DISCRETE MATHEMATICS AND ITS APPLICATIONS
Series Editor KENNETH H. ROSEN

HANDBOOK OF MATHEMATICAL INDUCTION

THEORY AND APPLICATIONS

David S. Gunderson

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

DISCRETE MATHEMATICS AND ITS APPLICATIONS
Series Editor KENNETH H. ROSEN

HANDBOOK OF MATHEMATICAL INDUCTION

THEORY AND APPLICATIONS

David S. Gunderson
University of Manitoba
Winnipeg, Canada

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

DISCRETE MATHEMATICS AND ITS APPLICATIONS

Series Editor

Kenneth H. Rosen, Ph.D.

Juergen Bierbrauer, Introduction to Coding Theory

Francine Blanchet-Sadri, Algorithmic Combinatorics on Partial Words

Richard A. Brualdi and Dragoš Cvetković, A Combinatorial Approach to Matrix Theory and Its Applications

Kun-Mao Chao and Bang Ye Wu, Spanning Trees and Optimization Problems

Charalambos A. Charalambides, Enumerative Combinatorics

Gary Chartrand and Ping Zhang, Chromatic Graph Theory

Henri Cohen, Gerhard Frey, et al., Handbook of Elliptic and Hyperelliptic Curve Cryptography

Charles J. Colbourn and Jeffrey H. Dinitz, Handbook of Combinatorial Designs, Second Edition

Martin Erickson, Pearls of Discrete Mathematics

Martin Erickson and Anthony Vazzana, Introduction to Number Theory

Steven Furino, Ying Miao, and Jianxing Yin, Frames and Resolvable Designs: Uses, Constructions, and Existence

Mark S. Gockenbach, Finite-Dimensional Linear Algebra

Randy Goldberg and Lance Riek, A Practical Handbook of Speech Coders

Jacob E. Goodman and Joseph O'Rourke, Handbook of Discrete and Computational Geometry, Second Edition

Jonathan L. Gross, Combinatorial Methods with Computer Applications

Jonathan L. Gross and Jay Yellen, Graph Theory and Its Applications, Second Edition

Jonathan L. Gross and Jay Yellen, Handbook of Graph Theory

David S. Gunderson, Handbook of Mathematical Induction: Theory and Applications

Darrel R. Hankerson, Greg A. Harris, and Peter D. Johnson, Introduction to Information Theory and Data Compression, Second Edition

Darel W. Hardy, Fred Richman, and Carol L. Walker, Applied Algebra: Codes, Ciphers, and Discrete Algorithms, Second Edition

Daryl D. Harms, Miroslav Kraetzl, Charles J. Colbourn, and John S. Devitt, Network Reliability: Experiments with a Symbolic Algebra Environment

Silvia Heubach and Toufik Mansour, Combinatorics of Compositions and Words

Leslie Hogben, Handbook of Linear Algebra

Titles (continued)

Derek F. Holt with Bettina Eick and Eamonn A. O'Brien, Handbook of Computational Group Theory

David M. Jackson and Terry I. Visentin, An Atlas of Smaller Maps in Orientable and Nonorientable Surfaces

Richard E. Klima, Neil P. Sigmon, and Ernest L. Stitzinger, Applications of Abstract Algebra with Maple™ and MATLAB®, Second Edition

Patrick Knupp and Kambiz Salari, Verification of Computer Codes in Computational Science and Engineering

William Kocay and Donald L. Kreher, Graphs, Algorithms, and Optimization

Donald L. Kreher and Douglas R. Stinson, Combinatorial Algorithms: Generation Enumeration and Search

C. C. Lindner and C. A. Rodger, Design Theory, Second Edition

Hang T. Lau, A Java Library of Graph Algorithms and Optimization

Elliott Mendelson, Introduction to Mathematical Logic, Fifth Edition

Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone, Handbook of Applied Cryptography

Richard A. Mollin, Advanced Number Theory *with Applications*

Richard A. Mollin, Algebraic Number Theory

Richard A. Mollin, Codes: The Guide to Secrecy from Ancient to Modern Times

Richard A. Mollin, Fundamental Number Theory with Applications, Second Edition

Richard A. Mollin, An Introduction to Cryptography, Second Edition

Richard A. Mollin, Quadratics

Richard A. Mollin, RSA and Public-Key Cryptography

Carlos J. Moreno and Samuel S. Wagstaff, Jr., Sums of Squares of Integers

Dingyi Pei, Authentication Codes and Combinatorial Designs

Kenneth H. Rosen, Handbook of Discrete and Combinatorial Mathematics

Douglas R. Shier and K.T. Wallenius, Applied Mathematical Modeling: A Multidisciplinary Approach

Alan Slomson and Reginald B. J. T. Allenby, How to Count: An Introduction to Combinatorics, Second Edition

Alexander Stanoyevitch, Introduction to Cryptography with Mathematical Foundations and Computer Implementations

Jörn Steuding, Diophantine Analysis

Douglas R. Stinson, Cryptography: Theory and Practice, Third Edition

Roberto Togneri and Christopher J. deSilva, Fundamentals of Information Theory and Coding Design

W. D. Wallis, Introduction to Combinatorial Designs, Second Edition

W. D. Wallis and John George, Introduction to Combinatorics

Lawrence C. Washington, Elliptic Curves: Number Theory and Cryptography, Second Edition

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20131021

International Standard Book Number-13: 978-1-4200-9365-0 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To my darling daughter, Christine.

This page intentionally left blank

Contents

I Theory

II Applications and exercises

III Solutions and hints to exercises

IV Appendices

Foreword

The idea of mathematical induction has been with us for ages, certainly since the 16th century, but was made rigorous only in the 19th century by Augustus de Morgan who, incidentally, also introduced the term ‘mathematical induction’. By now, induction is ubiquitous in mathematics and is taken for granted by every mathematician. Nevertheless, those who are getting into mathematics are likely to need much practice before induction is in their blood: The aim of this book is to speed up this process.

Proofs by induction vary a great deal. In fact, when it comes to finite structures or, more generally, *sequences* of assertions, *every* proof may be viewed as a proof by induction; when proving a particular proposition, we may as well assume that we have already proved every assertion which comes earlier in the sequence. For example, when proving the simple result that every graph with n vertices and more than $n^2/4$ edges contains a triangle, we may as well assume that this is true for graphs with fewer than n vertices. Thus, when a professor asks his class for ideas as to how to try to prove a result about finite groups and gets the suggestion *‘By induction!’*, he is right to dismiss this as being unhelpful, since we are always *free* to use induction, and in some sense we are always using it. Nevertheless, it is true that in some cases induction plays a major role, while in others we hardly make any use of it. And the question is not whether to use induction but, when using it, *how* to use it.

It would be impossible for this *Handbook of Mathematical Induction* to cover all aspects of mathematical induction and its variants for infinite sets, but there is plenty of exciting material here, selected with much care, with emphasis on some of the most elegant results. This book contains all the standard exercises on induction and many more, ranging from the trifle and the trivial to the significant and challenging.

There are numerous examples from graph theory, point set topology, elementary number theory, linear algebra, analysis, probability theory, geometry, group theory, game theory, and the theory of inequalities, with results about continued fractions, logical formulae, Latin rectangles, Hankel matrices, Hilbert’s affine cube, and the numbers of Fibonacci, Bernoulli, Euler, Catalan and Schröder, among others. Furthermore, the reader is guided through appropriate proofs of the theorems of Ramsey, Schur, Kneser, Hales and Jewett, Helly, Radon, Caratheodory, and many other results.

What prompts someone to write a book on mathematical induction? To share his passion for mathematics? Gunderson's passion for all of mathematics is evident. Perhaps this remarkable passion is due to the unusual road he has taken to mathematics. When I first met him, at Emory University in 1993, he was a graduate student. A rather 'mature' graduate student; as I learned later, in his youth he had flown aerobatics, and then had been a laborer and truck driver for ten years or so before starting in pure mathematics for the fun of puzzle solving. Although he has been in mathematics for over two decades, his physical prowess is still amazing: he has a penchant for ripping telephone books, and has not lost an arm-wrestling match since 1982.

This book is the first example that I know of which treats mathematical induction seriously, more than just a collection of recipes. It is sure to be an excellent student companion and instructor's guide for a host of courses.

Béla Bollobás

University of Cambridge and University of Memphis

Preface

Mathematical induction is a powerful proof technique that is generally used to prove statements involving whole numbers. Students often first encounter this technique in first or second year university courses on number theory, graph theory, or computer science. Many students report that their first exposure to mathematical induction was both scary yet simple and intriguing. In high school, formal proof techniques are rarely covered in great detail, and just the word “proof” seems daunting to many. Mathematical induction is a tool that seems quite different from anything taught in high school.

After just a few examples of proof by mathematical induction, the average student seems to gain an appreciation for the technique because the format for such a proof is straightforward and prescribed, yet the consequences are quite grand. Some students are further fascinated by the technique because of the erroneous conclusions available when the format is not followed precisely. It seems as if many students view mathematical induction as simply a necessary evil. Few beginning students in mathematics or computer science realize that all of mathematics is based on mathematical induction, nor do they realize that the foundations for the technique are of a completely different type than “scientific induction”, or the “scientific method” forms of “hypothesis, testing, and conclusion” arguments used in most sciences.

In part, because of the recent explosion of knowledge in combinatorics, computing, and discrete mathematics, mathematical induction is now, more than ever, critical in education, perhaps surpassing calculus in its relevance and utility. The theory of recursion in computing science is practically the study of mathematical induction applied to algorithms. The theory of mathematical logic and model theory rests entirely on mathematical induction, as does set theory. It may be interesting to note that even in calculus, mathematical induction plays a vital role. Continuous mathematics (like calculus or analysis) uses counting numbers, dimension of a space, iterated derivatives, exponents in polynomials, or size of a matrix, and so mathematical induction might one day be taught in all junior math courses. In fact, mathematical induction is absolutely essential in linear algebra, probability theory, modelling, and analysis, to name but a few areas. Mathematical induction is a common thread that joins all of mathematics and computing science.

This book contains hundreds of examples of mathematical induction applied in a vast array of scientific areas, as well as a study of the theory and how to find and write mathematical induction proofs. The presentation here is quite unlike that of a discrete mathematics book, as theory and examples took precedence over nice pictures, charts, and chapters intended for one or two lectures.

The inception of this book

As with many books in mathematics, the incipient version of this book was a collection of notes for students. Nearly a decade ago, I put together a few pages with some standard induction problems for discrete math students. To help their writing of inductive proofs, I then provided a template and a few pages of advice on writing up induction proofs, producing a small booklet for the students that I distributed in any course requiring induction.

Since there seemed to be no readily available books on induction (most were out of print), I originally had the idea to write something small that could be universally available as a supplement to courses in discrete mathematics, including linear algebra, combinatorics, or even geometry. My first goal was to have around a hundred of the standard exercises in induction, complete with solutions. I also wanted the solutions to be written in a format that students could follow easily and reproduce. When I began to collect and write up problems for this small planned booklet, I found so many examples and major theorems that engaged me, I couldn't wait to write them down and share them with anyone who would listen. I then tried to supplement this early collection to somehow give a fair treatment to all of the mathematical sciences, including computing science.

By that time, it was too late. As many collectors do, I became obsessed with finding different kinds of inductive proofs from as many areas as possible. Even after gathering many different types of questions, I continued to add to the collection, giving more examples of some types, and also including a healthy amount of set theory and foundations—in an attempt to give a "credible" or "scholarly" representation of the theory and applications surrounding induction. In a sense, I was constructing a tribute to one of the major proof techniques in mathematics.

After the book quickly burgeoned into a few hundred pages, people (including publishers) asked me, "for whom is this book?" or "can this book be used for any course?" I could only reply that this book will work well with nearly *any* mathematics or computing science course. Then I just kept adding problems! Only when the collection began to point north of 500 pages, did Chapman & Hall/CRC suggest that I put together an encyclopedia of induction, a handbook. So, I added a few hundred more pages, sampling from as many fields I had the courage and time for. This is the product.

Who is this book for?

This book is intended for anyone who enjoys a good proof, and for those who would like to master the technique of mathematical induction. I think that nearly every student and professor of mathematics and computer science can get a little something from this book.

Students may find inductive solutions for their practice or even their homework, they may learn how to write inductive proofs, and they may discover some interesting mathematics along the way. Most topics in this book include definitions and simple theory in order to deliver the exercises, and any student perusing these might acquire new interests in areas previously unexplored.

The professor may find examples to give as exercises, test questions, or contest practice questions. Some professors and high school teachers might appreciate sections here on writing mathematical induction proofs, both for themselves and in passing along such skills to their students.

The professor or student might also use this text for definitions and references, as well as many famous theorems and their proofs. This book is designed to be a source book for everyone interested in mathematics. When I was an undergraduate, I spent all of my spare money (and more) on reference books, including collections of worked exercises, and had this book been available back then, I would have most certainly purchased it—not only to help me with induction homework, but also as a resource of popular results and mathematical tricks.

This book may enhance nearly every course in mathematics—from freshman to graduate courses. At the university, mathematical induction is taught in many different courses, including those in discrete mathematics, graph theory, theoretical computer science, set theory, logic, combinatorics, linear algebra, and math education. Other areas, including courses in computing science, engineering, analysis, statistics, modelling, game theory, and economics now use induction as a standard tool. These and many other areas are treated generously.

Structure of this book

The book is essentially divided into three parts: "theory", "applications and exercises", and "solutions". These titles aren't completely accurate, as there are exercises and solutions in the theory part and there is theory in the exercises part. The theory part also contains far more than just theory, but a more appropriate title could not be found.

In the theory part, first a brief introduction is given. The introduction is not meant to be expository nor complete in a way that some discrete mathematics books might cover mathematical induction. The formal development of natural numbers from axioms is given by mathematical induction. Many readers will want to skip this section, as it can be a little dry, but this material can be understood and appreciated by most undergraduates in their second or third year. Having basic arithmetic skills in hand, different inductive techniques are discussed: well-ordered sets, basic mathematical induction, strong induction, double induction, infinite descent, downward induction, and variants of some of these.

Chapter 4 is about mathematical induction and infinity, including an introduction to ordinals and cardinals, transfinite induction, more on well-ordering, the axiom of choice, and Zorn's lemma. The material in Chapter 4 is intended for the senior math or computer science student, and can be omitted by the inexperienced reader. One reviewer suggested that this material be moved to much later in the book; however, I feel that it fits well from a logical perspective, perhaps just not from a pedagogical one when viewed by first-year students.

There are sections on the history of induction (Section 1.8) and the present state of literature on mathematical induction (Section 1.9). Fallacies and induction (Chapter 5) and empirical induction (Chapter 6) are also surveyed. Chapters 7 and 8 on doing and writing inductive proofs are given with the intention of helping the student and perhaps providing some guidelines that a teacher might use when teaching presentation skills. Much of these two chapters are directed at the student, and so the advanced reader can safely skip these.

Part II, "Applications and exercises", contains over 750 exercises, showcasing the different levels of difficulty of an inductive proof, the variety of inductive techniques available, and the scope of results provable by mathematical induction. Topics are grouped into areas, complete with necessary definitions, theory, and notation, so each chapter is nearly independent from all others. I tried to include some famous or fundamental theorems from most major fields. In many areas, I include some very specialized problems, if only because I enjoyed them. In general, exercises are not ranked according to difficulty, so expect surprises. Many advanced topics are covered here, so there are many examples appropriate for even graduate-level courses.

The number of published mathematical induction proofs is finite; however, one might get the impression that this number is infinite! There can be no comprehensive coverage. The present collection identifies results spanning many fields, and there seems to be no end of topics that I could continue to add. It seemed that whenever I researched some mathematical induction proof, I found yet another nearby. People have joked that, by induction, I could then find infinitely many examples. At some point, I had to (at least temporarily) wrap up the project, and this is the outcome.

In part, I feel like a travel guide commissioned to write a handbook about touring Europe; after staying in Budapest for a month but only driving through Paris, the "handbook" may seem like only a biased "guidebook". I have delved deeply into specialist areas, and only glossed over some more usual topics.

If this book survives to a second edition, many more topics will be developed. For example, the theory of Turing machines or Markov processes might make worthy additions. Additive number theory, computational geometry, the theory of algorithms and recursion might be developed. I welcome suggestions for possible future inclusion.

In Part III, solutions to most exercises are given. Solutions are most often written in a strict format, making them slightly longer than what might be ordinarily found in texts (and much longer than those found in journals). The extra structure does not seem to interfere with reading the proof and, in fact, it may sometimes help. I have also attempted to eliminate as many pronouns as possible, and have avoided the royal "we" that often occurs in mathematics.

Of the over 750 exercises, over 500 have complete solutions, and many of the rest have either brief hints or references.

For some unusual exercises presented here without solutions, I have tried to provide references. Many induction exercises are now "folklore" with origins difficult to trace, so citations often just direct the reader to at least one instance of where a problem has occurred previously. Readers are invited to inform me if I have missed some key citations.

There are nearly 600 bibliography references, and results are cross referenced and indexed thoroughly. I have given over 3000 index entries to assist in quick referencing. The bibliography is also back-referenced; bold face numbers following each entry indicate where in this book the entry is cited [*].

DSG

Winnipeg, Canada

Acknowledgements

I first learned of Peano's axioms in a course on elementary mathematics from an advanced viewpoint (PMAT 340) taught to education majors by the late Eric C. Milner in 1987 at the University of Calgary. Some of the ideas related here to Peano's axioms are adapted from those notes (together with Landau's presentation in [339]).

Thanks are owed to discrete math students at Mount Royal College, Calgary, Canada, in particular Nitin Puri, for encouraging me to assemble more problems for them. Gratitude is owed to Dr. Peter Morton of MRC for observations regarding Chapter 2.

Students at the University of Manitoba provided much appreciated help. Manon Mireault researched and organized early drafts. Trevor Wares also researched and proofread. Rob Borgersen helped with research, typos, and many challenges I had with LaTeX. Karen Johannson was incredible in finding (hundreds of) typos, doing research, helping with some proofs, and making pictures.

Colleagues at the University of Manitoba have been tremendous. Dr. Sasho Kalajdzievski was especially kind and careful with his detailed comments on the first few chapters, as was Dr. Tommy Kucera on the foundations and set theory.

Drs. Julien Arino, Rob Craigen, Michelle Davidson, Michael Doob, Kirill Kopotun, Shaun Lui, Ranganathan Padmanabhan, Craig Platt, Stephanie Portet, and Grant Woods have contributed problems or helped with proofreading.

Penny Gilbert and Professor George Grätzer shared some LaTeX secrets and advice on formatting and typesetting.

Professors Ben Li (CompSci), Bill Kocay (CompSci), and Brad Johnson (Stats) were also very kind with their help. Thanks are also owed to Jenö Lehel (University of Memphis).

I also thank Dr. Ann Levey (University of Calgary, Philosophy) for support and encouragement in the early stages of this book.

Many thanks are owed to Chapman & Hall/CRC, in particular, Bob Stern and Jennifer Ahringer for their kindness, almost infinite patience, and advice. They not only accepted my proposal, but asked for a more comprehensive treatment than I had originally planned and then helped to realize it, despite my many delays in submitting a final copy. The project editor, Karen Simon, was immensely helpful and patient while guiding the final preparation. The entire manuscript was proofread by Phoebe Roth, and Kevin Craig designed the cover. Many thanks are owed for a very thoughtful external review by Dr. Kenneth Rosen; his ideas have greatly improved this book.

This manuscript was typeset by the author in LaTeX, 11pt, book style.

About the author

David S. Gunderson obtained a B.Sc. and M.Sc. from the University of Calgary, and a Ph.D. in pure mathematics from Emory University under the supervision of Vojtech Rödl in 1995. He is currently an assistant professor and head of the mathematics department at the University of Manitoba. Previous positions include postdoctoral work at the University of Bielefeld, Howard University, and McMaster University. He is an elected fellow of the Institute of Combinatorics and Its Applications, and a member of many other mathematical societies.

His research interests are primarily in combinatorics, including Ramsey theory, extremal graph theory, combinatorial geometry, combinatorial number theory, and lattice theory. As a hobby, he has made many polyhedra from wood and other mathematical models for display and teaching (see his home page at `http://home.cc.umanitoba.ca/~gunderso/`).

Part I

Theory

This page intentionally left blank

Chapter 1

What is mathematical induction?

> *Induction makes you feel guilty for getting something out of nothing, and it is artificial, but it is one of the greatest ideas of civilization.*
>
> —Herbert S. Wilf,
>
> MAA address, Baltimore, 10 Jan. 1998.

