Principles of mathematics and logic
A course for Liberal Arts students

R.A.G. Seely
Mathematics Department
John Abbott College
Ste Anne de Bellevue, QC

Version 2020.9.15

Contents

Note to students

This book is intended to accompany the *Principles of Mathematics and Logic* course, given in the Liberal Arts program at John Abbott College. This course covers virtually all the material in the text; you should expect to read it cover to cover. Of course, there are also lectures which make up an important part of the course; you will find that often I emphasise things somewhat differently in class and in the book—the intention is that each should complement the other, rather than replace it. You should not skip class, expecting to make it up with the text (instead, attend class regularly), and similarly, you should not rely solely on your class notes (read the book for the extra examples and explanations). The most important part of the book is the exercises: it is a (true!) cliché that mathematics is a poor spectator sport, and to *learn* mathematics properly, you must *do* mathematics. Take this seriously: you will find it very hard to succeed unless you actually practice the ideas learned in class.

Generally, when you read this text, indeed any mathematics, it is important to engage the text actively, not passively. You should have pencil and paper beside you, and try to follow each statement, doing the suggested calculations or reasoning yourself. It is not a novel or short story, whose meaning will just flow over you, but a dialogue, only one side of which is on the page. You must provide the other side yourself!

In particular, you will find lots of examples with explanations; try to do the examples yourself (especially after the first one, or after seeing some in class). A good idea is to try to do an example without looking at the explanation, only turning to the text for hints as you go. When you've done the example, read through my explanation to see if you understand it all, and then go onto the next example. And of course, *do the exercises!*

There is a course webpage (`www.math.mcgill.ca/rags/jac.html`); I have put additional material there, including further readings (some intended to give you further explanation and examples of topics covered in class, some intended to go further in some topics than covered by the course, and some intended to interest you, without any intent at "examinable material"), further exercises (particularly practice tests to help you prepare for the class tests), and any other relevant information (for instance, your marks will be found there after tests). You should bookmark the webpage, and visit it often to see if I've put new material there for you.

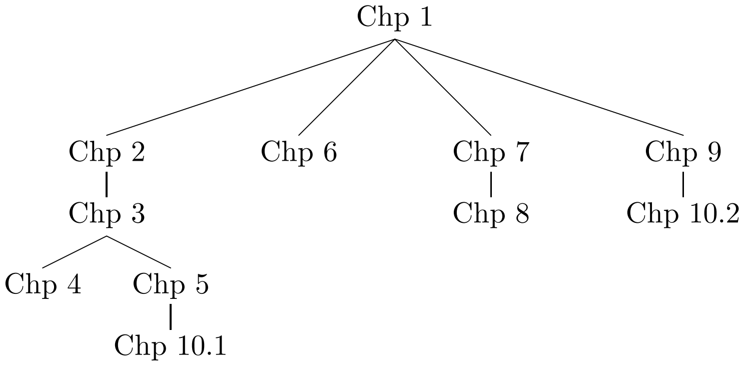
Acknowledgements

This book is based on many sources, in addition to my own work: I am most particularly indebted to the earlier text written by the late Gerald LaValley, used in this course for many years. A substantial portion of this text is inspired by Gerry's book, especially Chapters 1 and 6–8 which are heavily influenced by his approach. Gerry also carefully read early drafts of my text, and I am very grateful for his corrections, comments and suggestions. I have also benefited from notes used at the University of Ottawa, written by Phil Scott and Peter Selinger, and from *Formal Systems and Logic in Computing Science* by W.W. Armstrong, F.J. Pelletier, R.A. Reckhow, & P. Rudnicki. In addition, the section on Knights and Knaves (Chapter 1) relies heavily on the related book by Raymond Smullyan.

Many decades ago I learned from my brother John that the right way to combine the Natural Sciences with the Liberal Arts is with a conjunction, not with a disjunction. I hope to help the reader to get this message too: engage deeply with at least one aspect of the human story, but remain in touch with the rest of the tale. Otherwise you will not be as complete a person as you can be.

Major dependencies

The text is intended to be read (and covered in class) in order, but in fact some chapters only depend lightly on some prior chapters. So, with minor dependencies that can be "ignored", here is the essential dependency graph: you really need the chapters above a certain chapter to understand the latter.



Note: From Chapter 7 onwards, the language of set theory is used, but little of the "serious" content of Chapter 6 is needed. So, these later chapters do depend on Chapter 6, but only lightly. Section 10.2 is a bit unusual: it really does not use previous material in a serious way, but it will perhaps make more sense to a reader familiar with what's gone on earlier in the text. I've indicated this with a dependency on Chapter 9, but even that isn't really true(!).

Section 10.1 depends to some extent on understanding the basic structure of the natural numbers, so in a sense also depends on Chapter 8, but could be read without that. For example, mathematical induction (from Chapter 8) is mentioned, but not used in any essential manner.

For John, in memory

Chapter 1

Introduction to Logic

1.1 Mathematics and Science in A Liberal Arts Education

Historically the liberal arts included arithmetic, geometry, astronomy, fine arts and history, grammar, rhetoric and logic. As broad general education, Liberal Arts programs are alternatives to training in a trade or craft.

Some students think of Liberal Arts as the history of (Western) culture and the themes, styles and movements in literature and the fine arts—an encyclopedia of cultural facts; lists of historical particulars.

General (liberal arts) education has always been more than just particular truths in a narrow range of fields. It aims to reveal the relevance of these truths, the connections and relations among the particulars, and the subsumption of particulars under abstract general principles. Students should understand not just the truths but the search for truth; not just knowledge but the methods by which we acquire and confirm knowledge. Facts are important, but the interpretation of what the facts mean is crucial.

Logic is central to this understanding.

Western culture is the result of developments in mathematics and the physical and social sciences as much as it is a product of "merely historical" accidents or of changes in artistic or literary directions. Human creativity and awareness are as evident in logic, mathematics, and the sciences as they are in philosophy and the fine arts. For those who develop the understanding, sensitivity, and taste, the great logical and mathematical proofs and the deep and subtle theories of the sciences are as beautiful and as admirable as any product of the human spirit.

1.1.1 Logic and rationalization

"Logic" has been defined as the science of right reasoning.

Freud and Marxists and the existentialists encourage a common confusion about the relationship between logic, rationality, and rationalization. Their idea is that people use logic to "explain away" behaviours and attitudes whose real explanations are non-logical. Freudians claim that one's beliefs are not grounded on logical reasons but have their source in the sub-conscious; Marxists blame ideology; existentialists emphasize "bad faith".

The science of logic begins with a value-judgment: "what is right reasoning?". The identification of logic with rationalization ("explaining away") is based on a relativistic view of values. The claim is that there is no one standard of right reasoning, but that "right" (like any value) is a matter of taste. Standard canons of logic are decided by whatever social group (class, gender, etc.) has the power to impose its standards of "right reasoning" on the rest of society. So Marxists claim that logic is a bourgeois requirement. Some feminists claim that it is something that males impose on the world.

This text rejects such relativism.[1]

1.1.2 Many logics, not just one logic

May one admit different standards of "right reasoning" if one is considering different contexts? It seems perfectly reasonable that one might. However, admitting that there are various notions of "right reasoning" does not mean that one admits that the notion of "right reasoning" is merely a matter of taste. One thing we insist on is what philosophers of science call "reproducibility": if two reasonable observers observe the same phenomenon, they will make the same observation. We shall insist, then, that "right reasoning", and the logic that encodes it, must satisfy this requirement of "reproducibility"; logic is no mere matter of opinion.

This may well make developing a logic suitable for political purposes an impossible task! In fact, what logic is suitable for a specific purpose may well be a matter of opinion (and often is!). But that gets us into philosophical disputes that have lasted for centuries, and so takes us way beyond what we can cover in one semester.

So, it has to be said right at the start of the course that there is no single logic which encompasses the idea of capturing "right reasoning"; instead there are several candidates which have been studied in considerable detail in the past century. Each of these logics attempts to capture specific aspects of right reasoning, usually suitable for specific applications or circumstances. We shall emphasize (mainly for reasons of simplicity) the traditional "classical" logic, whose origins go back to the Greek philosophers (such as Aristotle), and which was what was principally meant by the term "logic" for centuries.

A very important variant of classical logic, which became a serious matter of philosophical and mathematical scrutiny only around the early 20th century, is "intuitionist logic", a logic intended to capture a more "constructive" aspect of logic; we shall discuss this at several junctures as we study classical logic. With the development of computing science, an increasing need grew for a logic which was more "resource sensitive"; this need is met by a collection of what are known as "sub-structural logics". These were not new to philosophers, as they had already been considering various "relevance logics", which were an attempt to address some seemingly paradoxical behaviour of the classical notion of *implication* (remarked upon near the end of Section 1.3.2); in Chapter 10 we shall briefly consider how relevance logic can avoid some of this. Related to such logics are various candidates for a logic of quantum phenomena, a logic suitable for underlying (for example) quantum computing, and more generally for understanding quantum physics. This is a field of current active research.

Other distinctions are possible: for instance there is a large family of "modal logics",[2] whose intention is to study the logical properties of notions like "possibility", "necessity", "belief", *etc.* Again, we shall discuss these, but only briefly. Another distinction is made between deductive and inductive logic (inductive logic is very commonly used in the natural sciences, less so in mathematics): this is essentially a distinction between a logic aiming at determinate, definite conclusions and a logic aiming at probabilistic conclusions. Inductive logic is often identified with statistics, but there is active current research into a suitable formal logic for such matters.

[1] Perhaps this is as good a time as any to mention that mid-way through the course, you will be asked to read the essay *On Bullshit* by Harry Frankfurt.

[2] In the wide sense, "modal logic" includes many logics such as temporal, deontic, and doxastic logic, for instance. You may Google these, if you wish, as we shall not go there!

Furthermore, even traditional logics (such as the classical logic we shall study) may be enhanced by adding other features, to allow other aspects of "right reasoning" to be captured. For instance, we may add to classical logic the ability to make sentences of the form "infinitely many objects have such-and-such a property" (in addition to the sentences of the form "all objects have such-and-such a property" and "some object has such-and-such a property", which are already allowed, and which we shall study). The list of variants seems almost endless (in a sense, it is!).

Many logics for many purposes—each has its own characteristics, its own properties. The study of most of these logics follows a similar plan, which is simplest in the case of classical logic, so we shall use that as an illustration of how a formal logic, a logic to capture the elusive notion of "right reasoning", may be developed in a scientific, perhaps we should say mathematical, fashion.

1.1.3 Pure mathematics and logic

In this course we mostly study pure (or formal or theoretical) mathematics and logic, more than applied mathematics and logic. We study maths and logic from a theoretical point of view. Practical applications (such as using logic to persuade somebody or using mathematics for utilitarian calculations) are secondary. The aim is to develop some understanding of what these two sciences are about, of their methods, and of their beauty and interest for their own sakes.

So we study the principles of these two fields of study, not how one applies them. In logic, the course does not aim to teach rhetoric. In mathematics, this means that our ability to perform calculations will not be emphasized. Unlike most college mathematics courses, this course does not emphasize applying the techniques of trigonometry and differential and integral calculus. We look at the basic assumptions behind the two fields, the way mathematicians and logicians arrive at their basic assumptions, and the way they arrive at conclusions based on those assumptions. In particular, we do *not* regard mathematics as "the science of quantity", or any such definition (if this challenges your preconceptions, so much the better!). Rather, we regard the essence of mathematics (including logic) as the study of *pattern*. The word "pattern" means many things of course, but one property that is intended by my usage is "reproducibility": whether or not a pattern is (say) beautiful is (probably) not a reproducible property (we may disagree on whether something is beautiful), but the pattern itself is—by the very nature of what a pattern is. Mathematicians study pattern in many contexts: among numbers, geometric shapes, and logic, to be sure, but also in other domains, including (but not limited to) music, natural language, computer programming languages and computer programs themselves, as well as more "useful" domains, like the movement of planetary bodies, and the performance of the stock exchange. In this course we'll see a few instances of this, at an elementary level: we'll consider patterns in logic, in numbers and sets, and (time permitting) in natural language.

In addition, since the early twentieth century, mathematics (especially pure mathematics) has tended to have a characteristic method or procedure, often referred to as "the axiomatic method" (we shall study some examples at the end of the course). In studying a subject or discipline, one begins with a set of undefined elements, properties and relations among these elements, and fundamental "truths" called postulates or axioms, which establish the basic facts of the subject. From these all other facts (theorems) should be derived by formal logic, without appeal to any external knowledge. Some commentators stress that the undefined elements should not be regarded as concrete entities, but rather as some sort of "variables", which may be interpreted in any way consistent with the axioms—in this way, the subject becomes merely the study of what consequences may be derived from the initial axioms. It might appear that this approach tends to identify mathematics with "applied logic", and indeed, there was a serious attempt to reduce mathematics to logic early in the twentieth century with the mammoth 3-volume set *Principia Mathematica*

(Cambridge University Press, 1910) by A.N. Whitehead and B. Russell.

But the reduction of mathematics to logic suffered serious blows, right from its inception as an idea. Firstly, when Whitehead and Russell tried to implement their idea, they found they needed a non-logical axiom of infinity in order to even capture simple arithmetic, in spite of many efforts to avoid such an extra assumption. (They could describe simple "natural numbers" like 1 and 2 in purely logical terms, and they could even, after several hundred pages, prove that $1 + 1 = 2$, but what they could not do was talk about *all* natural numbers without the axiom of infinity.)

Then, in the early 1930s, Kurt Gödel proved that in their system (or in any similar system for mathematics) there were statements (which were "obviously true" in some sense) which could never be proven nor disproven (unless their system was in fact inconsistent). Gödel explicitly saw this result of his as showing that mathematics could never be reduced to a merely formal or logical system, but that some other considerations, mathematical considerations, were an essential part of the story. Moreover, he believed that mathematics studied real phenomena, not merely intellectual creations: entities such as numbers, geometric shapes, *etc.*, may not exist as tangible objects like rocks, cats and dogs, or even MP3 files, but they have a reality nonetheless. This view is often called "Platonic" (for reasons I need not explain to this audience!). We may discuss these matters when we consider Gödel's theorem.

Furthermore, while the axiomatic method certainly describes the modern practice of a considerable body of mathematics, and it does indicate the close relationship between mathematics and logic (a relationship we shall see throughout the course), nonetheless one must remember to distinguish between 'method' and 'essence'. The method does describe part of what mathematicians do (and how they do it), but still it does not entirely address the essence of what mathematics *is*, a much more complicated and obscure matter. In particular, it ignores the question of what makes some collections of axioms more valuable as an object of study than others, for instance. That usually involves the consideration of what the mathematics is used for, whether for other parts of mathematics or for "real-life problems"; it may also involve matters of "taste" (and the less reputable, but equally compelling, notion of "fashion"), and often simply what captures the imagination and passion of the mathematicians working in a discipline. One key motivation is the love of beauty (which brings us back to "pattern").

1.1.4 Patterns in sciences and arts

For a Liberal Arts student, the main relation between the sciences on one hand and the fine arts and literature on the other is that both study patterns (there's that word again!). Empirical sciences study the patterns in nature. Logic studies the patterns of human reasoning. Mathematics studies the patterns to be found in patterns.

It shares many characteristics with music in this aspect of its nature, an observation made often by many writers and mathematicians. In a paper on Newton, the mathematician James Joseph Sylvester (himself a talented amateur musician) wrote

> May not Music be described as the Mathematics of sense, Mathematics as the Music of reason? Thus the musician feels Mathematics, the mathematician thinks Music—Music the dream, Mathematics the working life—each to receive its consummation from the other when the human intelligence, elevated to its perfect type, shall shine forth glorified in some future Mozart-Dirichlet or Beethoven-Gauss.[3]

[3] Quoted in Edward Rothstein, *Emblems of Mind: The Inner Life of Music and Mathematics*, Times Books 1995. Sylvester (1814-1897) was an English algebraist who spent his professional life in both the US and England. Mozart and Beethoven need no introduction; Dirichlet (1805-1859) was a German mathematician active in the field of analysis.

Æsthetics and appreciation of the arts involves feeling and responding to the patterns in nature and in works of art. Artistic creation is a matter of creating or reproducing or elaborating patterns. Empirical science is the discovery, description and analysis of patterns in nature. The formal sciences (logic and mathematics) study, describe, and create patterns of patterns. The beauty of structure and pattern is as central to the study of logic and mathematics as it is to literature, music or painting. Some recognition of this centrality is the main thing I hope students will develop through this course.

1.2 Introduction to Logic

1.2.1 Some history

A *very* brief, even superficial, history of logic may help put the content of this course into some context.[4]

The western scholarly study of logic goes back to Aristotle, who listed the variants of the syllogism. The law of non-contradiction and the law of the excluded middle are also credited to Aristotle. Most European students of logic followed the Aristotelian tradition until the mid nineteenth century.

However, another classical Greek school of logic, which may be loosely identified with the Stoics, went beyond the syllogistic tradition, and essentially had an understanding of propositional logic, in almost the modern sense. They formulated five basic "inference schemata":

- If the first, then the second; but the first; therefore the second.
- If the first, then the second; but not the second; therefore, not the first.
- Not both the first and the second; but the first; therefore, not the second.
- Either the first or the second [and not both]; but the first; therefore, not the second.
- Either the first or the second; but not the second; therefore the first.

