y\,n, f)\})\end{aligned}$$

Using this definition, theorems similar to the two reduction rules in the recursive definition of sumr are proved:

$$\begin{aligned}&\texttt{sumhr}\ (m, 0, f) = 0\\ &\texttt{sumhr}\ (m, n+1, f) = \ \texttt{if}\ n < m\ \texttt{then}\ 0\\ &\qquad\qquad\qquad\qquad\ \texttt{else}\ \texttt{sumhr}\ (m, n, f) + (\texttt{*fNat*}\ f)\,n\end{aligned}$$

The nonstandard extension with its possible infinite hypernatural limits, preserves the formal behaviour of finite summation. In fact, with the help of the theorems just introduced, the properties of the finite sum are directly transferred from sumr to sumhr. A few of the theorems proved in Isabelle are:

1) $\texttt{sumhr}\ (m, n.f) + \texttt{sumhr}\ (m, n, g) = \texttt{sumhr}\ (m, n, \lambda i.\ f\,i + g\,i)$

2) $\texttt{sumhr}\ (0, \Omega, \lambda i.\ 1) = \omega - 1$

3) $\texttt{sumhr}\ (m, n, \lambda i.\ r \cdot (f\,i)) = \widetilde{r} \cdot \texttt{sumhr}\ (m, n, f)$

4) $n < p \Longrightarrow \texttt{sumhr}\ (0, n, f) + \texttt{sumhr}\ (n, p, f) = \texttt{sumhr}\ (0, p, f)$

5) $\texttt{hrabs}\ (\texttt{sumhr}\ (m, n, f)) \leq \texttt{sumhr}\ (m, n, \lambda i.\ \texttt{rabs}\ (f\,i))$

6) $(\forall r.\ m \leq r \wedge r < n \longrightarrow f\,r = g\,r)$
$\Longrightarrow\ \texttt{sumhr}\ (\overline{m}, \overline{n}, f) = \texttt{sumhr}\ (\overline{m}, \overline{n}, g)$

7) $\texttt{sumhr}\ (0, 2\Omega, \lambda i.\ (-1)^{\texttt{Suc}\ i}) = 0$

8) $\texttt{sumhr}\ (0, 2\Omega - 1, \lambda i.\ (-1)^{\texttt{Suc}\ i}) = 1$

9) $\texttt{sumhr}\ (0, N, f) = (\texttt{*fNat*}\ (\lambda n.\ \texttt{sumr}\ (0, n, f)))N$

In theorem (2), Ω refers to the infinite hypernatural $[\langle 0, 1, 2, \ldots\rangle]$ defined in Section 3.8.1, while ω refers to the infinite hyperreal $[\langle 1, 2, \ldots\rangle]$ defined in Section 3.7. The sum involved in this theorem can thus be literally taken as infinite. It is proved by observing that, according to the definitions formalized in Isabelle,

$$\begin{aligned}\sum_{i=0}^{\Omega}{}^{*} 1 &= \left[\left\langle \sum_{i=0}^{0} 1, \sum_{i=0}^{1} 1, \sum_{i=0}^{2} 1, \ldots \right\rangle\right]\\ &= [\langle 0, 1, 2, \ldots\rangle]\\ &= [\langle 1, 2, 3, \ldots\rangle] - [\langle 1, 1, 1, \ldots\rangle]\\ &= \omega - 1\end{aligned}$$

Of the other theorems shown, (9) is perhaps the best illustration that sumhr is the nonstandard extension of sumr. It shows how the framework naturally extends any standard function (of a single variable), enabling it to take a nonstandard argument. This theorem is important to the derivation of results about

convergence of series. Theorems (7) and (8) illustrate the comment made above that the value of the infinite sum depends on the number of terms taken.

Following Harrison [43], a relation sums is defined to denote that an infinite series converges to some limit a as its sum. An infinite series $\sum_{i=0}^{\infty} f_i$ 'sums to' some real number a if and only if the sequence of *partial sums* $\sum_{i=0}^{n} f_i$ converges to a as its limit. This provides the following definition in Isabelle:

$$f \texttt{ sums } a \equiv (\lambda n.\ \texttt{sumr}\ (0, n, f)) \texttt{ ----> } a$$

Hence, it also follows that the infinite series is convergent if and only if the sequence $(\lambda n.\ \texttt{sumr}\ (0, n, f))$ is a Cauchy sequence.

In nonstandard terms, the definition of a convergent series is given by

DEFINITION 6.3.1. *The infinite series defined by the sequence $\langle f_n \rangle$ is said to converge if there exists some real number a such that for every infinite hypernatural number N,*

$$\sum_{i=0}^{N}{}^{*} f_i \approx a$$

In Isabelle, this definition becomes

$$f \texttt{ NSsums } a \equiv (\forall N \in \texttt{HNatInfinite}.\ \texttt{sumhr}\ (0, N, f) \approx \widetilde{a})$$

Form this definition, the following theorems are proved:

- A necessary and sufficient condition for an infinite series to converge is that for any two infinite hypernatural numbers M and N, we have

$$\sum_{i=0}^{M}{}^{*} f_i \approx \sum_{i=0}^{N}{}^{*} f_i$$

  Or equivalently in Isabelle,

$$\exists a.\ f \texttt{ NSsums } a \iff \forall M \in \texttt{HNatInfinite}.\ \forall N \in \texttt{HNatInfinite}.\ \texttt{sumhr}\ (0, M, f) \approx \texttt{sumhr}\ (0, N, f)$$

- The theorem above is also expressed in an alternative form using result (4) from the list of theorems given about sumhr:

$$\exists a.\ f \texttt{ NSsums } a \iff \forall M \in \texttt{HNatInfinite}.\ \forall N \in \texttt{HNatInfinite}.\ M < N \longrightarrow \texttt{sumhr}\ (M, N, f) \approx 0$$

As we have seen, NSA does indeed simplify the treatment of real sequences and infinite series. As a further benefit, the nonstandard extension of sums enables us to treat finite and infinite series in a homogeneous fashion. There is no need to use ∞ as a purely notational device in defining infinite series: it is now possible to take the sum to N terms, where N can be a natural number or an infinite hypernatural. In a sense, the ∞ symbol now stands for any member of HNatInfinite.

6.4 Some Elementary Topology of the Reals

We now survey the development of some basic topology on the reals in Isabelle. The aim of this formalization is to see the benefits that might be gained using nonstandard analysis when dealing with elementary topological notions such as open sets and neighbourhoods.

6.4.1 Neighbourhoods

We begin by giving the standard and nonstandard definitions of the **neighbourhood** of a point. For the standard definition, the concept of a **ball** is first defined. If a is any point in $\mathbb{R}$ and r is any real number, then the set of all real points x whose distance from a is less than r is defined as

$$\texttt{rBall}\ a\ r \equiv \{x.\ \texttt{rabs}\ (a - x) < r\}$$

DEFINITION 6.4.1. **Standard Neighbourhood.** *A set $M \subseteq \mathbb{R}$ is said to be a neighbourhood of point $a \in \mathbb{R}$ if and only if there exists some $r > 0$ such that*

$$\texttt{rBall}\ a\ r \subseteq M$$

Expressing this in Isabelle, we have

$$\texttt{isnbhd}\ a\ M \equiv \exists r.\ 0 < r \wedge \texttt{rBall}\ a\ r \subseteq M$$

The nonstandard formulation, on the other hand, is given by the following

DEFINITION 6.4.2. **Nonstandard Neighbourhood.** *A set $M \subseteq \mathbb{R}$ is said to be a neighbourhood of point a if and only if every hyperreal x infinitely close to a belongs to the nonstandard extension M^* of M.*

In Isabelle, this is formalized as

$$\texttt{isNSnbhd}\ a\ M \equiv \texttt{monad}\,(\widetilde{a}) \subseteq *\texttt{s}*\ M$$

As can be seen, the concept of a monad enables the definition to be expressed concisely. The monad is a set of hyperreals, formally defined by[5]

$$\texttt{monad}\ x \equiv \{y.\ x \approx y\}$$

The next step, as usual, is to prove the equivalence of the two definitions as a theorem in Isabelle. The proof is mechanized without much difficulty with the help of result 7 from Section 6.1.2. This lemma is necessary to prove that the NS definition implies the standard one. The formulations are next used to deal with the notion of open sets.