1.1 Introduction

In the sciences and in philosophy, essentially two types of inference are used, deductive and inductive. Deductive inference is usually based on the strict rules of logic and in most settings, deductive logic is irrefutable. Inductive reasoning is the act of guessing a pattern or rule or predicting future behavior based on past experience. For example, for the average person, the sun has risen every day of that person's life; it might seem safe to then conclude that the sun will rise again tomorrow. However, one can not prove beyond a shadow of a doubt that the sun will rise tomorrow. There *may* be a certain set of circumstances that prevent the sun rising tomorrow.

Guessing a larger pattern based upon smaller patterns in observations is called *empirical induction.* (See Chapter 6 for more on empirical induction.) *Proving* that the larger pattern always holds is another matter. For example, after a number of experiments with force, one might conclude that Newton's second "law" of motion $f = ma$ holds; nobody actually proved that $f = ma$ always holds, and in fact, this "law" has recently been shown to be flawed (see nearly any modern text in physics, *e.g.*, [56, p.76]).

Another type of induction is more reliable: *Mathematical induction* is a form of reasoning that *proves*, without a doubt, some particular rule or pattern, usually infinite. The process of mathematical induction uses two steps. The first step is the "base step": some simple cases are established. The second step is called the "induction step", and usually involves showing that an arbitrary large example follows logically from a slightly smaller pattern. Observations or patterns proved by mathematical induction share the veracity or assurance of those statements proved by deductive logic. The validity of a proof by mathematical induction follows from basic axioms regarding positive integers (see Chapter 2 for more on the foundations of the theory).

In its most basic form, mathematical induction, abbreviated "MI", is a proof technique used to prove the truth of statements regarding the positive integers. (The statements themselves are rarely discovered using mathematical induction.) In this chapter, mathematical induction is only briefly introduced, with later chapters spelling out a more formal presentation.

It is easy to get excited about introducing the proof technique called "mathematical induction", especially since no mathematical aptitude or training is necessary to understand the underlying concept. With only very little high school algebra (and sometimes none at all!), mathematical induction enables a student to quickly prove hundreds of fascinating results. What more can a teacher ask for—an easy to understand technique complete with an amazing array of consequences!

1.2 An informal introduction to mathematical induction

To demonstrate the claim that no mathematical sophistication is necessary to comprehend the idea of mathematical induction, let me share an anecdote. When my daughter Christine decided to keep a stray cat as a pet, the two of them soon became inseparable—until it was time to go to bed. Christine slept in the top of a set of large bunk beds, but the cat was not so eager about climbing this strange contraption we humans know as a ladder. The cat, named Jupiter, sat on the floor meowing until I lifted him to Christine's warm bed each night. (He could jump down without fear, however, via the dresser.)

So I tried to teach Jupiter how to climb the ladder. (The cat probably could climb a ladder without my help, however it seemed as if he was waiting for permission—so for the sake of this story, assume that he did not know how.) There seemed to be two separate skills that Jupiter needed to acquire. First, he was apprehensive about just getting on the ladder, so with a little guidance and much encouragement, he discovered that he could indeed get on and balance on the first rung. Second, he had to learn how to climb from one rung to the next higher rung. I put his front paws on the next step and then tickled his back feet; to escape the tickle, he brought up his hind legs to the next rung. I repeated this on the next rung; he quickly realized how to go up one more (or that it was okay to do so?), and almost immediately upon "learning" this second skill, he applied it a few more times, and a moment later was rewarded with a big hug from Christine at the top.

That's the basic idea behind what is called "the principle of mathematical induction": in order to show that one can get to any rung on a ladder, it suffices to first show that one can get on the first rung, and then show that one can climb from any rung to the next. This heuristic applies no matter how tall the ladder, or even how far up the "first" rung is; one might even consider the 0-th rung to be the floor.

1.3 Ingredients of a proof by mathematical induction

In mathematical jargon, let $S(n)$ denote a statement with one "free" variable n, where, say, $n = 1, 2, 3, \ldots$. For example, $S(n)$ might be "the cat can get on the n-th rung of the ladder" or say, "rolling n dice, there are $5n+1$ totals possible" (see next section). To show that for every $n \geq 1$, the proposition $S(n)$ is true, the argument is often in two parts: first show that $S(1)$ is true (called the "base step"). The second part (called the "induction step") is to pick some arbitrary $k \geq 1$ and show that if $S(k)$ is true, then $S(k+1)$ follows. In this case, $S(k)$ is called the "inductive hypothesis". Once these two parts have been shown, if one were then asked to demonstrate that $S(4)$ is true, begin with $S(1)$, then by repeating the second part three times,

$$S(1) \rightarrow S(2); \quad S(2) \rightarrow S(3); \quad S(3) \rightarrow S(4).$$

This method succeeds in reaching the truth of $S(n)$ for *any* $n \geq 1$, not just $n = 4$.

The base step above need not have been $n = 1$. Sometimes induction starts a little later. For example, the statement $S(n)$: $n^2 < 2^n$ is not true for $n = 1, 2, 3$, or 4, but is true for any larger $n = 5, 6, 7, \ldots$. In this case, the base step is $S(5) : 5^2 < 2^5$, which is verified by $25 < 32$. The inductive step is, for $k \geq 5$, $S(k) \rightarrow S(k+1)$ (which is not difficult: see Exercise 159).

So the principle of mathematical induction can be restated so that the base step can be any integer (positive or negative or zero): [This is stated again formally in Chapters 2 and 3.]

> **Principle of mathematical induction**: For some fixed integer b, and for each integer $n \geq b$, let $S(n)$ be a statement involving n. If
> (i) $S(b)$ is true, and
> (ii) for any integer $k \geq b$, $S(k) \rightarrow S(k+1)$,
> then for all $n \geq b$, the statement $S(n)$ is true.

The expression "principle of mathematical induction" is often abbreviated by "PMI", however in this text, simply "MI" is used. In the statement of the principle of mathematical induction above, (i) is the base step and (ii) is the induction step, in which $S(k)$ is the inductive hypothesis. A proof that uses mathematical induction is sometimes called simply "a proof by induction" when no confusion can arise.

For an assortment of reasons, mathematical induction proofs are, in general, easy. First, the general rule often does not need to be guessed, it is usually given. A great deal of work is often required to guess the rule, but an inductive proof starts after that hard work has been done. Another aspect of proving by mathematical induction that makes it easy is that there are usually clearly defined steps to take, and when the last step is achieved, the logic of the proof makes the answer undeniable. For some, the most challenging part of an inductive step is only in applying simple arithmetic or algebra to simplify expressions.

A proof by mathematical induction has essentially four parts:

1. Carefully describe the statement to be proved and any ranges on certain variables.

2. The base step: prove one or more base cases.

3. The inductive step: show how the truth of one statement follows from the truth of some previous statement(s).

4. State the precise conclusion that follows by mathematical induction.

For more on the structure of a proof by mathematical induction, see Chapters 2, 3; for the reader just learning how to prove by mathematical induction, see Chapter 7 for techniques and Chapter 8 for how to write up a proof by mathematical induction.

1.4 Two other ways to think of mathematical induction

Many authors compare mathematical induction to dominoes toppling in succession. If the b-th domino is tipped, (see Figure 1.1) then all successive dominoes also fall.

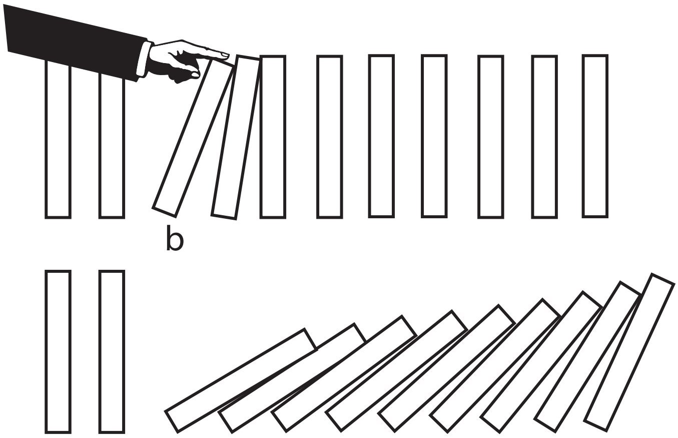


Figure 1.1: Dominoes fall successively

This comparison allows one to view mathematical induction in a slightly more general form, since all dominoes need not be in a single row for the phenomenon to work; as long as each "non-starting" domino has one "before it" which is close enough to topple it. So, in a sense, mathematical induction is not just done from any one integer to the next; induction can operate for many sequences of statements as long as for each non-initial case, there is a previous case by which one can use a rule to jump up from.

Another analogy for mathematical induction is given by Hugo Steinhaus n *Mathematical Snapshots* [508] [in the 1983 edition see page 299]: Consider a pile of envelopes, as high as one likes. Suppose that each envelope except the bottom one contains the same message "open the next envelope on the pile and follow the instructions contained therein". If someone opens the first (top) envelope, reads the message, and follows its instructions, then that person is compelled to open envelope number two of the pile. If the person decides to follow each instruction, that person then opens all the envelopes in the pile. The last envelope might contain a message "Done". This is the principle of mathematical induction applied to a finite set, perhaps called "finite induction". Of course, if the pile is infinite and each envelope is numbered with consecutive positive integers, anyone following the instructions would (if there were enough time) open all of them; such a situation is analogous to mathematical induction as it is most often used.

1.5 A simple example: Dice

Here is an example of a problem, a conjecture, and a proof of this conjecture by mathematical induction.

When rolling a single die, there are six possible outcomes: 1,2,3,4,5,6. When rolling two dice, there are 11 possible totals among two dice: 2,3, . . . , 12, and for three dice, the 16 possible totals are 3,4,. . . , 18. After a moment of reflection, one might guess that for $n \geq 1$ dice, the number of possible totals is $5n + 1$.

Proposition 1.5.1. *The number of possible totals formed by rolling $n \geq 1$ dice is $5n + 1$.*

Proof: (By mathematical induction on n) For each positive integer n, denote the statement

$$S(n): \quad \text{When rolling } n \text{ dice, there are } 5n + 1 \text{ possible totals.}$$

So $S(1)$, $S(2)$, $S(3)$, . . . form an infinite family of statements. (Using mathematical induction, all such statements are proved.)

Base step: The statement $S(1)$ is already verified as there are $6 = 5(1) + 1$ outcomes.

Inductive step: Fix $k \geq 1$ and suppose that $S(k)$ is true (the inductive hypothesis), that is, among k dice, there are $5k + 1$ possible outcomes. To complete the inductive step, one needs only to show that the subsequent statement

$$S(k+1): \quad \text{When rolling } k+1 \text{ dice, there are } 5(k+1)+1 \text{ possible totals}$$

is also true.

Consider $k+1$ dice, say $D_1, D_2, \ldots, D_k, D_{k+1}$. Among the first k dice there are (by the inductive assumption $S(k)$) $5k+1$ possible totals. Among these totals, the smallest possible is k (where each dice shows 1), and so the lowest total possible using all $k+1$ dice is $k+1$ (when D_{k+1} also shows 1). The highest possible total for all the first k dice is $6k$ (when each of $D_1, \ldots, D_k$ show a 6). Then using D_{k+1}, each of $6k+1, 6k+2, \ldots, 6k+6$ is a new possible total. Hence, there are six new possible totals, and one old possible total (k) which no longer occurs among $k+1$ dice. Hence, there are 5 more totals possible with $k+1$ dice than with k dice, giving $5k+1+5 = 5(k+1)+1$ outcomes as desired. This completes the inductive step.

Hence, one concludes by mathematical induction that for any $n \geq 1$, the statement $S(n)$ is true. This concludes the proof of Proposition 1.5.1. □

[The "□" indicates the end of a proof.]

1.6 Gauss and sums

It seems to be tradition in teaching induction that the first example demonstrating how well MI can work is in proving a formula for summing the first n positive integers.

There is a story about a young Carl Friedrich Gauss (1777-1855) that is often told. I first give the apocryphal version, which is an over-simplification of the supposed facts, because it so aptly creates a segue to the inductive proof. [The more historical version—which is even more unbelievable—is given after the proof of Theorem 1.6.1.]

Gauss was extremely quick as a child, and his teachers had a tough time keeping ahead of him. To keep Gauss busy, his teacher once asked him to sum the numbers from 1 to 100—to which Gauss almost immediately replied "5050". Perhaps he had discovered the following fact.

Theorem 1.6.1. *For each positive integer n,*

$$1+2+3+\cdots+n = \frac{n(n+1)}{2}.$$

Proof of Theorem 1.6.1 by MI: Let $S(n)$ be the statement

$$S(n): \quad 1+2+3+\cdots+n = \frac{n(n+1)}{2}.$$

Base step ($n = 1$): The statement $S(1)$ says that $1 = \frac{1(2)}{2}$, which is clearly true, so $S(1)$ holds.

Inductive step($S(k) \to S(k+1)$): Fix some $k \geq 1$, and suppose that

$$S(k): \quad 1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}$$

holds. (This statement is called the *inductive hypothesis.*) To complete the inductive step, it suffices to verify that the statement

$$S(k+1): \quad 1 + 2 + 3 + \cdots + k + (k+1) = \frac{(k+1)(k+2)}{2}$$

also holds. Beginning with the left-hand side of $S(k+1)$,

$$\begin{aligned} 1 + 2 + 3 + \cdots + k + (k+1) &= (1 + 2 + 3 + \cdots + k) + (k+1) \\ &= \frac{k(k+1)}{2} + (k+1) \qquad \text{(by ind. hyp.)}, \\ &= (k+1)\left(\frac{k}{2} + 1\right), \\ &= (k+1)\left(\frac{k+2}{2}\right), \end{aligned}$$

which is equal to the right-hand side of $S(k+1)$. Hence $S(k) \to S(k+1)$ is proved, completing the inductive step.

Conclusion: By the principle of mathematical induction, for each $n \geq 1$, the statement $S(n)$ is true. □

Many statements provable by mathematical induction are also provable in a direct manner. For example, here is one of many other proofs of the expression in Theorem 1.6.1:

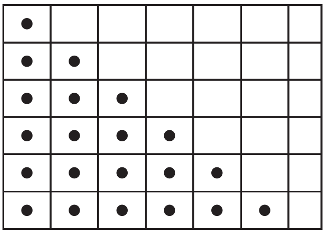
Direct proof of Theorem 1.6.1: (without explicit use of MI) Write the sum $s(n) = 1 + 2 + \cdots + n$ twice, the second time with the summands in reverse order, and add:

$$\begin{array}{rclccccccc} s(n) & = & 1+ & 2 & + & 3 & +\cdots+ & (n-1) & +n \\ s(n) & = & n+ & (n-1) & + & (n-2) & +\cdots+ & 2 & +1 \\ \hline \\ 2s(n) & = & (n+1)+ & (n+1) & + & (n+1) & +\cdots+ & (n+1) & +(n+1) \end{array}$$

The summand $(n+1)$ occurs n times, and so $2s(n) = n(n+1)$; division by 2 completes the proof. □

The numbers $T_n = 1 + 2 + 3 + \cdots + n$ are called the *triangular numbers.* One reason that they are called triangular might be because if one makes a diagram with n rows of dots, starting with one dot in the first row, and in subsequent rows putting one more dot, then the dots form a triangle, and T_n is the total number of dots.

Here is an example for $n = 6$:



To compute T_n of Theorem 1.6.1, put an n by $n+1$ box around such a triangle, and notice that T_n accounts for half of the box. See also Nelsen's wonderful little book *Proof without words* [403, p. 69], where the caption is "—"The ancient Greeks" (as cited by Martin Gardner)". Another similar "Proof without words" of the formula for T_n is given by Ian Richards [453] (also reprinted in [403, p. 70]). See also [404, p. 83]. One can also think of the triangle above as being equilateral. For other polygons, there are other "figurate numbers", for example, $n(3n-1)/2$ is a pentagonal number (the square numbers you already know). See the wonderfully illustrated [116, pp. 38ff] for more on polygonal (and polyhedral) numbers. [Polygonal numbers are also a rich source for induction problems as most are defined recursively, though few appear in this volume.]

For a moment, return to Gauss in the classroom. Expanding on the account given above, here is an excerpt from E. T. Bell's *Gauss, Prince of Mathematicians* [44] (also found in Newman's 1956 anthology [45]):

> Shortly after his seventh birthday Gauss entered his first school, a squalid relic of the Middle Ages run by a virile brute, one Büttner, ...
>
> Then, in his tenth year, Gauss was admitted to the class in arithmetic. As it was the beginning class none of the boys have heard of an arithmetical progression. It was easy then for the heroic Büttner to give out a long problem in addition whose answer he could find by a formula in a few seconds. The problem was of the following sort, $81297 + 81495 + 81693 + \cdots + 100899$, where the step from one number to the next is the same all along (here 198), and a given number of terms (here 100) are to be added.
>
> It was the custom, of the school for the boy who first got the answer to lay his slate on the table; the next laid his slate on top of the first, and so on. Büttner had barely finished stating the problem when Gauss

> flung his slate on the table: “There it lies,” he said—“*Liggit se*” in his pleasant dialect. Then, for the ensuing hour, while the other boys toiled, he sat with his hands folded, favored now and then by a sarcastic glance from Büttner, who imagined the youngest pupil in the class was just another blockhead. At the end of the period Büttner looked over the slates. On Gauss’ slate there appeared but a single number. To the end of his days Gauss loved to tell how the one number he had written was the correct answer and how all the others were wrong.

1.7 A variety of applications

One aspect of mathematical induction is that it can be found in the proofs of a broad spectrum of results. In this section a sample is given of areas that mathematical induction is found.

Hundreds of equalities and inequalities have proofs by induction. For example, Exercise 54 asks to show the well-known formula

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Trigonometric identities also can be proved by induction, as in Exercise 124 where for any real number x and $n \geq 1$,

$$\cos^{2n}(x) + \sin^{2n}(x) \geq \frac{1}{2^{n-1}}.$$

Many such identities (or inequalities) are proved in a manner very similar to that in Theorem 1.6.1. Some inequalities have mathematical induction proofs that are not so evident. For example, in Exercise 204, induction is applied to show that any positive integer n,

$$\sqrt{2\sqrt{3\sqrt{4 \cdots \sqrt{n}}}} < 3.$$

Suppose that a sequence of numbers is defined recursively, that is, a few initial values are given, and then a formula or rule shows how to get the nth number from earlier numbers. For example, define a sequence $a_0, a_1, a_2, a_2, \ldots$ by first setting $a_0 = 3$ and let $a_1 = 3$. Then for each $n \geq 2$, define $a_n = 2a_{n-2} + a_{n-1}$, each a combination of the two previous values. Working out the first few values are $3, 3, 9, 15, 33$. There is a method by which to come up with a formula for the general term a_n; however, one might also guess that for each $n \geq 0$,

$$a_n = 2^{n+1} + (-1)^n.$$

Mathematical induction can be used to prove that this guess is correct. In the theory of recursion, mathematical induction is indispensable in proving correctness of formulas or algorithms. See Chapter 16 for mathematical induction in the theory of recursion. Many popular algorithms are analyzed here by mathematical induction.

Induction can also solve problems that have no apparent equation associated with it. For example, on a circular track, put n cars (with engines off), and among all cars, distribute just enough gas for one car to go around a track. In Exercise 590, induction is used to prove that there is a car that can make its way around a lap by collecting gas from the other cars on its way.

In an election, a votes are cast for candidate A and $b < a$ votes cast for candidate B. In Exercise 764, one counts the number of ways $a + b$ votes can be ordered so that after each vote, candidate A is winning. Similar results have an impact in game theory, probability theory, and economics.

Various forms of mathematical induction can be used to prove very general and powerful results about infinite structures. For example, a special form of mathematical induction, called Zorn's lemma, is applied in Exercise 692, to show that every vector space has a basis.

An abundance of results in discrete math and graph theory are proved by induction. For example, if a graph on n vertices has more than $n^2/4$ edges, Exercise 509 shows that the graph always contains a triangle. Problems in geometry (see Chapter 20) have surprising solutions using induction, as well.

Many basic counting principles have proofs by mathematical induction; for example, both the pigeonhole principle and the inclusion-exclusion principle have proofs by induction (see Exercises 743 and 427, respectively).

Model theory, foundations of mathematics, and computing theory are highly reliant on inductive proof techniques. Most elementary properties of arithmetic are derived using induction.