We'll see how these fit into our presentation of logic at the end of Chapter 2.[5]

An early precursor to the mathematical tradition in logic was Leibniz, who envisioned a calculus of logic, a set of rules which would completely automate the reasoning process, so that disagreements might be settled by simple calculation. He even imagined one might build a machine to do these calculations mechanically. Needless to say, this idea (dare I call it a dream?) of Leibniz's was never realized in his lifetime.[6]

But in the nineteenth century, an algebraic approach to propositional logic was successfully designed by George Boole (we shall study Boolean algebras in the last section of the course). This

Curiously, given the nature of this quotation, he was Felix Mendelssohn's brother-in-law. Gauss (1777-1855) was one of the greatest mathematicians ever, who contributed significantly to just about every aspect of mathematics, as well as to physics (he was particularly famous during his lifetime for his contributions to the understanding of electricity and magnetism, and for his part in developing the telegraph).

[4] A slightly longer account, from the *Oxford Companion to Philosophy*, may be found on my webpage.

[5] But for now notice that in the 4th of these schemata the parenthetical phrase "and not both" renders the scheme rather redundant. The Stoics would not have had that phrase, and in effect, this really just says that (unlike modern logicians) they took "or" to be the "exclusive or", which we'll define very soon.

[6] Leibniz did get quite close to the idea behind the system Boole developed in one remark he made, using numbers to represent "thoughts" or properties: "if the term for an 'animate being' should be imagined as expressed by the number 2, and the term for 'rational' by the number 3, then the term for 'man' will be expressed by the number 2×3, that is 6". [Quoted in Kitty Ferguson, *The Music of Pythagoras*, Walker Books 2008.]

could be marked as the first step in modern logic. What was missing, however, was a mathematical account of predicates, and what we call quantifiers (Chapter 5 of this text). That was managed by Gottlob Frege later in the century, in a remarkable text called *Begriffsschrift*, whose unconventional notation makes this text an effort to penetrate for most readers. His ideas were quickly picked up by the mathematical-logical community, however, and with a notation very like the one we use in this text, became the basis for twentieth century logic. Very soon after, Peano gave an axiomatization of the theory of natural numbers, and although there are technical reasons why a complete mechanization of the rules of predicate logic cannot exist, one could claim the essence of Leibniz's dream was realized. (One of those technical reasons is Gödel's incompleteness theorems, which we shall study at the end of the course.)

In the late nineteenth, early twentieth century, logic went hand in hand with the attempt to put the foundations of mathematics on a sure foundation. A number of paradoxes were becoming apparent in mathematics, especially with the study of the infinite, and a need for a firm philosophical basis for mathematics and logic was thought to be necessary. Set theory was a central tool in this attempt, but a series of paradoxes in logic and set theory underlined just how conceptually tricky ("subtle" might be a more positive term) things were.

The paradoxes

Without getting too technical, let's consider some of the paradoxes that caused such a concern. One is very old, in fact: it is often called *The Liar* and finds its origins in classical Greece. Consider the statement: "This statement is false". If it's false, it's true, but if it's true, it's false. There are many variants of this; here is a simple one. Imagine a card with the following two statements, one on one side, the other on the other: "The statement on the other side of this card is false", "The statement on the other side of this card is true".

The work of Frege was interrupted by Bertrand Russell, who found an error in his system, which allowed sets to be formed if defined by some property. The paradox Russell found was this (if your knowledge of sets is insufficient, come back to this story after we've done chapter 6—in any case I'll give a simpler version in a moment): notice that some sets seem to contain themselves as an element, such as the set of abstract entities (it is itself an abstract entity), whereas other sets (most, in fact) do not, such as the set of words on this page, which is itself *not* a word on the page. We'll call those sets which do *not* contain themselves as an element "standard sets", and those which *do* contain themselves as an element "non-standard sets". Consider now the collection of all standard sets: is this set standard or not? If it's standard, then it's non-standard, but if it's non-standard, then it's standard.

If this is too technical, here's a non-technical variant. Consider a village, with just one (male) barber, who shaves every man in the village who does not shave himself, and no one else. Who shaves the barber? If he shaves himself, then he cannot shave himself, but if he doesn't shave himself, then he does shave himself.

Here's a numerical paradox: notice that some numbers may be described with only a few words ("one", "the hundredth prime"), and others take more words ("one million seven hundred and forty five thousand three hundred and twenty nine"). Generally (though there are exceptions), numbers that take lots of words tend to be large, and ones that can be described in fewer tend to be smaller. Here is an interesting number: the smallest number that cannot be described in less than thirteen words. What's the paradox? I just described it in twelve words.

There are oodles of other paradoxes—what they all have in common is that trying to understand them caused mathematicians and philosophers to think hard about mathematics and logic, and that resulted in a clearer understanding of what is going on in those domains. Many of these paradoxes still inspire thought and commentary (though for the practicing mathematician, they are more like pleasant entertainments these days).

One serious result of the philosophical ferment in the early decades of the twentieth century was the development of an alternate view of mathematics and logic, which goes by the general name of intuitionism. We shall consider some aspects of intuitionism later in the course, but for now, let's just say that it demanded a more constructive approach to mathematics and logic. For instance, if you asserted the existence of something, in intuitionist practice, you had to actually show the thing in question, or give a clear algorithm for its construction. Here is an example (though you may have to return to this after we study Chapter 7 to understand some of the terms): suppose you wonder if there are numbers x, y with the property that x, y are not rational (are not expressible as fractions of natural numbers), and yet x^y is rational. Here is one possible answer, one that would have been accepted by most mathematicians at the end of the nineteenth century (and would still be acceptable today by most—non-intuitionist—mathematicians): consider $\sqrt{2}^{\sqrt{2}}$. Either this is rational or it is not. If it is rational, then since $\sqrt{2}$ is not rational, you have your numbers ($x = y = \sqrt{2}$). If it isn't rational, then again you have your numbers: just take $x = \sqrt{2}^{\sqrt{2}}$, which you have supposed isn't rational, and $y = \sqrt{2}$, which isn't rational, but now $x^y = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = \sqrt{2}^{(\sqrt{2}\sqrt{2})} = \sqrt{2}^{2} = 2$ which is rational. What's wrong with this, according to the intuitionists? Simply that at the end of the argument, you still don't really know what values x, y are that have the required property. Just what is x? You don't know from this argument: it might be $\sqrt{2}$ or it might be $\sqrt{2}^{\sqrt{2}}$—the argument didn't specify which one it really was. Your proof wasn't constructive, in that it didn't put the necessary values at your fingertips.[7] During the past century, intuitionist logic has had a lot of study, and has become very important for practical reasons, for example in theoretical computer programming, where constructivity is a key ingredient, as well as for philosophical purposes.

What about logic today? The past century has been (and continues to be) a golden age in mathematical logic, with ever more impressive gains in understanding and in practical and theoretical applications of that understanding to many disciplines. There are several features of contemporary mathematical logic that distinguish it from the past practices. Perhaps the most striking is that one no longer thinks of “logic” as a single entity, but rather there are many different “logic**s**”, for many different purposes. Logic(s) is(are) studied with mathematical tools, and indeed, logics are mathematical entities in their own right. We shall see a simple example of this at the end of the course, when we consider a logic suitable for the analysis of sentence structure in linguistics.

1.2.2 Some vocabulary

Logic is the science of discursive reasoning.[8]

As a science, logic aims to discover general laws that apply to all discursive reasoning. Narrowly specific kinds of reasoning that are only relevant to some particular subject matter are the concern of the special sciences.

Discursive reasoning consists of arguments made up of statements.

A **statement** is made by a declarative sentence. A statement is either true or false (although its truth or falsity may not be known). No statement is both true and false.

An **argument** is a collection of one or more statements called **premises** and one statement called the **conclusion**. Premises are offered as grounds for the conclusion.

[7] Actually, one can give a constructive proof. For example, it's a fact that $\sqrt{5}$ and $\log_5 9$ are irrational (these facts follow from the results of Chp 8), but (using your high-school algebra!) $\sqrt{5}^{\log_5 9} = (5^{\log_5 9})^{1/2} = \sqrt{9} = 3$.

[8] This is a provisional definition—we shall improve on it as we go.

Premise-statements are grounds for a conclusion when the truth of those premise-statements gives some assurance that the conclusion-statement is true. When the premises (if true) really do give some assurance that the conclusion is true, we say that the premises support the conclusion, or that they **entail** the conclusion, or that the conclusion follows from the premises. In such cases, we say the argument is **valid**. If the conclusion does not follow from the premises, we call the argument **invalid**.

Whether a collection of statements is an argument depends on the intention of the person who makes the statements. It is an argument if she intends some of the statements as grounds for a statement that she offers as a conclusion. If the premises do not support the conclusion, it is still an argument, but it is a bad (invalid) argument. If the person never intended the premises to support or entail the conclusion, it is not an argument.

In logic, we do not use "valid" to describe a statement. *Statements* are true or false; *arguments* are valid or invalid.

Using this vocabulary, we refine the definition of "logic" given above. **Logic** is the science that studies the general principles or laws of valid arguments.

Deductive logic is the science of **deductive arguments**. In a good deductive argument (a valid deduction), the conclusion *cannot be false* if the premise(s) are true. The rules for deductive reasoning guarantee that one cannot get a false conclusion from true premises.

Inductive arguments offer less assurance than deductive arguments. In a good inductive argument (a valid induction), true premises assure us only that the conclusion is *probably true*. A valid inductive argument makes it *rational to believe* that its conclusion is true, while allowing that it might turn out to be false.[9]

Most of this course (and most of mathematics and logic) concerns deductive arguments. Inductive argument is touched on in the section on the mathematics of probability and statistics.

1.2.3 Beginning steps in deductive logic

What kinds of arguments *guarantee* that their conclusions cannot be false when their premise(s) are true?

An obvious example of such an argument is an argument based on definitions. For example, if we define "bachelor" as an adult unmarried human male, we could argue:

John is a bachelor.	(Premise)
(Therefore) John is unmarried.	(Conclusion)

The conclusion cannot be false if the premise is true, because "bachelor" means (among other things) "unmarried".

Another obvious example is the classic:

All men are mortal.	(Premise)
Socrates is a man.	(Premise)
(Therefore) Socrates is mortal.	(Conclusion)

This argument is an example of *predicate logic*.

Before looking at predicate logic, we study *propositional logic*. Propositional logic studies arguments whose conclusions depend on the way compound statements are composed of simple statements and special "connectives". The compound statement "Ivanhoe is safe and Rebecca is

[9] These definitions of "deductive" and "inductive" are (demonstrably) better than most definitions found in dictionaries—even *good* dictionaries. Dictionaries are not the ultimate arbiters of meaning.

relieved" consists of two *simple* statements ("Ivanhoe is safe", "Rebecca is relieved") linked with the *conjunction connective* "and". "Nero is not pleased" is the negation of the simple statement "Nero is pleased" that results from adding the *negation ("not") connective.*

Here is an example of a propositional logic argument. From the (compound) statement "Ivanhoe is safe and Rebecca is relieved" we can infer "Ivanhoe is safe". That is, in the argument:

Ivanhoe is safe and Rebecca is relieved. (Premise)
(So) Ivanhoe is safe. (Conclusion)

when the conjunction "Ivanhoe is safe and Rebecca is relieved" is true, "Ivanhoe is safe" *cannot* be false.

Notice that the validity of the Ivanhoe argument does not depend on the fact that it's about Ivanhoe and Rebecca. Look at the argument:

Mickey Mouse is safe and Minnie is relieved (Premise)
Therefore Mickey Mouse is safe. (Conclusion)

This is just as good (valid) as the Ivanhoe argument.

Frodo is an airhead and Arwen is neat. (Premise)
Frodo is an airhead. (Conclusion)

Here again the conclusion cannot be false when the premise is true. It's just as good as the Ivanhoe and Mickey arguments.

Dusty is silly but he's beautiful. (Premise)
Dusty is silly. (Conclusion)

This is an argument that behaves exactly like the others even though it uses "but" instead of "and" as the connective in the premise. A statement that results from linking two sentences with "but" works the same way as one that uses "and". Both kinds of statements are *conjunctions.* Another word that has the same logical use (*i.e.*, meaning) as "and" and "but" is "while" (as in "I'll wait in the car while you go in". This permits the valid deductive inference "I'll wait in the car"). Others are "whereas" and "as" and "at the same time as" and so on. These are all instances of the same logical connective, *conjunction*: they may have slightly different meanings in ordinary English, but they all have the same property that the truth of the compound sentence requires the truth of the first (indeed, both) constituent parts. This property characterizes conjunction.

As in any science, these observations enable us to *propose a conjecture.* We propose as a *law of deductive logic* (a principle of valid deductive reasoning) that, **given any conjunction as a premise, we may validly infer the first conjunct**. This means that whenever any conjunction is true, its first conjunct cannot be false. We can restate the conjecture as a proposal for a **valid argument form**. We propose that any argument of the form:

Statement 1 and/but/while/whereas ... Statement 2 (Premise)
Statement 1 (Conclusion)

is a valid argument.

You should convince yourself that a similar principle allows the conclusion of Statement 2 from the same premise: check each example to see how "obvious" this is.

Statement 1 and/but/while/whereas ... Statement 2 (Premise)
Statement 2 (Conclusion)

Another example of a valid deductive inference involves a **conditional**, as in the inference:

If you passed logic then I am delighted.	(Premise)
You passed logic.	(Premise)
(Therefore) I am delighted.	(Conclusion)

If the two premises are true, the conclusion cannot be false.

Other English words might serve the same logical purpose as the "if ... then ..." construction; other languages have other words to do the same job. The validity of the argument has nothing to do with your success in the course or with my happiness. We conjecture that this is another valid argument form, expressed as:

If Statement 1 then Statement 2	(Premise)
Statement 1	(Premise)
Statement 2	(Conclusion)

This process of generalization leads us to propose laws of valid arguments that do not depend on whether the premises are actually true. **The validity of an argument only depends on whether the conclusion follows from the premises.**

A deductive argument aims to provide true premises and it aims to provide assurance that the conclusion cannot be false when the premises are true. But a deductive argument might have true premises and a true conclusion even if the premises do *not* guarantee the truth of the conclusion. In such a case, the argument is (deductively) **invalid**. An argument can be valid even if the conclusion and premises are not true.

The proposed rules above did not specify anything about what Statement 1 or Statement 2 were about. The *content* of the statements was not relevant to the validity of arguments. The rules also were quite general about what particular words were used for the connective. In the first conjecture, *any* word in any language was acceptable, as long as it *did the same logical work as* the conjunction-connective "and". I used "and/but/while/whereas..." to indicate that it doesn't matter which of these words was used for the conjunction connective. Similarly, in the second conjecture, other verbal expressions might be used for "if ... then ...", as long as they do the same logical work.

The conjectured rules describe only the **forms** of valid arguments (which is why I said they were "proposals for valid argument **forms**"). We refine our definition of deductive validity[10], to read:

> **A valid deductive argument is an argument whose *form* is such that it is *impossible* to construct an argument *of that form* that has true premises and a false conclusion.**

From this definition, it follows that deductive invalidity can be defined as:

> A deductive argument is **invalid** when **it has a form such that one *could* construct another argument *of the same form* whose premises were true and whose conclusion was false.**

We also improve our definition of "deductive logic" to:

> **Deductive logic is the science of the rules of truth-preserving transformations on statements.**

[10] **Record this definition in your soul.** It is *central* to understanding logic.

All that matters for logic is the relation between the truth of our premise-statements and the truth or falsity of our conclusion-statements. Every logic we shall consider will have this property.

A **deductive argument** can be:

1. *Valid*, with *true* premises and a *true* conclusion;
2. *Invalid*, with *true* premises and a *true* conclusion;
3. *Invalid*, with *true* premises and a *false* conclusion;
4. *Valid*, with *false* premises and a *true* conclusion;
5. *Invalid*, with *false* premises and a *true* conclusion;
6. *Valid*, with *false* premises and a *false* conclusion;
7. *Invalid*, with *false* premises and a *false* conclusion.

The one thing it *cannot* be (by the definition of "valid") is "*valid*, with *true* premises and a *false* conclusion". Notice that you can have an *invalid* argument whose premises and conclusion are all *true*, and *valid* arguments whose conclusions are *false*.

One more definition may be useful in light of the above. We call an argument (**not** a statement or belief) *sound* when it satisfies the definition:

A sound deductive argument is a valid argument whose premise(s) are true.

From the definitions, what can you say about the conclusion of a sound deductive argument?

1.2.4 Exercise on arguments

For each of the following informal arguments[11], identify the premises and the conclusion of the argument made. Write these in "standard form", meaning list the premises first, and the conclusion last, each statement on a separate line. Some statements will be neither (they will be intermediate parts of the argument from the premises to the conclusion); you should not include those in your answers.

1. It is right that men should value the soul rather than the body; for perfection of soul corrects the inferiority of the body, but physical strength without intelligence does nothing to improve the mind. (Democritus)
2. There cannot be any emptiness; for what is empty is nothing, and what is nothing cannot be. (Melissus)
3. About the gods, I am not able to know whether they exist or do not exist, nor what they are like in form; for the factors preventing my knowledge are many: the obscurity of the subject, and the shortness of human life. (Protagoras)
4. In the beginning man was born from creatures of a different kind; because other creatures are soon self-supporting, but man alone needs prolonged nursing. For this reason he would not have survived if this had been his original form. (Anaximenes)

[11] Taken from *The Logic Book* by Bergmann, Moor, and Nelson.