6.4.2 Open Sets

A subset G of $\mathbb{R}$ is said to be **open** if and only if G is a neighbourhood of each of its points. This leads to the following direct formalization of the standard and nonstandard characterizations:

$$\texttt{isOpen}\ G \equiv \forall a \in G.\ \texttt{isnbhd}\ a\ G$$
$$\texttt{isNSOpen}\ G \equiv \forall a \in G.\ \texttt{isNSnbhd}\ a\ G$$

[5]The name monad was originally chosen as a tribute to Leibniz.

The equivalence proof follows trivially from that of neighbourhood. The theorems given next are all proved *automatically* since they are direct consequences of the results about boolean operations on nonstandard extensions of sets (Section 6.1.2):

1) $[|\texttt{isOpen}\ A; \texttt{isOpen}\ B|] \Longrightarrow \texttt{isOpen}\ A \cap B$

2) $[|\texttt{isOpen}\ A; \texttt{isOpen}\ B|] \Longrightarrow \texttt{isOpen}\ A \cup B$

3) $[|\texttt{isOpen}\ A|] \Longrightarrow \texttt{isOpen}\ (\bigcup A)$

4) $\texttt{isOpen}\ (\texttt{UNIV} :: \texttt{real set})$

5) $\texttt{isOpen}\ \emptyset$

By contrast, and as an example, a *standard* proof in Isabelle that open sets are closed under finite intersections requires several steps including an explicit instantiation of variables, a case split, and the use of the following lemma:

$$[|r_1 < r_2; x \in \texttt{rBall}\ a\ r_1|] \Longrightarrow x \in \texttt{rBall}\ a\ r_2$$

The gain from using Nonstandard Analysis seems once more obvious. In this development of elementary real topology, several other concepts such as closed sets, limit points, and derived sets are also introduced. Their various properties are formalized and in most cases the proofs are automatic. One of the main results to be formalized in this theory using a nonstandard approach is the Bolzano-Weierstrass theorem. Its nonstandard proof, as given by Hurd [49] is extremely short and simple compared to the standard proof.

6.5 Limits and Continuity

There are several notions of limits that share a number of common theorems (such as uniqueness, for example). It is clear that an efficient mechanization of standard analysis should seek to limit proof replication by developing a generic treatment of limits. Harrison uses the well known theory of convergence nets to prove a number of general theorems that can then be specialized to fit each notion of limit [43].

Since the development, however, involves standard theorems about limits using a nonstandard approach, we did not initially feel a need for such a streamlined generic treatment of limits. Moreover, this is only an initial investigation into the benefits gained from working in the hyperreals. So there is scope for further improvement. An interesting idea would be to seek a generalization for the nonstandard theory of limits as well. However, since we are already working with much simpler and more algebraic formulations than in the standard case, the gains might not be worth the trouble. After all, as we noticed in our development, having independent notions of sequential and pointwise limits does not represent a lot of extra work since the proofs of similar properties are all done automatically. Having said this, it is probably wise in any mechanization to favour the approach that cuts down on work. So, when trying to prove addition of limits, for example, we might like to have a general lemma, in the spirit of convergence nets, that can be specialized when needed. This would prevent us

from having two similar-looking theorems like the ones below that were used for sequential and pointwise limits respectively:

$$[|(*\texttt{fNat}*\ f)\,x \approx l; (*\texttt{fNat}*\ g)\,x \approx m|] \\ \Longrightarrow (*\texttt{fNat}*\ (\lambda y.\ f\,y + g\,y))\,x \approx l + m$$
$$[|(*\texttt{f}*\ f)\,x \approx l; (*\texttt{f}*\ g)\,x \approx m|] \\ \Longrightarrow (*\texttt{f}*\ (\lambda y.\ f\,y + g\,y))\,x \approx l + m$$

All this falls under the more general concept of preservation of properties across nonstandard extensions. We would like to prove general properties that hold for all nonstandard extensions of functions rather than deal with specific cases like those above. Textbooks usually state the properties that we presented in Section 6.1.4 as general results that apply to all extensions. In our case, since we extend each type of function explicitly, we need to prove similar properties each time.

Let us now return to the standard and nonstandard characterizations of the notions of pointwise limits. A function f is said to have a limit l as x approaches a point a if and only if for any given $\epsilon > 0$, there exists a $\delta > 0$ such that for every value of x satisfying the inequality $0 < |x - a| < \delta$, we have $|f(x) - l| < \epsilon$. This is the standard $\epsilon - \delta$ definition for the limit of a function at a given point. The conventional notation used when this condition is satisfied is $\lim_{x \to a} f(x) = l$. We will, however, use a relational approach in this case as well and denote the condition by f -- a --> l for the standard case, and by f -- a --NS> l for the nonstandard one. In Isabelle, we have:

$$f \texttt{ -- } a \texttt{ --> } l \equiv \forall \epsilon.\ 0 < \epsilon \longrightarrow (\exists \delta.\ 0 < \delta \wedge \\ (\forall x.\ 0 < \texttt{rabs}\ (x - a) \wedge \texttt{rabs}\ (x - a) < \delta \\ \longrightarrow \texttt{rabs}\ (f\,x - l) < \epsilon))$$

The nonstandard definition, once again, is more concise and captures the intuition behind the notion:

$$f \texttt{ -- } a \texttt{ --NS> } l \equiv \forall x.\ x \neq \widetilde{a} \wedge x \approx \widetilde{a} \longrightarrow (*\texttt{f}*\ f)\,x \approx \widetilde{l}$$

The equivalence of the two definitions is not too difficult to prove and has the same structure as the other similar proofs. We make use of a few lemmas such as

- $\forall n.\ \texttt{rabs}\ (X\,n - x) < \texttt{rinv}\ (n) \Longrightarrow$
  $(\texttt{Abs\_hypreal}\ (\texttt{hyprel}\hat{}\hat{}\{X\}) - \widetilde{x}) \in \texttt{Infinitesimal}$

which enables us to define a hyperreal infinitely close to a real number x given a real sequence converging towards that number.

We prove properties analogous to those presented in Section 6.3.4. In this part of the formal investigation, however, we decided to prove some of the properties twice: first using only the standard approach and then using the nonstandard approach. The aim was to examine more closely the gains from using nonstandard analysis in terms of the number of steps required to complete each proof, instantiations of variables, and theorems used. If we consider, for example, the formalization of the addition property,

$$[|f \texttt{ -- } a \texttt{ --NS> } l; g \texttt{ -- } a \texttt{ --NS> } m|] \Longrightarrow (\lambda x.\ f(x) + g(x)) \texttt{ -- } a \texttt{ --NS> } l + m$$

a few interesting remarks can be made:

- The nonstandard proof expands the definitions and is completed automatically in one step (0.08 seconds):

```
Goalw [NSLIM_def]
"[| f -- x --NS> l; g -- x --NS> m |]
 ==> (%x. f(x) + g(x)) -- x --NS> (l + m)";
 by (auto_tac (claset() addSIs [starfun_add_inf_close],
               simpset() addsimps [hypreal_real_add]));
```

  while the standard proof, with our direct formalization, takes some 15 steps.