Mathematical induction is often associated with discrete counting; however, it can be used to prove many results in calculus and analysis. For example, starting with the simple product rule $(fg)' = f'g + fg'$, by induction one can prove (see Exercise 611) an extended version:

$$(f_1 f_2 \cdots f_n)' = f_1' f_2 f_3 \cdots f_n + f_1 f_2' f_3 \cdots f_n + \cdots + f_1 f_2 \cdots f_{n-1} f_n'.$$

This example hints at a theme.

Very loosely speaking, there are countless examples in mathematics where a concept is generalized or extended from one dimension to two; then from two to three; if a pattern becomes obvious in these first jumps, the pattern often describes a recursion, one that can serve as a model for an induction step taking the concept to any finite dimension required. The same is true for linear algebra and matrix theory; in fact, it might appear that most concepts in linear algebra "grow by induction" from smaller ones. See Exercises 637–668 for what might seem to be most of the major results in matrix theory, including a few applications, all proved by mathematical induction.

After only a brief perusal of the exercises in this book, one might conclude that most of mathematics is tied to induction. To many, this comes as no surprise, because counting numbers and basic rules of arithmetic and algebra are either developed or proved true using induction. Hence nearly all of discrete mathematics is based on induction, in a sense.

The first part of Chapter 2 establishes some useful notation and terminology, and the latter parts of that chapter are for those interested in the theory behind induction. To continue the introduction to mathematical induction, Chapter 3 gives examples of the many different inductive techniques and examples of each. If the reader is just beginning to learn induction and how to write proofs, I recommend also reading Chapters 7 and 8.

1.8 History of mathematical induction

> *I have read somewhere or other, in Dionysius of Halicarnassus, I think, that History is Philosophy teaching by examples.*
>
> —Henry St. John (Viscount Bolingbroke) (1678–1751),
>
> *On the study and use of history.*

A usual (modern) development of the principle of mathematical induction begins with Peano's axioms. In this book, too, this approach is adopted. This perspective is admittedly a bit naive, since there were many other key players in establishing the present confidence held in the concept.

It is not clear who first used mathematical induction, but in Bussey's 1917 article [91], he reported that Blaise Pascal (1623–1662) recognized that an Italian named D. Franciscus Maurolycus (1494–1575) (also spelled *Francesco* or *Francesko Maurolico* or *Maurolyci*) used induction in his book [376] published in 1575. In that book, (actually, in Book I) he proved by induction that the odd numbers are formed by successively adding 2 to the first odd number, 1. Maurolycus used what is now called "induction" to prove that the sum of the first n odd numbers is n^2. These and many other ideas were learned by Blaise Pascal, in the mid 1600s, with Pascal perhaps being the first to apply induction for a formula for the sum of the first n natural numbers. In Struik's [515, p. 106] *A Concise History of Mathematics*, two works ([201] and [448]) are cited as evidence that "He [Pascal] was the first to establish a satisfying formulation of the principle of complete induction."

Maurolycus' proof of the formula for the sum of the first n numbers was non-inductive, although Georg Cantor (1845–1918) claimed that Pascal got his inductive proof from Maurolycus; Bussey refutes this claim. Cantor (Georg Ferdinand Ludwig Philip Cantor) once claimed that Pascal was the originator of mathematical induction, but later withdrew his claim after he was informed by someone named G. Vacca about Maurolycus (see [544]). So it seems, Pascal learned induction from Maurolycus.

It might be interesting to note that Bussey's article was published while Cantor was still alive. Cantor quit teaching at the University of Halle in 1905, was very ill late in life, and died in a mental hospital in Halle in 1918, so perhaps he never saw the article. Cantor is now credited with being the founder of set theory, particularly, the theory of infinite sets.

In George Pólya's (1887–1985) 1962 book *Mathematical Discovery* [435], mathematical induction is credited to Pascal as well, but in Bourbaki's *The Set Theory* [69] (1965), "Maurolico F." receives credit. [Bourbaki was not a person, but a group of sometimes 20 persons, at various times including C. Chevally, J. Delsarete, J. Dieudonne, and A. Weil—they had to retire from the group at age 50.]

It seems odd that such a simple technique was only learned in the 16th century. In fact, it would not be a surprise if Euclid (*ca.* 330–275 BC) used mathematical induction, though there does not seem to be any explicit instance of it. It might be worth noting that Euclid's result that states that there are infinitely many primes can be easily proved by induction; see Exercise 207. This has led some authors to the opinion that Euclid used, if even tacitly, induction. The debate as to whether or not Euclid knew of mathematical induction has gone as far as to interpret induction less formally. For more on Euclid and induction, see [175], [197], [541], [542], and [557]. It has been suggested [523] that Pappus (*ca.* 300AD) also knew of induction, though I have not yet seen the evidence. Even Plato might have known of the principle (see [3]).

The method of mathematical induction has been compared to the "method of exhaustion", due to Eudoxus (408–355 B.C.) [about a century before Euclid] and used by Archimedes (287–212 B.C.) in his derivation of many formulas (for areas and volumes), and his "method of equilibrium"—which often uses the method of slicing—called the *method of indivisibles* by Cavalieri (1598–1647), a technique still used in modern integral calculus. The method of exhaustion begins with an assumption that magnitudes can be divided an infinite number of times. For example, the method can be used to prove that the formula $A = \pi r^2$ for the area of a circle is correct by finding larger and larger polygons that fit inside a circle. (See [180, 11-3] for a details.) What this method has in common with mathematical induction is that a formula must first be guessed, and the proof is an iteration of (perhaps) infinitely many steps, often based on some kind of recursion depending on earlier steps. Some proofs by the method of exhaustion can be translated into proofs by induction, however the method of induction does not seem to be used explicitly by any of these masters from (nearly) ancient times.

Internet sources suggest that Ralbag (Rabbi Levi Ben Gershon) gave proofs that used induction in the 13th century. One such correspondence was from Boaz Tzaban, Bar Ilan University, Israel; another was from Ed Sandifer at Western Connecticut State University, Danbury, CT. They reported on a talk given by Shai Simonson of Stonehill College in Massachusetts, a scholar of Ben Gershon's work. It is not clear that Gershon formalized the concept, but there seems to be some agreement that he used mathematical induction. For more support on these ideas, see [444]. Many other authors report on the use of induction or inductive techniques by al-Karaji (sum of cubes formula, around 1000 A.D.), al-Haytham (sum of fourth powers), and al-Samawal (binomial theorem). [I have not yet directly seen these references, however, more information is available in [307].]

According to Quine [443, p. 243], "Mathematical induction was used and explicitly recognized by Pascal in 1654 ... and Fermat in 1659 ... But the principle of mathematical induction retained the status of an ultimate arithmetic axiom until 1879 when Frege defined the ancestral and by its means the class of natural numbers." Quine also says that "...Such inference, called *mathematical induction*, is afforded by the following metatheorem" and then uses very careful (and barely readable) logical notation to give the metatheorem.

Grimaldi reports in his textbook on discrete mathematics [238], that it was Augustus DeMorgan (1806–1871) who, in 1838, carefully described the principle and gave it its present name "mathematical induction". The reference Grimaldi gave for this fact was Bussey's paper [91], however, a quick look at Bussey's paper does not seem to confirm this. In fact, on the website *Earliest Known Uses of Some of the Words of Mathematics* [384], it is reported[1]:

> The term INDUCTION was first used in the phrase *per modum inductionis* by John Wallis in 1656 in *Arithmetica Infinitorum.* Wallis was the first person to designate a name for this process; Maurolico and Pascal used no term for it (Burton, page 440).

and

> The term MATHEMATICAL INDUCTION was introduced by Augustus de Morgan (1806-1871) in 1838 in the article *Induction (Mathematics)* that he wrote for the *Penny Cyclopedia.* De Morgan had suggested the name *successive induction* in the same article and only used the term *mathematical induction* incidentally. The expression *complete induction* attained popularity in Germany after Dedekind used it in a paper of 1887 (Burton, page 440; Boyer, page 404).

The references for the above citations are Boyer [70] and Burton [89]. See also [92] for more on the history of the name "mathematical induction". One might note that the method of mathematical induction still is occasionally referred to as "complete induction" (*e.g.*, in [556]) or "full induction."

Near the end of the 19th century, David Hilbert (1862–1943) was writing a book [269], attempting to establish geometry based not on "truths", but on axioms. Gottlob Frege (1848–1925) had been studying mathematical logic and communicated regularly with Hilbert. Much debate arose about what axioms were, what they "should" be, and what "truth" in mathematics is. (See [451] for an account of

[1] Used with kind permission from Jeff Miller

the discussions between Frege and Hilbert regarding axioms.) Frege was essentially trying to reduce mathematical reasoning to purely logical reasoning. For some kinds of reasoning, a “second-order” kind of logic was necessary, but Frege wanted (perhaps) to rephrase mathematical induction that did not rely on second-order logic. To this end, he used terms like “ancestors” (well, in German, he must have used “Vorfahren” or something similar) and “ancestor induction”. The basic idea was to extend reasoning of the form: “Ole is an ancestor of John, and John is an ancestor of David, so Ole is an ancestor of David.” [These, inasmuch as my parents tell me, are accurate statements.]

In [128] [thanks to Dr. Peter Morton for supplying this reference] Demopoulos mentions that Crispin Wright presented an argument that *Hume’s principle* [the number of elements in a set S is equal to the number of elements T if and only if there is a one-to-one correspondence between elements of S and T] implies one of Peano’s axioms: “...in the context of the system of second-order logic of Frege’s *Bereffsschrift*, Peano’s second postulate [every natural number has a successor] is derivable from Hume’s principle.” Demopoulos continues to mention “...that Frege discovered that, in the context of second-order logic, Hume’s principle implies the infinity of the natural numbers, *Frege’s theorem*.” (If the reader wants another perspective, readable but confusing, on these matters, see [556].)

Ernst Zermelo (1871–1953), Richard Dedekind (1831–1916), Bourbaki, Bertrand Russell (1872–1970), and many others continued the debate regarding assumptions about the natural numbers. Concepts like “well-ordering” and “Axiom of Choice” were also introduced in an attempt to logically legitimize what students of mathematics all “know” to be “true” about natural numbers. For present purposes, assume that all the necessary groundwork has been done to establish that present assumptions (or Peano’s assumptions) are reasonable. For more facts and debates regarding the history of induction, see [175], [197], [300], [523], [541], [542], and [581].

The interested reader may pursue these discussions from a model theoretic perspective as well; the mathematical logician Leon Henkin [265] examines Peano models in contrast to induction models (those with only the induction axiom). Classifying algebraic systems according to the set of axioms that generate the system, and examining which functions arise from “primitive recursion”, is too deep a subject to entertain here. The reader is recommended to see some of the popular literature that is referred to in Section 2.2. The theory can get quite complex; it is hard to say what the *best* approach is.

Instead of being drawn into further discussions regarding epistemology and philosophy, this discussion is concluded with a quotation from Ernst Mach, as found in [433], regarding Jacques Bernoulli (1667–1748):

> Jacques Bernoulli’s method is important to the naturalist. We find what seems to be a property A of the concept B by observing cases $C_1, C_2, C_3, \ldots,$. We learn from Bernoulli’s method that we should not

attribute such a property A, found by incomplete, non-mathematical induction, to the concept B, unless we *perceive* that A is *linked* to the characteristics of B and is *independent* of the variation of the cases. As in many other points, mathematics offers here a model to natural science.

1.9 Mathematical induction in modern literature

One of the chiefest triumphs of modern mathematics consists in having discovered what mathematics really is.

—Bertrand Russell

International Monthly, 1901.

In any mathematics textbook that contains a section on induction, there is usually a collection of problems, a handful of which are now used repeatedly in nearly every such text. There are perhaps about a hundred problems *in toto* that might, due to their frequency, be called "standard"; virtually all problems appearing in modern texts are adaptations of these. A few books have been devoted exclusively to induction. This chapter contains a brief overview of books specifically on induction, articles about induction, and typical books containing chapters or sections on induction, primarily from the last century; for articles concerning mathematical induction before that, see Section 1.8 on the history of mathematical induction. This overview hopefully contains most major works and a few less well-known. Aside from references given here, there are likely thousands more articles concerning induction, so parts of this review can never hope to be comprehensive. On 11 February 2009, MathSciNet showed 395 matches to "mathematical induction", 74 of which were in the title. There were 1436 titles containing simply the word "induction", most in well respected refereed journals. The number of books or articles that use induction in them is probably in the hundreds of thousands.

My own introduction to induction in high school was from *Mathematical Induction and Conic Sections* [550], a booklet excerpt from a textbook. That booklet has only a few pages on induction, but it lists 39 exercises. There have been a few other books specifically on induction, most of which I only recently became aware of, and none of which seem to be in print any more.

In 1958, a 63-page book [388] by Mitrinović on mathematical induction appeared in Serbo-Croatian, the last chapter of which contains a short history of induction. The translated title was *The Method of Mathematical Induction*. A dozen years later, the same author came out with another book [389], about half of which is reportedly devoted to problems solvable by mathematical induction, (also in Serbo-Croatian), however I have not seen either.

In 1956, I. S. Sominskii's Russian text [498] on induction was already enjoying its fourth edition. In 1959, he published *Metod Matematicheskoii Induktsii*; this was translated into English and published in 1961 as *The Method of Mathematical Induction* [499], a 57-page collection of theorems and 52 problems; most appear with helpful, complete solutions. A reviewer named N. D. Kazarinoff reviewed that book for *Math Reviews* [27 5669] and wrote "In addition to a high school training in these subjects, the reader must have good eyesight: symbols in formulas are often about the size of periods." This book has enjoyed dozens of editions in various languages, including Russian (*e.g.*, [498], 4th ed., 48 pages), German (*e.g.*, [501], 13th ed., 55 pp., [287], 120 examples, 183 pp., with two other authors), Spanish ([502], 2nd ed., 62 pp.), and Turkish (*e.g.*, [500], viii+72 pp.).

In 1964, a 55 page booklet, *Mathematical Induction* [582], by Bevan K Youse [note: there is no period after the "K" in his name] appeared, repeating many of the problems in Sominskii's book, but with a few interesting additions. Youse's book has 72 problems, most of which now commonly appear in today's texts without reference. There are only 29 complete solutions.

In 1979, the 133-page book *Induction in Geometry* [220], published in Moscow, contains inductive proofs of many difficult theorems in geometry (only a few of which are covered in this volume). This book is no longer in print and is hard to find [thanks to R. Padmanabhan for giving me his copy], but, in my opinion, well worth an effort to locate.

Another, more recent book is *Manuel d'Induction Mathématique* (Handbook of mathematical induction) by Luís Lopes [350]; this book has 100 problems complete with solutions (in French), many of which are also standard and easy; however, the author does not shy away from some really challenging solutions. The exercises occupy just over a dozen pages, with the bulk of the 127 pages being solutions.

The principles behind mathematical induction are studied in almost every logic text or set theory text (for example, in [95], [289]). There are numerous articles on mathematical induction from different points of view in logic, language, model theory, universal algebra, or philosophy (*e.g.*, [200] on predicate synthesis, [265] on model theory, [383] on formal theory of finite sets, [145] on variable free algebra and Galois connections, [111] on material implication, [139] on predicates on any well-founded set, [471] on ramified type theory as an adequate formalization of predictive methods).

More general works, like [181], [274], [400], and [556] give broad historical perspective in the modern foundations of mathematics and induction. History of mathematics texts almost always describe how induction arrived on the mathematical scene (*e.g.*, [180]) and how it relates to other areas of mathematics.

Hundreds of references have been used in assembling the collection of exercises here. Many problems using mathematical induction are now part of the folklore, but unusual problems are referenced. Here are a few kinds of books that deal explicitly with mathematical induction.

Many texts in discrete mathematics have sections on induction (*e.g.*, see [10], [8], [33], [38], [52], [55], [83], [147], [195], [222], [238], [292], [299], [355], [363], [373],

[375], [431], [462], [464], and [535]). Of these volumes, [238], [292], and [462] are very popular in North America, probably because of the colossal amount of mathematics (including induction, of course) contained in each.

Closely related are books on combinatorics, many with a prodigious array of applications of induction (for example, see [6], [77, 78], [94], [255], [266], [455], or [506]). Lovász's now classic compilation, *Combinatorial Problems and Exercises* [354] is also an abundant source of wonderful inductive proofs, many highly non-trivial. Also, for induction in advanced combinatorics, see [58].

One might be delighted to know that even some calculus books (for example, the classic book by Apostol [20], and the more modern text by Trim [534]) devote a section to induction. Books on programming cover induction, as well (see, *e.g.*, [483]). Texts that concentrate on mathematical problem solving often contain sections on induction and are a rich source of problems. In particular, Engel's book *Problem-solving Strategies* [161] contains a chapter on induction in which 39 exercises and solutions are discussed; hundreds of solutions using induction also occur throughout the book. [Some solutions are little on the brief side, but considering the plethora of problems that are actually solved, Engel's work might be considered as one of the richest sources for problem solving available today.] Three more references of this type that come to mind are [47], [124] and [461]. Such texts are an invaluable resource for mathlete training. Other works concentrate on aspects of teaching induction (*e.g.*, [194], [382], [490], and [516] to name but a few).

For anyone wanting a general insight into how to conjecture and prove mathematical statements, particularly by induction, one might be pleasantly rewarded with a look at Pólya's books [433], [434], [435]. A fairly recent collection of non-trivial problems over a broad range of fields, many of which employ induction, quickly became one of my favorites: *The Art of Mathematics: Coffee Time in Memphis* [61], by Béla Bollobás.

Leo Zippin's classic monograph *Uses of Infinity* [589] shows off induction in various settings, most notably in proving limits. In *What is Mathematics?* [120, §§1.2.1–1.2.7, pp. 9–20] by Richard Courant and Herbert Robbins one finds a particularly easy-to-read discussion of mathematical induction. (Zippin, [589, p. 106] also refers the reader to the Courant and Robbins book.) Another, more recent delightful problem book (which has a section on induction, and various induction problems throughout) is *Winning Solutions*, by Edward Lozansky and Cecil Rousseau [357], a collection of contest problems and their solutions that might complement any library.

Some books on recreational mathematics and popular science include discussion of mathematical induction. One of the most noteworthy of these is Martin Gardner's *Penrose Tiles to Trapdoor Ciphers*, [214, Ch. 10, pp. 137–149], a chapter called "Mathematical induction and colored hats". Another, [560], discusses Penrose's non-computability of thought, consciousness, self-referencing, and discusses mathematical thinking viz-a-viz Gödel's theorem, Poincaré, and Galois, and somehow manages to tie in mathematical induction.

There has been some work on computer programs designed to produce inductive proofs (also called “automated induction”). See, for example, [28], [68] (using SPIKE), [87], [187], [303], [398], [543], [557] (using LISP), and [588]. There is a great deal more literature on this subject, as proofs by mathematical induction are central in many computer science and AI applications. An older article [139] highlights the importance of mathematical induction in termination of programs and properties of programming languages.

A special kind of mathematical induction, called “transfinite induction” (see Section 4.2) is closely related to complexity theory in [489]. The invariance theorem [whatever that is] and induction are studied in [278]. Induction and program verification and modelling are also closely related and many books and articles discuss this relation (see, *e.g.*, [302], [361], [452]).

Many texts with “finite mathematics” in the title contain sections on mathematical induction, as induction is often taught in high school and beginning university math courses. Various other subject areas (for example, number theory, algebra, and graph theory) use induction quite heavily, and some related texts contain sections on induction (*e.g.*, [150], [566]).

One can find numerous articles on induction in various popular journals, too; for example, see [82] or [265]. The article by Dragos Hrimiuc [280] is short (3 pages!) and easy to read, yet is a substantial introduction to the subject. Some are from a historical perspective (*e.g.*, [91], [175], [197], [300], [523], [541], [542], [581]). There are a variety of journal articles on induction in general (*e.g.*, [138], [175], [237] (in Spanish), [262] (in Japanese), [290] (in Chinese), and [504]).

Induction is not only applied in discrete situations. Analysis and induction are more closely related than one might think (see [155] for some classical connections). In fact, there is a kind of *continuous*, or *non-discrete* induction at play. Some of the first (and most referred to) articles in this area seem to be by Pták [441, 442] (with the Banach fixed point theorem, Banach algebra, closed graph theorem, Newton’s process, and more); see also [26], [25], [27], [578], [579]. For those who can read Russian and are interested in differential equations, see [318].

Induction is ubiquitous. In fact, in any volume of a mathematics journal (popular or specialized) it seems rare *not* to find at least one proof by induction!