5. Let us reflect in another way, and we shall see that there is great reason to hope that death is good; for one of two things—either death is a state of nothingness and utter unconsciousness, or as men say, there is a change and migration of the soul from this world to another. Now if you suppose that there is no consciousness, but a sleep like the sleep of him who is undisturbed even by dreams, death will be an unspeakable gain ... for eternity is then only a single night. But if death is the journey to another place, and there, as men say, all the dead abide, what good, O my friends and judges, can be greater than this? ... Above all, I shall then be able to continue my search into true and false knowledge; as in this world, so also in the next; and I shall find out who is wise, and who pretends to be wise, and who is not. (Socrates)

1.3 Truth-Functional Connectives

Expressions used to link sentences to create a new *compound* sentence are called "**connectives**". "Not" is a connective, even though it is used with a single sentence rather than connecting two sentences.

Connectives actually link or modify *statements*, rather than *sentences.*[12] Many sentences, in many languages make the *same* statement. All that matters for logic is whether the conclusion of a deductive argument can be false when the premise(s) are true. We don't care about the particular words or language in which the premise-statement(s) and conclusion-statement are expressed. We don't care about the *sentences* used to make the *statements.*

The connectives themselves are not language-specific. The conjunction-connective can be expressed many ways, even in English (as we saw). "Et" and "und" and "y" and a whole bunch of words in other languages do the same logical job of conjoining two statements. Similarly "not" has equivalents in English ("He is not going" is equivalent to "It is not the case that he is going" and so on) and in other languages. We treat all connectives that do the same logical job as the same.

Some connectives are unimportant to logic.[13] We shall only investigate "truth-functional connectives", *viz* connectives with the following property.

By "**truth-functional connective**" we mean a connective which links statements or modifies a statement in such a way that the truth or falsity of the resulting compound statement (the original statement(s) plus the connective) is a function of (*i.e.*, depends only on) the truth or falsity of the (original) component statement(s).

A connective that is *not* truth-functional is the phrase "I believe that...". You can stick that phrase on the front of a sentence (*e.g.*, "Arwen is somewhat neat".) and get a new sentence ("I believe that Arwen is somewhat neat"). But the connective is *not truth-functional.* The truth or falsity of the new statement (made by the sentence "I believe that Arwen is somewhat neat") cannot be known just by knowing whether "Arwen is somewhat neat" is true or false. Other examples might include "... because ..." (for instance "John skipped class because Susan loves Jo"); the

[12] Although one should keep this distinction clear, there may be times when any of us might use "sentence" when really meaning "statement". You should be able to decide, based on the context, which is intended: is it the actual words used (the sentence), or is it their essence (the statement) that is at issue?

[13] Well, not exactly—what I really mean here is "unimportant to the classical logic we shall study". Many connectives that fall outside our survey are considered by other logics, for example modal logic, which studies non-truth-functional operators like "necessarily" and "possibly".

truth (or otherwise) of such a statement does not depend on the truth of the individual components. It is not truth-functional.[14]

On the other hand, the "not" connective *is* truth-functional. If it is true that Arwen is somewhat neat, then "Arwen is not somewhat neat" must be false.

Propositional logic is the logic of the truth-functional connectives. The laws of valid deductive reasoning in propositional logic are based on the meanings (rules for the correct use) of the truth-functional connectives, and on nothing else.

1.3.1 The symbolism of propositional logic

The formal sciences (logic and mathematics) develop their own languages. These artificial languages[15] are more precise and clear than natural languages. More precise and clearer language leads to more precise and clearer thought and concepts. Learning mathematics and modern logic involves learning its special language—the symbolism.

The main advantages of symbolic logic (using an artificial language of symbols) are: (1) we avoid writing a lot of stuff that doesn't matter for logic; (2) we emphasize the universality of logic; and (3) it makes it easier to recognize the *form* of a compound statement or of an argument.

The English sentence "Frodo is an airhead" makes a statement. Other sentences (in English or in other languages) make the *same* statement. We can refer to the statement they all make as "the airhead statement" or "what you said about Frodo". In propositional logic we use symbols, usually letters, to stand for particular simple statements. For example, the letter "A" can represent the statement made by the sentence "Frodo is an airhead". "B" could be symbolic shorthand for the statement made by "Arwen is somewhat neat".

Then we could use "A and B" as shorthand for the statement made by the compound sentence "Frodo is an airhead and Arwen is somewhat neat". "Not B" could stand for "It is not the case that Arwen is somewhat neat".

Often we refer to *any* statement or *a* statement without having any particular statement in mind. For example, we might want to say, "The negation of a true statement is false". When mathematicians want to say something about particular numbers they use symbols (called constants) like 12 or 12364. They use variable-symbols like x or y to stand for some number, or an unknown number, or any number. In propositional logic, we use letters like p and q to represent some (unspecified) statement, or any statement. We could symbolize a conjunction of two (unspecified) statements as "p and q". That way, we can talk about a conjunction without worrying about what statements are conjoined. We can discuss conjunctions in general, or the form of a conjunction. Such letters act as *statement variables.*

We call the truth or falsity of a statement its *truth-value.* Every statement can have one of two possible truth-values (true or false) and every meaningful statement has one of those values (even when we don't know what value it has).[16]

We define the truth-functional connectives in terms of the truth-value of the statement that results from using that connective with one or more statements.

[14] We shall see a variant of "because", namely "material implication", which is truth functional. The distinction comes from the fact that with material implication we drop all suggestion of causality, which is what stops "because" from being truth-functional.

[15] Called "artificial" to distinguish them from "natural" languages like English, French, *etc.*

[16] There are alternate logics which do in fact allow more truth values, but we shall not study them in this course.

1.3.2 The connectives

You must *understand* and memorize the four fundamental truth tables (for $\neg$, $\wedge$, $\vee$, and $\longrightarrow$) of this section. They form the basis of our approach to logic, and they define how these four basic logical operators behave.

Negation

The simplest truth-functional connective is the negation connective. We use the symbol "$\neg$" to stand for the word "not" and its logical equivalents. So we symbolize "Frodo is not an airhead" by $\neg A$.

Using our statement-variable symbols, we can define the negation operator as: "$\neg p$ is false when p is true, and $\neg p$ is true when p is false". The variable p tells us that this applies to any statement that you might substitute for p. So it tells us that $\neg A$ is false if A is true, and $\neg A$ is true if A is false, for any constant (representing a particular statement) A. It also tells us that $\neg B$ is false if B is true, and so on.

We **define** the $\neg$ connective with a **truth table**. The truth table for $\neg$ is:

p	$\neg p$
$\top$	$\bot$
$\bot$	$\top$

In a truth table, "$\top$" indicates the case where the statement-form p is replaced by a statement that is guaranteed to have the truth-value "true," (for instance the statement $1 = 1$), and "$\bot$" indicates the case where the statement-form p is replaced by a statement that is guaranteed to have the truth-value "false" (for instance the statement $1 = 0$). In other words, $\top$ is a generic true statement, and $\bot$ is a generic false statement. The truth table tells you exactly what the "$\neg$" connective *does* to the truth-value of any compound statement. **The whole logical meaning of "$\neg$" consists of what it does to the truth-value of a statement.** It is *truth-functional* because *it operates on a truth-value and produces a new truth-value according to a rule.* The rule is that it produces a true statement from a false one, and *vice versa.*

> A compound statement whose main connective[17] is the negation connective is a *negation.* The negation $\neg p$ is false when p is true and true when p is false.

Negation is a **unary** connective. It works with only *one* statement (which could be a compound statement). The following connectives are all **binary** connectives. Each works with two statements (which could be compound statements).

Conjunction

A conjunction is a compound statement of the form "p and q". Suppose you replace p with a true statement and replace q with a false statement. Is the conjunction of the true statement with a false statement true? A false conjunct makes the whole conjunction false. If *both* conjuncts are false, the conjunction is false. We would only say that a conjunction is true if **both** conjuncts were true.

> A compound statement that results from linking two (simple or compound) statements p and q with the conjunction connective $\wedge$ is a **conjunction**, and p and q are its

[17] The notion of a "main connective" will be developed in section 1.3.3, on formation rules.

conjuncts. A conjunction is true if and only if both conjuncts are true. It is false if and only if at least one conjunct is false.

We symbolize the conjunction connective with “$\wedge$”.[18] The truth table defining $\wedge$ is:

p	q	$p \wedge q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$
$\bot$	$\bot$	$\bot$

Since “$\wedge$” connects two statements, we have to define it *for all possible combinations of truth-values* of p and q. Start with the rightmost of the simple statement forms (q, in this case) and go down the column, alternating $\top$ and $\bot$. Then go down the column of the next simple form to the left (p), alternating *pairs* of $\top$s and $\bot$s. If there were three simple forms, the next column to the left would consist of four $\top$s followed by four $\bot$s, and so on. This technique ensures that every combination is included.

The English word “but” has exactly the same truth-functional meaning as “and”. The two words are *truth-functionally equivalent*: if we construct a truth table for “and” and for “but”, we would end up with the same table. That is, “Frodo is an airhead but Arwen is somewhat neat” is true or false in exactly the same circumstances as “Frodo is an airhead and Arwen is somewhat neat”. The difference in meaning is not truth-functional, but reflects attitudes propositional logic does not attempt to capture. Symbolize “Frodo is an airhead but Arwen is somewhat neat” as $A \wedge B$, just as you would “Frodo is an airhead and Arwen is somewhat neat”.

Disjunction

Another common English connective is “or”. We symbolize this as “$\vee$”.[19] It is truth-functional, but defining it presents a slight problem, since there are at least two meanings (one more common than the other) of the word “or”.

Consider “Either you do all the assignments or you fail the course” ($D \vee F$). When is that statement false? Suppose you do all the assignments (D is true) and pass (don’t fail) the course (F is false). You’d say the statement was true. Suppose you don’t do all the assignments and you fail the course (D is false, but F is true). Still true. If you don’t do all the assignments and don’t fail the course (D is false and F is false), the statement is false. But what if you do all the assignments (D is true) and you fail anyway (F is also true)? Did the statement imply that both these eventualities could not occur? The answer is “yes and no, depending on what ‘or’ means”.

The English word “or” is *ambiguous*: it has two (at least) possible meanings: **inclusive-or** and **exclusive-or**. “D or F” could mean “either D or F *or both*” (inclusive) or it could, more commonly, mean “either D or F *but not both*” (exclusive). When a highwayman holds up a stage and shouts “Your money or your life”, the passengers hope that he intends to take their lives *or* their money *but not both.* But when a teacher says, “Do the work or flunk the course”, he usually intends that either *or both* could happen: he is not *guaranteeing* a pass to everyone who works. (But notice something: if his statement is true, then he *is* guaranteeing a failure to anyone who doesn’t do the work. In other words, if someone does not do the work *and* does not fail the course, then his statement was false. You will see this in the following truth table.)

[18] Different authors favour different symbols; & is not uncommon, since it’s simple to find on a keyboard.

[19] This looks like the letter “v”, but in a different typeface. It comes from the Latin “vel”, meaning “or”.

Such ambiguity is intolerable in logic. We have to decide which "or" we want to symbolize by "$\vee$". Logicians and mathematicians use the **inclusive** sense of "or", where it means "either or both". The truth table is, then:

p	q	$p \vee q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\top$
$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\bot$

> A compound statement whose main connective is $\vee$ ("vel") is called a **disjunction**, and its two component statements are called its **disjuncts**. A disjunction is true if and only if either or both of its disjuncts is/are true. It is false only if both disjuncts are false.

We could use another symbol (and truth table) for the exclusive sense of "or" ("exclusive-or") but we don't need it. "And" means the same (logically) as "but", so "either A or B but not both" is just $(A \vee B) \wedge \neg(A \wedge B)$. Later we'll use truth tables to show that this expression captures the sense of exclusive-or. It's an amusing exercise (see **BAFact 5** in Chapter 9) to define inclusive-or in terms of exclusive-or (and other connectives) in a similar fashion; primarily we use inclusive-or as the basic form not only because inclusive-or has very nice properties (*e.g.* it is dual to conjunction), but also because in mathematics (and science), we usually want the inclusive-or, and only rarely seem to need exclusive-or—one way in which mathematicians differ from highwaymen. It's important to remember this, since the inclusive-or is less common in daily non-mathematical usage.

Material implication

Consider the English **implication** or *conditional*[20] "If you marry me (then) I'll do all the cooking". When the **premise** (or *antecedent*) of the implication (the part between "if" and "then") is true (in this case, you do marry me) and the **conclusion** (or *consequent*) (the part after the "then") is false (I don't do all the cooking), we would say that the conditional statement is false. In the case where you marry me and I *do* do all the cooking, we would call it true.

What if the marriage does not take place? The premise of the implication is false. Should we say that the implication is true or that it is false?

It doesn't seem to matter whether the conclusion is true or false. In a sense "all bets are off!": you didn't marry me, so any promise I made on that basis no longer holds. I won't be breaking my word, regardless of what meals I do or don't cook. Our problem does not arise out of ignorance as to whether or not I do the cooking.

We might like to say that the implication is neither true nor false when the premise is false. But propositional logic requires that *every meaningful statement must have a truth-value*—must be either true or false, and not both. If the truth-value of an "if ... then ..." statement does not depend on the truth-values of its components (the premise and the conclusion), then "if ... then ..." is not a truth-functional connective. This would be a disaster for propositional logic, where implication is centrally important.

It would appear then that the ordinary-language "if ... then ..." conditional connective is (logically) ambiguous. We shall remove this ambiguity by *defining* "if ... then ..." truth-functionally. We call the resulting connective **material implication** or **the material conditional**, symbolized

[20] "Conditional" and "hypothetical" are nouns or adjectives logicians use for "if ... then ..." sentences. Often, especially in mathematical usage, we also call them "implications", and say "... implies ...".

using an arrow ($\longrightarrow$).[21]

The definition is:

p	q	$p \longrightarrow q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$

In the first two rows, this truth table confirms our natural tendencies about how to handle implication. The last two rows reflect the fact that we have *defined* the implication to be true whenever the premise is false. Generally, those last two rows correspond to situations where normal language probably doesn't consider the "if ... then ..." situation, so our definition doesn't interfere too badly with everyday usage (but more on that below!).

You will have noticed we use terminology for implications very reminiscent of the terminology for arguments. It is important to keep the following distinction clear: an implication or conditional is a **single** statement, with two components (think of one sentence with two subordinate clauses), but an argument is a collection of **several** statements. A premise of an argument is a complete statement, whereas the premise of an implication is only a sub-statement.

However, our terminology has the advantage of reflecting the similarity between a conditional statement and an argument. In a sense, a conditional statement collects the separate statements of an argument into a single statement with the same structure. Remember that an argument is valid if it is impossible (from the form of the argument) for a conclusion to be false if the premises are true. False premises do not make an argument invalid. This is similar to the structure of $\longrightarrow$ as given in the table above. There is a technical way to make this idea precise; you will find that in Remark 1.3.14 later.

You *must* remember the behaviour of "if ... then ..."; it is central to propositional logic, (just as is the related notion of a valid argument).

> The compound statement that results from linking two simple or compound statements with the "if ... then ..." connective $\longrightarrow$ is a **material implication** or *conditional.* The component statement that states the condition (in the "if ..." clause) is put before the $\longrightarrow$ and is called the **premise** or *antecedent.* The other component statement is put after the $\longrightarrow$ and is called the **conclusion** or *consequent.* The material implication is false if and only if the premise is true and the conclusion is false. It is therefore true in all other cases.

Logicians call the conclusion of a material implication the **necessary condition** for whatever the premise says, and they say that the premise is a **sufficient** condition for whatever the conclusion says. Think this statement over very carefully—it might seem counter-intuitive, but if $A \longrightarrow B$ is true, then the truth of A is sufficient to guarantee the truth of B, whereas the truth of B is necessary for A to be true (for if B were false and $A \longrightarrow B$ true, then A could not possibly be true).

Warning: In everyday English usage, there is a tendency to "misinterpret" implication, to assume that when one says "if A then B", what is meant is "A if and only if B", or in other words, that A and B are "equivalent", *i.e.* that they are either both true or both false. This is not the meaning of $A \longrightarrow B$ at all. Be very clear about this; it will probably require you to unlearn a meaning that seems very natural to you.

[21] This is a very important notion, and so it's perhaps no surprise that there are many different notations for it, including a double shafted arrow ($\Rightarrow$), a hook ($\supset$), a less-than symbol ($<$), and various other "weird arrows" ($\multimap$ is a favorite of mine), among others.

The implication $B \longrightarrow A$ is the **converse** of the implication $A \longrightarrow B$. The **contrapositive** of $A \longrightarrow B$ is $\neg B \longrightarrow \neg A$. The *converse* of a true implication could be false (or true), but the *contrapositive* of a true implication must be true. Notice that the contrapositive is just another way of saying the implication: $\neg B \longrightarrow \neg A$ has the same meaning, the same truth table, as $A \longrightarrow B$. These statements are "equivalent". (Check this for yourself!) By contrast, the converse is quite independent of the implication: $B \longrightarrow A$ and $A \longrightarrow B$ may both be true, both false, or one can be true and the other false, depending on the particulars of what A and B say. (Again, check this for yourself, with various examples for A and B.)

Remark: Is material implication paradoxical? Before leaving implication, let's consider the following seeming paradox. According to our definition of (material) implication, the following is a true statement: "If you are a purple unicorn, then you will win the lottery today". It is true simply because in fact, you are *not* a unicorn (purple or otherwise), and so whether or not you win the lottery has no bearing on the truth of this statement. Similarly, this is also a true statement: "If you are 50 years old, then $1+1=2$". This is true simply because $1+1=2$ is true (regardless of your age), and our definition of the (material) implication has the property that if the conclusion is true, the implication is true also, regardless of the truth of the premise. (Check it out: $\bot \longrightarrow q$ and $p \longrightarrow \top$ are always $\top$, regardless of what p and q are. Keep in mind *always* that the *only* way a implication is $\bot$ is $\top \longrightarrow \bot$.)