- We need to give instantiations of variables in a large number of steps for the standard proof — the level of automation is thus fairly low and requires the user paying attention to a lot of details. Moreover, there is the added difficulty of deciding what the instantiation should be and dealing with a three-way case split arising from the linearity of the reals.

- The standard proof requires theorems about transitivity of the ordering relation, absolute function (triangle inequality theorem), monotonicity of the ordering relation under addition, etc., while the nonstandard proof only needs a theorem about monotonicity of the infinitely close relation under addition, and one about the preservation of the addition operation by the embedding function for the reals. Both of these are supplied to Isabelle's automatic tactic as just shown.

Therefore, we notice that the nonstandard proof offers a clear gain in terms of automation. The user is freed from some of the more difficult and tedious steps through the use of the simpler formalization.

In addition, theorems such as

- f `--` a `-->` $l \Longleftrightarrow (\lambda h.\, f(a+h))$ `--` 0 `-->` l

are simple to prove using the nonstandard formulation. This is a useful lemma that can be used to simplify theorems about continuity and differentiability, for example. Next, the standard notion of continuity is examined. A standard real function f is continuous at a point a when $f(x)$ tends to $f(a)$ as x tends to a. In Isabelle,

$$\texttt{isCont}\ f\ a \equiv (f\ \texttt{--}\ a\ \texttt{-->}\ f\ a)$$

We give again the nonstandard definition of continuity that we mentioned in Section 6.2. A standard real function f is *continuous* at the point a if and only if $f^*(x)$ is infinitely close to $f(a)$ for every hyperreal x infinitely close to a. Expressed formally in Isabelle,

$$\texttt{isNSCont}\ f\ a \equiv (\forall x.\, x \approx \widetilde{a} \longrightarrow (\texttt{*f*}\ f)\, x \approx \widetilde{f(a)})$$

Once again, the formalization makes it explicit that the definition is referring to the embedded copies of a and $f(a)$ in the hyperreals. The equivalence of the two definitions follows immediately from that of standard and nonstandard limits. A number of useful theorems are proved immediately. Examples are:

1) `isNSCont` $f\ a \Longleftrightarrow (f$ `--` a `--NS>` $f\ a)$

2) $\texttt{isCont}\ f\ a \Longleftrightarrow (\lambda h.\ f(a+h)\ \texttt{--}\ 0\ \texttt{-->}\ f\ a)$

We also have two distinct ways of proving the usual theorems about continuous functions:

i) As results of the corresponding theorems for pointwise limits. This is a conventional approach and, though (some of) the limit theorems themselves might have been proved using NSA, the process is wholly standard.

ii) As simple algebraic consequences of the nonstandard formulation of continuity. This approach bypasses the limit results — one of the main achievements of NSA in general — and provides alternative simple proofs. Moreover, it has an added power. It can prove at least one elementary result — the composition of continuous functions — that does not follow from limit theorems. This is examined next.

We prove that the sum, product, and division of continuous functions are also continuous. These are results that can be proved by either of the two ways above. We also prove that the composition of continuous functions is continuous:

$$[|\texttt{isCont}\ f\ a; \texttt{isCont}\ g\ (f\ a)|] \Longrightarrow \texttt{isCont}\ (g\ o\ f)\ a$$

Proof: If $x \approx \widetilde{a}$ then $f^*(x) \approx \widetilde{f(a)}$, and so it follows that $g^*(f^*(x)) \approx g(\widetilde{f(a)})$. □

This result is proved automatically by Isabelle's `auto_tac`. Contrast with Harrison's corresponding one in HOL, which is longer and required instantiating $\epsilon - \delta$ properties. In a sense, this also hints at another powerful aspect of nonstandard techniques in mechanical theorem proving: their simple algebra enables them to deal uniformly with a wide range of theorems. The standard approach, on the other hand, required Harrison to go back to a direct formalization in HOL because the theorem does not follow from any of the results about limits. An analogous difficulty occurs if the standard treatment is used to formalize the chain rule of differentiation.

Using the nonstandard framework, it is an interesting exercise to prove more involved theorems such as the following topological characterization of continuity. A function f is continuous on $\mathbb{R}$ if and only if the inverse image $\{x \in \mathbb{R}.\ f(x) \in A\}$ of any open set A is itself always an open set. In Isabelle, the following theorem is proved without any difficulties:

$$(\forall x.\ \texttt{isCont}\ f\ x) \Longleftrightarrow (\forall A.\ \texttt{isOpen}\ A \longrightarrow \texttt{isOpen}\ \{x.\ f(x) \in A\})$$

The formalization of all the theorems about limits, topological notions and so on only needs to use the theorems about the free ultrafilter directly when we are proving the equivalence of the standard and nonstandard definitions. All the other theorems are proved at the higher, more intuitive algebraic level. The equivalence theorems are essential because the standard formulations are the ones that are in widespread use. With the success and widening acceptance of NSA, it might be that in a few decades the so-called nonstandard definitions will become the established ones.

Using the various continuity theorems, nonstandard proofs are produced for some important results of real analysis. These include:

Intermediate Value Theorem. If f is continuous on the closed interval $[a,b]$ and $f(a) < d < f(b)$ for some d, then there exists a c between a and b with $f(c) = d$. The proof formalized in Isabelle is given by Hurd [49] and proceeds through nonstandard arguments. It considers the points $x_k = a + k(a-b)/n, 0 \leq k \leq n$ and the values of f at x_k. The proof then proceeds through a Skolemization step and a $*$-transform to get to the result directly.

Extreme Value Theorem. If f is continuous on the closed and bounded interval $[a,b]$, then there exists a c between a and b so that $f(x) \leq f(c)$ for all x between a and b. The proof is also provided by Hurd and proceeds elegantly and succinctly using arguments similar to the ones above. The points $x_{n,k} = a + k(b-a)/n, 0 \leq k \leq n$ are considered this time.

More theorems about elementary analysis are developed by considering the important notion of differentiability. Its nonstandard treatment is examined next.

6.6 Differentiation

The development of the theory of differentiation which follows using results of the previous section can now be surveyed. The standard formulation states that a function f has a derivative d at a point x if $(f(x+h) - f(x))/h \to d$ as $h \to 0$. In Isabelle, we formalize the relational definition $\texttt{DERIV}(x)\ f :> d$ meaning 'the derivative of f at x is d' as

$$\texttt{DERIV}(x)\ f :> d \equiv (\lambda h.\ (f(x+h) - f(x)) \cdot \texttt{hrinv}\ h)\ \texttt{--}\ 0\ \texttt{-->}\ d$$

The notation $\texttt{DERIV}(x)$ can be regarded as a variation of the Leibniz notation and standing for d/dx. We prove this equivalent form of the standard definition:

$$\texttt{DERIV}(x)\ f :> d \Longleftrightarrow (\lambda z.\ (f(z) - f(x)) \cdot \texttt{rinv}\ (z - x))\ \texttt{--}\ x\ \texttt{-->}\ d \qquad (6.2)$$

The nonstandard definition is stated as

$$\begin{aligned}\texttt{NSDERIV}(x)\ f :> d \equiv\ & \forall h \in \texttt{Infinitesimal} - \{0\}. \\ & ((\texttt{*f*}\ f)(\widetilde{x} + h) - \widetilde{f(x)}) \cdot \texttt{hrinv}\ h \approx \widetilde{d}\end{aligned}$$

We first prove that this nonstandard definition can also be given in terms of limits exactly as the standard definition. The proof does not cause much difficulty and, from it, we have immediately that the two definitions of derivative are equivalent. In addition, using Theorem 6.2, we provide a second useful nonstandard characterization for the differentiability of a function f at a point x:

$$\begin{aligned}\texttt{NSDERIV}(x)\ f :> d \Longleftrightarrow\ & \forall y.\ y \approx x \wedge y \neq x \longrightarrow \\ & ((\texttt{*f*}\ f)(y) - \widetilde{f(x)}) \cdot \texttt{hrinv}\ (y - \widetilde{x}) \approx \widetilde{d}\end{aligned}$$

We then proceed to prove standard results in an extremely simple fashion. For example, we prove that a function f differentiable at a point x is continuous at that point,

$$\texttt{NSDERIV}(x)\ f :> d \Longrightarrow \texttt{isNSCont}\ f\ x$$

This is a simple algebraic theorem using the nonstandard formulation since $f^*(\widetilde{x}+h) - \widetilde{f(x)} \approx \widetilde{d} \cdot h$ for all $h \approx 0$, and so $f^*(\widetilde{x}+h) \approx \widetilde{f(x)}$ i.e. f is continuous at x.