Incidentally, it might come as a bit of a surprise that the word “induction” does not seem to be mentioned in George Gamow’s classic book *One Two Three … Infinity*—one can be comforted, though, by the knowledge that Gamow [205, pp. 19–23] explains well two problems that are solved inductively.

Finally, there is the internet. In September 2005, a *Google* search for “mathematical induction” produced “about 2,610,000” hits! For some reason, this number dropped to 436,000 as of January 2009. Any ranking of these sites is hopeless, however, many seem to be rather well done. The sites seem to range from the very elementary to some collections of somewhat challenging problems.

Chapter 2

Foundations

> *The reasoning of mathematicians is founded on certain and infallible principles. Every word they use conveys a determinate idea, and by accurate definitions they excite the same ideas in the mind of the reader that were in the mind of the writer. When they have defined the terms they intend to make use of, they premise a few axioms, or self-evident principles, that every one must assent to as soon as proposed. They then take for granted certain postulates, . . . , and from these plain, simple principles they have raised most astonishing speculations, and proved the extent of the human mind to be more spacious and capacious than any science.*
>
> —John Adams,
>
> *Diary.*

This chapter attempts to put mathematical induction (MI) on a sound logical ground, and the principle of mathematical induction is described more formally. The usual starting point is a set of axioms called “Peano’s axioms”, the last of which is, essentially, the principle of mathematical induction. Using these axioms one can prove many of the basic properties of natural numbers, perhaps a reasonable place to start in mathematics.

2.1 Notation

The notation used in this text is fairly standard. If S is a set, “$x \in S$” denotes that x is an element of S. The notation “$x, y \in S$” is a common shorthand for “$x \in S$ and $y \in S$”. Use “$T \subset S$” or “$T \subseteq S$” to denote that T is a *subset* of S, that is, every element of T is an element of S; in either notation, T can be equal to S. If $T \neq S$, yet $T \subseteq S$, then T is a *proper subset* of S (denoted by $T \subsetneq S$, if necessary).

Though they have yet to rigorously defined, let $\mathbb{N} = \{1, 2, 3, \ldots\}$ denote the set of natural numbers. The empty set is denoted by $\emptyset$ (this is not a computer 0).

Note: Many authors, especially combinatorists, set theorists, and those trained in the British system, include the number 0 in the natural numbers; here 0 is *not* included, and so where ever confusion can arise, different notation is used. In some schools, the set $\mathbb{W} = \{0, 1, 2, 3, \ldots\}$ is called the set of *whole numbers*, though the expression "non-negative integers" is used here. [I was taught to remember the difference by observing that the whole numbers had an extra "hole".] To avoid confusion, one might also say "positive integers" rather than "natural numbers".

There is, however, good reason to include 0 in the natural numbers (as one might witness with ordinal numbers and the Zermelo hierarchy—set theoretic interpretations of counting numbers). The tradition of natural numbers without 0 is a tradition followed in many North American schools. [I deliberated for some time on this choice of notation, and I am still not sure that I have made the correct choice; from a mathematical perspective, it seems to make more sense to include 0.]

The symbols $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ denote the sets of integers, rationals, reals, and complex numbers, respectively. The notation $\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$ is often used to indicate the set of the positive integers; this notation is somewhat universal, and hence is occasionally used instead of $\mathbb{N}$ to avoid confusion (regarding the inclusion of 0). Throughout, unless otherwise noted, all variables in this text are non-negative integers. For statements p and q, use the shorthand $p \to q$ to abbreviate "if p then q", or "p implies q", and $p \Rightarrow q$ for "p logically implies q". In mathematics, one often confuses their meanings; the "$\to$" is implication in the object language, and "$\Rightarrow$" is in the metalanguage. Many mathematicians use the double arrow to mean simply "implies", perhaps to differentiate from the single arrow used for functions.

The symbol $\forall$ means "for all" and the symbol $\exists$ means "there exists"; as handy as these quantifiers are, their use is limited in this text since they tend to make simple statements unreadable to some non-mathematics students. The symbols "$\wedge$" and "$\vee$" are occasionally used to represent "and" and "or" respectively. If a paragraph is followed by "□", then this indicates the end of a proof. The expression "iff" is an abbreviation for "if and only if".

2.2 Axioms

There are many statements in mathematics that are not proven, but are simply assumed to be true. For example, in Euclidean geometry, it is assumed that *for any pair of distinct points in the plane, there is a unique line that contains them.* Some people find this to be a reasonable assumption, however, might have difficulty proving such an assumption (whatever that might mean).

A statement that is assumed to be true (but not necessarily provable) is called an *axiom* or *postulate.* To state an axiom, one often requires that certain terms are accepted without meaning. For example, undefined terms might include "element", "set", "point", "line", "plane", "is incident with", and "is in between". Having established the terms, one might agree on rules of logic (where the default is usually to simply accept standard Boolean logic, with or without quantifiers, for example, and the standard connectives). With these in place, one can state an axiom, either a property of a term, (*e.g.*, "there exists something called a point") or a relation between terms (*e.g.*, "there exists a set which does not contain any elements").

A *theorem* is a statement that then follows deductively from the axioms, either directly or indirectly using other theorems. A *lemma* is a "lesser" theorem, often used to help prove a more significant theorem. (The plural for "lemma" is "lemmata" or more simply, "lemmas".) A *corollary* is a statement that is a consequence of a theorem; usually a corollary follows from a theorem in a fairly obvious way.

When speaking of the validity of a particular result, one actually only refers to whether or not the result follows from axioms. In Edmund Landau's book *Grundlagen der Analysis* [339], he begins with axioms and derives most of the foundations of arithmetic. The approach here is similar, beginning with the same set of axioms.

Any discussion in set theory, logic, geometry, number theory, or even mechanics, usually presumes that a set of axioms has been agreed upon. How would a scientist decide on a list of absolute truths (axioms) from which to develop a particular system? Of any collection of axioms a scientist might assemble, there are two properties of the collection that may be desirable:

First, insist that the list is as short as possible. Perhaps most importantly, it would not be desirable to want so many axioms that from any (or all) of the axioms, one could derive a *contradiction* (that is, both a statement and its negation). If one can not deductively derive any contradictions from a particular collection of axioms, the collection is called *consistent*, and the system that rests upon these axioms is also called consistent or *sound*.

If a particular system is sound, it might be very difficult to prove such a fact. Even proving the inconsistency of a system by exhibiting a contradiction might be an impossible task.

One reassurance of soundness is to find a *model* or *interpretation* that realizes all of the axioms. In fact, depending on your assumptions about the world, finding a model is sometimes *proof* that a set of axioms is consistent—as it is in many mathematical situations. For example, the naive image of standard Euclidean geometry seems to be a model that satisfies the postulates in Euclid's *Elements of Geometry* (written around 300 B.C.). If a collection of axioms is consistent, any subcollection is also consistent. Different models for geometries have been found that realize all but the fifth of Euclid's postulates, (*e.g*, elliptic or hyperbolic geometries). See [274, pp. 88–93] for a lively, easy to read discussion of the discoveries that led to various "non-Euclidean" geometries.

Any attempt to construct a set of consistent axioms might start by selecting a very large set of axioms, deriving some contradiction, then throwing out one possibly offending axiom, and trying again, continuing until no contradictions are derivable. It is yet another problem, however, to show that contradictions can not arise at any one stage. To support a claim that one particular axiom is consistent with a given set of axioms, one might assume its negation and try to prove a contradiction. Enough about consistency for the moment.

The second desirable property for a collection of axioms to satisfy is that the collection of axioms is large enough so as to be able to derive all truths in the system from the axioms. Such an axiomatic system is called *complete*. In Euclid's thirteen books of *Elements of Geometry* is a set of five postulates, however it seems that Euclid's postulates are not complete (see [264, p. 1636]) for what is now called "Euclidean geometry". Hilbert's set of axioms for geometry [269] arose out of efforts to find "completeness", efforts which were destined for failure as well.

There has been much discussion about what sort of minimal collection of axioms "should be" agreed upon so that one can do, say, set theory, geometry, or arithmetic. In this text, to describe the natural numbers, a set of axioms (now commonly thought to be not too problematic), Peano's axioms, is the starting point. The standard axioms of ZFC are implicitly assumed here. (See appendix IV for a list of ZFC axioms and further discussion about consistency and completeness.)

2.3 Peano's axioms

In the 19th century, Giuseppe Peano (1858–1932), a professor from the University of Turin (Italy), published (in *Formulario Matematico*, 1889), a collection of axioms for the natural numbers $\mathbb{N}$, defined here to be $\mathbb{Z}^+$.

Peano received the axioms from Dedekind in a letter, and he even recognized this in his publication, however, the term "Peano's axioms" has survived to refer to Dedekind's axioms (*e.g.*, see Pollock's book [432], though Pollock does not give references). This fact doesn't seem to be widely cited in other textbooks. Peano's axioms are generally now accepted by the mathematical community as a starting point for arithmetic.

To describe these axioms, common function notation is used: The cartesian product of sets S and T is $S \times T = \{(s,t) : s \in S, t \in T\}$. A *function* f from a domain S to T (written $f : S \longrightarrow T$) is a subset $f \subset S \times T$ so that for every $s \in S$, there is exactly one $t \in T$ so that $(s,t) \in f$. In this case, write $f(s) = t$. (See Section 18.2 for more details on functions.)

Peano's axioms are usually given as a list of five, yet one more appears in his writings, one roughly equivalent to "$\mathbb{N}$ is a class of things called numbers." (See [340, p. 1872] for a translation; many other wonderful articles regarding axioms are also found in the same collection.) His fifth axiom is really the principle that is now known as "mathematical induction".

Peano's axioms:

P1 $1 \in \mathbb{N}$.

P2 There is a function $\delta : \mathbb{N} \to \mathbb{N}$ where for each $x \in \mathbb{N}$, $\delta(x) = x' \in \mathbb{N}$ is called the successor of x.

P3 For every $x \in \mathbb{N}$, $x' \neq 1$.

P4 If $x' = y'$, then $x = y$.

P5 If $S \subset \mathbb{N}$ is such that
(i) $1 \in S$, and
(ii) for every $x \in \mathbb{N}$, $x \in S \to x' \in S$,
then $S = \mathbb{N}$.

A proof employing P5 is said to be "inductive" or "is by induction". The step P5(i) is called the *base step* and P5(ii) is called the *inductive step.* Some philosophers call these two parts the *basic clause* and the *inductive clause* (for example, see [29, p. 468]). The antecedent "$x \in S$" in P5(ii) is called the *inductive hypothesis* (or sometimes *induction hypothesis.*)

2.4 Principle of mathematical induction

This section contains a very brief formulation of what is called the "principle of mathematical induction" as it is applied to various statements, instead of just for sets. Applications and various forms of this principle are discussed again in Chapter 3.

There are many forms of mathematical induction—weak, strong, and backward, to name a few. In what follows, n is a variable denoting an integer (usually non-negative) and $S(n)$ denotes a mathematical statement with one or more occurrences of the variable n. The following is the standard presentation of mathematical induction, also called "weak mathematical induction". Observe that $\delta(x) = x' = x + 1$ is a successor function satisfying P2, P3, and P4 (it is shown in Theorem 2.5.4 that this is the only successor function on natural numbers).

Theorem 2.4.1 (Principle of Mathematical Induction (MI))**.**
If $S(n)$ is a statement involving n and if
(i) $S(1)$ holds, and
(ii) for every $k \geq 1$, $S(k)$ implies $S(k+1)$,
then for every $n \geq 1$, the statement $S(n)$ holds.

The two stages (i) and (ii) in a proof by MI are still called the *base step* (in which the *base case* is proved), and the *inductive step*, respectively. In (ii), $S(k)$ is called the *inductive hypothesis* (also called the *induction hypothesis*). Depending on the definition of the natural numbers used by different authors, the base step might also be $S(0)$.

Proof of MI from Peano's axioms: Define $A = \{n \in \mathbb{N} : S(n) \text{ is true}\}$. Then by (i), $1 \in A$. By (ii), if $k \in A$, then $k+1 \in A$. So by P5, $A = \mathbb{N}$, proving MI. □

2.5 Properties of natural numbers

The next few results (proved from Peano's axioms) will enable one to talk about $\mathbb{N}$ in more familiar terms.

First observe that for any successor function $\delta(x) = x'$, to each x there is a unique x', and hence $[x = y] \to [x' = y']$.

Lemma 2.5.1. *For any $x, y \in \mathbb{N}$, $[x \neq y] \to [x' \neq y']$.*

Proof: This is just the contrapositive of P4. [If a statement is of the form "if P, then Q, the contrapositive of the statement is "if not Q, then not P". The two statements are logically equivalent.] □

Theorem 2.5.2. *If $x \in \mathbb{N}$ then $x' \neq x$.*

Proof: (By induction) Let $A = \{x \in \mathbb{N} : x' \neq x\}$.
BASE STEP: By P3, $1 \in A$.

INDUCTIVE STEP: Assume that $y \in A$, that is, $y' \neq y$. Lemma 2.5.1 then implies $(y')' \neq y'$, and so $y' \in A$.

Hence, by P5, $A = \mathbb{N}$. □

The next result shows that predecessors are unique.

Theorem 2.5.3. *If $x \in \mathbb{N}$ and $x \neq 1$, then there is a unique y so that $x = y'$.*

Proof: (By induction) Let

$$A = \{x \in \mathbb{N} : x = 1 \text{ or there exists } y \in \mathbb{N} \text{ so that } \ x = y'\}.$$

BASE STEP: $1 \in A$ by definition.
INDUCTIVE STEP: Suppose that $x \in A$. Then either $x = 1$ or $x = y'$ for some $y \in \mathbb{N}$. To be shown is that $x' \in A$. If $x = 1$, then $x' \in \mathbb{N}$; if $x = y'$, then $x \in \mathbb{N}$ by P2. Hence, in any case, $x \in \mathbb{N}$, and by the definition of A, $x' \in A$.

Therefore, by P5, $A = \mathbb{N}$. Thus, for any $x \neq 1$, there is some $y \in \mathbb{N}$ so that $x = y'$. The uniqueness of y follows from P4. □

The next theorem shows that the successor function is what one might expect it to be, namely $x' = x + 1$. In this theorem, a function is defined from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$, that is, it takes ordered pairs and returns natural numbers. For such a function f, it is standard to write $f(x, y)$ instead of the more proper $f((x, y))$.

Theorem 2.5.4. *There exists a unique function* $f : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{N}$ *so that for all* $x, y \in \mathbb{N}$,

(a) $f(x, 1) = x'$;

(b) $f(x, y') = (f(x, y))'$.

Proof: There are two things to show, existence and uniqueness.

(Existence) A function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ can be described by an infinite matrix:

$$\begin{bmatrix} f(1,1) & f(1,2) & f(1,3) & f(1,4) & \dots \\ f(2,1) & f(2,2) & f(2,3) & f(2,4) & \dots \\ f(3,1) & f(3,2) & f(3,3) & f(3,4) & \dots \\ f(4,1) & f(4,2) & f(4,3) & f(4,4) & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

The idea in this existence part of the proof is to create this matrix row by row. It will suffice to show that the first row can be constructed so that (a) and (b) hold, and then to show that an arbitrary row can be constructed from a previous one.

Define B to be the set of all $x \in \mathbb{N}$ so that one can find a set of function values $\{f(x, i) : i \in \mathbb{N}\}$ so that for all y, both (a) and (b) hold (for the fixed x). To be shown is that $B = \mathbb{N}$.

Base step ($x = 1$): For every $y \in \mathbb{N}$, define $f(1, y) = y'$. By definition, $f(1, 1) = 1'$, and so (a) holds with $x = 1$. Also, by definition, $f(1, y') = (y')' = (f(1, y))'$, and so (b) holds with $x = 1$.

Inductive step: Suppose that $x \in B$. Then $f(x, y)$ is defined for all $y \in \mathbb{N}$ so that (a) and (b) hold. Define $f(x', y) = (f(x, y))'$. Then, by definition, $f(x', 1) = (f(x, 1))' = (x')'$ and so (a) holds with x' in place of x. Also

$$\begin{aligned} f(x', y') &= (f(x, y'))' && \text{(by definition)} \\ &= ((f(x, y))')' && \text{(by (b) since } x \in B) \\ &= (f(x', y))' && \text{(by definition),} \end{aligned}$$

and so (b) holds for x'. Thus, $x' \in B$, completing the inductive step.

By induction, $B = \mathbb{N}$, finishing the existence part of the proof.

(Uniqueness) Suppose that f is defined so that for all $x, y \in \mathbb{N}$ both (a) and (b) hold and also suppose that g is a function satisfying the corresponding equalities:

(a') $g(x, 1) = x'$, and

(b') $g(x, y') = (g(x, y))'$

Let $x \in \mathbb{N}$ be fixed and define the set $A_x = \{y \in \mathbb{N} : f(x, y) = g(x, y)\}$. Induction is used to first to show that $A_x = \mathbb{N}$.

Base step: $1 \in A_x$ since $f(x, 1) = x' = g(x, 1)$.

Inductive step: Suppose that $y \in A_x$. that is, $f(x, y) = g(x, y)$. Then by (b) and (b'), $f(x, y') = (f(x, y))'$ and $g(x, y') = (g(x, y))'$. Hence, by P4, $f(x, y') = g(x, y')$. So $y' \in A_x$.

Hence by P5, $A_x = \mathbb{N}$. Since x was arbitrary, this completes the uniqueness part of the proof, and hence the entire proof. □

The function f above is better known by its common notation, $f(x, y) = x + y$, and hence the successor function is $x' = x + 1$ (as one might expect). One can now freely use the result of the previous theorem, namely, the existence of the unique function f defined so that

(a) $f(x, 1) = x'$;
(b) $f(x, y') = (f(x, y))'$;
(c) $f(1, y) = y'$;
(d) $f(x', y) = (f(x, y))'$,

where (c) and (d) are from the way f was defined in the existence part of the proof; translating (a)–(d) into common notation using the "+" sign,

(a') $x + 1 = x'$;
(b') $x + y' = (x + y)'$;
(c') $1 + y = y'$;
(d') $x' + y = (x + y)'$.

The expression "$x + y$" is called the sum of x and y, and the process of computing $x + y$ is called addition.

Theorem 2.5.5. *Addition of natural numbers is associative, that is, for every $x, y, z \in \mathbb{N}$,*

$$(x + y) + z = x + (y + z).$$

Proof: Let x and y be fixed natural numbers and put

$$A = \{z \in \mathbb{N} : (x + y) + z = x + (y + z)\}.$$

Base step: $1 \in A$ because

$$\begin{aligned}(x + y) + 1 &= (x + y)' && \text{(by (a'))}\\ &= x + y' && \text{(by (b'))}\\ &= x + (y + 1) && \text{(by (a')).}\end{aligned}$$

Inductive step: Suppose that $z \in A$. Then

$$(x + y) + z' = ((x + y) + z)' \qquad \text{(by (b'))}$$

$$\begin{aligned} &= (x + (y + z))' && \text{(because } z \in A\text{)} \\ &= x + (y + z)' && \text{(by (b'))} \\ &= x + (y + z') && \text{(by (b')),} \end{aligned}$$

and so $z' \in A$, completing the inductive step. Hence, by P5, $A = \mathbb{N}$. □

Since addition is associative, it matters not which adjacent terms are added first, and hence parentheses are not needed.

For natural numbers $x_1, x_2, x_3, \ldots$, define inductively

$$x_1 + x_2 + \ldots + x_n = (x_1 + x_2 + \ldots + x_{n-1}) + x_n.$$

To abbreviate the left side, one uses so-called *sigma notation*:

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i.$$

Such notation extends in the obvious way, for example,

$$\sum_{j=3}^{7} y_j = y_3 + y_4 + y_5 + y_6 + y_7.$$

For later reference, a formal definition of the sigma notation is given:

Definition 2.5.6. Let $x_1, x_2, x_3, x_4, \ldots$ be a sequence of natural numbers. Define $\sum_{i=1}^{1} x_i = x_1$ and recursively define for each $n > 1$,

$$\sum_{i=1}^{n} x_i = \left(\sum_{i=1}^{n-1} x_i \right) + x_n.$$

Generalizing this slightly, for any $j \in \mathbb{N}$, define $\sum_{i=j}^{j} x_i = x_j$ and recursively define for each $n > j$,

$$\sum_{i=j}^{n} x_i = \left(\sum_{i=j}^{n-1} x_i \right) + x_n.$$

Finally, define the sum over an empty set of indices to be zero.

According to [556], the next theorem is due to H. Grassman, from *Lehrbuch der Arithmetik*, 1861 (though I have not seen the original proof).