To most English-speakers, this seems ... well, silly! (though I'll say "paradoxical" to sound more impressive). Surely there ought to be some connection between the premise and the conclusion of an implication of this sort. But these examples play fast and loose with that expectation. And that's the nub of the matter: the paradox, if there is any, is merely one of *expectation.* We expect the English words "if ... then ..." to behave, in the formal setting of propositional logic, as they might in the informal setting of everyday English. *They do not.* It's as simple as that. The connectives $\neg$, $\wedge$, $\vee$, and $\longrightarrow$ are not words in everyday English, even if we pronounce them "not", "and", "or", and "if ... then ...". They are technical terms, whose meanings are strictly defined (by their truth tables), and on that there is no room for a difference of opinion. Sorry! If you think "or" should mean the exclusive variety, tough luck—that is not what $\vee$ means, and if it bugs you to call it "or", give it another name, like "vel". Similarly for $\longrightarrow$: call it "implies" or some other word you don't use very much. You cannot change the meanings of these connectives, however much you want to, without changing the very subject we are studying. If it makes you feel better, there are other logics which attempt to do just that, and you can turn to study them as a personal project *after* the semester is over! For example, "relevance logic" is a generic name for a family of logics which attempt to define an implication/conditional which requires some causal connection between antecedent and consequent. For now, however, we shall stay with classical propositional logic, with the connectives defined as above.

The biconditional—a derived connective

When an implication $p \longrightarrow q$ and its converse $q \longrightarrow p$ are both true, we say "p if and only if q". This kind of compound statement is important in mathematics. To represent this, we define another connective, the biconditional, written $\leftrightarrow$ or "$\equiv$" or even "iff", as in $p \leftrightarrow q$, in terms of the implication and its converse:

$$p \leftrightarrow q \ := \ (p \longrightarrow q) \wedge (q \longrightarrow p)$$

Whenever one sees $\equiv$ or $\leftrightarrow$, one should replace it as defined above.

In this definition, we had to use parentheses so that it was clear that the conjunction had implications as its two conjuncts. The ":=" symbol indicates that the expression on the right *defines* (*i.e.*, is the **definiens** for) the expression on the left (the **definiendum**). In logic, whenever we see the definiens, we can replace it with the definiendum, and *vice versa*. Since the two expressions have the same meaning, it is impossible for one statement to be false when the other is true. Replacing the definiens with the definiendum can never lead to an invalid (false conclusion from true premises) logical step. The truth table for $(p \longrightarrow q) \wedge (q \longrightarrow p)$ (exercise: construct it!—or look at section 1.3.9.) shows that this whole expression is true whenever p and q have the same truth-value (either both true or both false) and is false when p and q have different truth-values (one true and the other false).

1.3.3 Formation rules

The rules of the syntax of our symbolic language are the **formation rules**.

WFF is an abbreviation for *Well-Formed Formula*, often simply called a "formula". A simple statement (*e.g.*, A, or B, or Dusty is a cat) is a WFF. As we've defined the "$\neg$" connective, $\neg A$ is a WFF. But $A\neg$ is meaningless. It is not "well-formed". It is not a WFF. Neither is $A\neg C$. Neither is $\wedge A$, but $B \wedge A$ is a WFF. So is $A \wedge A$ (nothing in the definition of "$\wedge$" says that p and q have to be different statements). We shall adopt the convention that a statement variable is a WFF (since it obviously becomes a WFF whenever it is replaced by a simple statement or by a WFF).

The components of a compound statement may be compound. As we saw when we defined the biconditional, the conjuncts of a conjunction can be implications. But we may need parentheses to clarify the logical meaning of the resulting compound.

Is $\neg A \wedge B$ the conjunction of a negation and a simple statement or is it the negation of a conjunction? Is "$A \wedge B \vee C$" the conjunction of a simple statement A with a disjunction $B \vee C$ or the disjunction of a conjunction $A \wedge B$ and a simple statement C? We add **parentheses** and a **precedence rule** to the syntax of our symbolic language.

> **Parentheses Rule**: When a binary compound statement is used as a component of a compound statement, it must be surrounded by parentheses.

The example of conjoining (making a conjunction of) A and $B \vee C$ requires that we put parentheses around $B \vee C$ before using it as a conjunct. We get $A \wedge (B \vee C)$. Now it is clear that this is a conjunction and that its second conjunct is a disjunction.

> **Precedence rule**: the $\neg$ connective takes precedence over any other connective.

This means that $\neg A \wedge B$ must be interpreted as a conjunction whose first conjunct is a negation. It works like $(\neg A) \wedge B$. If we want to negate a conjunction, we must use parentheses to get $\neg(A \wedge B)$.

> **WFF rules**: A symbolic expression is a WFF if and only if:
>
> 1. it is a simple statement or a statement variable. Thus, A is a WFF. p is a WFF, $\top$ is a WFF, as is $\bot$. The statement represented by the sentence "It is raining" is a WFF. And so on. We call such WFFs "atomic formulas", "atomic propositions", or simply "atoms".
> 2. it is any WFF (in parentheses if it is a binary compound WFF) preceded by $\neg$. Any WFF formed according to rule (2) is a **negation**. Its **main connective** is the negation connective $\neg$.
> 3. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\wedge$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (3) is a **conjunction** and its **main connective** is the

conjunction connective $\wedge$. The two WFFs linked by the connective are called **conjuncts**.

4. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\vee$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (4) is a **disjunction** and its **main connective** is the disjunction connective $\vee$. The two WFFs linked by the connective are called **disjuncts**.
5. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\longrightarrow$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (5) is an **implication** or conditional, and its **main connective** is the conditional connective $\longrightarrow$. The WFF before the connective is called the **premise** of the implication or the **sufficient condition**. The WFF after the connective is the **conclusion** of the implication, also called the **necessary condition**.

By rule (2), $\neg A$ is a WFF, because A is a WFF. Since $\neg B$ is a WFF by rule (2), so is $\neg\neg B$. So is $\neg\neg\neg\neg C$. So is $\neg(A \longrightarrow B)$ (since $A \longrightarrow B$ is a WFF according to rule (5)). Rule (3) says that $A \wedge B$ is a WFF. So is $A \wedge \neg B$. So is $\neg\neg B \wedge \neg(A \longrightarrow B)$. And so on. Rule (4) permits $A \vee B$ as a WFF. So is $A \vee \neg B$. So is $\neg\neg B \vee \neg(A \longrightarrow B)$. And so on. By Rule (5), $A \longrightarrow B$ is a WFF. So is $A \longrightarrow \neg B$. So is $\neg\neg B \longrightarrow \neg(A \longrightarrow B)$. And so on.

There is a second precedence rule we shall frequently use:

> **Second Precedence Rule**: $\longrightarrow$ binds less strongly than $\wedge$ and $\vee$ (which bind equally strongly as each other).

So, we can unambiguously interpret $A \longrightarrow B \wedge C$ to mean $A \longrightarrow (B \wedge C)$, and not $(A \longrightarrow B) \wedge C$, which must be bracketed as shown.

Remark concerning parentheses. Although one needs to be careful about parentheses, one has also to keep a sense of proportion about them. Their purpose is simply to avoid ambiguity in logical expressions, nothing more. We shall feel free to use different styles of bracketing when it helps make the expressions clearer. For example, contrast the following two expressions (intended to be regarded as identical):

$$(((A \wedge p) \longrightarrow (q \vee B)) \longrightarrow ((\neg A \wedge q) \vee \neg(B \longrightarrow \neg p))) \longrightarrow r$$

$$([(A \wedge p) \longrightarrow (q \vee B)] \longrightarrow [(\neg A \wedge q) \vee \neg(B \longrightarrow \neg p)]) \longrightarrow r$$

The use of brackets [ ] instead of parentheses ( ) helps keep track of what groups go together.

In fact, parentheses may be totally avoided if we adopt a different technical presentation (known as "reverse Polish notation"), where the connective *follows* the two formulas it joins, as in $AB\wedge$. With this presentation, the formula above would look like this:

$$Ap \wedge qB\vee \longrightarrow A\neg q \wedge Bp\neg \longrightarrow \neg\vee \longrightarrow r \longrightarrow$$

but for most folks, the current ("in-fix", *i.e.* connectives in between their arguments) presentation is clearer, even if it means fiddling with parentheses.

If we ignore parentheses for the moment, we can summarize the WFF rules in the following compact form:

$$P := A \mid \neg P \mid P \wedge P \mid P \vee P \mid P \longrightarrow P$$

meaning a proposition (WFF) is either an atomic proposition (a constant or a variable), the negation of a proposition, the conjunction of two propositions, the disjunction of two propositions, or the implication of two propositions.

1.3.4 Parsing complex symbolic expressions

"Parsing" is the process of analyzing an expression to discover (1) if it is well-formed (syntactically correct, grammatical) and (2) supposing it is well-formed, what kind of statement it symbolizes (is it a negation, a conjunction, a disjunction, or an implication—in other words, what is its main connective?).

To parse a compound expression, start by counting the left parentheses and the right parentheses. If there is not the same number of each, and if they are not balanced, the expression is not a WFF.

You probably remember "balanced parentheses" from high-school maths, but if not, here is a simple test you can use to check for them. Start at the left end of the expression, and count parentheses, beginning with 1 and adding "+1" for each "(" and "−1" for each ")". Here's an example, with the counting numbers indicated as superscripts:

$$({}^1({}^2({}^3A \wedge p)^2 \longrightarrow ({}^3q \vee B)^2)^1 \longrightarrow ({}^2({}^3\neg A \wedge q)^2 \vee \neg({}^3B \longrightarrow \neg p)^2)^1)^0 \longrightarrow r$$

You should never get a negative number, and you should end up with 0, if the parentheses are balanced. Note that each matched pair of parentheses carry indices n, $n-1$.

If the expression passes the parentheses-check, proceed as follows:

Every simple statement symbol or statement variable in the expression is a WFF, by Rule (1). Underline or highlight the simple statement symbols and statement variables. Then follow the algorithm (where "highlight" means "highlight or underline"):

1. If the only thing that is not highlighted is a single connective, go to step 5.
2. If everything enclosed by a pair of parentheses is highlighted, highlight the parentheses and go to step 1. Otherwise, go to step 3.
3. If any highlighted component is immediately preceded by a ¬, extend the highlight to include the ¬ and go to step 1. Otherwise, go to step 4.
4. If there are two highlighted components separated by a binary connective (either $\wedge$, $\vee$, or $\longrightarrow$), extend the highlighting on the components to include the connective and go to step 1. Otherwise, go to step 2.
5. The connective that is not highlighted is the main connective. If the main connective is ¬, the expression is a negation; if $\wedge$, it is a conjunction; if $\vee$, the expression is a disjunction; if $\longrightarrow$, it is an implication.

Suppose we see an expression like

$$(D \vee \neg H) \longrightarrow (R \wedge (S \vee T)).$$

Here there are three left and three right parentheses and they are balanced, so this expression passes the parentheses-test. Highlighting the simple statement symbols, we get

$$(\underline{D} \vee \neg\underline{H}) \longrightarrow (\underline{R} \wedge (\underline{S} \vee \underline{T}))$$

There are several connectives that are not highlighted, so we go on to step 2. No highlighted thing has parentheses on both sides of it, so go to step 3. In step 3 of the algorithm, we get

$$(\underline{D} \vee \underline{\neg H}) \longrightarrow (\underline{R} \wedge (\underline{S} \vee \underline{T})).$$

We go to step 1. Several connectives are not highlighted. Nothing happens in steps 2 or 3. Step 4 gives

$$(\underline{D \vee \neg H}) \to (\underline{R} \wedge (\underline{S \vee T}))$$

(Notice that we did not highlight the $\wedge$ connective between R and $(S \vee T)$ because there was more than just a connective between the highlighted R and the highlighted $S \vee T$. There was also a left-parenthesis.) Back to step 1. There are still connectives that are not highlighted, so step 2 gives

$$\underline{(D \vee \neg H)} \to (\underline{R} \wedge \underline{(S \vee T)}).$$

Back to step 1. There are two connectives that are not highlighted, so we go to steps 2 and 3. There is no $\neg$ immediately preceding a highlighted component, so we go to 4. In $\underline{R} \wedge \underline{(S \vee T)}$ we have two highlighted components separated by a binary connective, so we extend the highlighting, getting

$$\underline{(D \vee \neg H)} \to (\underline{R \wedge (S \vee T)}).$$

One connective and two parentheses are not highlighted, so we go to step 1, then 2, and extend the highlighting on the right side to include the parentheses and back to 1. The highlighted expression now looks like

$$\underline{(D \vee \neg H)} \to \underline{(R \wedge (S \vee T))}.$$

The only thing that is not highlighted is the $\to$. We go to step 5. $\to$ is the main connective. The expression is an implication. Its premise is $D \vee \neg H$, which is a disjunction (parse it and see). Its conclusion is $R \wedge (S \vee T)$, a conjunction whose second conjunct is a disjunction.

OK—take a breath!! You will find with a little practice that in fact this takes about 5 seconds to do, really. Scan the displays above, noting how the underlines grow out from the simple statements to engulf more and more of the structure, by identifying the role each bit plays in the whole expression, ending up with two underlined bits joined by the main connective. In a way, every time a connective is put in parentheses it is "buried" lower in the structure, and the "topmost" connective is the main connective. It's usually pretty obvious which one that is—just match the parentheses and check that each connective joins exactly two (one for $\neg$) expressions.

1.3.5 Examples

1. Parse $\neg(C \wedge D) \vee (A \vee M)$. Two left and two right parentheses: check. Start with $\neg(\underline{C} \wedge \underline{D}) \vee (\underline{A} \vee \underline{M})$. Go to step 2, then 3: there is no highlighted component preceded by $\neg$ (there is an un-highlighted parenthesis after the $\neg$). Step 4: $\neg(\underline{C \wedge D}) \vee (\underline{A \vee M})$. Step 2 gives: $\neg\underline{(C \wedge D)} \vee \underline{(A \vee M)}$. From step 3: $\underline{\neg(C \wedge D)} \vee \underline{(A \vee M)}$. There is now only one un-marked connective. The expression is a disjunction. The left disjunct is a negation of a conjunction. The right disjunct is a disjunction.

2. Parse $\neg(\neg B \wedge \neg A)$. Following the algorithm, we get: $\neg(\neg\underline{B} \wedge \neg\underline{A})$, leading to $\neg(\underline{\neg B} \wedge \underline{\neg A})$, then $\neg(\underline{\neg B \wedge \neg A})$, and finally $\neg\underline{(\neg B \wedge \neg A)}$. The expression is a negation (of a conjunction, each of whose conjuncts is a negation).

1.3.6 Parsing and WFF exercise

1. Each of the following expressions is a WFF. Parse it and say what kind of WFF (negation, conjunction, disjunction, or implication) it is:

   (a) $\neg(A \vee \neg A)$

   (b) $A \wedge \neg(B \vee C)$

   (c) $\neg(A \to B) \to ((A \wedge \neg B) \vee \neg A)$

   (d) $\neg(A \wedge B) \vee \neg(\neg C \to \neg D)$

2. Make the longest WFF you can using only the symbols given (plus any number of parentheses you need). You don't have to use all the symbols shown, but you cannot use any symbol more often than explicitly shown. (For example, in (a), you can use two As, one B, one $\wedge$, one $\vee$, and two $\neg$s.) Check (parse) each WFF you construct and say what kind of WFF it is.

   (a) $A\ A\ B\ \wedge\ \vee \neg\neg$ (b) $A\ A\ \neg$ (c) $G\ H\ W\ N\ \neg\neg\ \vee\ \wedge\ \longrightarrow$

   (d) $D\ E\ F\ \neg\ \longrightarrow$ (e) $A\ B\ C\ \longrightarrow\longrightarrow\ \vee$ (f) Make up some of your own.

3. *Remark*: Technically, any expression containing a subexpression of the form $p \leftrightarrow q$ is not a WFF, since one must expand the subexpression according to the definition of $\leftrightarrow$. However, it is easy to show that we could add rules for $\leftrightarrow$ to the WFF formation and parsing rules, analogous to the rules for $\longrightarrow$, which would allow us to treat $\leftrightarrow$ as if it were a connective in the usual way. With such new rules, any expression would be a WFF if and only if the expression with the $\leftrightarrow$ expanded (using its definition) is a WFF, so nothing is a WFF with the new $\leftrightarrow$ rule that shouldn't be a WFF. Verify this claim. (As an exercise, this is optional, but you should understand the result.)

1.3.7 Substitution instances

> **A particular statement S is a substitution instance of a statement form F if S is the result of replacing every simple statement variable in F with a *simple or compound* statement constant.** None of the connectives in F may be altered or eliminated. If any statement variable occurs more than once in F, every occurrence must be replaced by the same statement constant in S.

We can replace a *simple* variable with a *compound* constant. $\neg((A \wedge B) \longrightarrow C)$ is a substitution instance of $\neg q$, because it results from replacing every distinct simple statement variable (*i.e.*, q) in $\neg q$ with the compound statement constant $(A \wedge B) \longrightarrow C$. The same statement $\neg((A \wedge B) \longrightarrow C)$ is also a substitution instance of $\neg(q \longrightarrow r)$, replacing the statement variable q with the constant $A \wedge B$ and the variable r with the constant C (and following the parentheses rule). So a single statement may be a substitution instance of many statement forms. It is also true that many statements may be substitution instances of the same statement form; for instance, $\neg(A \longrightarrow C)$ is another substitution instance of $\neg q$.