A functional form is also defined for the derivative using the standard part function and the non-zero infinitesimal ϵ defined previously:

$$\texttt{nsderiv}(x)\ f \equiv \texttt{st}\ (((\texttt{*f*}\ f)(\widetilde{x}+\epsilon) - \widetilde{f(x)}) \cdot \texttt{hrinv}\ \epsilon)$$

The right hand side of the definition is sometimes known as the *slope* of the real function f [52].

6.6.1 Standard Properties of Derivatives

We prove the familiar rules about the differentiation of simple functions and their combination:

- $\texttt{NSDERIV}(x)\ (\lambda x.\ k) :> 0$
- $\texttt{NSDERIV}(x)\ f :> d \Longrightarrow \texttt{NSDERIV}(x)\ (\lambda y.\ c \cdot f(y)) :> c \cdot d$
- $\texttt{NSDERIV}(x)\ f :> d \Longrightarrow \texttt{NSDERIV}(x)\ (\lambda y.\ -f(y)) :> -d$
- $[|\texttt{NSDERIV}(x)\ f :> d; \texttt{NSDERIV}(x)\ g :> e|] \Longrightarrow \texttt{NSDERIV}(x)\ (\lambda y.\ f(y) + g(y)) :> d + e$
- $[|\texttt{NSDERIV}(x)\ f :> d; \texttt{NSDERIV}(x)\ g :> e|] \Longrightarrow \texttt{NSDERIV}(x)\ (\lambda y.\ f(y) \cdot g(y)) :> d \cdot g(x) + e \cdot f(x)$

The absence of any explicit notions of limits makes many of the standard results about derivatives straightforward to derive. The properties follow from simple algebraic manipulations of infinitesimals. As a result, the simplifier of Isabelle plays an important part in these proofs to do the tedious term manipulation and cancellation. To achieve this, we might need to add rules for associative-commutative rewriting, for example. However, there are cases when we need to prove lemmas explicitly to help the simplifier re-arrange terms. For example, to prove the theorem about the derivative of product, we need the following lemma:

$$(a \cdot b) - (c \cdot d) = b \cdot (a - c) + c \cdot (b - d)$$

6.6.2 Chain Rule

One of the important theorems about differentiation is the **chain rule**. In his formalization of differentiation in HOL, Harrison reports on the problems that arise when proving this theorem directly. The main difficulty is that, when using the standard definition, the theorem does not follow directly from any limit results. Indeed, unlike continuity, limits are not compositional. To deal with this problem, Harrison had to formalize an alternative, rather different characterization of differentiability in HOL, the so-called Carathéodory derivative. In our case, however, due to the nonstandard formulation, the chain rule admits an entirely straighforward derivation. The Isabelle theorem is given by

$$[|\texttt{NSDERIV}(a)\ g :> d; \texttt{NSDERIV}((g\,a))\ f :> e|] \Longrightarrow \texttt{NSDERIV}(a)\ (f \circ g) :> d \cdot e$$

Proof: This follows immediately from

$$\frac{f^*(g^*(x)) - f^*(\widetilde{g(a)})}{x - \widetilde{a}} = \frac{f^*(g^*(x)) - f^*(\widetilde{g(a)})}{g^*(x) - \widetilde{g(a)}} \frac{g^*(x) - \widetilde{g(a)}}{x - \widetilde{a}} \approx d \cdot e$$

This nonstandard proof, unlike its standard counterpart, reflects nicely and directly the intuition behind the Leibnizian notation for the rule:

$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx}$$

It is to be noted that Ballantyne and Bledsoe's prover [6] could not prove the chain rule automatically. In our case, we use a simple lemma to help set up the required product of fractions:

- $y \neq 0 \Longrightarrow x \cdot z = (x \cdot \mathtt{hrinv}\ y) \cdot (y \cdot z)$

The main proof is directly formalized, though we have to do some manipulations explicitly — for example, we need to use one of Isabelle's instantiation tactics with the lemma above to set the variable y in it to the correct binding. The level of automation could be made higher by building stronger routines in the simplifier to deal with division. For example, the recent addition of generic simplification procedures for subtraction have been helpful to many algebraic proofs. This is a case where development of new theories can call for more support from the prover. This ultimately benefits many other theories.

Coming back to our development, we prove the theorems about the inverses and quotients of functions using the chain rule and the fact that, for non-zero x, the derivative of $f(x) = 1/x$ is $-1/x^2$. The proofs remain simple and algebraic. Stated as theorems of Isabelle, these various extra results (shown in terms of the equivalent standard notation) are formalized like this:

- $x \neq 0 \Longrightarrow \mathtt{DERIV}(x)\ (\lambda x.\ \mathtt{rinv}\ x) :>\ -\mathtt{rinv}\ (x^2)$
- $[|\,\mathtt{DERIV}(x)\ f :> d; f(x) \neq 0|]$
  $\Longrightarrow \mathtt{DERIV}(x)\ (\lambda x.\ \mathtt{rinv}\ (f\,x)) :>\ -d \cdot \mathtt{rinv}\ (f(x)^2)$
- $[|\,\mathtt{DERIV}(x)\ f :> d; \mathtt{DERIV}(x)\ g :> e; g(x) \neq 0|]$
  $\Longrightarrow \mathtt{DERIV}(x)\ (\lambda z.\ f(z) \cdot \mathtt{rinv}\ (g\,z)) :> (d \cdot g(x) - e \cdot f(x)) \cdot \mathtt{rinv}\ (g(x)^2)$

6.6.3 Rolle's Theorem

More classic theorems of analysis are proved. These include Rolle's theorem that involves notions from both continuity and differentiability:

> **Rolle's Theorem.** If f is defined and continuous on the finite closed interval $[a,b]$ and differentiable at least on the open interval (a,b), then there exists x_0 between a and b such that $f'(x_0) = 0$.

The formalized proof is from Hoskins [48] and proceeds through a case analysis on the values that f can take in the interval between a and b. The argument

is once again nonstandard and yields a direct formalization. In Isabelle, the theorem is given by

$$\begin{array}{l}[|\ a < b; f(a) = f(b);\\ \ \forall x.\, a \leq x \wedge x \leq b \longrightarrow \texttt{isNSCont}\ f\ x;\\ \ \forall x.\, a < x \wedge x < b \longrightarrow f\ \texttt{NSdifferentiable}\ x;\\ |] ==> \exists x0.\, a < x0 \wedge x0 < b \wedge \texttt{NSDERIV}(x0)\ f :> 0r\end{array}$$

where the nonstandard infix predicate `NSdifferentiable` stands for 'the real function f is differentiable at x' and is defined by

$$f\ \texttt{NSdifferentiable}\ x \equiv \exists d.\, \texttt{NSDERIV}(x)\ f :> d$$

In the previous sections, we have presented the initial investigation of analysis using a nonstandard treatment. There are several important aspects of elementary analysis that still need to be formalized including Taylor and power series and the theory of Integration. A nonstandard approach promises to be useful for these as well. We next give a brief overview of one of the major theorems from which much of the power of NSA stems. We look at a version of the theorem specialized for the purpose of real analysis.