Theorem 2.5.7. *Addition in natural numbers is commutative, that is, for every $x, y \in \mathbb{N}$,*

$$x + y = y + x.$$

Proof: Let $x \in \mathbb{N}$ be fixed, and put $A = \{y \in \mathbb{N} : x + y = y + x\}$.

Base step: $1 \in A$ because by (a') and (c') respectively,

$$x + 1 = x' = 1 + x.$$

Inductive step: Suppose that $y \in A$. Then

$$\begin{aligned} x + y' &= (x + y)' && \text{(by (b'))} \\ &= (y + x)' && \text{(because } y \in A\text{)} \\ &= y' + x && \text{(by (d')),} \end{aligned}$$

and so $y' \in A$, completing the inductive step.

Hence, by P5, $A = \mathbb{N}$, finishing the proof of the theorem. □

Theorem 2.5.8. *For every $x, y \in \mathbb{N}$, $x + y \neq x$.*

Proof: Let A be the set of all those $x \in \mathbb{N}$ such that for every $y \in \mathbb{N}$, $x + y \neq x$.

Base step: By P3 and property (c'), for every $y \in \mathbb{N}$, $1 + y \neq 1$, and so $1 \in A$.

Inductive step: Assume that $x \in A$, that is, x is such that for any $y \in \mathbb{N}$, $x + y \neq x$. If for some y, $x' + y = x'$ holds, then by property (b'), it follows that $(x + y)' = x'$, and so by P4, $x + y = x$, contradicting that $x \in A$; hence conclude that $x' + y \neq x'$, and thus $x' \in A$.

By P5, $A = \mathbb{N}$. □

The next sequence of exercises establishes the properties for the operation known as "multiplication" of natural numbers; they are proved in a very similar manner to those above. The content of the exercises in this chapter are really theorems whose proofs are perhaps boring or repetitive and are not intended as the first exercises regarding induction that a student might see.

Exercise 1. *Prove that there exists a unique function $g : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{N}$ so that for all $x, y \in \mathbb{N}$*

(e) $g(x, 1) = x$;
(f) $g(x, y') = (x + g(x, y))$.

Replace the notation $g(x, y)$ by $x \cdot y$, the multiplication of x and y, and then abbreviate $x \cdot y$ by xy.

Oddly enough, it helps to first prove distributivity before associativity of multiplication.

Exercise 2. *Prove that the distributive laws hold, that is, prove that for any $x, y, z \in \mathbb{N}$,*

$$x(y+z) = xy + xz,$$

and

$$(x+y)z = xz + yz.$$

Exercise 3. *Prove that the general distributive laws hold for natural numbers, that is, for $x_1, x_2, \ldots, x_n, c \in \mathbb{N}$,*

$$c\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} cx_i.$$

Using one of the basic distributive laws, associativity comes fairly easily.

Exercise 4. *Prove that multiplication of natural numbers is associative, that is, prove that for any $x, y, z \in \mathbb{N}$,*

$$(xy)z = x(yz).$$

Definition 2.5.9. The notation $\prod_{i=1}^{n} x_i$ is defined recursively by $\prod_{i=1}^{1} x_i = x_1$, and for $n \geq 1$,

$$\prod_{i=1}^{n+1} x_i = \left(\prod_{i=1}^{n} x_i\right) x_{n+1}.$$

Since multiplication of natural numbers is associative, if the x_i's are natural numbers, the meaning of

$$\prod_{i=1}^{n} x_i \;=\; x_1 x_2 \cdots x_n$$

is unambiguous. Finally, define the product over an empty set of vertices to be equal to one, that is,

$$\prod_{i \in \emptyset} x_i = 1.$$

Note that when all x_i's are equal, the first simple definition of exponentiation is given (for positive integers): define $x_1 = x$, and for $n > 1$, having defined x^{n-1}, define $x^n = x^{n-1} \cdot x$.

The discussions above are just a beginning to thoroughly define the real numbers, or to check all of the properties of or operations on the natural numbers. A few of these are given as exercises.

2.6 Well-ordered sets

Given a set S, a *binary relation on* S is a subset of the cartesian product $S \times S = \{(a,b) : a \in S, b \in S\}$.

A binary relation R on S is

- *reflexive* iff for every $x \in S$, $(x,x) \in R$.
- *symmetric* iff for every $x, y \in S$, $[(x,y) \in R] \rightarrow [(y,x) \in R]$.
- *antisymmetric* iff for every $x, y \in S$, $[((x,y) \in R) \wedge (x \neq y)] \rightarrow [(y,x) \notin R]$.
- *transitive* iff for every $x, y, z \in S$, $[((x,y) \in R) \wedge ((y,z) \in R)] \rightarrow [(x,z) \in R]$.

A binary relation R on S is a *partial order* if and only if R is reflexive, antisymmetric, and transitive. If R is a partial order on S, the set (S,R) is called a *partially ordered set* , abbreviated, *poset*.

One can write $x \leq_R y$ if $(x,y) \in R$, and if also $x \neq y$, write $x <_R y$ and say that x is less than y. The notation xRy is also quite common. If the relation R is implicitly understood, simply write $x \leq y$ or $x < y$, rather than $x \leq_R y$ or $x <_R y$ (or xRy). (Some texts define a partial order without reflexivity, and so a total order is then always written with "<" rather than "≤"; such notation is often practiced regardless of whether or not reflexivity is insisted on in the definitions, since a relation without reflexive property determines precisely one with reflexivity.)

A *least element* in a partially ordered set $(P, \leq)$ is an element $x \in P$ so that for every $y \in P$, $x \leq y$. For example, the poset $\{(a,c),(b,c)\}$ has no least element (instead it has two minimal elements: a and b) but the poset $\{(x,y),(x,z)\}$ has a least element x. If a least element exists, then it is unique. For a subset $Q \subseteq P$, a *lower bound* for Q is an element $u \in P$ so that for every $q \in Q$, $u \leq q$; if u is a lower bound for Q, write $u \leq Q$. Similarly define *greatest element* and *upper bound*. A *least upper bound* for $Q \subset P$ is a least element in the set of all upper bounds; note that if a least upper bound for Q exists, it is unique. Similarly define *greatest lower bound*. Sometimes the notation $x < Q$ denotes that for every $q \in Q$, $x < Q$.

A partial order R on a set S is called a *total order* (or *linear order*) if for every $x, y \in S$, either $(x,y) \in R$ or $(y,x) \in R$ holds; in this case, the ordered set (S,R) is called a *totally ordered set*.

The standard order on $\mathbb{N}$ is often defined by $x < y$ if and only if there exists $n \in \mathbb{N}$ so that $y = x+n$. (Note that one can not yet really say in this definition "...if and only if there exists $n \geq 1$ so that...", since the order $\geq$ is being defined!) As one might expect, this standard order on $\mathbb{N}$ is indeed a linear order or total order. One first step in proving this is to show that any two elements in $\mathbb{N}$ are *comparable*, that is, for any $x, y \in \mathbb{N}$, one of $x < y$, $x = y$, or $y < x$ holds. The following "Law of Trichotomy" says precisely that.

Exercise 5 (Law of Trichotomy)**.** *For any $x, y \in \mathbb{N}$, exactly one of $x < y$, $x = y$, or $y < x$ holds. Prove this result by induction.*

This law also confirms that since $\leq$ means $<$ or $=$, the relation $\leq$ as defined is antisymmetric. It also follows that $<$ defines a total order. Again, by induction, addition preserves order:

Exercise 6. *For any natural numbers x, y, p,*

$$x < y \quad \textit{if and only if} \quad x + p < y + p.$$

There are different *kinds* of total orderings. For example, $\mathbb{N}$, the integers $\mathbb{Z}$, the rationals $\mathbb{Q}$, and the reals $\mathbb{R}$ all have no largest element. Of these, only $\mathbb{N}$ has a smallest element. Also, both $\mathbb{Q}$ and $\mathbb{R}$ are dense (between any two there is another), yet of these two, only $\mathbb{R}$ contains all its limit points.

Definition 2.6.1. A *well-ordering* on a set W is a total order $\leq$ (or $<$) on W so that for any non-empty $S \subset W$, S contains a least element. Any ordered set $(W, \leq)$ where $\leq$ is a well-ordering is called *well-ordered.*

As well [pun intended] noted in [95], the term "well-ordering" might very well be replaced with "good-ordering", because "well" in this instance is an adjective, not an adverb, however this usage has survived to become standard these days.

Peano's axioms imply that every non-empty subset of natural numbers indeed has a least element:

Theorem 2.6.2. *The standard order on $\mathbb{N}$ is a well-ordering.*

Proof: Let $S \subset \mathbb{N}$. First observe that if any least element in S exists, then it is unique, since if there were two least elements, say m_1 and m_2, then one would have both $m_1 \leq m_2$ and $m_2 \leq m_1$. Consequently, by the Law of Trichotomy, $m_1 = m_2$.

Assume that S is without a least element; to finish the proof, it suffices to show that $S = \emptyset$. Let

$$A = \{m \in \mathbb{N} : \text{ no number less than } m \text{ belongs to } S\}.$$

By P3, $1 \in A$. Suppose that $k \in A$. If $n < k + 1$, then either $n < k$ (in which case $n \notin S$ since $k \in A$) or $n = k$ (in which case $n \notin S$, for if $n \in S$, then n would be least in S). In any case, such an n is not in S. Hence $k + 1 \in A$. Thus by P5, $A = \mathbb{N}$, and so $S = \emptyset$. □

Exercise 7. *Let $(X, \leq)$ be a well-ordered set and let $Y \subseteq X$. Show that if $f : X \longrightarrow Y$ is an isomorphism, then for all $x \in X$, $f(x) \geq x$.*

Theorem 2.6.3. *A linearly ordered set $(W, <)$ is well-ordered if and only if there is no infinite decreasing sequence in W.*

Proof: Suppose that W is well-ordered. If $w_1 > w_2 > \ldots$ is an infinite decreasing sequence in W, put $S = \{w_1, w_2, \ldots\}$. By well-ordering, let w_k be the least element of S; then $w_k > w_{k+1} \in S$, contradicting the minimality of w_k.

Assume that W is a linearly ordered set with no infinite decreasing sequence. Fix any non-empty $S \subseteq W$, and let $t \in S$. If t is not the least element of S, pick $w_1 \in S$, $w_1 < t$. If w_1 is not the least element of S, pick $w_2 \in S$ with $w_2 < w_1$. Continue choosing successively smaller elements. Since W contains no infinite decreasing sequence, the same is true for S, so this process must stop after finitely many steps, and at that time, the least element of S is produced. □

Note: Jech [289, p. 18] states that the direction “no infinite decreasing subset implies well-ordered” in Theorem 2.6.3 follows from the Axiom of Choice (see Section 4.5); however, AC does not seem to be needed in the above proof.

Definition 2.6.4. For totally ordered sets $(W_1, \leq_1)$ and $(W_2, \leq_2)$ a function $f : W_1 \to W_2$ is *order preserving* (o.p.) iff $x \leq_1 y$ implies $f(x) \leq_2 f(y)$. Well-ordered sets A and B are *similar*, written $A \sim B$, iff there is an order preserving bijection between them.

An order preserving bijection is also called a *similarity*. [Caution: “similarity” is also a term used in geometry for functions that preserve ratios of distances.] Some authors use the term “isomorphism” to describe a similarity; “isomorphism” is often used for a bijection that preserves algebraic or relational structure; in this case the structure is only the order.

Theorem 2.6.5. *2.6.3 Let $(W, <)$ be a well-ordered set and let $Y \subseteq W$. If $f : W \to Y$ is an order preserving bijection, then for all $w \in W$, $f(w) \geq w$.*

There are two ways to present this proof, one by induction, and the other by contradiction; the difference is subtle.

First proof of Theorem 2.6.3: Let $M = \{w \in W : f(w) < w\}$. If M is non-empty, pick some $m \in M$; then $f(m) < m$ and f being order preserving imply $f(f(m)) < f(m)$, so $f(m) \in M$ as well. Continue applying f, by induction, giving an infinite decreasing sequence $m, f(m), f(f(m)), \ldots$ in M. But since M is a subset of the well-ordered set W, M has a least element, so the assumption that M is non-empty must be abandoned. Thus conclude that $M = \emptyset$. □

Second proof of Theorem 2.6.3: With the same notation as above, if $M \neq \emptyset$, since M is a subset of a well-ordered set, M contains a least element y_0. Then $f(y_0) < y_0$ implies that $f(f(y_0)) < f(y_0)$ and hence $f(y_0) \in M$, contradicting that y_0 is least in M. So $M = \emptyset$. □

The first proof above can be thought of as a proof that uses *downward* induction to produce an infinite decreasing sequence, a sequence which contradicts a previously established fact. This very same technique is used in Fermat’s method of infinite descent (see Section 3.6). Thus, many proofs by contradiction might be considered as proofs by induction.

2.7 Well-founded sets

A partial order $(P, \leq)$ is called *well-founded* if for every non-empty subset $X \subseteq P$, X contains a minimal (with respect to $\leq$). In a well-founded partial order, for every element $x \in P$, there is a well-ordered set containing both x and a minimal element of P. Just as mathematical induction is used on well-ordered sets, so too is mathematical induction valid for well-founded sets. This kind of induction might be called *generalized induction.* For example, suppose that one wants to prove a sequence of statements $P(m, n)$ that depend on two variables, say, for finitely many m. Suppose also that one knows $P(m, n) \rightarrow P(m, n+1)$. An inductive proof could start with the base cases as $P(m_i, 0)$, and from each base case, ordinary induction can be applied to reach all statements of the form $P(m_i, n)$. Generalized induction is most often used for induction on two variables, called “double induction”, discussed in the next chapter. Generalized induction also includes the notion of “alternative induction”, also in the next chapter.

This page intentionally left blank

Chapter 3

Variants of finite mathematical induction

> *Mathematics is either Pure or Mixed..... And as for Mixed Mathematics, I may only make this prediction, that there cannot fail to be more kinds of them, as nature grows further disclosed.*
>
> —Francis Bacon,
>
> *Advancement of Learning.*

There are many forms of mathematical induction—weak, strong, and backward, to name a few. In what follows, n is a variable denoting an integer (usually non-negative) and $S(n)$ denotes a mathematical statement with one or more occurrences of the variable n.

3.1 The first principle

For convenience, the standard presentation of mathematical induction is repeated here. Sometimes this standard version of induction is called the "first principle of mathematical induction", and is also called "weak mathematical induction" (as opposed to "strong" induction, a modification appearing in Section 3.2). Recall that the notation $P \to Q$ is short for "P implies Q".

Theorem 2.4.1 [Principle of Mathematical Induction (MI)]
Let $S(n)$ be a statement involving n. If
(i) $S(1)$ holds, and
(ii) for every $k \geq 1$, $S(k) \to S(k+1)$,
then for every $n \geq 1$, the statement $S(n)$ holds.

By Theorem 4.5.5, an inductive step can also be accomplished indirectly by showing that the set of integers for which $S(n)$ fails has no least element, contradicting the well-ordering of $\mathbb{N}$. Such an example occurs in the following Section 3.2 on strong mathematical induction. (See also one solution to Exercise 477.)

Note that the base step in an inductive proof is essential, since, for example, if one were to attempt to prove that for any positive integer n, the statement $S(n)$: $\sum_{i=1}^{n}(2i-1) = n^2 + 5$ holds, it is not hard to show that $S(n) \to S(n+1)$, however, $S(1)$ does not even hold, and so one may not conclude that $S(n)$ holds for all $n \geq 1$. Another such statement (where n is a positive integer) is "$n^2 + 5n + 1$ is even", for which the inductive step works, but the statement is in fact never true!

The base case for MI need not be 1 (or 0); in fact, one may start at any integer. Here is a slightly more general (but equivalent) form of the principle of induction:

Theorem 3.1.1 (Principle of Mathematical Induction (MI)).
Let $S(n)$ denote a statement regarding an integer n, and let $k \in \mathbb{Z}$ be fixed. If
(i) $S(k)$ holds, and
(ii) for every $m \geq k$, $S(m) \to S(m+1)$,
then for every $n \geq k$, the statement $S(n)$ holds.

Proof: Let $T(n)$ be the statement $S(n + k - 1)$, and repeat the above proof, instead with T replacing every occurrence of S. Then the base case becomes $T(1) = S(1 + k - 1) = S(k)$ as desired. □

3.2 Strong mathematical induction

While attempting an inductive proof, in the inductive step one often needs only the truth of $S(n)$ to prove $S(n+1)$; sometimes a little more "power" is needed, and often this is made possible by strengthening the inductive hypothesis. The following version of mathematical induction can be viewed as contained in the principle of transfinite induction (see Section 4.2).

Theorem 3.2.1 (Strong Mathematical Induction).
Let $S(n)$ denote a statement involving an integer n. If
(i) $S(k)$ is true and
(ii) for every $m \geq k$, $[S(k) \wedge S(k+1) \wedge \cdots \wedge S(m)] \to S(m+1)$
then for every $n \geq k$, the statement $S(n)$ is true.

The principle of strong induction is also referred to by some as *course-of-values induction* (*e.g.*, see [42]). A few professionals use "full induction" or "complete induction" to denote strong induction; these terms have long been accepted as meaning simply "mathematical induction" (as opposed to empirical induction). [See Section 1.8 on history of induction.]

In Theorem 2.6.2, it was shown that Peano's axioms imply the well-ordering of $\mathbb{N}$. This well-ordering is used (below) to prove strong induction, and hence, to show that strong induction also follows from P5. Notice that Theorem 4.5.5 also shows that both forms of induction follow from well-ordering.

Proof of strong induction principle from weak: Assume that for some k, the statement $S(k)$ is true and for every $m \geq k$, $[S(k) \wedge S(k+1) \wedge \cdots \wedge S(m)] \rightarrow S(m+1)$. Let B be the set of all $n > m$ for which $S(n)$ is false. If $B \neq \emptyset$, $B \subset \mathbb{N}$ and so by well-ordering, B has a least element, say ℓ. By the definition of B, for every $k \leq t < \ell$, $S(t)$ is true. The premise of the inductive hypothesis is true, and so $S(\ell)$ is true, contradicting that $\ell \in B$. Hence $B = \emptyset$. □

Strong induction also implies weak induction.

Proof of weak induction from strong: Assume that strong induction holds (in particular, for $k = 1$). That is, assume that if $S(1)$ is true and for every $m \geq 1$, $[S(1) \wedge S(2) \wedge \cdots \wedge S(m)] \rightarrow S(m+1)$, then for every $n \geq 1$, $S(n)$ is true.

Observe (by truth tables, if you will), that for $m + 1$ statements p_i,

$$[p_1 \rightarrow p_2] \wedge [p_2 \rightarrow p_3] \wedge \ldots \wedge [p_m \rightarrow p_{m+1}] \Rightarrow [(p_1 \wedge p_2 \wedge \ldots \wedge p_m) \rightarrow p_{m+1}],$$

itself a result provable by induction (see Exercise 456).

Assume that the hypotheses of weak induction are true, that is, that $S(1)$ is true, and that for arbitrary t, $S(t) \rightarrow S(t+1)$. By repeated application of these recent assumptions, $S(1) \rightarrow S(2)$, $S(2) \rightarrow S(3)$, ..., $S(m) \rightarrow S(m+1)$ each hold. By the above observation, then

$$[S(1) \wedge S(2) \wedge \cdots \wedge S(m)] \rightarrow S(m+1).$$

Thus the hypotheses of strong induction are complete, and so one concludes that for every $n \geq 1$, the statement $S(n)$ is true, the consequence desired to complete the proof of weak induction. □

Hence it has been demonstrated that weak and strong forms of mathematical induction are equivalent. For remarks on this relationship, see [477].

Here is an example where strong induction is used. Recall that a prime number (or simply, a prime) is one whose only divisors are itself and 1 (and convention says that 1 is not a prime); the first few primes are

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, \ldots.$$

Theorem 3.2.2. *Any positive integer $n \geq 2$ is a product of primes.*

Proof: Let $S(n)$ be the statement "n is a product of primes."

BASE STEP ($n = 2$): Since $n = 2$ is trivially a product of primes (well, actually only one prime), $S(2)$ is true.