Note 1: The definition of "substitution instance" permits us to obtain a substitution instance from a statement form by replacing two *different* statement variables with the *same* statement constant. $A \vee A$ is a substitution instance of $p \vee q$.

Note 2: The definition does *not* permit replacing a compound statement form with a simple statement constant. $A \wedge B$ is not a substitution instance of $(p \vee q) \wedge r$, because A is not a substitution instance of $p \vee q$. The $\vee$ connective and the third distinct statement variable are lost in that substitution.

Note 3: The definition requires that two simple forms that are represented by the same variable be replaced by the same statement constants, so $A \vee \neg B$ is not a substitution instance of $p \vee \neg p$ (one p is replaced with A and the other with B).

Note 4: The definition of substitution instance is strictly *syntactic* (it depends on exactly what symbols are used and how they appear), not *semantic* (it does not in any way involve equivalence of statements). So, even though $\neg(A \vee B)$ is equivalent to $\neg A \wedge \neg B$, *i.e.* even though they always have the same truth value, the first is a substitution instance of $\neg p$, but the second is not, and the second is a substitution instance of $p \wedge q$, but the first is not. The notion of substitution instance is not a matter of truth values, but rather one of the actual symbols used.

Note 5: You should notice the role of parsing here: a statement S can only be a substitution instance of a statement form F if both S and F have the same main connective, and this must hold *recursively* as you go "deeper" into the structure of F, so for example, if both S and F are conjunctions, then the first and second conjuncts of each must also have the same main connective. This only ceases to be a condition when you reach variables in F. I won't make this statement too technical: look at the examples and you should see what is meant here. (This is really the main reason we spent time on parsing: recognising substitution instances is crucial for the next chapter, but we'll never look at parsing for its own sake again.)

Any simple statement is a substitution instance of the simple statement form p. So is any compound statement, like $(W \longrightarrow A) \longleftrightarrow ((A \wedge B) \vee W)$ (it results from replacing p with that whole long thing). And so is $\neg\top$. But A is *not* a substitution instance of $\neg p$ (we lost the negation connective). $(A \vee B) \wedge (W \longrightarrow \neg\bot)$ is a substitution instance of $p \wedge q$ because (as parsing shows) it is a conjunction, and $p \wedge q$ is the form of a conjunction. $A \wedge \neg A$ is also a substitution instance of $p \wedge q$. So is $H \wedge H$. $(W \wedge X) \longrightarrow J$ is not (it's an implication, and the form $p \wedge q$ is the form of a conjunction).

This may seem complicated and unfamiliar, but it is really quite simple, once you get the idea—it should help if you practice the exercises. Identifying substitution instances is crucial to the ability to show that an argument is valid when it is represented in the symbols of truth-functional logic.

1.3.8 Exercise on statement forms and substitution instances

For each statement form in the left-hand column, say which of the statement constants in the right-hand column are substitution instances of that form.

a. p
b. q
c. $\neg p$
d. $p \longrightarrow q$
e. $\neg p \longrightarrow q$
f. $\neg(p \longrightarrow q)$
g. $\neg(\neg p \longrightarrow q)$
h. $(p \vee q) \longrightarrow r$
i. $(p \vee p) \longrightarrow \neg r$
j. $(\neg p \vee q) \longrightarrow r$
k. $\neg(\neg p \vee q) \longrightarrow r$
l. $\neg(p \vee q) \longrightarrow r$
m. $\neg(\neg(\neg p \vee q) \longrightarrow r)$
n. $\neg((p \vee q) \longrightarrow r)$
o. $\neg(\neg(p \vee q) \longrightarrow r)$

1. A
2. $A \longrightarrow B$
3. $(A \vee B) \longrightarrow C$
4. $(\neg A \vee B) \longrightarrow C$
5. $\neg(A \vee B) \longrightarrow C$
6. $\neg(\neg A \vee B) \longrightarrow C$
7. $\neg((A \vee B) \longrightarrow C)$
8. $\neg(\neg(A \vee B) \longrightarrow C)$
9. $\neg(\neg(\neg A \vee B) \longrightarrow C)$
10. $\neg((\neg A \vee B) \longrightarrow C)$

1.3.9 Truth tables

To make a truth table of a compound statement type, we must look at the form of that type of statement and construct a truth table that applies to every statement of that form. For this reason, we only use statement variables in the WFFs for which we construct truth tables. (We allow one exception: we may usefully construct truth tables for WFFs which contain the constants $\top$ or $\bot$, and as many variables as we wish: for example, we might construct the truth table for the WFF $(p \vee \bot) \longrightarrow (\bot \wedge \neg q)$. Why not do so now as an exercise?)

Earlier we saw that we don't need an exclusive-or connective, because $(p \vee q) \wedge \neg(p \wedge q)$ did the same truth-functional job (and therefore has the same meaning) as "p exclusive-or q". This is illustrated by its truth table. (We shall explain the meaning of the various columns in a moment.)

p	q	$(p \vee q)$	$\wedge$	$\neg$	$(p \wedge q)$
$\top$	$\top$	$\top$	$\bot$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
			*		

To construct this truth table, we start by making columns for all the simple statement-forms we'll need. The truth-values for the right-most simple form q are assigned alternating $\top$s and $\bot$s. The next column to the left gets alternating pairs of $\top$s and $\bot$s. If there is a third column, it gets four $\top$s followed by four $\bot$s, and so on. Then we do the columns for the simplest compound statement-forms—the next ones that we underline or highlight when parsing. We look at the truth table for the connective in the compound form (*e.g.*, in the example above we look first at the truth table for disjunction). That truth table determines what we put under the $p \vee q$, based on the values of the components p and q. We then do the column for the next-simplest component (the simple conjunction), and then the column for its negation. Finally we do the column for the most complex compound statement-form. The column $\bot \ \top \ \top \ \bot$ under this last form (the column that contains the main connective of the whole statement) is called the "**main column**" of the truth table. We have put an asterisk under the main column to draw your attention to it. We use the truth table for the main connective ($\wedge$) to decide whether to put a $\top$ or a $\bot$ into the main column.

We defined the biconditional connective as the conjunction of a conditional and its converse: any expression of the form $p \leftrightarrow q$ is defined to mean the same as an expression of the form $(p \longrightarrow q) \wedge (q \longrightarrow p)$. The truth table for the biconditional then is:

p	q	$(p \longrightarrow q)$	$\wedge$	$(q \longrightarrow p)$
$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

To enter the values for the $q \longrightarrow p$ column, remember that an implication is only false when its premise is true and its conclusion is false. Look for cases where q has value $\top$ and p has value $\bot$ in the first two columns. We find it only on the third row, so we enter $\bot$ in the third row of the column, and put $\top$ on the other rows. The main connective is $\wedge$, which is only true when both conjuncts are true. We look for the cases where both $p \longrightarrow q$ and $q \longrightarrow p$ are true (the first and last rows) and put a $\top$ there, and put $\bot$ on the other rows. From this truth table we see that the truth table for $p \leftrightarrow q$ is:

p	q	$p \leftrightarrow q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$
$\bot$	$\bot$	$\top$

As a shortcut, we may now use this truth table whenever we meet the biconditional in a WFF, instead of expanding the biconditional *via* its definition.

Looking at these tables, we can notice some properties of the WFF forms they represent. For instance, we see that the main column of the exclusive-or truth table $(p \vee q) \wedge \neg(p \wedge q)$ is exactly the opposite of the main column for the biconditional. That means that we could get the exclusive-or result either by a WFF of the form $(p \vee q) \wedge \neg(p \wedge q)$ or by a WFF of the form $\neg(p \leftrightarrow q)$. This also tells us that we could have defined the biconditional as $\neg((p \vee q) \wedge \neg(p \wedge q))$ (the negation of exclusive-or). (In the exercises below you will see this is also equivalent to $(\neg p \wedge \neg q) \vee (p \wedge q)$; in words: "$p$ is equivalent to q iff either both p and q are false, or they are both true". The truth table makes this very clear.)

Caution: Looking at a WFF like $\neg(A \vee B)$, some students do what you would do with an arithmetic expression like $-(a+b)$, which is to "multiply through by minus one". But $\neg$ is not a minus sign, and this kind of move would be a mistake. Exercises 4, 5 below let you look at the truth tables for the forms of similar-looking expressions that have different truth tables (and so, different meanings). There is an algebra of truth values, which we shall see in Chapter 9, at the end of the semester, but although it has some similarities with ordinary high-school algebra, there are very important differences as well, and you should not confuse them.

1.3.10 Exercise on truth tables

1. Construct the truth table for $\neg((p \vee q) \wedge (q \longrightarrow \neg p))$.

2. Construct the truth table for $(p \wedge q) \vee (\neg p \wedge \neg q)$.

3. Compare the main columns of the truth tables you constructed in 1 and 2 with the main column for the biconditional truth table. What do you observe? Since a truth table defines the meaning of an expression, these three expressions have the same meaning for logic. We say that they are equivalent (see the next section).

4. Compare the truth tables for $\neg p \wedge q$, $\neg(p \wedge q)$, $\neg p \wedge \neg q$ and $\neg p \vee \neg q$. Draw conclusions about the similarities or differences between them.

5. Compare the truth tables for $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$; do the same for $p \vee (q \wedge r)$ and $(p \vee q) \wedge (p \vee r)$. You should find that each pair has the same truth table, so we have two more equivalences. The first equivalence pair looks rather like the high-school algebra distributive law: $a \times (b + c) = (a \times b) + (a \times c)$. However, under this analogy, the second equivalence would seem to say $a + (b \times c) = (a + b) \times (a + c)$, which certainly is not true of high-school algebra. (Try it with numbers: *e.g.* is $1 + (2 \times 3) = (1 + 2) \times (1 + 3)$?) As we said above, the algebra of propositions is quite different from the algebra of numbers, in spite of some similarities. Don't try to use your high-school algebra here!

1.3.11 Truth-functional equivalence

When the main columns of the truth tables for two expressions are the same, we say that the two expressions are **truth-functionally equivalent**. Using one expression rather than the other makes no difference in truth-functional logic.

Remark: In view of the nature of the truth table for $\leftrightarrow$, it should be clear that two expressions P and Q are truth-functionally equivalent (or simply "equivalent") if the truth table for $P \leftrightarrow Q$ has the property that its main column has only $\top$ as truth value. (We call such an expression a "tautology", as you will see soon.) Informally, you may remember this as "P and Q are equivalent if each implies the other".

Sometimes when you translate a statement from a natural language into the language of the symbolism, you may discover two (or more) alternative translations. Which is correct? A truth table will show whether the different-seeming translations are equivalent. If they are, the best translation is the one that best captures the "feel" of the original statement. If not, choose the translation that best captures the truth-functional meaning of the natural-language statement.

We can symbolize the same natural-language compound statement using different connectives. We could have used fewer connectives than the four (plus the defined biconditional connective) that we have. Just one carefully chosen connective can be used to do everything that we do with our five. It is more difficult to translate from natural language to the symbolism when we use fewer connectives. Our set of connectives is a compromise between simplicity (why we left out the exclusive-or connective) and ease of translation.

1.3.12 Some equivalences

Here is a list of useful equivalences; try constructing the truth tables for some of these to verify that they are indeed equivalences. Think of the "intuitive meaning" for each statement, and try to see why it must be an equivalence.

1. Commutativity:
   - (a) $(p \wedge q) \leftrightarrow (q \wedge p)$
   - (b) $(p \vee q) \leftrightarrow (q \vee p)$
2. Associativity:
   - (a) $((p \wedge q) \wedge r) \leftrightarrow (p \wedge (q \wedge r))$
   - (b) $((p \vee q) \vee r) \leftrightarrow (p \vee (q \vee r))$
3. Distributivity:
   - (a) $((p \wedge q) \vee r) \leftrightarrow ((p \vee r) \wedge (q \vee r))$
   - (b) $((p \vee q) \wedge r) \leftrightarrow ((p \wedge r) \vee (q \wedge r))$
4. De Morgan Laws:
   - (a) $\neg(p \wedge q) \leftrightarrow (\neg p \vee \neg q)$
   - (b) $\neg(p \vee q) \leftrightarrow (\neg p \wedge \neg q)$
5. Others:
   - (a) $(p \longrightarrow q) \leftrightarrow (\neg p \vee q)$
   - (b) $\neg(p \longrightarrow q) \leftrightarrow (p \wedge \neg q)$

1.3.13 Tautologies, contradictions, and contingencies

The truth table for $p \wedge \neg p$ is

p	p	$\wedge$	$\neg p$
$\top$		$\bot$	$\bot$
$\bot$		$\bot$	$\top$

which shows that a statement of the form $p \wedge \neg p$ is false whether p represents a true or a false statement.

> A *statement-form* (made of variables and connectives) is a **contradiction** if and only if its truth-value is $\bot$, no matter what the truth-values of its component statements. A *particular statement* is a contradiction (it is **contradictory**, or self-contradictory) if it is a substitution instance of a contradictory form.

So any statement that is a substitution instance of the form $p \wedge \neg p$ is a contradiction. Contradictions are **trivially false.** "Trivial" means that their truth doesn't depend on the truth of any particular statements, and therefore it doesn't depend on the way the world is.

Now look at $(p \wedge q) \longrightarrow p$.

p	q	$(p \wedge q)$	$\longrightarrow$	p
$\top$	$\top$	$\top$	$\top$	
$\top$	$\bot$	$\bot$	$\top$	
$\bot$	$\top$	$\bot$	$\top$	
$\bot$	$\bot$	$\bot$	$\top$	

Statements of the form $(p \wedge q) \longrightarrow p$ are true, no matter what truth-values p and q have.

> A *statement-form* is a **tautology** (is tautologous) if and only if its truth-value is always $\top$. By extension, a *particular statement* is a tautology if it is a substitution instance of a tautologous statement-form.

Tautologies are **trivially true**. $(p \wedge q) \longrightarrow p$ is a tautology (it is **tautologous**). A common example of a tautologous statement-form is $p \vee \neg p$, whose truth table is:

p	p	$\vee$	$\neg p$
$\top$		$\top$	$\bot$
$\bot$		$\top$	$\top$

Finally, some (indeed, most) statement-forms have neither of these properties, and take on both $\top$ and $\bot$ as truth values, depending on the truth values of the constituent parts. We call these contingencies.

> A statement-form is a **contingency** if and only if it is neither a tautology nor a contradiction. We call a particular statement **contingent** when it is not a substitution instance of any tautologous or contradictory form. Such statements are **non-trivial** (although they may not be important).

One way to show that a statement is a tautology or a contradiction is to construct a truth table of its statement form. We re-write the statement using variables, if necessary, replacing each distinct simple statement constant (other than $\top$ and $\bot$) with a distinct simple variable, and using the same simple variable for every occurrence of the same simple constant. The statement form should exactly duplicate all connectives and parentheses that are in the original compound statement.

For example, given the statement $((A \longrightarrow M) \wedge (M \longrightarrow L)) \longrightarrow (A \longrightarrow L)$, we could either treat A, M, L as variables, or we could construct the truth table for the form $((p \longrightarrow q) \wedge (q \longrightarrow r)) \longrightarrow (p \longrightarrow r)$. The form is exactly like the statement we're checking, except every A in the statement is replaced with a p, every M is replaced with a q, and every L is replaced with an r.

The statement form contains three simple variables, p, q and r. We start with a column for each of these. Beginning with the rightmost column (r), we alternate $\top$ and $\bot$ values. Moving to the next column to the left (q), we write alternating pairs of $\top$s and $\bot$s. Moving one more column to the left (p), we write alternating sets of four $\top$s and $\bot$s. This mechanical procedure ensures that our truth table will contain every possible combination of truth-values for the three variables. To the right of those three columns we make a column for each compound form that is a component of the whole compound form: this gives us columns for $p \longrightarrow q$, for $q \longrightarrow r$, and $p \longrightarrow r$, then finally for $(p \longrightarrow q) \wedge (q \longrightarrow r)$, and the main connective, *i.e.* for the whole expression. (Since this is our first example of a truth table with three variables, we have indicated the order in which we fill the columns with little numbers above them; as before, we indicate the main column with an asterisk.)

			(1)	(4)	(2)	(5)	(3)
p	q	r	$((p \longrightarrow q)$	$\wedge$	$(q \longrightarrow r))$	$\longrightarrow$	$(p \longrightarrow r)$
$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\bot$
$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$
						*	

Every row of the main column contains a $\top$. That means that an expression of the form $((p \longrightarrow q) \wedge (q \longrightarrow r)) \longrightarrow (p \longrightarrow r)$ is true no matter what the truth-values of its component simple parts (p, q and r). It is a tautologous form. Any statement that is a substitution instance of this form is a tautology. Our original statement is a substitution instance of this form, so it is a tautology.

1.3.14 Remark

Earlier we remarked on the similarity between material implications and argument-forms. For example, an argument of the form "Premise: P; Conclusion: C" is valid if and only if the statement $P \longrightarrow C$ is a tautology. If the argument has several premises: "Premise: P_1; Premise: P_2; Premise: P_3; Conclusion: C" for instance, then we can "internalize" this (represent it as a single statement) by $(P_1 \wedge P_2 \wedge P_3) \longrightarrow C$. The argument is valid if and only if the corresponding conditional sentence is a tautology. In general, the **validity** of any argument-form can be expressed as whether or not the corresponding conditional statement is tautological. It is in this sense that material implication "internalizes" valid argument.