6.7 On the Transfer Principle

We now expand some more on the Transfer Principle, on which we remarked briefly in Section 6.3.3. Consider the statement, true in $\mathbb{R}$, stating that the set of natural numbers $\mathbb{N}$ is unbounded as a subset of $\mathbb{R}$: the Archimedean property holds for the reals. Formalized in Isabelle, this is expressed by

$$\forall x{::}\texttt{real}.\, \exists n{::}\texttt{nat}.\, x < \texttt{\&}n$$

Using the definitions of hyperreals, hypernaturals, and the properties of the free ultrafilter, we can then deduce the theorem that the set of hypernaturals $\mathbb{N}^*$ is unbounded as a subset of the hyperreals $\mathbb{R}^*$. Stated in Isabelle HOL, with explicit typing information shown, we have[6]

$$\forall x{::}\texttt{hypreal}.\, \exists n{::}\texttt{hypnat}.\, x < \texttt{\&\&}n$$

This second statement about the hyperreals thus appears to be, in some sense, a transform of the original statement about the reals. One can go from one to the other, as this example illustrates, by making certain specific changes about the types of the terms (and the embedding functions) appearing in each. The crux of Nonstandard Analysis is that transformation of statements along these lines can be carried out generally. It is this general idea that is captured by the Transfer Principle [48]:

THEOREM 6.3. ***Transfer Principle for real analysis.*** *There exists a set $\mathbb{R}^*$ such that*

1) $\mathbb{R}$ is a proper subset of $\mathbb{R}^*$.

[6]The notations `&n` and `&&n` stand for embeddings in the reals and hyperreals respectively

2) to each function $f : \mathbb{R} \to \mathbb{R}$ there corresponds a function $f^* : \mathbb{R}^* \to \mathbb{R}^*$ which agrees with f on $\mathbb{R}$.

3) to each n-place relation P on $\mathbb{R}^*$ there corresponds a n-place relation P^* on $\mathbb{R}^*$ which agrees with P on $\mathbb{R}$.

Further, every well-formed statement φ formulated in terms of

- *particular real numbers $r_1, r_2, \ldots, r_m$,*
- *particular functions $f_1, f_2, \ldots, f_m$,*
- *particular relations $P_1, P_2, \ldots, P_m$,*
- *logical connectives and quantifiers, with variables ranging over $\mathbb{R}$*

is true with respect to $\mathbb{R}$ if and only if the statement φ^* obtained from φ by replacing each f_k by f_k^* and each P_k by P_k^*, and by allowing variables to range over $\mathbb{R}^*$, is true with respect to $\mathbb{R}^*$.

In the current work, proving the equivalence of standard and nonstandard formulation has involved working with sequences and checking whether certain sets belong to the ultrafilter or not each time a new property is introduced. By implementing some form of the transfer principle, one should be able to capture much of the power that NSA derives from the use of such metatheorems. This has not been investigated thoroughly — we have formalized (1), and particular cases of (2) and (3) above though — and so, producing an effective form of the principle provides scope for further research. Our work has shown, though, that a powerful theory is still possible if one is willing to transfer properties by separate proofs. General automatic tactics to check whether supersets, intersections, or complements of sets belonged to the free ultrafilter have been coded that enabled many of the goals to be simplified greatly, and in quite a few cases to be proved automatically.

6.8 Related Work and Conclusions

The automated theorem proving community does not seem to have shown much interest in NSA, even though its importance has grown in many fields such as physics, analysis and economics, where it has successfully been applied. Ballantyne and Bledsoe [6] implemented a prover using nonstandard techniques in the late seventies. Their work basically involved substituting any theorem in the reals $\mathbb{R}$ by its analogue in the extended reals $\mathbb{R}^*$ and proving it in this new setting. Even though the prover had many limitations, and the work was just a preliminary investigation, the authors argued that through the use of nonstandard analysis, they had brought some new and powerful mathematical techniques to bear on the problem.

Despite this rather promising work, there does not seem to have been much done over the last two decades. Chuaqui and Suppes [22] have proposed an axiomatic framework for doing proofs in NSA, and Bedrax has implemented a prototype for a simplified version of the Suppes-Chuaqui system called Infmal [8]. Infmal is implemented in Common Lisp and contains the various axioms

(logical, algebraic and infinitesimal) required by the deduction system and extensions to the usual arithmetic operations. Unfortunately, Infmal is a simple experiment and though interactive, is rather limited in the proofs it can carry out. There has also been some work carried out by Beeson [9] who developed a restricted axiomatic version of NSA using the logic of partial terms. The properties of the infinitely close relation (c.f. Section 3.7.3), standard parts (c.f. Section 3.7.4), infinitesimals and so on, are asserted as axioms leading to a theory similar in spirit to the one that could be developed starting from the axioms of Section 3.2. Beeson uses NSA to ensure the correctness of applications of calculus in a system called *Mathpert* which combines computer algebra with theorem-proving.

We have verified the various basic axioms asserted by Beeson in his approach in our development. Moreover, we have also verified, through our strictly definitional approach, the axioms about properties of the hyperreals that were built into Ballantyne and Bledsoe's prover.

In summary, this work describes an initial and rigorous investigation of the mechanization of analysis using nonstandard techniques. As shown by the extensive development of analysis in HOL by Harrison, the need for abstraction leading to general theories is important since it saves a lot of similar proofs from being repeated. Our main aim has been to show that there are advantages to be gained by using nonstandard analysis as the framework for real analysis. We feel that the simplicity of the formulations and relative ease with which many different results are proved have amply justified the promises held by the approach.

Chapter 7

Conclusions

The various chapters of this thesis have covered the development of a range of theories in the theorem prover Isabelle. The main unifying theme has been the use of infinitesimals as valuable tools that are once again respectable in mathematics. Infinitesimals have been around for over two thousand years and generally had a bad press. Their use has at times been free and viewed as a blessing (throughout the 18th century, for example), and at others viewed as heresy and banned (from the 19th century till the middle of this century). What is undeniable is that they have been valued as intuitive tools at all times by generations of mathematicians who used them solve problems and carry out proofs. Even those trying to get rid of them sometimes lapsed into infinitesimal reasoning. The great insight of Robinson, leading to the creation of NSA, has been praised over and over again. In a sense, his work vindicates the (sometimes blind) faith of influential mathematicians such as the Marquis de L'Hospital in the use of infinitesimals. The major achievement of Robinson and his followers was not only to rehabilitate the infinitesimal as a sound mathematical concept but also to bring along new types of numbers that added to the power of the nonstandard approach.

In the next few sections, we reflect on the aspects of this thesis we feel are most important. We recapitulate some of the points that were made in the previous chapters and add concluding remarks. We also give an indication of a few areas that might yield interesting further work.

7.1 Geometry, Newton, and the *Principia*

Geometry is one of the oldest and most trusted areas of mathematics. Indeed, Euclidean geometry was believed for over two thousand years to be the faithful model of reality and space. This trust was based on human perception and experience of the world. This explains partly the drive of mathematicians before, and even for some time after, the seventeenth century to carry out mathematical demonstrations using geometric arguments. Such an approach was, in any case, *de rigueur* for any mathematical result to gain peer acceptance and not be ridiculed.