Inductive step: Fix some $m \geq 2$, and assume that for every t satisfying $2 \leq t \leq m$, the statement $S(t)$ is true. To be shown is that

$$S(m+1): \quad m+1 \text{ is a product of primes,}$$

is true. If $m+1$ is prime, then $S(m+1)$ is true. If $m+1$ is not prime, then there exists r and s with $2 \leq r \leq m$ and $2 \leq s \leq m$ so that $m+1 = rs$. Since $S(r)$ is assumed to be true, r is a product of primes; similarly, by $S(s)$, s is a product of primes. Hence $m+1 = rs$ is a product of primes, and so $S(m+1)$ holds. So, in either case, $S(m+1)$ holds, completing the inductive step.

Thus, by mathematical induction, for all $n \geq 2$, the statement $S(n)$ is true. □

The so-called "Fundamental theorem of arithmetic" says that any integer $n \geq 2$ is a product of primes in exactly one way, that is, the prime factorization is unique—another result provable by induction (see Exercise 206).

3.3 Downward induction

Suppose that you are trying to prove a statement $S(n)$ and a forward inductive argument is difficult for every n. Here is another strategy: first prove the statement for infinitely many n (for example, when n is a power of 2—either directly or by an inductive step of the form $S(k) \to S(2k)$, say) and then prove $S(n)$ for the gaps between. The proof for the gaps can either be by forward induction, or backward induction. For example, in the case where one has the truth of $S(n)$ for all powers of 2, one can then fill in the gaps with an inductive argument for each fixed k of the form $S(2^k + t) \to S(2^k + t - 1)$ for each t satisfying $1 \leq t \leq 2^k$.

Downward (also called "backward") inductive arguments have been around a long time; many authors, including Cauchy (1759–1857) and Weierstrass (1815–1897) (see [259, p.19]) have used them. The term "backward induction" can also be used in game theory where players reason "working backward from the last possible move in a game to anticipate each other's rational choices."[114]. What has recently become known as "downward induction" defined below might be more appropriately called "upward-downward" induction.

Downward induction: Let $S(n)$ be a statement involving n. If
(i) $S(n)$ is true for infinitely many n, and
(ii) for each $m \geq 2$, $S(m) \to S(m-1)$
then for every $n \geq 1$, the statement $S(n)$ is true.

Proof of downward induction from MI: Assume the hypotheses (i) and (ii) hold and let $n_1, n_2, n_3, \ldots$ be an infinite sequence so that for each $i \in \mathbb{Z}^+$, $S(n_i)$ holds. Fix some $k \in \mathbb{Z}^+$, and prove $S(k)$ holds as follows: Fix i so such that $n_{i-1} < k \leq n_i$.

For $j = 0, 1, \ldots, n_i - n_{i-1}$, define the statement $T(j) = S(n_i - j)$. It suffices to prove that $T(n_i - k) = S(k)$; this is done by induction on j.

Base step ($j = 0$): $T(j) = T(0) = S(n_i - 0) = S(n_i)$, which was assumed to be true by (i).

Inductive step: Suppose that for some $j \geq 0$, $T(j) = S(n_i - j)$ holds. By (ii), $T(j+1) = S(n_i - j - 1)$ holds, completing the inductive step $T(j) \rightarrow T(j+1)$.

Therefore, by MI, $T(j)$ holds for all $j \geq 0$, in particular, $T(n_i - k) = S(k)$ holds, finishing the proof of downward induction. □

There are different proofs of the so-called "theorem of arithmetic and geometric means"; for example, there is one downward induction proof appearing in [259] and another simpler one also suggested there. The simpler one is presented here. Another proof follows from Jensen's inequality on convex functions—see Exercise 602 or 603—both provable by downward induction. After giving the proof by downward induction, one more simple, but tricky proof by ordinary induction due to Kong-Ming Chong [102] is presented. For other proofs prior to 1976 (mentioned in [102]) of the AM-GM inequality, see, *e.g.*, [5, pp. 200–224], [43, §5 pp. 4–5; §11 pp. 9–10], [104, p.46], [259], [135], and [397].

Theorem 3.3.1 (AM-GM inequality)**.** *Let $a_1, \ldots, a_n$ be non-negative real numbers. Then*

$$(a_1 a_2 \cdots a_n)^{1/n} \leq \frac{a_1 + a_2 + \cdots + a_n}{n},$$

with equality holding if and only if all a_i's are equal.

Proof: Let $S(n)$ be the statement that for any $a_1, \ldots, a_n$,

$$(a_1 a_2 \cdots a_n)^{1/n} \leq \frac{a_1 + a_2 + \cdots + a_n}{n},$$

with equality holding if and only if all a_i's are equal. The first part of the proof is to show that $S(n)$ holds whenever n is a power of 2. This requires a form of strong induction, one with two base cases.

Base step $n = 1$: The statement $S(1)$ reduces to $a_1 = a_1$, which is true.

Base step $n = 2$: To show $S(2)$, let $a_1 = a$ and $a_2 = b$; then

$$\begin{aligned} ab &= \left(\frac{a+b}{2}\right)^2 - \left(\frac{a-b}{2}\right)^2 \\ &\leq \left(\frac{a+b}{2}\right)^2, \end{aligned}$$

with equality holding if and only if $a = b$.

Upward inductive step ($S(k) \to S(2k)$): For some $k \geq 2$ assume that $S(k)$ holds, that is, assume that for non-negative $c_1, c_2, \ldots, c_k$,

$$c_1 c_2 \cdots c_k \leq \left(\frac{c_1 + c_2 + \cdots + c_k}{k}\right)^k,$$

with equality if and only if the c_i's are all equal. To show that $S(2k)$ follows:

$$\begin{aligned}
&a_1 a_2 \cdots a_k b_1 b_2 \cdots b_k \\
\leq\ & \left(\frac{a_1 + \cdots + a_k}{k}\right)^k \left(\frac{b_1 + \cdots + b_k}{k}\right)^k \quad \text{(by } S(k) \text{ twice)} \\
=\ & \left(\frac{(a_1 + \cdots + a_k)(b_1 + \cdots + b_k)}{k^2}\right)^k \\
\leq\ & \left(\left(\frac{a_1 + \cdots + a_k + b_1 + \cdots b_k}{2}\right)^2 \frac{1}{k^2}\right)^k \quad \text{by } S(2), \\
=\ & \left(\frac{a_1 + \cdots + a_k + b_1 + \cdots b_k}{2k}\right)^{2k}
\end{aligned}$$

and inequalities are strict unless all a_i's and b_j's are equal. Hence $S(k) \to S(2k)$, completing this inductive step.

By induction, for all n that are powers of 2, the statement $S(n)$ holds.

Downward inductive step ($S(m) \to S(m-1)$): For some $n \geq 2$, assume that $S(m)$ holds, and let $a_1, a_2, \ldots, a_{m-1}$ be non-negative, not all equal, and put

$$A = \frac{a_1 + a_2 + \ldots + a_{m-1}}{m-1}.$$

Then

$$\begin{aligned}
a_1 a_2 \cdots a_{m-1} A &< \left(\frac{a_1 + a_2 + \cdots + a_{m-1} + A}{m}\right)^m \qquad \text{(by } S(m)\text{)}, \\
&= \left(\frac{(n-1)A + A}{m}\right)^m \\
&= A^m,
\end{aligned}$$

and hence $a_1 a_2 \cdots a_{m-1} < A^{m-1}$, thus showing $S(m-1)$. This completes the proof of the downward induction step, and hence the proof. □

Note: Theorem 3.3.1 has a more direct proof, based on Exercise 199, the solution of which is a fairly easy inductive proof; see comments after the solution to Exercise 199.

As mentioned above, here is an outline of Chong's simple (but tricky) inductive proof of the AM-GM inequality. Suppose that the base case $n = 2$ is done, and for some $k \geq 3$, suppose that $S(k-1)$ is true, in particular, suppose that for any choice of $b_1, a_2, \ldots, a_{k-1}$ not all equal,

$$\frac{b_1 + a_2 + a_3 + \cdots + a_{k-1}}{k-1} > (b_1 a_2 a_3 \cdots a_{k-1})^{\frac{1}{k-1}}.$$

To be shown is that $S(k)$ holds (in the case when all numbers are not equal). Let $a_1 \leq a_2 \leq \cdots \leq a_n$ be not all equal, that is, $a_1 < a_n$, and let $A = (a_1 + a_2 + \cdots + a_n)/n$ be their arithmetic mean. Then $a_1 < A < a_n$, which implies that

$$A(a_1 + a_k - A) - a_1 a_k = (a_1 - A)(A - a_k) > 0,$$

and so

$$a_1 + a_k - A > \frac{a_1 a_k}{A}. \tag{3.1}$$

Let $b_1 = a_1 + a_k - A$; then

$$\frac{b_1 + a_2 + \cdots a_{k-1}}{k-1} = \frac{(\sum a_i) - A}{k-1} = \frac{kA - A}{k-1} = A.$$

Thus, by induction hypothesis,

$$\begin{aligned} A &> (b_1 a_2 a_3 \cdots a_{k-1})^{\frac{1}{k-1}} \\ &> (\frac{a_1 a_k}{A} a_2 \cdot a_k)^{\frac{1}{k-1}} \qquad \text{by eqn (3.1)} \end{aligned}$$

which yields

$$\begin{aligned} A^{k-1} &> \frac{a_1 a_2 \cdots a_k}{A}, \\ A^k &> a_1 a_2 \cdots a_k, \end{aligned}$$

showing that $S(k)$ is true, completing the (upward) inductive step, and hence Chong's proof. □

There are other inductive proofs of the AM-GM inequality; one inductive step begins by assuming that for any a_i's satisfying $a_1 a_2 \cdots a_n = 1$ then $a_1 + \cdots + a_n \geq n$. Then assume that $b_1 b_2 \cdots b_n b_{n+1} = 1$; without loss of generality, let $b_n < 1$ and $b_{n+1} > 1$. Then $b_1 + \cdots + b_{n+1} \geq n+1$ by setting $a_1 = b_1$, $a_2 = b_2$, $\ldots$, $a_{n-1} = b_{n-1}$ but $a_n = b_n b_{n+1}$. Then by inductive hypothesis, $b_1 + b_2 + \cdots + b_{n-1} + b_n b_{n+1} \geq n$. To finish the inductive step, it suffices to show

$$b_n + b_{n+1} \geq b_n b_{n+1} + 1,$$

or

$$(1 - b_n)(1 - b_{n+1}) \leq 0,$$

which is true by the initial assumption on b_n and b_{n+1}, finishing the inductive step. □

See [146, pp. 37–40] for yet another solution by induction based on the following lemma:

Lemma 3.3.2. *For real numbers w, x, y, and z, if $w+x=y+z$, then the largest of the two products wx and yz is formed by the pair with the smallest difference.*

Proof: Let $w+x=y+z$. Note the following two identities:

$$(w+x)^2-(w-x)^2=4wx,$$
$$(y+z)^2-(y-z)^2=4yz.$$

Since $w+x=y+z$, also $(w+x)^2=(y+z)^2$, so the left-hand side of the two identities is made largest when the second term is smallest. □

It might be interesting to note that the case $n=2$ in Theorem 3.3.1 can be used to prove that no chord of a circle is longer than the diameter. Let three points A, B, C form a straight line segment with distances $|AB|=a$ and $|BC|=b$ units (see Figure 3.1.

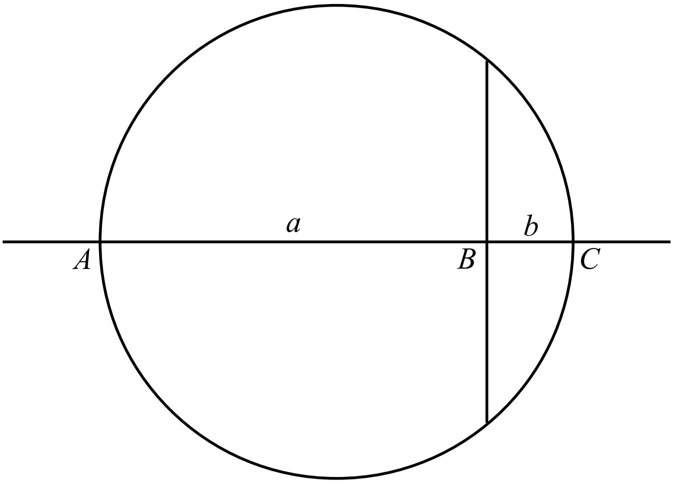


Figure 3.1: Chords are shorter than diameters

Using the line segment AC as a diameter, form the circle whose diameter is $|AC|=a+b$ units. Form a chord of that circle perpendicular to AC through B. Then with a simple application of Pythagoras' theorem, one finds that the length of that chord is $2\sqrt{ab}$. By Theorem 3.3.1, $\sqrt{ab}\leq\frac{a+b}{2}$, and so the length of the chord is not longer than the diameter of the circle.

3.4 Alternative forms of mathematical induction

There are many ways to apply inductive reasoning. For example, if $S(0)$ and $S(1)$ are true, and if for any $n\geq 0$, $S(n)\rightarrow S(n+2)$ holds, then for all $n\geq 0$, $S(n)$ is true, since actually, two separate inductive proofs are combined in one (one for the even cases, and one for the odd cases). Here is an example of a situation where three inductive proofs are rolled into one. [Note: This very example is listed again as Exercise 311.] Other applications of this alternative form of mathematical induction appear throughout the exercises, *e.g.*, in Exercises 113 and 275.)

Theorem 3.4.1. *For any integer $n \geq 14$, n is expressible as a sum of 3's and/or 8's.*

Proof: Let $S(n)$ be the statement: n is expressible as a sum of 3's and/or 8's.

BASE CASES ($S(14), S(15), S(16)$): Since $14 = 3 + 3 + 8$, $15 = 3 + 3 + 3 + 3 + 3$, and $16 = 8 + 8$, the base steps are shown.

INDUCTIVE STEP ($S(k) \to S(k+3)$): Assume that for some $k \geq 14$, $S(k)$ holds, that this, there exist $\alpha, \beta \in \mathbb{Z}$ so that $k = \alpha \cdot 3 + \beta \cdot 8$. Then $k + 3 = \alpha \cdot 3 + \beta \cdot 8 + 3 = (\alpha + 1) \cdot 3 + \beta \cdot 8$, that is, $k + 3$ is expressible as a sum of 3's and/or 8's, showing $S(k + 3)$ holds, completing the inductive step.

By MI, for all $n \geq 14$, the statement $S(n)$ is true. (Actually, there are three separate proofs by MI rolled into one, one proving the statement for the sequence $n = 14, 17, 20, \ldots$, one for $n = 15, 18, 21, \ldots$, and another for $n = 16, 19, 22, \ldots$.) □

An inductive proof might be also encountered when both $S(2)$ and $S(3)$ hold and for $k \geq 2$, $[S(k) \wedge S(k + 1)] \to S(k + 2)$, then $S(n)$ holds for all $n \geq 2$; such a proof might be classified somewhere between weak and strong induction. In Section 12.2 on Fibonacci numbers, there are many exercises where such a technique is required.

Many mathematical induction proofs use more than one base case, and such proofs can fall into the category of "generalized induction" on well-founded sets. The technique relies on the fact that in any well-founded set (a partial order with minimal elements) and a statement S about elements of that set, for each x in the set, there is an inductive argument for $S(x)$ that has as its base case one of the minimal elements. For example, consider the set $X = \{2, 3, 4, 5, 6, \ldots\}$. If X is ordered according to divisibility, the proof of any statement S about members of X might start with the base cases being a proof about each prime, the primes being the set of minimal elements in the partially ordered set X.

3.5 Double induction

A special kind of inductive argument is called "double induction"; some texts refer to double induction as an inductive step that requires, say, $S(n)$ and $S(n+1)$ to prove $S(n + 2)$. Another kind of "double induction" is where two statements involving n are proved simultaneously (see, for example, Exercise 320 or Exercise 122 where the inductive step consists of two proofs, one for each of two statements). In this section, however, "double induction", means an induction on two variables simultaneously.

Many mathematical statements involve two (or more) variables, each of which may vary independently over, say, $\mathbb{N}$. Some such statements can be proved by induction in different ways. Let $S(m,n)$ be a statement involving two positive integer variables m and n. One method to prove $S(m,n)$ for all $m \geq 1$ and $n \geq 1$ is to first prove $S(1,1)$, then by induction prove $S(m,1)$ for each m, and then for each fixed m_0, inductively prove $S(m_0,n)$ for all n. Here is a rather simple example of the technique.

Theorem 3.5.1. *Let positive integers m and n be given.*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(i+j) = \frac{mn(m+n+2)}{2}.$$

Proof: Let $S(m,n)$ be the equality in the statement of the theorem.

First it is proved that for all $m \geq 1$, $S(m,1)$ is true.
BASE STEP: The statement $S(1,1)$ is true since $1+1 = 1\cdot 1(1+1+2)/2$.

INDUCTIVE STEP (inducting on m): For some $k \geq 1$, assume that $S(k,1)$ is true, that is, $\sum_{i=1}^{k}(i+1) = k(k+3)/2$. Beginning with the left-hand side of $S(k+1,1)$,

$$\begin{aligned}\sum_{i=1}^{k+1}(i+1) &= \left(\sum_{i=1}^{k}(i+1)\right) + (k+1)+1\\ &= \frac{k(k+3)}{2} + k + 2 \quad (\text{by } S(k,1))\\ &= \frac{k^2+5k+4}{2}\\ &= \frac{(k+1)(k+1+1+2)}{2},\end{aligned}$$

the right-hand side of $S(k+1,1)$. Hence $S(k,1) \rightarrow S(k+1,1)$, and so by mathematical induction, for all $m \geq 1$, $S(m,1)$ is true.

Fix an arbitrary m_0. Then $S(m_0,1)$ is true and so this is a base step for proving that for all n, $S(m_0,n)$ holds.

INDUCTIVE STEP (inducting on n): This step is of the form $S(m_0,\ell) \rightarrow S(m_0,\ell+1)$. Let $\ell \geq 1$ be fixed and assume that $S(m_0,\ell)$ is true, that is,

$$\sum_{i=1}^{m_0}\sum_{j=1}^{\ell}(i+j) = \frac{m_0\ell(m_0+\ell+2)}{2}.$$

Beginning with the left side of $S(m_0,\ell+1)$,

$$\sum_{i=1}^{m_0}\sum_{j=1}^{\ell+1}(i+j) = \sum_{i=1}^{m_0}\left(\left(\sum_{j=1}^{\ell}(i+j)\right) + (i+\ell+1)\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m_0}\left(\sum_{j=1}^{\ell}(i+j)\right) + \sum_{i=1}^{m_0}(i+\ell+1) \\
&= \frac{m_0\ell(m_0+\ell+2)}{2} + \sum_{i=1}^{m_0}(i+\ell+1) \qquad \text{(by } S(m_0,\ell)\text{)}, \\
&= \frac{m_0\ell(m_0+\ell+2)}{2} + \frac{m_0(m_0+1)}{2} + m_0(\ell+1) \\
&= \frac{m_0\ell(m_0+\ell+2)}{2} + \frac{m_0(m_0+1+2(\ell+1))}{2} \\
&= \frac{m_0(\ell m_0+\ell^2+2\ell+m_0+1+2(\ell+1))}{2} \\
&= \frac{m_0(\ell+1)(m_0+\ell+1+2)}{2},
\end{aligned}
$$

which is the right-hand side of $S(m_0, \ell+1)$. Hence, by induction, for each fixed m_0 and all $n \geq 1$, $S(m_0, n)$ is true, completing this inductive step.

Since m_0 was arbitrary, by induction, for all $m \geq 1$ and $n \geq 1$ the statement $S(m,n)$ is proved. □

Sometimes the inductive proofs contained in each stage of the double induction require multiple base cases and alternative forms of induction (or strong induction). See Exercise 304 for such a situation, where the alternative form of induction in the second stage requires two proofs by induction in the first stage.

Another way to apply a double induction argument would be to use $S(1,1)$ as a base step, then show that both $S(m,n) \to S(m+1,n)$, and $S(m,n) \to S(m,n+1)$. This would prove $S(m,n)$ for all m and n. One must be careful, however, for only the step $S(m,n) \to S(m+1,n+1)$ would not prove the statement for all m and n, only the cases where $m = n$.