As an example, *via* the previous truth table we have just established a form of valid argument, namely if one has premises of the form $p \longrightarrow q$ and $q \longrightarrow r$, and a conclusion of the form $p \longrightarrow r$, then the argument is valid. Every tautology corresponds to a valid form of argument in such a manner.

The method of showing that a statement is a contradiction is exactly parallel. If a truth table constructed according to the method above results in a main column that contains both $\top$s and $\bot$s, the statement is contingent.

1.3.15 Exercise on tautology, contradiction and contingency

For each of the following statements, determine (show and say) whether it is a contingency, a tautology, or a contradiction. In each case write an English interpretation of the statement.

1. $(A \longrightarrow B) \longrightarrow \neg A$
2. $(A \wedge B) \wedge \neg A$
3. $A \wedge \neg(B \vee A)$
4. $(A \longrightarrow B) \leftrightarrow (\neg B \longrightarrow \neg A)$
5. $(A \longrightarrow B) \longrightarrow (B \vee \neg A)$
6. $(A \longrightarrow B) \vee (B \longrightarrow A)$

1.4 Translation

Translating from one natural language (like English) to another (*e.g.*, French) is difficult. You have to be familiar with both languages and with all sorts of subtle nuances of expression and idiom. Artificial languages like symbolic logic are much simpler than natural languages. Logical symbolism is less ambiguous or vague than a natural language. This greater precision can be a problem when translating from informal natural languages into the formal language of the symbolism.

Some people think that this shows that the symbolism is an inadequate language. They are wrong. It is the natural languages that are inadequate—for logic. Natural languages are far more powerful and expressive than any artificial language. But they were invented for use in a wider range of human activities than just doing logic. For the restricted uses of logic, natural language is too vague and ambiguous.

The problem is that logic[22] needs to restrict itself to the truth-functional aspect of language. Many natural-language expressions do several jobs over and above the truth-functional-connecting job. The non-truth-functional components of the meanings[23] of these expressions obscure the truth-functional meaning. It can be difficult to dig out just the truth-functional part.

For example, look again at the words "and" and "but". Truth-functionally, they have exactly the same meaning. For logic, there is no difference between "Rosemary is an attractive woman and she's a lawyer" and "Rosemary is an attractive woman but she's a lawyer". The two sentences have somewhat different meanings, but there is no difference for logic. The difference is not truth-functional.

In natural languages, words that sometimes do one (truth-functional) job may also play another role in the language. For example, "and" may be used to connect two sentences in a conjunction-sentence, or it may be used to conjoin two expressions to make the compound *subject* (or object) of a *simple* sentence. "Monica and Steffi are tennis players" makes a compound statement, whose logical meaning is "Monica is a tennis player and Steffi is a tennis player". The "and" is the truth-functional conjunction connective we symbolize with "$\wedge$". But the sentence "Monica and Steffi are rivals" does *not* make the compound statement "Monica is a rival and Steffi is a rival". It makes a *simple* statement about a relation between two people—"Monica is a rival of Steffi".

Natural languages require subtle changes in the way a statement is expressed for *grammatical* reasons that have nothing to do with the *logical* meaning of the sentence. The result may be that two sentences that appear quite different may both have the same logical meaning (the differences being merely grammatical or stylistic).

Sometimes translation is difficult because the statement of an argument in a natural language may include one or all of the premises and the conclusion in a single sentence. In such cases, you have to figure out whether the conclusion is (1) the whole single compound statement or (2) a part of the sentence, so that other parts will be translated as one or more premise-statements.

When translating from natural language into the symbolism of propositional logic, the trick is to check and re-check that the truth-value of the translation behaves exactly like the truth-value of the form of the original sentence. If the translated statement would be true or false in exactly the same circumstances as the original, the translation has the same logical meaning as the original.

An additional requirement for good translation is that the symbolic representation should reflect as closely as possible the simplicity or complexity of the original sentence. For example, "If you love me and I'm able, I'll return to you" makes the same statement as "If you love me then, if I'm able, I'll return to you". The first could be symbolized as $(L \wedge A) \longrightarrow R$ and the second as $L \longrightarrow (A$

[22] At least, the kind of logic that we study in this course.

[23] By "meaning" in this text, I generally mean "use" or "rule for the correct use" of some piece of language.

$\longrightarrow R$). These have identical logical meaning (their forms have the same truth tables—check it!). But the first translation more closely reflects the logical "feel" of the English sentence.

Once an argument has been represented in the symbolism, the problems of natural language (for logic) disappear. Translating a natural-language sentence or argument into the symbolism clarifies the logic of the sentence or argument. It is then easier to construct good arguments and to recognize bad (invalid) arguments. The symbolism is a more logical language than any natural language. It is pathetically poor for writing love songs.

Warning: There is one thing that might confuse you in making translations. You will know that "if p then q" will become $p \longrightarrow q$. Other statements that also will become $p \longrightarrow q$ include "q if p", "p implies q", and (likely only in a mathematical context) "p is a sufficient condition for q". But what about an expression of the form "p only if q"? Be careful: this is *not* $q \longrightarrow p$, but is instead $p \longrightarrow q$. This is one place where the statement following the "if" is *not* the premise, but instead is the conclusion. Another way to say this is that "q is a necessary condition for p" is also translated $p \longrightarrow q$. (We did discuss this when introducing the material implication in section 1.3.2.) The reason for this is that if we say "p only if q", then we are saying it is impossible for p to be true unless q is also true, so that if q is false, so is p: that is, $\neg q \longrightarrow \neg p$, or in other words, $p \longrightarrow q$. The best way to remember this is probably to think of "only if" as meaning "implies". Don't let the position of the word "if" confuse you!

There is a lot of useful advice in section 3.3 of the Alberta Notes—you might want to refer to that as well if you have trouble with the next set of exercises.

1.4.1 Translation exercise

Symbolize the following sentences (use appropriate abbreviations for the various statements, such as H for "Harry will run for class president", *etc.*):

1. Harry and Judith will both run for class president.
2. Either Harry will run for class president or Judith won't.
3. If Harry runs for class president, then Judith won't run, but if Harry doesn't run, then Judith will.
4. If Harry and Judith both run, then George won't run.
5. It won't happen that Harry and Judith both run.
6. If Judith runs, then either Harry won't run or George will.
7. Harry will run if and only if Judith runs.
8. Neither Judith nor Harry will run.
9. George will run, and if Judith runs then Harry will also.
10. Harry will run only if Judith doesn't.
11. Harry will run unless Judith runs.
12. On the assumption that Harry will run if George does, it follows that Judith won't run.
13. Supposing that George runs provided that Judith does, it follows that Harry will run if Judith doesn't.

14. Alcohol and marijuana are drugs.
15. Alcohol and Benzedrine are a deadly combination.
16. Though he loved her, he left her.
17. Cigarettes and whiskey and wild, wild women will drive me crazy.
18. If we reduce pollution and population doesn't increase, our standard of living will not decline, but if we fail to reduce pollution, or if the population increases, then our standard of living will decline.
19. If we fail to reduce population or if the population increases, then our standard of living will decline and we'll have only ourselves to blame.

1.4.2 More translation exercises

For each of the following, construct a truth-functional paraphrase, and symbolize it in propositional logic. Use the following abbreviations:

A Albert jogs regularly.
B Bob jogs regularly.
C Carol jogs regularly.
L Bob is lazy.
M Carol is a marathon runner.
H Albert is healthy.

1. If Bob jogs regularly, he is not lazy.
2. If Bob is not lazy, he jogs regularly.
3. Bob jogs regularly if and only if he is not lazy.
4. Carol is a marathon runner only if she jogs regularly.
5. Carol is a marathon runner just in case she jogs regularly.
6. If Carol jogs regularly, then if Bob is not lazy he jogs regularly.
7. If both Carol and Bob jog regularly, then Albert does too.
8. If either Carol or Bob jogs regularly, then Albert does too.
9. If either Carol or Bob does not jog regularly, then Albert doesn't either.
10. If neither Carol nor Bob jogs regularly, then Albert doesn't either.
11. If Albert is healthy and Bob is not lazy then both jog regularly.
12. If Albert is healthy, he jogs regularly just in case Bob does.
13. Assuming Carol is not a marathon runner, she jogs regularly if and only if Albert and Bob both jog regularly.

14. Although Albert is healthy he does not jog regularly, but Carol does jog regularly if Bob does.
15. If Carol is a marathon runner and Bob is not lazy and Albert is healthy, then they all jog regularly.
16. If Albert jogs regularly, then Carol does provided that Bob does.
17. If Albert jogs regularly if Carol does, then Albert is healthy and Carol is a marathon runner.
18. If Albert is healthy if he jogs regularly, then if Bob is lazy he doesn't jog regularly.
19. If Albert jogs regularly if either Carol or Bob does, then Albert is healthy and Bob isn't lazy.

Now the reverse translation process: using the same abbreviations above, construct natural English sentences whose meaning is given by the following sentences of propositional logic.

1. $A \vee (B \vee C) \longrightarrow A \wedge (B \wedge C)$
2. $C \longrightarrow (A \wedge \neg B)$
3. $B \leftrightarrow (\neg L \wedge A)$
4. $\neg A \longrightarrow (\neg B \longrightarrow \neg C)$
5. $\neg A \wedge (B \leftrightarrow \neg L)$

1.5 Knights and Knaves

And now for something completely different . . .[24]

A story: There is an island far off in the Pacific, called the island of Knights and Knaves. On this island, there are people called **knights** (who always tell the truth, meaning everything that a knight says must be true) and **knaves** (who always lie, meaning everything a knave says must be false). They may be either male or female. The people of this island are often called "**knavghts**".[25] So some knavghts are knights, some are knaves, and every knight or knave is a knavght.

Another peculiarity of the knavghts: they seem to speak English, but with a small difference, in that they use the connectives of propositional logic in the strict sense we have defined earlier. So they only use "or" in the inclusive sense, and they only use "if . . . then . . . " (and similar expressions) to mean material implication. There is no ambiguity in their use of these propositional connective words.

One day you visit the island: you are then the only non-knavght on the island (so everyone else is either a knight or knave). Suppose you meet two knavghts, and one says, pointing to the other, "He said he was a knave". What can you conclude from this?

Well, clearly the speaker is a knave. You can figure this out this way: no knight could say "I am a knave", for that would be false, and knights always tell the truth. But no knave could say "I am a knave" either, for such a statement would then be true, but a knave always lies. So, no

[24] The material in this section comes (sometimes slightly modified) from the wonderful logic puzzle books of Raymond Smullyan, in particular his book *What is the name of this book?.* This book is, I believe, presently out-of-print, but if you ever come across a copy, I really recommend you buy it—it's a great collection of amusing logic puzzles, of which the following is merely a sample.

[25] "Knavght" is usually pronounced "knot", although folks from Brooklyn sometimes pronounce it "knate".

knavght could ever say "I am a knave", and anyone who tells you otherwise must be telling you a lie. So the speaker told a lie (made a false statement): he must be a knave therefore.

You can see from this analysis that it's often possible to deduce facts about knavghts from statements they make, facts that aren't explicitly part of their statements. In the situation above, the knavght made a statement about his companion, but really he was telling you something about himself (we still don't know whether the companion was a knight or knave).

Let's consider another situation: A knavght man was asked (about his wife, who was also a knavght, and himself) which, if either, was a knight and which, if either, was a knave. He answered "We are both knaves"; what are they?

See if you can figure out the answer yourself before you read the next paragraph!

He cannot be a knight, since a knight couldn't say he was a knave. So he is a knave. Now you might wonder about that, since a knave also couldn't say he's a knave. But that's not really what he said: he said "We are both knaves", which can in fact be a legitimately false statement (such as all knaves always make), provided his wife is not a knave. Careful now: if his wife were a knave, then "We are both knaves" would be true, and so an impossible utterance by a knave. So the only possibility is that he is a knave, his statement is false, and so his wife is a knight.

Here are some other situations; see if you can answer the questions posed. I have given you the answers, with some hints as to how they may be obtained. But try them yourself first.

1. Another knavght man was asked, of his wife and himself, "Are you both knaves?". He answered "At least one of us is"; what are they?

   (Ans: He cannot be a knave, because if he were a knave, his statement would be true, which is impossible for a knave. So he's a knight, and so his statement is true, so his wife must be a knave.)

2. Same situation: this time the man answers "If I am a knight, then so is my wife". What are they?

   (Ans: Assume he's a knight. Then it would follow that his wife is a knight too, since that's just what he said, and if he's a knight, his statement must be true. But look what we have here: we just showed that if he's a knight, then so is his wife. This is exactly what he claimed, and we've just seen this statement is true. Since he said a true statement, he must be a knight, and so therefore his wife must be too. There is a general principle at work here, which I'll summarize below, but see if you can guess what it must be.)

3. Same situation: this time his answer is "My wife and I are of the same type" (meaning either both knights or both knaves). What are they?

   (Ans: You cannot determine the husband's type, he could be knight or knave, but since we know he cannot claim he's a knave, his wife couldn't be a knave since that would in effect mean his statement would be claiming he's a knave too—so she's a knight. You can verify this by cases if you like. Again, there's a general principle working here too.)

There are some general principles which one can see in looking at these situations and their analyses.

1. No knavght can say "I am a knave"; every knavght must claim "I am a knight".

2. For any statement P, if a knavght says "If I am a knight, then P", then the knavght is in fact a knight, and P is true.

3. If a knavght says "If P then I am a knave", then P must be false and the knavght is in fact a knight. (Exercise: this is essentially the same as the previous principle.)

4. If a knavght says "I am a knight if and only if P", then P must be true (but the knavght could be either knight or knave).

5. If a knavght is asked "Is the statement you are a knight equivalent to the statement P?", then a "yes" answer means P is true, and a "no" answer means P is false.

6. Remember that sometimes it's simpler to transform the sentence by standard equivalences, such as $A \longrightarrow B \leftrightarrow \neg A \vee B$, $\neg(A \longrightarrow B) \leftrightarrow A \wedge \neg B$, $\neg(A \vee B) \leftrightarrow \neg A \wedge \neg B$, $\neg(A \wedge B) \leftrightarrow \neg A \vee \neg B$.

   So, for example: instead of thinking "If P then I am a knave", think "either not P or I am a knave". (Whatever makes it easier for you to deconstruct the sentence.)

You may use these principles in analysing other scenarios, in particular, in solving the following exercises. (If you have trouble, you might like to look at the next section, 1.5.2.)

1.5.1 Knights and knaves exercises

1. We have three people A, B, and C on the Island of Knights and Knaves. Suppose A and B say the following:

   A: All of us are knaves.
   B: Exactly one of us is a knave.

   Can it be determined what B is? Can it be determined what C is?

2. Suppose A says, "I am a knave but B isn't." What are A and B?

3. We again have three inhabitants, A, B and C, each of whom is a knight or a knave. Two people are said to be of the same type if they are both knights or both knaves. A and B make the following statements:

   A: B is a knave.
   B: A and C are of the same type.

   What is C?

4. Again three people A, B and C. A says "B and C are of the same type." Someone then asks C, "Are A and B of the same type?" What does C answer?

5. We have two people A, B, each of whom is either a knight or a knave. Suppose A makes the following statement: "If I am a knight, then so is B." Can it be determined what A and B are?

6. Someone asks A, "Are you a knight?" He replies, "If I'm a knight, then I'll eat my hat!" Prove that A has to eat his hat.

7. A says, "If I'm a knight, then two plus two equals four." Is A a knight or a knave?

8. A says, "If I'm a knight, then two plus two equals five." What would you conclude?

9. Given two people, A, B, both of whom are knights or knaves. A says, "If B is a knight then I am a knave." What are A and B?

10. Two individuals, X and Y, were being tried for participation in a robbery. A and B were court witnesses, and each of A, B is either a knight or a knave. The witnesses make the following statement:

    A: If X is guilty, so is Y.
    B: Either X is innocent or Y is guilty.

    Are A and B necessarily of the same type? (i.e. either both knights or both knaves.)

11. On the island of knights and knives, three inhabitants A,B,C are being interviewed. A and B make the following statements:

    A: B is a knight.
    B: If A is a knight so is C.

    Can it be determined what any of A, B, C are?

12. Another three inhabitants, A, B, C, make these statements:

    A: B is a knave.
    B: A is a knave.
    C: Both A and B are knaves.

    Can it be determined what any of A, B, C are?

13. Suppose the following two statements are true: (1) I love Betty or I love Jane. (2) If I love Betty then I love Jane. Does it necessarily follow that I love Betty? Does it necessarily follow that I love Jane?

14. Suppose that I am a knight, and someone asks me, "Is it really true that if you love Betty then you also love Jane?" I reply, "If it is true, then I love Betty." Does it follow that I love Betty? Does it follow that I love Jane?

15. This problem, though simple, is a bit surprising. Suppose it is given that I am either a knight or a knave. I make the following two statements:

    (a) I love Linda.
    (b) If I love Linda then I love Kathy.

    Am I a knight or a knave?

16. Is There Gold on This Island? On a certain island of knights and knaves, it is rumored that there is gold buried on the island. You arrive on the island and ask one of the natives, A, whether there is gold on this island. He makes the following response: "There is gold on this island if and only if I am a knight." Our problem has two parts:

    (a) Can it be determined whether A is a knight or a knave?
    (b) Can it be determined whether there is gold on the island?