Proofs in geometry are usually hard. Newton's use of geometric arguments to prove the complex results of the *Principia* is a marvel of intellectual achievement

and insight. For over three centuries, thousands of mathematicians, physicists, historians, and philosophers have contributed a large volume of work analysing and discussing the influence, thought processes, and reasoning of Newton. Many people though, with the exception of historians of science, are familiar with the various results of the *Principia* expressed algebraically. The proofs are done using differential calculus and vectors in most cases.

In this work, we aimed at mechanizing Newton's work by respecting as much as possible the geometric reasoning. Our goal was not just to derive the theorems by any means possible, as for that we could have used our theory of vectors and reproduce the modern proofs found in most textbooks on mechanics and dynamics. Instead, the aim was to capture the proofs as Newton intended them and to that effect formal concepts had to be devised that would capture his reasoning. This was especially important in the cases where Newton's procedures departed from the usual Euclidean arguments. In a sense, we have developed a calculus that provides formal rules enabling us to capture Newton's reasoning.

The pleasing aspect is the way infinitesimals have blended nicely with geometric concepts to give a theory that remains intuitive, but is nevertheless more powerful than pure Euclidean geometry. In it, one can prove the type of theorems that occur in the *Principia* but also deal with ordinary theorems of Euclidean geometry, by incorporating infinitesimal arguments. The proofs can be viewed as moving into the hyperreal space, just as it is possible to move into complex space when dealing with proofs in analytic geometry.

It is hoped that the mechanization of some of the important theorems of the *Principia* will help point out the intricate nature of Newton's geometry and also its rigour. The discovery of finite witnesses that enable many steps to be fitted within the NSA framework shows that Newton was aware of the problems that arise when dealing with ratios of infinitesimals. The flaw that we found, we believe, is the exception rather than the rule and caused much surprise. That there is a step that cannot be carried out exactly as Newton intended, and for explicit reasons borne out by nonstandard analysis, is a vindication of the rigour of our framework, and of infinitesimals in particular. We have checked numerous textbook proofs reproducing Newton's reasoning about the *Propositio Kepleriana*: all carry out the final multiplication step without any remarks. In fact, none have any formal devices or notation to handle ultimate situations as systematically as this work does.

7.2 Hyperreal Analysis

In the light of the impressive gains that the nonstandard treatment can bring to the process of mathematical analysis, it is somewhat surprising to note the suspicions still harboured by the mathematics community towards it. The historic legacy of Weierstrass and Cauchy is the enduring belief that the $\epsilon-\delta$ approach is the *only* acceptable, or even natural, approach to analysis. Yet, its inherent difficulties and complexity are striking when compared with the intuition and simplicity of nonstandard methods. Of course, such an attitude is merely an echo of the behaviour of mathematicians towards new concepts throughout the centuries. Terms such as "irrational" and "imaginary", for instance, reflect attitudes as new and unfamiliar notions were introduced that extended existing numbers.

This work showed, we hope, that, as far as number constructions are concerned, going beyond the real numbers is the next evident progression. Gödel, as quoted from the preface of Robinson's *Non-standard Analysis* [72], eloquently sums this up for us:

> Arithmetic starts with the integers and proceeds by successively enlarging the number system by negative and rational numbers, irrational numbers, etc. But the next quite natural step after the reals, namely the introduction of infinitesimals, has simply been omitted. I think, in incoming centuries it will be considered a great oddity in the history of mathematics that the first exact theory of infinitesimals was developed 300 years after the invention of the differential calculus.

Our work also shows the benefits that NSA brings to mechanization. The nonstandard approach can be viewed as adding greater computational abilities to the theorem prover. It enables, in that sense, a better mixture of logic and computation which is needed for a more straightforward, and powerful, formalization of mathematics in mechanical systems. Moreover, NSA provides intuitive methods while preserving mathematical rigour. The pressing need is for wider recognition of the many advantages that are being ignored due to the rigidity of current mathematical practice.

7.3 Further Work

Some of the possible ways of extending and applying the work done in this thesis are now outlined. We believe that there is scope for future work in areas ranging from mathematical analysis to diagrammatic reasoning and problem solving in physics. It should be mentioned at the outset that any areas where theorem proving in the reals have applications are also valid for the hyperreals. Several of these are mentioned in Harrison's thesis [43] to which we refer the interested reader. Some of the areas we mention next are those where the richness of the extended number systems brings additional tools or benefits.

7.3.1 Geometry Theorem Proving

One of the important roles that mechanical geometry theorem proving is set to play in the future is in education. Cognitive research has shown that traditional geometry problem solving is hard [54]. It is crucial that more research is done in the field to come up with systems that can be used effectively as educational tools. More generally, there has not been much interest in the United Kingdom for research in GTP despite this being one of the fields where automated reasoning has been the most successful so far. There are numerous powerful GTP techniques [82, 84] that can potentially be applied in other fields such as geometric constraints solving. These are of considerable commercial interest since they are essential to computer aided modelling systems, for example.[1] We believe that some of the ideas introduced in this work can be used to produce

[1]Some of these important aspects were only recently presented and discussed at the workshop on Automated Deduction in Geometry. This was held by the leading centre for automated GTP research, the Chinese Institute of Systems Science, Academia Sinica, Beijing.

new techniques in automated GTP, especially in dealing with conditions that are nearly degenerate.

7.3.2 Numerical Software Verification

Notions of infinitesimals from NSA have application in floating point error analysis. There have been theoretical techniques — known as asymptotic methods [47] — developed for formal verification of mathematical software. These deal with numerical error without quantifying it and model the behaviour of mathematical programs by considering their overall accuracy when that of their sequences of operations tends to infinity. In nonstandard terms, each sequence of operations represents an internal function whose result can be viewed as being infinitely close to the ideal value the program should yield. Asymptotic methods can expose the same problems as classical error analysis but require less effort and expertise. They deal well with cases where small numerical values (modelled by infinitesimals) lead to numerically ill-defined computations. With the framework established in Isabelle, this is an interesting and promising area of application for nonstandard concepts.

7.3.3 Physics Problem Solving

Novak has implemented several systems for problem solving in classical physics with the help of diagrams [65]. Although his work falls into the field of diagrammatic reasoning rather than GTP, it does require the implicit applications of geometry theorems to derive relations between various physical quantities represented geometrically in the diagrams. This work also shows that it is possible to closely relate physical and geometric principles through diagrams.

The use of geometry to represent physical situations is another potential area of applications for our techniques. Indeed, the geometric and infinitesimal tools developed for Newton's *Principia* can be applied to the study of the rich model built on Newton's exposition of the physical world. Infinitesimals are often introduced in geometric diagrams of physical systems, for example to characterise the small virtual displacement of a pendulum bob, of a body sliding down an inclined slope, or of light going through a lens. Infinitesimal arguments have been used informally by physicists for a long time to explain and predict the behaviour of physical systems. Physicists often make statements such as "when θ is small, $\sin(\theta) = \theta$" and rely on intuition to explain their approximation. The substitution of a quantity by an infinitely close one generally requires care and should be justified formally. Once the justification is made, the advantages gained through the use of infinitesimals are enormous in terms of clarity and intuitive meaning.

7.4 Concluding Remarks

Reading the *Principia* and making sense of the reasoning of Newton is a difficult but rewarding task. As is common with proofs using geometric tools, once the hard task of constructing the diagram and proof is done, the result that follows usually looks simple and intuitive. We have shown that, though Newton does not provide a set of rules for carrying out his proofs, the reasoning is rigorous and can

be mechanized. We have formally defined and proved the ultimate properties Newton tried to demonstrate. This provides a means of validating, through mathematical logic, what have often been regarded as informal arguments by Newton. We have in effect bridged the gap between intuition and formality.