A slightly trickier double induction occurs in Exercise 380, where the induction step shows $S(n-2,k-1) \wedge S(n-1,k) \to S(n,k)$. In this case, one needs to prove two families of base cases, those of the forms $S(0,k)$, $S(1,k)$ and those of the form $S(n,0)$, $S(n,1)$. Then, for example, to prove $S(6,3)$, one proceeds as follows:

$$
\begin{aligned}
S(0,1) \wedge S(1,2) &\to S(2,2) \\
S(1,1) \wedge S(2,2) &\to S(3,2) \\
S(2,1) \wedge S(3,2) &\to S(4,2) \\
S(3,1) \wedge S(4,2) &\to S(5,2) \\
\\
S(0,2) \wedge S(1,3) &\to S(2,3)
\end{aligned}
$$

$$\begin{aligned}
S(1,2) \wedge S(2,3) &\quad\rightarrow\quad S(3,3)\\
S(2,2) \wedge S(3,3) &\quad\rightarrow\quad S(4,3)\\
S(3,2) \wedge S(4,3) &\quad\rightarrow\quad S(5,3)\\
S(4,2) \wedge S(5,3) &\quad\rightarrow\quad S(6,3).
\end{aligned}$$

So in this situation, one fixes $k-1$ and k, inducts on n, then one repeats the process for k and $k+1$, and so on. Given that the base cases and the inductive step are proved, the interested reader can try to write up the proof formally.

3.6 Fermat's method of infinite descent

One of the more famous applications of Fermat's method of infinite descent showed that any right angle triangle with sides having rational lengths could not have integral area. His technique was to first show that the truth of the theorem follows from the special case for right triangles with integer lengths, that is, for *Pythagorean triangles.* Then he showed that if one could find such a Pythagorean triangle with integer area, one could then (using the number-theoretic properties of the lengths of the sides from the first one) produce a smaller Pythagorean triangle with the same property. From the smaller one, applying precisely the same argument, one would find yet a smaller one. By induction, one gets an infinite sequence of consecutively smaller triangles with the desired property. Clearly there is no infinite descending sequence of Pythagorean triangles (by the well-ordering of natural numbers)—a contradiction. So one must abandon the assumption that one found a Pythagorean triangle with integer area.

The above discussion hints at the possibility that there are proofs that are inductive, but not in any straightforward way. To demonstrate the beauty of Fermat's technique, the next theorem is given with a proof by infinite descent; in many respects, it duplicates the proof alluded to above for Pythagorean triangles.

Theorem 3.6.1. *The equation*

$$x^4 + y^4 = z^2 \tag{3.2}$$

has no solution in non-zero integers x, y, and z.

Proof by infinite descent: If any triple of integers (x, y, z) satisfy (3.2), then so do any of $(\pm x, \pm y, \pm z)$; thus to show the theorem, it suffices to show that are no *positive* integer solutions to (3.2).

The proof is accomplished by showing that *if* some solution x, y, z to (3.2) exists, then from that solution one can create another "smaller" solution x', y', z', where "smaller" means that $z' < z$. Since the positive integers are well-ordered, this process can not continue forever, and so one must abandon the original assumption some solution exists.

Hypothetically, suppose that x, y, z is a solution to (3.2).

Considering the equation (3.2) modulo 4, one observes that x and y can not both be odd (any odd number squared is congruent to 1 modulo 4, and since then z has to be even, z^2 is 0 modulo 4).

Let x and y be even, say $x = 2k$, $y = 2\ell$; then 16 divides z^2, and so 4 divides z, say $z = 4m$. Then $(2k)^4 + (2\ell)^4 = (4m)^2$, and division by 16 yields $k^4 + \ell^4 = m^2$, a smaller solution. Similarly, if any prime p divides both x and y, write $x = pk$, $y = p\ell$, and $z = p^2m$. Division by p^4 shows that $x' = k$, $y' = \ell$, and $z' = m$ is another smaller solution to (3.2). Hence, it suffices to assume that x and y are relatively prime.

So suppose that exactly one of x and y is even, the other odd (and x and y are relatively prime). Without loss of generality, suppose that x is even and y is odd. Then x^2 and y^2 are relatively prime, and so the triple x^2, y^2, z form a *fundamental Pythagorean triple* (a triple of positive integers a, b, c, each pair relatively prime, satisfying $a^2 + b^2 = c^2$). It is well known (*e.g.*, see [150]) that a fundamental Pythagorean triple is a triple of the form $2mn$, $m^2 - n^2$, and $m^2 + n^2$, where m and n are relatively prime positive integers with exactly one of m, n odd.

Fix such an m and n, and write $x^2 = 2mn$, $y^2 = m^2 - n^2$, and $z = m^2 + n^2$. Since $y^2 + m^2 = n^2$, the triple y, m, n is a Pythagorean triple. Since m and n are relatively prime, so are y and m, with y odd and n even, and so y, m, n is a fundamental triple. Hence, there are relatively prime p and q, so that $m = 2pq$ and $n = p^2 + q^2$.

Since $x^2 = 2mn = 4pq(p^2 + q^2)$ and p and q are relatively prime, each of p, q, and $p^2 + q^2$ are all relatively prime and hence each must be a perfect square, say $p = \alpha^2$, $q = \beta^2$, and $p^2 + q^2 = \gamma^2$. Then $\alpha^4 + \beta^4 = \gamma^2$ with $\gamma \leq m < z$, giving a smaller solution to (3.2). □

Since z^4 is a perfect square, Theorem 3.6.1 implies that $x^4 + y^4 = z^4$ has no non-zero integer solutions, a special case of what is now called "Fermat's last theorem": for each integer $n \geq 3$, the equation $x^n + y^n = z^n$ has no non-zero solutions. (This was a conjecture until Andrew Wiles *et al.* finally proved it in 1995 (see [569]). In 1753, Euler gave an incorrect proof for $n = 3$, later corrected by Gauss; both ideas were using descent, however Gauss failed to notice that unique factorization did not hold in his "proof". The case $n = 5$ was solved with infinite descent by Dirichlet (1805–1859) and Legendre (1752–1833) in 1825 (both proofs were based on a result by Sophie Germain); Dirichlet also managed $n = 14$ in 1832. Lamé settled the case $n = 7$ in 1839. [*Added note*: I forget the reference, but I recall reading that it was proved that the method of descent would not work for $n > 17$. Also, Kummer proved Fermat's last theorem for all "regular primes".]

Some authors prove that $\sqrt{2}$ is irrational by the method of infinite descent. The argument is by contradiction and begins by assuming that $\sqrt{2}$ is rational, say, $\sqrt{2} = \frac{a}{b}$ for some positive integers a and b. Squaring each side, $2 = \frac{a^2}{b^2}$ and so $2a^2 = b^2$. Then 2 divides b^2, and hence 2 divides b, so write $b = 2k$. Then replacing this in the previous equation gives $2a^2 = 4k^2$, yielding $a^2 = 2k^2$. Again, this shows that now 2 divides a, and so $a = 2\ell$, for some positive integer $\ell > 1$. Replacing a in the previous equation now gives $4\ell^2 = 2k^2$, and hence $2 = \frac{k^2}{\ell^2}$. Thus two smaller integers k and ℓ are found with $\sqrt{2} = \frac{k}{\ell}$. This process of finding smaller integers to represent $\sqrt{2}$ as a fraction can continue forever, contradicting the well-ordering of the natural numbers.

The above proof of the irrationality of $\sqrt{2}$ does not have to take the form of infinite descent if one merely assumes at the outset that a and b are relatively prime. The contradiction is then quickly arrived at since the above proof then delivers that 2 is a common factor to both a and b. See, for example, [216] or [260] for more on proving the irrationality of $\sqrt{2}$.

Exercise 8. *Using infinite descent, prove that for each positive integer n, $\sqrt{4n-1}$ is not a rational number.*

For more results provable by descent, see (among others) Exercises 214, 222, 223, 224, and 225.

3.7 Structural induction

Computer scientists refer to mathematical induction, when applied to a recursively defined structure, as “structural induction”. Apparently, the term originally came from model theory (although I cannot find the origin) where various properties of models are proved by using chains of models, and some kind of induction on each chain. The discussion here is far less serious. In the rest of mathematics, the term “structural induction” is rarely used outside of computer science applications—as a friend once said, “it’s all just induction”.

Assume that $\mathcal{S}$ is a class of structures (it is not important what kind of structure) with some partial order $\leq$ relating individual structures. Suppose $\mathcal{S}$ contains minimal elements, and for every structure $S \in \mathcal{S}$ there is a well-ordered set of structures beginning with a minimal element in $\mathcal{S}$ and culminating in S (in other words, $\mathcal{S}$ is well-founded). Let P be some proposition about elements of $\mathcal{S}$. Then to prove the truth of $P(S)$, it suffices to prove inductively along the chain leading to S, where each inductive step is maintained by some property of the recursion used to generate structures. Then such a proof might be called a “structural induction” proof.

The most common way in which structural induction is implemented is on recursively defined structures that have some kind of “rank”—a measure of how many recursions are necessary to construct a structure from minimal structures. The typical example to help make things clear is that of rooted trees (see Section 15.2 for terminology). The rank of a rooted tree is its height, and any finite rooted tree of height h can be constructed recursively from trees of height $h-1$ by simply adding a new root. The inductive step for structural induction is usually proved by some simple property that follows from a recursive definition for the structure.

Structural induction is also used to prove properties with many base cases (as in generalized induction on well-founded sets) and can even be applied with transfinite induction (see Chapter 4).

Structural induction appears throughout this book. For examples using permutations, see the proof of (12.5). For examples regarding well-formed formulae, see Exercises 465, 466, 467. For examples using trees, see Exercises 483, 484, or 485. Graph theory uses structural induction frequently; as just one example, see Exercise 513, where structures are partite graphs, and r-partite graphs are constructed from $(r-1)$-partite graphs recursively. Other examples in graph theory where structural induction is used include theorems for amalgamation (see *e.g.*, Theorem 21.5.1 as a restricted form of amalgamation) because certain classes of graphs can be constructed by recursively gluing together two graphs on some common subgraph(s).

Hadamard matrices might be the structures concerned, and a simple tensor product construction creates recursively larger and larger Hadamard matrices (see Exercise 659). A similar notion is encountered when constructing latin squares recursively from latin rectangles (see Exercise 666). Functions form a large class of structures, and one can recursively define a function by its behavior on larger and larger domains (see, *e.g.*, Exercise 426). Colorings of objects are themselves functions, and so, for example, Exercise 731 is solved with structural induction. Certain classes of geometric objects can be considered as structures, in which case many exercises in Chapter 20 are by structural induction.

The instances of structural induction in this book are too numerous to list here. The index points to a few more examples of structural induction.

This page intentionally left blank

Chapter 4

Inductive techniques applied to the infinite

But of all other ideas, it is number, which I think furnishes us with the clearest and most distinct idea of infinity we are capable of.

—John Locke,

An essay concerning human understanding.

So far, mathematical induction has only been applied to one type of infinity, namely that of the counting numbers. In fact, mathematical induction can be performed on many other kinds of sets that have some kind of order defined on them, in particular, to sets that have a larger cardinality than that of $\mathbb{Z}^+$. These different forms of induction often depend on the axiom system decided upon. In the most common axiom systems, forms of induction for infinite sets are used to prove very powerful theorems. For example, the fact every vector space has a basis is easily proved by one of these forms.

4.1 More on well-ordered sets

Theorem 4.1.1. *There is at most one order-preserving bijection between any two well-ordered sets.*

Proof: Let $(A, <)$ and $(B, \prec)$ be well-ordered sets. Suppose that both f and g are order-preserving bijections from a A onto B. Then $g^{-1} \circ f$ is an order preserving bijection from A to itself. By Theorem 2.6.3, for all $a \in A$, $a \leq g^{-1}(f(a))$, and applying g to each side, $g(a) \leq f(a)$. Similarly, applying $f^{-1} \circ g$ to A, for each $a \in A$, $f(a) \leq g(a)$. Combining these two facts shows that for all $a \in A$, $f(a) = g(a)$. □

Definition 4.1.2. For a well-ordered set $(W, <)$ and $t \in W$, define the *initial segment* of $(W, <)$ up to t by

$$\text{seg}_{(W,<)}(t) = \{w \in W : w < t\}.$$

Define a *closed initial segment* by

$$\overline{\text{seg}(t)} = \{w \in W : w \leq t\} = \text{seg}(t) \cup \{t\}.$$

When no confusion can arise, the notations $\text{seg}_W(t)$, $\text{seg}_<(t)$, or $\text{seg}(t)$ denote $\text{seg}_{(W,<)}(t)$. If a subset S of W satisfies $a \in S, b \leq a \Rightarrow b \in S$, then either S is an initial segment of W or $S = W$.

A closed initial segment $\overline{\text{seg}(t)}$ is also an initial segment, for if ℓ is the least element of $W \backslash \text{seg}(t)$, then $\overline{\text{seg}(t)} = \text{seg}(\ell)$. However, an initial segment need not be closed; for example, consider the well-ordered set $X = \omega + 1 = \{0, 1, 2, 3, \ldots, \omega\}$. Then $\text{seg}_X(\omega) = \omega$, which is not a closed initial segment in X.

A well-ordered set is similar (or isomorphic) to the collection of all its initial segments:

Theorem 4.1.3. *Let $(W, <)$ be a well-ordered set, and put $S = \{seg(w) : w \in W\}$. Then $(W, <) \sim (S, \subset)$.*

Proof outline: It is not difficult to verify that the function $f(x) = \text{seg}(x)$ is the desired order preserving bijection. □

Lemma 4.1.4. *Let $(P, <_1)$ and $(Q, <_2)$ be well-ordered sets with $a, b \in P$ and $s, t \in Q$ and let $g : seg_P(a) \to seg_Q(s)$ and $h : seg_P(b) \to seg_Q(t)$ be order preserving bijections. If $a < b$, then $h \mid_{seg_P(a)} = g$.*

Proof: Let $a < b$. Suppose the conclusion fails, that is, suppose $h \mid_{\text{seg}_P(a)} \neq g$. Because the set of all those $y \in \overline{\text{seg}_P(a)}$ with that $g(y) \neq h(y)$ is a subset of a well-ordered set, fix the least element $y_0 \in \text{seg}_P(a)$ such that $g(y_0) \neq h(y_0)$. There are a number of ways to derive a contradiction. If $g(y_0) < h(y_0)$, then for every $x, z \in \text{seg}(a)_P$ with $x < y_0 < z$, since h is order-preserving, $h(x) = g(x) < g(y_0) < h(y_0) < h(z)$ shows that $g(y_0)$ is not in the range of h, contradicting h being onto an initial segment. Similarly, $h(y_0) < g(y_0)$ implies that g is not onto an initial segment. □

Theorem 4.1.5. *Let $(W, <)$ be a well-ordered set. For any $w \in W$, there is no order preserving bijection from W to $seg(w)$.*

Proof: If $f : W \to \text{seg}(w)$ is any *function*, then $f(w) < w$, so by Theorem 2.6.3, such a function can not be an order preserving bijection. □

Exercise 9. *Show that an arbitrary union of initial segments in a well-ordered set $(W, <)$ is either another initial segment of W or is W itself. Similarly, the union of closed initial segments will always be an initial segment, the closure of an initial segment, or W itself.*

4.2 Transfinite induction

The principle of mathematical induction, as seen so far, applies to only sets that have a well-ordering identifiable with the well-ordered set $\mathbb{N}$. In short, transfinite induction works just like the principle of mathematical induction, however applies to any well-ordered set, particularly, to infinite ordinals other than ω. In fact, transfinite induction is a generalization of "strong induction" (see Theorem 3.2.1).

Principle of transfinite induction: Let A be a subset of a well-ordered set X with x_0 being the least element of X. If

(i) $x_0 \in A$, and

(ii) for every $x \in X$, $[\mathrm{seg}_X(x) \subseteq A] \to [x \in A]$

then $A = X$.

In fact, in the statement of the principle of transfinite induction, one can even dispense with part (i), since $\emptyset \subset A$ and if (ii) holds, $\emptyset = \mathrm{seg}(x_0)$ implies that $x_0 \in A$.
Proof of the transfinite induction principle: Suppose that it fails, that is, suppose the condition (ii) holds, but $A \neq X$. Put $C = X \backslash A$. Since X is well-ordered and $C \subset X$, C has a least element, say $c \notin A$. Then $\mathrm{seg}_X(c) \subseteq A$, and by (ii), $c \in A$, a contradiction. □

Transfinite induction is suited to proving theorems about initial segments of well-ordered sets. The same principle can be adapted to other statements, however, caution is needed regarding what axioms are being used (see [160] and the comments at the end of Section 4.4). Loosely speaking, if some process or construction is based on transfinite induction over a well-ordered set, the process is called *transfinite recursion*. For example, if W is a well-ordered set, one can define a function f on W to any set X by defining inductively the map f, by defining each $f(w) \in X$ according to how f is defined before w. By transfinite induction, the resulting map on all of W is again a function.

Lemma 4.2.1. *For any two well-ordered sets P and Q, either they are similar or one is similar to an initial segment of the other.*

There are many proofs of Lemma 4.2.1; the proof below can be found in, *e.g.*, [160]. Given any function g and set A contained in the domain of g, the shorthand $g[A] = \{g(a) : a \in A\}$ is used for the image of A under g; the function g restricted to A is denoted by $g|_A$. If either of P or Q is finite, then the smaller one is similar to an initial segment of the other; hence any proof need only be applied when P and Q are infinite. One may interpret the idea in the proof given below as an attempt to construct (inductively) an order-preserving injection f from P onto an initial segment of Q; if this process fails for some $P' \subseteq P$, then f takes P' onto Q (then f^{-1} is a bijection from Q onto P').

Proof of Lemma 4.2.1: Let P and Q be well-ordered sets. Fix some symbol x not in Q. By transfinite recursion, define the function $f : P \to Q \cup \{x\}$ by

$$f(p) = \begin{cases} \text{the least element of } Q \setminus f[\text{seg}(p)], & \text{if } Q \setminus f[\text{seg}(p)] \neq \emptyset \\ x, & \text{otherwise.} \end{cases}$$

Set $Q' = Q \cap f[P] = f[P] \setminus \{x\}$ and $P' = f^{-1}[Q] = f^{-1}[Q']$ and define $f' : P' \to Q'$ by $f' = f|_{P'}$. Then f' is onto Q'.

To see that f' is one-to-one, let $a, b \in P'$ with $a < b$; since $a \in \text{seg}(b)$ and $f(b) \in Q \setminus f[\text{seg}(b)] \subseteq Q \setminus \{f(a)\}$, $f(b) \neq f(a)$.

To see that f' is order preserving, let $a, b \in P'$ with $a \leq b$. Then since $f[\text{seg}(a)] \subseteq f[\text{seg}(b)]$, $Q \setminus f[\text{seg}(a)] \supseteq Q \setminus f[\text{seg}(b)]$ and so $f(a) = f'(a) \leq f'(b) = f(b)$.

Claim: Either $P' = P$ or $Q' = Q$.

Proof of claim: If $P' \neq P$, then there is $a \in P$ so that $f(a) \notin Q$, in which case $f(a) = x$; this means that $Q \subseteq f[\text{seg}(a)] \subseteq f[P]$ and hence $Q' = Q$, proving the claim.

Also, P' is an initial segment of P or $P' = P$ and Q' is an initial segment of Q or $Q' = Q$. If $a, b \in P$ with $a \leq b$ and $b \in P'$, then $\emptyset \neq Q \setminus f[\text{seg } b] \subseteq Q \setminus f[\text{seg}(a)]$. Thus $f(a) \in Q$ and hence $a \in P'$. Given $z, w \in Q$ with $z < w$ and $w \in Q'$, let $p \in P$ be such that $f(p) = w$. Then w is the least element of $Q \setminus f[\text{seg } p]$ and since $z < w$, $z \notin Q \setminus f[\text{seg}(p)]$ and so $z \in f[\text{seg}(p)] \subseteq f[P]$ and hence $z \in Q'$.

Therefore, f' is an order-preserving bijection either from P onto an initial segment of Q or else from an initial segment of P onto Q. □

Transfinite induction can be applied with any well-founded sets (not just well-ordered), including in proofs by structural induction (the term "structural induction" likely originated in model theory); see Section 3.7.

4.3 Cardinals

This section is a very brief introduction to cardinals to establish some terminology. If there is an injection from a set A into a set B, write $|A| \leq |B|$. If there is a bijection from A to B, write $|A| = |B|$, and say that A and B have the same *cardinality*, or are *equinumerous*. This definition is due to Cantor.

To define a cardinal, one needs to give an interpretation for $|A|$. Given two sets A and B, if there is a bijection from A to B, write $A \approx B$. It is easily seen that the relation $\approx$ is an equivalence relation on the collection of all sets. (Note that one does not say "an equivalence relation on the set of all sets", for this leads to Russell's paradox.) Although the following definition leaves open just what an element of an equivalence class is, it is convenient:

Definition 4.3.1. A *cardinal number*, or simply a *cardinal*, is an equivalence class for $\approx$.