17. Suppose, instead of A having volunteered this information, you had asked A, "Is the statement that you are a knight equivalent to the statement that there is gold on this island?" Had he answered "Yes," the problem would have reduced to the preceding one. Suppose he had answered "No." Could you then tell whether or not there is gold on the island?

18. The First Island. On the first Island he tried, he met two natives A, B, who made the following statements:

    A: B is a knight and this is the island of Maya.
    B: A is a knave and this is the island of Maya.

    Is this the island of Maya?

19. The Second Island. On this Island, two natives A, B, make the following statements:

    A: We are both knaves, and this is the island of Maya.
    B: That is true.

    Is this the island of Maya?

20. The Third Island. On this island, A and B said the following: A: At least one of us is a knave, and this is the island of Maya. B: That is true. Is this the island of Maya?

21. Here is a bit of an offbeat question. One day, on the island of Knights and Knaves, you see an inhabitant. You go up to her and ask: “Are you a knight or are you a knave?” She says: “I won’t tell you” and walks away. Is it possible to decide if she is a knight or a knave?

1.5.2 What’s it all about?—more general principles

There are several serious points about the knights and knaves story (sorry! it isn’t all fun and games after all!), which have to do with how negation acts with the various connectives. We shall see this in several contexts, but here are a few comments to go on.

Notice there is a difference between a knavght saying one sentence: “p and q.” and a knavght saying two sentences: “p.” “q.” We saw that early on: consider the difference between a knave saying “We are both knaves.” (referring to himself and his wife), and a knave saying “I am a knave. My wife is a knave.” The second utterance is impossible, since he cannot say “I am a knave” (it would be a true statement uttered by a knave, an impossibility). But the first statement is possible: “We are both knaves.” (which is the same as “I am a knave and my wife is a knave.”) could be said by a knave, provided his wife is a knight. This reflects the fact that the negation of a conjunction is a disjunction: $\neg(p \wedge q) \leftrightarrow \neg p \vee \neg q$. In the language of knights and knaves, a knave saying “$p \wedge q$” means at least one of p, q is false (maybe both, maybe not). It does not mean both p and q *must* be false. But that’s just what a knave saying “p. q.” amounts to: both p and q would then have to be false. This distinction would not hold for knights, for they always tell the truth, and “$p \wedge q$” is true precisely if both p and q are true.

We can summarize this sort of thing as follows.

If a knight says	then	if a knave says	then
$\neg p$	p is $\bot$	$\neg p$	p is $\top$
$p \wedge q$	both p and q are $\top$	$p \wedge q$	at least one of p or q is $\bot$
$p \vee q$	at least one of p or q is $\top$	$p \vee q$	both p and q are $\bot$
$p \longrightarrow q$	either p is $\bot$ or q is $\top$	$p \longrightarrow q$	p must be $\top$ and q must be $\bot$

Be sure you understand this—it will help in doing the exercises of course, but it also should help firm up your understanding of how the connectives work.

[Optional:] Translation into propositional logic

Finally, it is actually possible to translate a knight/knave problem into pure propositional logic (I don't really suggest you do this to solve knights and knaves problems, but it is an interesting observation).

The crucial point is that A can make a statement P if and only if the statement " 'A is a knight' is equivalent to P" is true.

To see why this is so, consider first what it means for A to assert P: if A is a knight, P must be true, and if P is true, then (since A said P, *i.e.* the truth) A must be a knight. On the other hand, if " 'A is a knight' is equivalent to P" is true, then if A is in fact a knight, P must be true, and so A can say P, whereas if A is in fact a knave, then P must be false, and so again, A can say P.

Let's abbreviate "A is a knight" by simply A, and "A is a knave" by $\neg A$ ("A is not a knight"), so that the statement "A says P" is equivalent to the statement $A \leftrightarrow P$. This is notationally dubious, since we are using the same letter A to mean two totally different things: a person (the knavght making the assertion), and a statement (that he is a knight). It is convenient, however, for we have just shown that with this abbreviation, we can read $A \leftrightarrow P$ as "A says P" as well as "A is equivalent to P", making these propositional formulas a bit easier to read.

This allows us to translate statements about the statements of knavghts into statements in propositional logic. For example, consider our second example, where a knavght man was asked about his knavght wife and himself which, if either, is a knight and which, if either, is a knave, and he answered "We are both knaves". Translating this, we would get the following: $A \leftrightarrow (\neg A \wedge \neg B)$, where A is the man and B is his wife. The solution we outlined essentially amounted to showing this implies $\neg A \wedge B$. In other words, we have to show $[A \leftrightarrow (\neg A \wedge \neg B)] \longrightarrow [\neg A \wedge B]$ is a tautology, which is a standard exercise in truth tables. (In fact, the $\longrightarrow$ can be strengthened to $\leftrightarrow$, and we'd still have a tautology. In other words, if A is a knave and his wife is a knight, then if asked what they are, A could[26] reply "We are both knaves". Check this by analysing the possibilities.)

Optional exercise: Verify all this, and translate some of the other examples and statements of principle.

By the way: the usual classic knights-and-knaves puzzle is this: going to the town of EverlastingDelights, you come to a fork in the road, where you meet a knavght. Not sure which way to go, you want to ask him which direction will get you there; what single question, with only "yes" or "no" as possible answers, could you put to him which will allow you to know what direction to go?

There are some variants of this: here are two.

Two knavghts are standing at a fork in the road. By asking one yes/no question to one of them, can you determine the direction to the town of EverlastingDelights? And, by asking one yes/no question to one of the knavghts, can you determine whether he is a knight?

[26] There are other replies he could also make—find as many as you can.

1.6 Answers to the exercises

Exercises 1.2.4

There is considerable room for variant answers—if you have questions about these answers, ask them!
BTW: these are clearly not arguments in propositional logic (or rather, to make them so, many more premises would have to be added, premises having to do with attitudes and so on).
In each case, I have put the premises (numbered) above the horizontal line, and the conclusion below it.

1.

1. Perfection of soul corrects the inferiority of the body
2. Physical strength without intelligence does nothing to improve the mind

---

It is right that men should value the soul rather than the body

2.

1. What is empty is nothing
2. What is nothing cannot be

There cannot be any emptiness

3.

1. The subject of the gods' existence and form is obscure
2. Human life is short

About the gods, I am not able to know whether they exist or do not exist, nor what they are like in form

4.

1. Other creatures are soon self-supporting
2. Man alone needs prolonged nursing

In the beginning man was born from creatures of a different kind

5.

1. Either death is a state of nothingness and utter unconsciousness, or there is a change and migration of the soul from this world to another

---

Death is good

Exercise 1.3.6

1: (a) negation (b) conjunction (c) implication (d) disjunction
(I'll leave 2, 3 to you.)

Exercise 1.3.8

(a) everything (b) everything (c) 7, 8, 9, 10 (d) 2, 3, 4, 5, 6 (e) 5, 6 (f) 7, 8, 9, 10
(g) 8, 9 (h) 3, 4 (i) none (j) 4 (k) 6 (l) 5, 6 (m) 9 (n) 7, 10 (o) 8, 9

Exercise 1.3.10

It will be clear from number 3 that the main column for numbers 1, 2 will be the same as the main column for the biconditional ($\top$, $\bot$, $\bot$, $\top$). Use this to check your answers.
Here's number 4: (The column with the compound formula's truth values is indicated by a $*$.)

p	q	$\neg p$	$\wedge$	q
$\top$	$\top$	$\bot$	$\bot$	
$\top$	$\bot$	$\bot$	$\bot$	
$\bot$	$\top$	$\top$	$\top$	
$\bot$	$\bot$	$\top$	$\bot$	
			*	

p	q	$\neg$	$(p \wedge q)$
$\top$	$\top$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\bot$
		*	

p	q	$\neg p$	$\wedge$	$\neg q$
$\top$	$\top$	$\bot$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

p	q	$\neg p$	$\vee$	$\neg q$
$\top$	$\top$	$\bot$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

Clearly the only two that are equivalent are $\neg(p \wedge q)$ and $\neg p \vee \neg q$.

Exercise 1.3.15

1.

A	B	$(A \to B)$	$\to$	$\neg A$
$\top$	$\top$	$\top$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

2.

A	B	$(A \wedge B)$	$\wedge$	$\neg A$
$\top$	$\top$	$\top$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$	$\bot$	$\top$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$
			*	

3.

A	B	A	$\wedge$	$\neg$	$(B \vee A)$
$\top$	$\top$	$\top$	$\bot$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
			*		

4.

A	B	$(A \to B)$	$\leftrightarrow$	$(\neg B$	$\to$	$\neg A)$
$\top$	$\top$	$\top$	$\top$	$\bot$	$\top$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$
			*			

5.

A	B	$(A \to B)$	$\to$	$(B$	$\vee$	$\neg A)$
$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\bot$	$\bot$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\bot$	$\top$	$\top$
			*			

6.

A	B	$(A \to B)$	$\vee$	$(B \to A)$
$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

And here is the rest:
(1) "If A implies B then A is false" (contingency)
(2) "A, B and $\neg A$ are all true" (contradiction)
(3) "A is true, but not 'B or A'" (contradiction)
(4) "'A implies B' is equivalent to 'not B implies not A'" (tautology)
(5) "If A implies B then either B is true or A is false" (tautology)
NOTE: It is worth noticing that in fact "A implies B" is actually *equivalent* to "either B or not A".
(6) "Either A implies B or B implies A" (tautology)

Exercise 1.4.1

Variants are possible—check with me if you're not sure.

(1) $H \wedge J$ (2) $H \vee \neg J$ (3) $(H \longrightarrow \neg J) \wedge (\neg H \longrightarrow J)$ (4) $(H \wedge J) \longrightarrow \neg G$ (5) $\neg(H \wedge J)$
(6) $J \longrightarrow (\neg H \vee G)$ (7) $H \leftrightarrow J$ (8) $\neg H \wedge \neg J$ (9) $G \wedge (J \longrightarrow H)$ (10) $H \longrightarrow \neg J$
(11) $H \vee J$ (12) $(G \longrightarrow H) \longrightarrow \neg J$ (13) $(J \longrightarrow G) \longrightarrow (\neg J \longrightarrow H)$ (14) $A \wedge M$ (15) C
(16) Loved $\wedge$ Left (17) C (18) $(R \wedge \neg I \longrightarrow \neg D) \wedge (\neg R \vee I \longrightarrow D)$
(19) $(\neg R \vee I) \longrightarrow (D \wedge B)$

Exercises 1.4.2

There are possible variants, but I've generally given the one "closest" to the English.

1. $B \longrightarrow \neg L$
2. $\neg L \longrightarrow B$
3. $B \leftrightarrow \neg L$
4. $M \longrightarrow C$
5. $C \leftrightarrow M$ (though possibly $M \longrightarrow C$ or $C \longrightarrow M$)[27]
6. $C \longrightarrow (\neg L \longrightarrow B)$
7. $(C \wedge B) \longrightarrow A$
8. $(C \vee B) \longrightarrow A$
9. $(\neg C \vee \neg B) \longrightarrow \neg A$
10. $(\neg C \wedge \neg B) \longrightarrow \neg A$
11. $(H \wedge \neg L) \longrightarrow (A \wedge B)$
12. $H \longrightarrow (A \leftrightarrow B)$ (though possibly $H \longrightarrow (A \longrightarrow B)$ or $H \longrightarrow (B \longrightarrow A)$)[24]
13. $\neg M \longrightarrow (C \leftrightarrow (A \wedge B))$
14. $H \wedge \neg A \wedge (B \longrightarrow C)$
15. $(M \wedge \neg L \wedge H) \longrightarrow (C \wedge B \wedge A)$
16. $A \longrightarrow (B \longrightarrow C)$
17. $(C \longrightarrow A) \longrightarrow (H \wedge M)$
18. $(A \longrightarrow H) \longrightarrow (L \longrightarrow \neg B)$
19. $(C \vee B \longrightarrow A) \longrightarrow (H \wedge \neg L)$

Translations from propositional logic (there are lots of correct variations as well).

1. If any of Albert, Bob, or Carol jog regularly, then they all do.

[27] There is a possible dispute about the meaning of "p just in case q"; on reflection, I lean to it meaning $p \leftrightarrow q$, though an argument could be made to support $p \longrightarrow q$ or even $q \longrightarrow p$.

2. If Carol jogs regularly, then Albert does but Bob doesn't.
3. Bob jogs regularly if and only if he's not lazy and Albert jogs regularly.
4. If Albert doesn't jog regularly, then Carol doesn't if Bob doesn't.
   (*This is equivalent to:* Carol doesn't jog regularly if neither Albert nor Bob does.)
5. Albert doesn't jog regularly, and Bob jogs regularly if and only if he's not lazy.

Exercise 1.5.1

On a test, you would have to provide your reasons for the answers; here however I have usually merely given the conclusion, with a hint in a few cases. To abbreviate things a bit, I have adopted the following notation: if A is a knavght: $\top(A)$ means "A is a knight"; its negation, $\neg\top(A)$, also denoted $\bot(A)$, means "A is a knave". I hope it's clear why I use this notation: $\top(A)$ not only means "A is a knight", it also means "everything A says is $\top$", and similarly for $\bot$. $?(A)$ means "we do not know the type of A".

1. $\bot(A), ?(B), \top(C)$
2. $\bot(A), \bot(B)$
3. $\bot(C)$ (but $?(A), ?(B)$)
4. "yes"
5. $\top(A), \top(B)$ (Use principle 2)
6. $\top(A)$, so he must eat his hat (Use principle 2)
7. $\top(A)$ (Use principle 2)
8. This statement cannot be made by any knavght—so I must be a knave(!).
9. $\top(A), \bot(B)$ *via* principle 3
10. same type
11. all $\top$
12. $\bot(C)$; A and B are not both the same type (so either $\top(A), \bot(B)$ or $\bot(A), \top(B)$).
13. I love Jane (but ?Betty)
14. I love Betty (but ?Jane)
15. $\top(me)$
16. There is gold on the island (but $?(A)$)
17. No gold on the island (*via* principle 5)
18. $\bot(B), \bot(A)$ so not Maya
19. $\bot(A), \bot(B)$ so not Maya
20. $\bot(A), \bot(B)$ so not Maya
21. She's a knight.

1.7 X-treme Knights and Knaves—getting a bit blood-thirsty!

A visit to Transylvania

More problems from Smullyan's *What is the name of this book?*

Preliminaries

In Transylvania, the population is made up of humans (who always intend to tell the truth) and vampires (who always intend to lie); what complicates the matter is that half the population is insane: they believe every true statement is really false, and *vice versa*. So, sane humans and insane vampires always tell the truth, but insane humans and sane vampires always lie. (For example, an insane vampire intends to lie, but since he thinks true statements are false and false statements are really true, he ends up actually telling the truth.)

So, let's explore some of the consequences of this odd situation. For example, if a Transylvanian says "I am not a sane human", what can you conclude? He cannot actually be a sane human (for if so, he'd say so), nor can he be an insane human (for, being a human, he'd want to tell the truth, but couldn't, so he'd say he was not an insane human, or equivalently, he was either sane or a vampire—he wouldn't say he was not a sane human, for that would be equivalent to saying he was either insane or a vampire, which is a true statement in his case). Check for yourself he cannot be a sane vampire either, so he must be an insane vampire.

Another example (showing an alternate way to look at such statements): suppose he said "I am human, or I am sane". If this statement is false, he must be an insane vampire, so his statement must be true, which is a contradiction. So the statement is true, and so he's either human or sane, but also he must be (because he told the truth) either a sane human, or an insane vampire. Only a sane human fits both conditions. So he must be a sane human.

Here are some to try for yourself (the answers are in the footnotes). If he said "I am an insane human", what is he?[28] If he said "I am a vampire", what can you conclude?[29] If he said "I am insane", what can you conclude?[30]

Here's an interesting principle: If a Transylvanian believes that he believes something, then that something must be true. If he does not believe that he believes something, then that something must be false. (Note that his merely believing something doesn't tell you about its truth or falsehood—it's the believing that he believes it that is crucial here!) Try to convince yourself of this principle.

Here is an even more important principle: If a Transylvanian says "I believe X", where X is some statement, then if he is human, X must be true, whereas if he is a vampire, then X must be false. (Convince yourself of this!)

Problem 1: I meet two Transylvanians, A and B. I ask A "Is B human?", and A replies "I believe so." I ask B "Do you believe A is human?" B answered yes or no; which was it: "yes" or "no"?[31]

Another principle, an old one this time: Let's call sane humans and insane vampires "knightlike", and insane humans and sane vampires "knavelike" (for the obvious reasons). Then, if a knightlike individual says "If I am knightlike, then X" (for some statement X), then he must be knightlike, and X must be true.

[28] He's a sane vampire.

[29] He's insane (but could be human or vampire).

[30] He's a vampire (but could be sane or insane).

[31] "Yes"

By the way: if you asked a Transylvanian "Are you knightlike?" what would his answer be?[32] If you asked a Transylvanian "Do you believe you are knightlike?" what can you conclude from his answer?[33]

Is Dracula alive and well in Transylvania?

Any tourist to Transylvania is bound to ask himself this question; suppose you asked a Transylvanian about this, and he replied "If I am human, then Count Dracula is still alive", then what can you conclude? ... Well, think it over: you should realize that you still won't know what you want, even if you asked a knightlike Transylvanian (he could be a sane human, so Dracula would be alive, or he could be an insane vampire and Dracula might be alive—or dead!). Check that the same indeterminacy holds if you get the answer "If I am sane, then Count Dracula is still alive", or even if you get the answer "If I am a sane human, then Count Dracula is still alive". However, if you get the answer "If I am either a sane human or an insane vampire, then Count Dracula is still alive", then you will definitely know Dracula is really alive (because then he is saying he is knightlike, which only knightlike individuals can do—we saw a similar principle when we were doing ordinary knights and knaves problems).