We have developed a formal theory of infinitesimals by constructing a succession of number systems leading to the hyperreals. This approach has been adopted since it provides a clear and consistent way of introducing new types. The axiomatic approach for the development of NSA has been avoided as infinitesimals are always tricky and can lead to paradoxes if not used properly. The conservative definition of types in Isabelle HOL guarantees the soundness of the infinitesimal theory.

The introduction of infinitesimal elements in the geometry is an exciting aspect that can lead to the discovery of interesting properties that cannot be seen ordinarily. We have outlined the scope for more research that exists in several fields as a result of our investigation. These applications show that techniques involving Nonstandard Analysis, geometry or a combination of both have powerful practical significance.

Bibliography

[1] J. R. Abrial and G. Laffitte. Towards the mechanization of the proofs of some classical theorems of set theory. *Preprint*, 1993.

[2] S. Albeverio. *Nonstandard methods in stochastic analysis and mathematical physics.* Academic Press, 1986.

[3] Apollonius. *Conics*, volume 11 of *Great Books of the Western World.* Encyclopedia Britannica, 1939. Translation by R. Catesby Taliaferro.

[4] D. S. Arnon. Geometric reasoning with logic and algebra. *Artificial Intelligence*, 37:37–60, 1988.

[5] E. Artin. *Geometric Algebra.* Interscience, 1957.

[6] A. M. Ballantyne and W. W. Bledsoe. Automatic proofs of theorems in analysis using nonstandard analysis. *Journal of the Association of Computing Machinery*, 24(3):353–374, 1977.

[7] C. Ballarin and L. C. Paulson. Reasoning about coding theory: The benefits we get from computer algebra. In J. Calmet and J. Plaza, editors, *Proceedings of the International Conference on Artificial Intelligence and Symbolic Mathematical Computation*, volume 1476 of *Lecture Notes in Artificial Intelligence*, pages 55–66. Springer-Verlag, 1998.

[8] T. Bedrax. Infmal: Prototype of an interactive theorem prover based on infinitesimal analysis. Master's thesis, Pontifica Universidad Catolica de Chile, 1993. Liciendo en Mathematica con Mencion en Computation Thesis.

[9] M. Beeson. Using nonstandard analysis to ensure the correctness of symbolic computations. *International Journal of Foundations of Computer Science*, 6(3):299–338, 1995.

[10] G. Berkeley. *The Analyst: A discourse addressed to an infidel mathematician*, volume 1 of *The World of Mathematics, 1956.* Allen and Unwin, 1734.

[11] J. B. Brackenridge. Newton's mature dynamics and the *Principia*: A simplified solution to the Kepler Problem. *Historia Mathematica*, 16:36–45, 1989.

[12] J. B. Brackenridge. *The key to Newton's dynamics: The Kepler Problem and Newton's Principia.* University of California Press, 1995.

[13] B. Buchberger, G. E. Collins, and B. Kutzler. Algebraic methods for geometric reasoning. *Annual Review of Computational Science*, 3:85–119, 1988.

[14] J. C. Burkill and H. Burkill. *A Second Course in Mathematical Analysis.* Cambridge University Press, 1970.

[15] E. Cerutti and P. J. Davis. FORMAC meets Pappus: Some observations on elementary analytic geometry by computer. *American Mathematical Monthly*, 76:895–905, 1969.

[16] S. C. Chou, X. S. Gao, and D. S. Arnon. On the mechanical proof of geometry theorems involving inequalities. In C. Hoffmann, editor, *Issues in Robotics and Nonlinear Geometry*, pages 139–181. JAI Press, 1992.

[17] S. C. Chou, X. S. Gao, and J. Z. Zhang. The area method and affine geometries over any fields. Preprint, 1993.

[18] S. C. Chou, X. S. Gao, and J. Z. Zhang. Automated geometry theorem proving by vector calculation. In *ACM-ISSAC*, pages 284–291, Kiev Ukraine, July 1993.

[19] S. C. Chou, X. S. Gao, and J. Z. Zhang. Automated generation of readable proofs with geometric invariants, I. multiple and shortest proof generation. *Journal of Automated Reasoning*, 17:325–347, 1996.

[20] S. C. Chou, X. S. Gao, and J. Z. Zhang. Automated generation of readable proofs with geometric invariants, II. theorem proving with full-angles. *Journal of Automated Reasoning*, 17:349–370, 1996.

[21] S. C. Chou, W. F. Schelter, and J. G. Yang. Characteristic sets and Gröbner bases in geometry theorem proving. In H. Aït-Kaaci and M. Nivat, editors, *Resolution of Equations in Algebraic Structures*, pages 33–92. Academic Press, 1989.

[22] R. Chuaqui and P. Suppes. Free-variable axiomatic foundations of infinitesimal analysis: A fragment with finitary consistency proof. *Journal of Symbolic Logic*, 60(1), March 1995.

[23] A. Church. A formulation of the simple theory of type. *Journal of Symbolic Logic*, 5:56–68, 1940.

[24] G. E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. *Lecture Notes in Computer Science*, 33:134–165, 1975.

[25] J. H. Conway. *On Numbers and Games.* Academic Press Inc. (London) Ltd., 1976.

[26] P. J. Davis and R. Hersh. *The Mathematical Experience.* Harmondsworth, Penguin, 1983.

[27] D. Densmore. *Newton's Principia: The Central Argument.* Green Lion Press, Santa Fe, New Mexico, 1996.

[28] J. Dieudonné. *Linear Algebra and Geometry.* Hermann, 1969. Translated from the original French text *Algèbre linéaire et géométrie élémentaire.*

[29] E. W. Elcock. Representation of knowledge in a geometry machine. *Machine Intelligence*, 8:11–29, 1977.

[30] Euclid. *Elements.* Dover Publications Inc., 1956. Translation by T. L. Heath.

[31] W. M. Farmer, J. D. Guttman, and F. J. Thayer. Little theories. In D. Kapur, editor, *Automated Deduction—CADE-11*, volume 607 of *Lecture Notes in Computer Science*, pages 567–581. Springer-Verlag, 1992.

[32] D. Fearnley-Sander. The idea of a diagram. In H. Aït-Kaaci and M. Nivat, editors, *Resolution of equations in algebraic structures*, volume 1, pages 127–150. Academic Press, 1989.

[33] G. Fisher. Veronese's non-Archimedean linear continuum. In P. Ehrlich, editor, *Real Numbers, Generalizations of the Reals, and Theories of Continua*, volume 242 of *Synthese Library.* Kluwer Academic Publisher, 1994.

[34] J. D. Fleuriot and L. C. Paulson. A combination of geometry theorem proving and nonstandard analysis, with application to Newton's *Principia.* In C Kirchner and H. Kirchner, editors, *Automated Deduction - CADE-15*, volume 1421 of *Lecture Notes in Artificial Intelligence*, pages 3–16. Springer-Verlag, July 1998.

[35] G. Galileo. *Two New Sciences.* University of Wisconsin Press, 1638. Translation by S. Drake, 1974.

[36] F. De Gandt. *Force and Geometry in Newton's Principia.* Princeton University Press, 1995.

[37] H. Gelernter. Realization of a geometry theorem-proving machine. *Computers and Thought*, pages 134–152, 1959.

[38] C. Gilmore. An examination of the geometry theorem machine. *Artificial Intelligence*, 1:171–187, 1970.