Denote the cardinal number containing A by $|A|$. Then two sets have the same cardinality iff $|A| = |B|$.

A set A is called *countable* iff A is either finite or is equinumerous with $\mathbb{N}$, and *uncountable* otherwise. Standard proofs show that $\mathbb{Z}$ and $\mathbb{Q}$ are countable, yet $\mathbb{R}$ is uncountable. Even the set of algebraic numbers in $\mathbb{R}$ is countable. It is known that a union of countably many countable sets is again countable (see Exercise 434).

Cardinal numbers are well-ordered, so induction is often carried out on cardinalities of sets.

Cardinal numbers say something about the "size" of a set. The cardinal number (or *cardinality* of) for a well-ordered set says something about its size; to differentiate between well-ordered sets of the same size, something called *ordinals* are introduced.

4.4 Ordinals

There are different ways to define "ordinals", all equivalent. Recall that for well-ordered sets A and B (in fact, for any linearly ordered sets) A is similar to B, written $A \sim B$, if and only if there exists an order preserving bijection $f : A \to B$.

Definition 4.4.1. An *ordinal* is an equivalence class under $\sim$.

Ordinals are sometimes called *ordinal numbers*. Different ordinals have different "shape". If α is an ordinal and $A \in \alpha$, then A is said to be of type α. An ordinal α is often identified with any set of type α. For two ordinals α and β, say that α precedes β if and only if there exist $A \in \alpha$ and $B \in \beta$ so that $A \subseteq B$. The order on ordinals is given by $\alpha < \beta$ iff α precedes β. Thus, with an abuse of notation, $\alpha < \beta$ can be written $\alpha \subsetneq \beta$, or simply $\alpha \subset \beta$.

For an ordinal β, one can identify each element $\alpha \in \beta$ with its initial segment $\text{seg}(\alpha)$, the set of predecessors of α. Given this identification, some define an *ordinal*, to be a well-ordered set $(X, <)$ with the property that every element $\alpha \in X$ is equal to its initial segment. One can (and some do) *define* natural numbers (and 0) as ordinals: Put $0 = \emptyset$; $1 = \{\emptyset\}$=$\{0\}$; $2 = \{\emptyset, \{\emptyset\}\} = \{0, 1\}$; $3 = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0, 1, 2\}$, and in general, $n = \{0, 1, \ldots, n-1\}$.

The ordinal number $\omega = \{0, 1, 2, \ldots\}$ (with the usual well-order) is the first infinite ordinal, which is really the set of all natural numbers together with 0. (This is a "good" reason why some texts use 0 in the definition of natural numbers—so that they can identify ω with $\mathbb{N}$.) In ordinal arithmetic, if any one of $\alpha < \beta$, $\alpha \in \beta$, or $\alpha \subset \beta$ hold, then all hold.

An ordinal $\beta \neq \emptyset$ is one of two types:

- β is called a *limit ordinal* if $\beta = \cup_{\alpha<\beta}\alpha$, and
- β is called a *successor ordinal* if $\beta = \alpha \cup \{\alpha\}$ for some ordinal α (in this case, β is the smallest ordinal larger than α, sometimes denoted by $\beta = \alpha + 1$ or α^+).

Recall that two sets have the same *cardinality* (think "size") iff there is a bijection between them (see Exercises 430 and 431). A *cardinal number* is an ordinal β whose every initial segment has a different cardinality than β. Every infinite cardinal number is a limit ordinal.

Here are (without proof) a sequence of lemmas that can be used to prove the subsequent theorem. Some of these facts have proofs that rely on facts already proved for well-ordered sets in general; most proofs are simple, and can be considered as exercises. (For details, see [95, pp. 42–43].) The subsequent theorem is used later to give a simple proof by transfinite induction, a proof that could otherwise be very complicated.

Lemma 4.4.2. *Every initial segment of an ordinal is again an ordinal.*

Lemma 4.4.3. *If α and γ are ordinals with $\alpha \subset \gamma$, then α is an initial segment of γ.*

Lemma 4.4.4. *For any distinct ordinals α and β, one is an initial segment of the other. Thus any collection of distinct ordinals is linearly ordered by inclusion. This order is indeed a well-order.*

Lemma 4.4.5. *The union of a* set *of ordinals is again an ordinal.*

Theorem 4.4.6 (Burali-Forti paradox)**.** *The collections of all ordinals is not a set.*

Proof outline: Let C be the collection of all ordinals. Then C itself is an ordinal, greater than each of its members, a contradiction. □

Suppose that $P(\alpha)$ is a statement involving an ordinal α, perhaps infinite. If $P(\alpha_0)$ holds and $P(\alpha) \rightarrow P(\alpha+1)$ then by transfinite induction, one can conclude that $P(\beta)$ holds from α_0 up to any ordinal below the next limit ordinal. As is, however, one can not 'jump' to the limit ordinal.

For a limit ordinal β the statement $P(\beta)$ is proved by showing that for every $\alpha \in \beta$, $P(\alpha)$ holds. This allows one to prove P "across limit ordinals". Ordinary induction can be thought of as 'pushing up' from n to $n+1$; transfinite induction can be considered more as 'reaching down and pulling up'. Proofs by transfinite induction are often divided into three cases, one for the base case, one for successor ordinals, and one for limit ordinals. In many instances, only two steps are required, since the process for limit ordinals usually works for successor ordinals, too. In fact, by the comment above, often only one case is necessary.

There are many proofs in set theory that rely on constructing functions by transfinite induction. One might find different proofs in the literature for the same result, both using some kind of transfinite induction, some relatively short, some very long (and horrid). Thanks to an explanation by Prof. Kucera [331] the reason for difference in complexity lies in a subtlety not usually a concern for *non-foundationalists* (translation: mere mortals). Here is a very rough account of that subtlety: In the development of set theory, some authors prefer proofs that don't invoke a "replacement axiom" (see Appendix IV for statement) unless necessary. Avoiding unnecessary assumptions can make a proof very difficult.

As Kucera commented, the replacement axioms are not necessary if transfinite induction is used to construct functions from ordinals to ordinals, however for functions from ordinals to the class of all sets, often the extra axioms are required. (See [160, pp. 178–9] for more intelligent discussion on this matter.) Others implicitly assume these axioms (or assume them at the onset, many pages before the proof in question). For example, (these abbreviations are defined below) compare proofs of AC implying WO, [160, Thm 6M] or [347, p. 182] with [289, Thm 15], or proofs of AC implying ZL, [416, p.531] with [289, p. 40].)

4.5 Axiom of choice and its equivalent forms

In set theory, one begins not with Peano's axioms, but with axioms that apply to sets in general, not just $\mathbb{N}$. (See Appendix IV for such a collection of axioms.) The most famous of such axioms is the "Axiom of Choice" (AC). There are, in fact, many axioms that have been shown to be equivalent to the Axiom of Choice. In this section, a few of these forms are given together with a sequence of proofs showing them all to be equivalent. In any such sequence, it seems that there is always at least one step that is difficult, especially if one restricts the tools available. For more on such equivalences, the reader might look at the reference standards by Herman and Jean Rubin [472, 473]; see also a more recent book [281] by P. Howard and Jean Rubin on consequences of the Axiom of Choice.

To state the Axiom of Choice, a definition is helpful. If $\mathcal{F}$ is a family of sets, a choice function for $\mathcal{F}$ is a function

$$\gamma : \mathcal{F} \to \cup_{F \in \mathcal{F}} F$$

so that for every $F \in \mathcal{F}$, $\gamma(F) \in F$. So a choice function picks an element from every set in the family.

Axiom of Choice (AC): If $\mathcal{F}$ is a non-empty family of non-empty sets, then $\mathcal{F}$ has a choice function.

A standard example (some say it goes back to Bertrand Russell (1872–1970)) used to demonstrate what AC says is: among infinitely many pairs of shoes, it would be easy to pick one shoe from each pair—pick the left one of each pair. If, however, there were infinitely many pairs of socks, AC guarantees that there is still a choice function that picks one sock from each pair. This doesn't seem very surprising, and in fact, most would argue that this goes without saying. The subtlety might lie in the fact that the Axiom of Choice says that all these socks can be picked *at once* even though there is no way to differentiate between socks of a pair.

It might be interesting to note that the following apparently weaker "Axiom of Choice" is indeed equivalent to the Axiom of Choice:

Zermelo's Postulate: For every non-empty family of disjoint non-empty sets $\mathcal{S}$ there exists a choice function $f : \mathcal{S} \to \cup_{S\in\mathcal{S}} S$.

Theorem 4.5.1. *The Axiom of Choice is equivalent to Zermelo's Postulate.*

Proof: The Axiom of Choice clearly implies Zermelo's Postulate, so it suffices to prove only the other direction. Assume that Zermelo's Postulate is true and let $\mathcal{S} = \{S_i : i \in I\}$ be a non-empty family of sets, not necessarily disjoint. From $\mathcal{S}$, create a disjoint family as follows. For each $i \in I$, set $S_i^* = S_i \times \{i\} = \{(s,i) : s \in S_i\}$. Then $\mathcal{S}^* = \{S_i^* : i \in I\}$ is a disjoint family of non-empty sets. By Zermelo's postulate, fix a choice function $f* : \mathcal{S}^* \to \cup_{i\in I} S_i^*$, and for each $i \in I$, set $f*(S_i^*) = (s_i, i)$. Then the function $f : \mathcal{S} \to \cup_{i\in I} S_i$ defined by $f(S_i) = s_i$ is a choice function for $\mathcal{S}$. □

Note: In the above proof, the fact was used that the family of sets was indexed. If a given family of sets is not indexed, how can one create an index set for this family? One has to look more closely at what at an indexed set is. An *indexing* of a family of sets $\mathcal{F}$ by a set I is a bijection

$$\eta : I \to \mathcal{F}.$$

In this case, write $\eta(i) = F_i$ and $\mathcal{F} = \{F_i\}_{i\in I}$. If one chooses η to be the identity function on $\mathcal{F}$, a family of sets can itself act as the index set! Hence, any family of sets can be indexed.

The Axiom of Choice can be stated for indexed families of sets, but is often done using product notation. Since product notation can be a bit confusing for infinite products, the reader might be forgiven for erring on the side of being too pedantic in the following explanation.

Recall that the cartesian product of two sets is written

$$X_1 \times X_2 = \{(a,b) : a \in X_1, b \in X_2\}.$$

To generalize this to a product of infinitely many sets, reinterpret the product $X_1 \times X_2$ as follows: Each $(a,b) \in X_1 \times X_2$ can be considered as the image of a function, $\gamma : \{X_1, X_2\} \to X_1 \cup X_2$, where $\gamma(X_1) = a \in X_1$ and $\gamma(X_2) = b \in X_2$. As a trivial example, if $X_1 = \{5, 10\}$ and $X_2 = \{3, 5\}$, since there are four ordered pairs in $X_1 \times X_2$, there are four different functions, say γ, δ, ϕ, ξ, to be considered:

$$\begin{aligned} \gamma(X_1) &= 5, & \gamma(X_2) &= 3; \\ \delta(X_1) &= 5, & \delta(X_2) &= 5; \\ \phi(X_1) &= 10, & \phi(X_2) &= 3; \end{aligned}$$

$$\xi(X_1) = 10, \qquad \xi(X_2) = 5.$$

Notice that the meaning of the function γ is clear without the order mattering: "$\gamma(X_1) = 5$ and $\gamma(X_2) = 3$" means precisely the same thing as "$\gamma(X_2) = 3$ and $\gamma(X_1) = 5$". The meaning of the ordered pair $(3, 5)$ captures this if one remembers that the 3 came from X_1 and the 5 came from X_2. Then

$$X_1 \times X_2 = \{f : \{X_1, X_2\} \to X_1 \cup X_2 : f(X_1) \in X_1, f(X_2) \in X_2\}.$$

Dealing with indices only, write

$$X_1 \times X_2 = \{f : \{1, 2\} \to X_1 \cup X_2 : f(1) \in X_1, f(2) \in X_2\}.$$

If the sets were indexed by something other than numbers, the meaning of their product would then not depend on order at all:

$$X_\clubsuit \times X_\triangle = \{f : \{\clubsuit, \triangle\} \to X_\clubsuit \cup X_\triangle : f(\clubsuit) \in X_\clubsuit, f(\triangle) \in X_\triangle\}.$$

Recall that $A \times B \neq B \times A$, because the first is ordered pairs of the form (a, b) while the second consists of ordered pairs of the form (b, a). The difference is only in the order in which one writes them down. In fact, either would be fine, if only one had some way of knowing which of the ordered pair came from which set. Usually, the first coordinate is to mean that the element came from the first set listed in the collection A, B. If there are infinitely many sets, however, and no order imposed on the list, then what does one do? The answer is simple: go back to the function interpretation of the product.

Definition 4.5.2. For a family of indexed sets $\{F_i\}_{i \in I}$, define the infinite product

$$\prod_{i \in I} F_i = \left\{ f : I \to \bigcup_{i \in I} F_i : \text{ for each } i \in I,\ f(i) \in F_i \right\}.$$

Any function $f : I \to \cup_{i \in I} F_i$ for which each $f(i) \in F_i$, in fact determines a choice function. Thus, the Axiom of Choice can be restated as follows:

Axiom of Choice (indexed version): Let $\{A_i\}_{i \in I}$ be a family of non-empty sets. Then $\prod_{i \in I} A_i \neq \emptyset$.

Another axiom that is often a starting point in set theory is called the *well-ordering principle*. By Theorem 2.6.2, the natural numbers can be (or are) well-ordered; can any set be well-ordered? No one has been able to *prove* otherwise, so the following might seem like a reasonable axiom:

Well-ordering principle (WO): Any set can be well-ordered.

In 1904, Zermelo wrote in a letter to Hilbert that the Axiom of Choice implied the well-ordering principle (see [586]). However, Eves [181, p. 297] says that after Zermelo proved the well-ordering principle, it was Emile Borel who was searching for a flaw in Zermelo's proof, and discovered that it relied on the Axiom of Choice, and pointed out that the Axiom of Choice (AC) is equivalent to the well-ordering principle (WO) (that is, if AC is true, then WO is true, and if WO is true, then AC is true). A modified proof of AC implying WO was then published by Zermelo [587] in 1908. Some authors call the well-ordering principle the *well-ordering theorem* (since it can be derived it from the Axiom of Choice). The reverse direction is easy:

Theorem 4.5.3. *The well-ordering principle implies the Axiom of Choice.*

Proof: Suppose that WO holds and that $\mathcal{F}$ is a family of non-empty sets. Since WO holds, each $F \in \mathcal{F}$ can be well-ordered. Since every well-ordered set F contains a minimal element, say, $\min F$, then the function $f : \mathcal{F} \to \cup_{F\in\mathcal{F}} F$ defined by $f(F) = \min F$ is a choice function. □

The other direction (AC implying WO) is not as simple. Two proofs are given here. The first is an adaptation of that found in [347, p. 182] combined with notes on a lecture given by R. Aharoni, (at University of Calgary, 1986). This proof apparently does not rely on replacement axioms. A second proof, occupying only one paragraph, is from Jech [289, p. 39] and is vastly simpler, relying on the stronger form of transfinite induction.

Theorem 4.5.4. *The Axiom of Choice implies the well-ordering principle.*

First proof of Theorem 4.5.4: Let X be a set let $f : 2^X\backslash\{\emptyset\} \to X$ be a choice function.

Look at pairs of the form $(W, <)$, where $W \subseteq X$ and $<$ is a well-ordering of W. Define a pair $(W, <)$ to be *f-compatible* iff for every $t \in W$,

$$f(X\backslash\text{seg}_{(W,<)}(t)) = t.$$

Such f-compatible sets exist by the following: let $x_0 = f(X)$, $x_1 = f(X\backslash\{x_0\})$, and $x_2 = f(X\backslash\{x_0, x_1\})$. It is not difficult to verify that $W = \{x_0, x_1, x_2\}$ with the ordering $x_0 < x_1 < x_2$ is indeed f-compatible.

[*Comment:* If $(W, <)$ is f-compatible, then $(W, <)$ was created according to the rule: choose a next element to be $f(X\backslash$elements chosen so far)].

Fact 0: For any f-compatible sets $(W_1, <_1)$ and $(W_2, <_2)$, either they are equal or one is an initial segment of the other.

Proof of Fact 0: Let $(W_1, <_1)$ and $(W_2, <_2)$ be f-compatible. Since $(W_1, <_1)$ and $(W_2, <_2)$ are well-ordered sets, by Lemma 4.2.1, either they are similar or one is similar to an initial segment of the other. Without loss, let $\alpha : W_1 \to W_2$ be a similarity from W_1 onto either W_2 or an initial segment of W_2.

Let $W_1^* = \{w \in W_1 : \alpha(w) \neq w\}$, everything that α moves. If $W_1^* = \emptyset$, then either $W_1 = W_2$ or W_1 is an initial segment of W_2 (since α is either a similarity onto W_2 or onto an initial segment of W_2). Thus, Fact 0 is proven if it can be shown that $W_1^* = \emptyset$.

Suppose, in hopes of contradiction, that $W_1^* \neq \emptyset$, and let t_0 be the least element of W_1^*. Then

$$t_0 = f(X\backslash\text{seg}_{(W_1,<_1)}(t_0)) = f(X\backslash\text{seg}_{(W_2,<_2)}(\alpha(t_0))) = \alpha(t_0),$$

contradicting $t_0 \in W_1^*$. Hence $W_1^* = \emptyset$, proving Fact 0.

So by Fact 0, without loss assume that $(W_1, <_1)$ and $(W_2, <_2)$ are such that W_1 is an initial segment of W_2 (or $W_1 = W_2$). If $a, b \in W_1 \subset W_2$ then $a <_1 b$ iff $a <_2 b$ (that is, the orders are "compatible" for f-compatible well-orderings).

Let $V = \{x \in X : \text{for some } f\text{-compatible w.o. } (W, <), x \in W\}$. Define $(V, \prec) = \cup\{(W, <) : (W, <) \text{ is } f\text{-compatible}, W \subset X\}$.

By the compatibility of the f-compatible well-orderings, the next fact follows:

Fact 1: $(V, \prec)$ is a totally ordered set.

Fact 2: $(V, \prec)$ is a well-ordered set.

Proof of Fact 2: Let $T \subset V$, $T \neq \emptyset$, and let $t \in T$. Then for some well-ordered set $(W, <)$, $t \in W$. Since W is well-ordered, $W \cap T$ has a least element, call it x.

[*Aside*: If $x \in V$ and W is a w.o. such that $x \in W$, then for any $y \in V$, if $y \prec x$, then $y \in W$ because for some w.o. W', $y \in W'$, but one of W or W' is an initial segment of the other, so $y \in W \cap W'$.]

Thus for any $s \in T$, if $s \prec x$, then $s \in W$. But then $s \prec x$ and $s \in W \cap T$, contradicting x being the least in $W \cap T$. Therefore, x is the least element of T, proving Fact 2.

Fact 3: $(V, \prec)$ is f-compatible.

Proof of Fact 3: Pick $v \in V$ with $(W_1, <_1)$ an f-compatible set such that $v \in W_1$.

The next thing to show is that $\text{seg}_{(W_1,<_1)}(v) = \text{seg}_{(V,\prec)}(v)$. Since $(W_1, <_1) \subseteq (V, \prec)$, it follows that $\text{seg}_{(W_1,<_1)}(v) \subseteq \text{seg}_{(V,\prec)}(v)$. Let $x \in \text{seg}_{(V,\prec)}(v)$; then for any $(W_2, <_2)$ with $V \in W_2$, (by the aside above) $x \in W_2$, so $x \in \text{seg}_{W_1,<_1)}(v)$. Therefore, $\text{seg}_{(W_1,<_1)}(v) = \text{seg}_{(V,\prec)}(v)$.

So $f(X\backslash\text{seg}_{(V,\prec)}(v)) = f(X\backslash\text{seg}_{W_1,<_1)}(v)) = v$ since W_1 is f-compatible.

Fact 4: $V = X$.

Proof of Fact 4: Suppose not, that is, suppose that $X\backslash V \neq \emptyset$, and since f is a choice function, put $z = f(X\backslash V)$. Put $V' = V \cup \{z\}$ and extend the order $\prec$ to $\prec'$ by defining for every $v \in V$, $v \prec' z$. Then $\text{seg}_{(V',\prec')}(z) = V$ and so $f(\text{seg}_{(V',\prec')}(z)) = f(X\backslash V) = z$. So $(V', \prec')$ is f-compatible, and so $z \in V' \subset V$