[39] A. M. Gleason. *Fundamentals of Abstract Analysis.* Series in Mathematics. Addison-Wesley, 1966.

[40] D. L. Goodstein and J. R. Goodstein. *Feynman's Lost Lecture.* Vintage, 1997. Richard P. Feynman's Lecture on the motion of planets around the sun.

[41] M. Gordon and T. Melham. *Introduction to HOL: A theorem proving environment for Higher Order Logic.* Cambridge University Press, 1993.

[42] John Harrison. Constructing the real numbers in HOL. In L. J. M. Claesen and M. J. C. Gordon, editors, *Proceedings of the IFIP TC10/WG10.2 International Workshop on Higher Order Logic Theorem Proving and its Applications*, volume A-20 of *IFIP Transactions A: Computer Science and Technology*, pages 145–164, IMEC, Leuven, Belgium, 1992. North-Holland.

[43] John Harrison. *Theorem Proving with the Real Numbers.* Springer-Verlag, 1998. Also published as technical report 408 of the Computer Laboratory, University of Cambridge, 1996.

[44] J. M Henle and E. M. Kleinberg. *Infinitesimal Calculus.* The MIT Press, 1979.

[45] D. Hestenes and G. Sobczyk. *Clifford Algebra to Geometric Calculus.* Reidel Publishing Company, 1984.

[46] D. Hilbert. *The Foundations of Geometry.* The Open Court Company, 1901. Translation by E. J. Townsend.

[47] D. G. Hoover and D. M. McCullough. Verifying launch interceptor routines with the asymptotic method. Technical Report STARS-AC-A023/006/00, Odyssey Research Associates under contract to Unisys Corporation, 1993. Software Technology for Adaptable, Reliable Systems (STAR) Technical Report.

[48] R. F. Hoskins. *Standard and Nonstandard Analysis.* Mathematics and its Applications. Ellis Horwood Limited, 1990.

[49] A. E. Hurd and P. A. Loeb. *An Introduction to Nonstandard Real Analysis*, volume 118 of *Pure and Applied Mathematics.* Academic Press Inc., 1985.

[50] L. S. Jutting. *Checking Landau's "Grundlagen" in the Automath system.* PhD thesis, Eindhoven University of Technology, 1977.

[51] D. Kapur. Using Gröbner bases to reason about geometry problems. *Journal of Symbolic Computation*, 2:399–408, 1986.

[52] H. J. Keisler. *Foundations of Infinitesimal Calculus.* Prindle, Weber & Schmidt, 1976.

[53] M. Kline. *Mathematical Thought from Ancient to Modern Times.* Oxford University Press, 1972.

[54] K. R. Koedinger. Reifying implicit planning in geometry: Guidelines for model-based intelligent tutoring system design. *Computers as Cognitive Tools*, 1993.

[55] B. Kutzler and S. Stifter. On the application of Buchberger's algorithm to automated geometry theorem proving. *Journal of Symbolic Computation*, 2:389–397, 1986.

[56] L. Lamport. How to write a proof. Technical Report 94, Digital Systems Research Center, 1993.

[57] E. Landau. *Foundations of Analysis.* Chelsea, 1951.

[58] D. Laugwitz. Infinitely small quantities in Cauchy's textbooks. *Historia Mathematica*, 14:258–274, 1987.

[59] H. Li and M. Cheng. Clifford algebraic reduction method for automated theorem proving in differential geometry. *Journal of Automated Reasoning*, 21:1–21, 1998.

[60] T. Lindstrøm. An invitation to nonstandard analysis. In N. Cutland, editor, *Nonstandard Analysis and Its Applications*, volume 10 of *London Mathematical Society Students Text.* Cambridge University Press, 1988.

[61] T. Matsuyama and T. Nitta. Geometric theorem proving by integrated logical and algebraic reasoning. *Artificial Intelligence*, 75:93–113, 1995.

[62] E. Nelson. Internal set theory: A new approach to nonstandard analysis. *Bulletin American Mathematical Society*, 83, 1977.

[63] A. J. Nevins. Plane geometry theorem proving using forward chaining. *Artificial Intelligence*, 6:1–23, 1975.

[64] T. Nipkow and D. von Oheimb. Java$_{light}$ is type-safe — definitely. In *Proc. 25th ACM Symp. Principles of Programming Languages*, pages 161–170. ACM Press, New York, 1998.

[65] S. Novak Jr. Diagrams for solving physical problem. In Janice Glasgow, N. Hari Narayana, and B. Chandrasekaram, editors, *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, pages 753—774. AAAI Press/MIT Press, 1995.

[66] L. C. Paulson. *Isabelle: A Generic Theorem Prover*, volume 828 of *Lecture Notes in Computer Science.* Springer, 1994.

[67] L. C. Paulson. The inductive approach to verifying cryptographic protocols. *Journal of Computer Security*, pages 85–128, 1998.

[68] L. C. Paulson. Isabelle's object-logics. Technical Report 286, Computer Laboratory, University of Cambridge, February 1998.

[69] L. C. Paulson and K. Grąbczewski. Mechanizing set theory: Cardinal arithmetic and the axiom of choice. *Journal of Automated Reasoning*, 17:291–323, 1996.

[70] H. Poincaré. Review of Hilbert's foundations of geometry (1902). In P. Ehrlich, editor, *Real Numbers, Generalizations of the Reals, and Theories of Continua*, volume 242 of *Synthese Library.* Kluwer Academic Publisher, 1994.

[71] S. Rashid. *Economies with many agents: an approach using nonstandard analysis.* Johns Hopkins University Press, 1987.

[72] A. Robinson. *Non-Standard Analysis.* North-Holland, 1980.

[73] B. Russell. *The Autobiography of Bertrand Russell:1872–1914.* Allen & Unwin, 1967.

[74] E. Schechter. *Handbook of Analysis and Its Foundations.* Academic Press, 1997.

[75] A. P. Simpson. The Infidel is innocent. *The Mathematical Intelligencer*, 12(3), 1990.

[76] K. D. Stroyan and W. A. J. Luxemburg. *Introduction to the Theory of Infinitesimals.* Academic Press, 1976.

[77] A. Tarski. *A Decision Method for Elementary Algebra and Geometry.* University of California Press, 1951.

[78] R. Vesley. An intuitionistic infinitesimal calculus. In F. Richman, editor, *Constructive Mathematics*, volume 873 of *Lecture Notes in Mathematics.* Springer, 1983.

[79] D. Wang. Geometry machines: From AI to SMC. In *Proceedings of AISC'96 Third International Conference on Artificial Intelligence and Symbolic Mathematical Computation*, volume 1138 of *Lecture Notes in Computer Science*, pages 213–239. Springer-Verlag, 1996.

[80] D. Wang. Clifford algebraic calculus for geometric reasoning, with application to computer vision. In D. Wang, R. Caferra, L. Fariñas del Cerro, and H. Shi, editors, *Automated Deduction in Geometry, ADG'96*, volume 1360 of *Lecture Notes in Artificial Intelligence*, pages 115–140. Springer, 1997.

[81] D. T. Whiteside. The mathematical principles underlying Newton's *Principia Mathematica*. Technical Report 138, Glasgow University, 1970.

[82] W. Wu. On the decision problem and the mechanization of theorem in elementary geometry. In *Automated Theorem Proving: After 25 years*, volume 29 of *Contemporary Mathematics*, pages 213–234. A. M. S., 1984.

[83] W. Wu. Basic principles of mechanical theorem proving in elementary geometries. *Journal of Automated Reasoning*, 2:221–252, 1986.

[84] W. Wu. Mechanical theorem proving of differential geometries and some of its applications in mechanics. *Journal of Automated Reasoning*, 7:171–191, 1991.