Can you think of a statement you might receive as an answer that would convince you that (a) Dracula is alive and (b) the statement itself is false. How about an answer-statement which would convince you that Dracula is alive but for which you couldn't determine if the statement is true or false?[34]

Problem 2: Suppose a Transylvanian made these statements:

(1) I am sane.
(2) I believe that Count Dracula is dead.

Can you determine whether Dracula is alive?

Problem 3: Suppose instead that the Transylvanian made these statements:

(1) I am human.
(2) If I am human then Count Dracula is alive.

Can you determine whether Dracula is alive?

Problem 4: Here are some quickies: Find a single question you can ask a Transylvanian which will determine whether he is a vampire or not. Now find one to determine if he is sane or not. Next, find one which will force him to answer "Yes", regardless of what sort of individual he is. And finally, find a question which will determine if Count Dracula is still alive.[35]

Dracula's Castle

Now things get interesting: the upper aristocracy in Transylvania use the old traditional language for some words, and in particular, they don't use "yes" and "no", but instead "bal" and "da"—the problem is, you don't know which means which! So, when one day you find yourself

[32] "Yes", regardless of what type of Transylvanian he is.

[33] "Yes" would mean he was sane; "no" would mean he was insane.

[34] "I am knavelike and Dracula is dead"; "I am knightlike if and only if Dracula is still alive". There are other possible answers. Show that another answer for the second situation is "I believe that if someone asked me whether Dracula was alive, I would answer 'Yes' ".

[35] **2.** Dracula is dead. **3.** Dracula is alive. **4.** "Are you sane?"; "Are you human?"; "Do you believe you are human?" or "Are you knightlike?"; "Is the statement that you are knightlike equivalent to the statement that Dracula is alive?" or "Do you believe that the statement that you are human is equivalent to the statement that Dracula is alive?"

invited to Dracula's Castle, which is inhabited by aristocrats (possibly including Dracula himself!), you have to deal with the situation that not only do you not know which type of individual everyone is, but you cannot really tell what they're answering when they say "bal" or "da". But think about it a bit: it is possible to ask a single question (which will get you a "bal/da" answer) which will tell you whether the individual you are speaking to is a vampire: namely "Is 'Bal' the correct answer to the question 'Are you sane?'?" (Check this for yourself!)

Now find similar questions which will determine whether you are speaking to an individual who is sane or not; to determine what "Bal" means; to force him to answer "Bal" to your question; and to find out whether or not Dracula is alive.[36]

Dracula's Challenge

Finally, you get to meet Dracula. Normally, this wouldn't be a *good thing*, but he offers you a chance to save your life: he points out that although you've been very clever to get all those questions to sort out who's who, you've missed a general underlying principle that solves all such questions. If you can find that principle, he will let you go away unharmed (but if you cannot, then he'll turn you into a vampire!). There is one single sentence S with the miraculous property that it can help you determine the truth of any other sentence X merely by asking any individual "Is S equivalent to X?". For if they answer "Bal", X must be true, but if they answer "Da", then X must be false. (For example, to find out if Dracula is alive, you'd just have to ask any Transylvanian aristocrat "Is S true if and only if Dracula is alive?") To save your life, you must tell me what the sentence S is. (!)

Find a good candidate for S, to save your life.[37]

[36] "Is 'Bal' the correct answer to 'Are you human?'?" (If he answers "Bal" then he's sane.) "Do you believe you are human?" (Everybody must answer "Yes" to that, so whatever he says must mean "yes".) "Is 'Bal' the correct answer to the question 'Are you knightlike?'?" (Another question would be "Are you knightlike if and only if 'Bal' means 'Yes'?" Both these will force an answer "Bal".) And finally, "Do you believe that 'Bal' is the correct answer to the question 'Is the statement that you are human equivalent to the statement that Dracula is alive?'?", or similarly "Is 'Bal' the correct answer to the question 'Is the statement that you are knightlike equivalent to the statement that Dracula is alive?'?"

[37] Call a Transylvanian "Bal-ish" if he answers 'Bal' to the question 'Is $1 + 1 = 2$?'?" (or any similarly always-true question). The point is: if "Bal" means "Yes", then Bal-ish Transylvanians are knightlike, and if "Da" means "Yes", then Bal-ish Transylvanians are knavelike. So, $S =$ "You are Bal-ish" does the trick: If the answer to "Is X equivalent to the statement that you are Bal-ish?" is "Bal", then X must indeed be true, and X must be false if the answer is "Da". Check this yourself!

Chapter 2

Formal Proof—Form and Substitution

2.1 Derivations

We show that an argument is valid by showing that every step in the inference from premises to conclusion is justified by a derivation rule. **Derivation rules** are rules for deriving a new statement from one or more other statements.

> A **derivation** or proof is a sequence of statements, each of which is either (1) a premise or (2) a statement derived from one or more previous statement(s), where the last statement in the sequence is the conclusion.

When logicians draw up a list of derivation rules (laws of derivation) their goal is a list that is short enough to remember, but that is both complete and consistent.

> A list of derivation rules is **complete** if and only if its rules permit us to derive every conclusion that validly follows from any set of premises.

The definition of "valid deductive argument" says that a conclusion validly follows from a set of premises when the form of the argument is such that the conclusion could not be false when the premises are true. If our list of derivation rules is incomplete, there will be valid deductive arguments whose conclusion could not be derived from the premises. That would mean that the argument is valid, but it could not be **shown** to be valid.

> A list of derivation rules is **consistent** if and only if the rules do not permit the derivation of a contradiction from premises that do not contain a contradiction.

If our rules permit the derivation of a contradiction from true premises, then we could derive a false conclusion (every contradiction is necessarily false) from true premises. "Valid" *means* that true premises cannot yield a false conclusion. Our derivation-rules have to be consistent.

In Chapter 1 we proposed the conjecture that "given any conjunction as a premise, we may validly infer the first conjunct". Using the symbolism, we re-state the conjecture as a derivation rule: "any argument of the form

$$\frac{A \wedge B}{A}\ (\wedge E)$$

is a valid argument". Here we list all premises above the horizontal line, separated by spaces, and the conclusion below it. We have also given this rule a name: $(\wedge E)$, read "and elimination"—we shall explain our naming convention later, but for the moment, it might help to think of the "E" as standing for "elimination", and to notice that in the rule, passing from top to bottom does seem to eliminate an occurrence of the symbol $\wedge$.

A more compact way to say the same thing is to say that the form

$$p \wedge q \vdash p$$

is a valid argument form. This notation lists all the premise-forms (separated by commas if there are more than one) to the left of the $\vdash$ ("entailment") symbol, and the conclusion-form on its right.

It is fairly obvious that the argument

$$\frac{A \wedge B}{A}$$

is an argument of the specified form. It may be less obvious that

$$\frac{(\neg B \longrightarrow C) \wedge (C \longrightarrow D)}{\neg B \longrightarrow C}$$

is also an argument of that form. Is this an argument "whose form is such that its conclusion cannot be false when all of its premises are true"? How can we tell?

If we parse $(\neg B \longrightarrow C) \wedge (C \longrightarrow D)$, we see that the main connective is a conjunction, so this statement is a conjunction. The argument has a conjunction as its premise. The conclusion is exactly the same as the first conjunct (without the parentheses). So this particular argument is an example, a **substitution instance**, of the general form $p \wedge q \vdash p$.

It is pretty obvious that there is a variant of the $(\wedge E)$ rule, which we shall also call $(\wedge E)$ since, in a sense, it is "morally the same rule":[1]

$$\frac{A \wedge B}{B}\ (\wedge E)$$

or in words, from a conjunction, one may validly conclude the second conjunct. In compact notation, $p \wedge q \vdash q$.

There is also a quite different derivation rule involving conjunction, which may be expressed this way:

$$\frac{A \quad B}{A \wedge B}\ (\wedge I)$$

This says that if you have two premises, A and B, then a valid conclusion from them is their conjunction, *viz* the statement $A \wedge B$, for it is impossible that the premises A, B, could both be true and yet the conclusion $A \wedge B$ be false. We name this $(\wedge I)$ "and introduction".

In our more compact notation, this is the rule

$$p, q \vdash p \wedge q$$

Again, the issue is to recognise an instance of this rule when one sees it, which amounts to recognising substitution instances of the rule.

Here is an example:

$$\frac{A \longrightarrow \neg C \quad (\neg A \vee D) \longrightarrow B}{(A \longrightarrow \neg C) \wedge ((\neg A \vee D) \longrightarrow B)}$$

[1] Of course, it is not at all the same rule, for it mentions a different formula in its conclusion—and some folks do give the two rules different names, such as $(\wedge E)_l$ and $(\wedge E)_r$. You may use such names if you prefer, but I find no confusion results from "overloading" the single name $(\wedge E)$ with the two meanings. A similar remark might be made about the rule $(\vee I)$, which you will meet soon.

You should see that this is a substitution instance of the ($\wedge I$) rule by parsing the conclusion and showing it has the right conjunctive form.

We state derivation rules by using statement variables (p, q, r, *etc.*) to describe the form(s) of the premise(s) and of the conclusion. A particular argument made with particular statements is justified by a derivation rule if (and only if) the particular argument is a substitution instance of the form of the rule.

So, is the argument

$$\frac{A \to \neg C \quad (\neg A \vee D) \to B}{(A \to \neg C) \wedge ((\neg A \vee D) \to B)}$$

justified by the ($\wedge I$) rule $p, q \vdash p \wedge q$? Our conclusion is a conjunction, as the rule requires. Its two conjuncts are precisely the two premises, as the rule requires. So this is justified by the ($\wedge I$) rule.

Another example: is the argument

$$\frac{(\neg B \to C) \wedge (C \to D)}{\neg(B \to C)}$$

justified by the ($\wedge E$) rule $p \wedge q \vdash p$? Once again, we parse the premise, seeing that it is a conjunction, and we compare each conjunct with the conclusion. The first conjunct is an implication, whereas the conclusion is a negation—these do not match. The second conjunct isn't even close, so clearly this argument is *not* a substitution instance of the ($\wedge E$) rule, and so is not justified by that rule.

2.1.1 Substitution instance of an argument form or derivation rule

By now, the following definition is pretty obvious; when an argument *does* have the same form as a derivation rule, we say that the argument is a **substitution instance** of the rule:

> A particular argument is a **substitution instance of a derivation rule** if the particular argument is the result of replacing every distinct *simple* statement variable in the rule with a *simple or compound* statement. None of the connectives in the rule may be altered or eliminated. If any statement variable occurs more than once in the rule, every occurrence must be replaced by the same (simple or compound) statement in the particular argument.

Examples

1. Is the argument

   $$\frac{A \wedge B \quad (A \wedge B) \to \neg C}{\neg C}$$

   a substitution instance of the argument form $p \to q, p \vdash q$?

   The second premise is a conditional (parse it to check). Its antecedent is the compound WFF $A \wedge B$. The first premise is $A \wedge B$. The rule requires a conditional and its antecedent as premises, and that's what we've got. (The order of the premises does not matter.) The conclusion is $\neg C$, the consequent of the conditional, as the rule requires. Our argument is the result of substituting $A \wedge B$ for p and $\neg C$ for q in the rule, so our argument is a substitution instance of the rule.

2. Is the argument

   $$\frac{A \to B \quad \neg\neg A}{B}$$

a substitution instance of the argument form $p \longrightarrow q, p \vdash q$?

The first premise is a substitution instance of the first premise form $p \longrightarrow q$. Is the second premise a substitution instance of p? No, because we substituted A for p in the first premise, and the second premise does not make the same substitution. It substitutes $\neg\neg A$ for p. A is not the same WFF as $\neg\neg A$. You may object that they "mean the same" because the double-negation of a statement and the statement itself are truth-functionally equivalent. But they are not the same, syntactically. They are different: for instance one takes three symbols to write, the other only one.

3. The argument

$$\frac{\neg\neg A \longrightarrow B \quad \neg\neg A}{B}$$

is a substitution instance of the form $p \longrightarrow q, p \vdash q$.

4. Is the argument

$$\frac{A \vee B \quad \neg B}{A}$$

a substitution instance of the form $(p \wedge q) \vee r, \neg r \vdash p \wedge q$?

It is not. Although the argument does consistently substitute B for r, it tries to substitute A for $p \wedge q$. One of the connectives (and one statement) in the form is missing from the argument. However, the argument *is* a substitution instance of the form $p \vee r, \neg r \vdash p$.

5. Is the argument

$$\frac{A \vee B \quad \neg C}{A}$$

a substitution instance of the form $p \vee r, \neg r \vdash p$?

It is not. In one premise it substitutes B for r, but it substitutes C for r in the second.

6. Is the argument

$$\frac{(J \vee K) \longrightarrow N \quad F \vee (J \vee K) \quad F \longrightarrow N}{N \vee N}$$

a substitution instance of the form $p \longrightarrow q, r \longrightarrow s, p \vee r \vdash q \vee s$?

It is. It substitutes N for both q and s, but careful reading of the definition and restrictions on substitution instances of argument forms shows that every distinct different simple statement form in the form need not have a distinct *different* simple statement as its substitution. The rest of the substitution is F for p and $J \vee K$ for r (but not the other way round, for then you would have the mismatch of $r \vee p$ instead of $p \vee r$—equivalent, but not the same).

2.1.2 Exercise on argument form and substitution instance

For each of the argument forms in the left-hand column, say which of the arguments in the right-hand column are substitution instances of that form. Give reasons to justify your answers.

a. $p \rightarrow q, \neg q \vdash \neg p$

b. $p \rightarrow q, q \rightarrow r \vdash p \rightarrow r$

c. $p \rightarrow q, r \rightarrow s, p \vee r \vdash q \vee s$

d. $\neg(p \wedge q) \vdash \neg p \vee \neg q$

e. $p \vdash \neg\neg p$

f. $p \vdash p \vee \neg q$

g. $p \wedge (q \vee r) \vdash (p \wedge q) \vee (p \wedge r)$

1. $\dfrac{\neg B}{\neg\neg\neg B}$

2. $\dfrac{A \wedge (A \vee B)}{(A \wedge A) \vee (A \wedge B)}$

3. $\dfrac{J \rightarrow K \quad L \rightarrow K}{J \rightarrow L}$

4. $\dfrac{(A \vee B) \rightarrow (J \rightarrow K) \quad (M \wedge N) \rightarrow (F \rightarrow G) \quad (A \vee B) \vee (M \wedge N)}{(J \rightarrow K) \vee (F \rightarrow G)}$

5. $\dfrac{\neg(K \wedge B)}{\neg K \wedge \neg B}$

6. $\dfrac{A \rightarrow (J \rightarrow K) \quad (J \rightarrow K) \rightarrow B}{A \rightarrow B}$

7. $\dfrac{Y}{Y \vee \neg(W \rightarrow X)}$

8. $\dfrac{\neg A \rightarrow (F \rightarrow G) \quad \neg(F \rightarrow G)}{\neg\neg A}$

2.2 Basic Derivation Rules

In this section we shall consider the set of derivation rules we shall have available to create valid arguments. The basic context is one you should remember: we want to have a set of rules, so that (1) each is a valid argument form, and (2) every valid argument form may be obtained from the set of rules we give, as a derivation based on those rules. (This means we want our rules to be consistent and complete.)

Of course, we could (in principle) just give all valid argument forms as our rules, but that would not be particularly useful, as remembering all these forms would be rather difficult. We want our rules to be as simple as possible, and as natural as possible, so that it is easy to remember them, as well as easy to use them to construct derivations for other valid argument forms. We shall accomplish that by using what is called a "natural deduction" set of rules.

The rules we shall present in this (and the next) chapter, are in fact a complete and consistent set of derivation rules for propositional logic. (We shall not actually prove them to be complete, but I hope that fact will at least seem plausible by the end of the chapters. A completeness proof will be outside the scope of this course.)

In a natural deduction system, the derivation rules are structured in a very simple way. Using $\star$ to represent any of our connectives ($\wedge$, $\vee$, $\rightarrow$, $\neg$), for each connective there will be two rules, a $\star$-introduction rule ($\star I$) and a $\star$-elimination rule ($\star E$). The purpose of the $\star$-introduction rules will be to introduce the connective, *i.e.* to introduce a new formula whose main connective is $\star$, so that the rule is useful when one wants to prove a $\star$-formula, *i.e.* to produce a $\star$-formula as a conclusion. The purpose of the $\star$-elimination rules will be to use ("eliminate") a formula whose main connective is $\star$ to derive something else, so that the rule will be useful in handling $\star$-formulas as premises. In each case, these rules will be a natural reflection of the truth-functional meaning of the connective; so as long as you remember that, you should have no trouble in remembering the rules. (That is why we spent time on truth tables!)

To begin with, we shall describe these rules somewhat informally, using the format with the premises above a horizontal line, the conclusion below. In the next chapter we shall introduce a somewhat more streamlined way of writing these, which will make it a bit easier to write longer derivations using many of the rules together. In each case, you should focus on the *meaning* that lies behind the notation, and try to keep the simple essence in mind, so you can understand these rules easily.

2.2.1 Conjunction

We have already seen the two conjunction rules.

$$\frac{p \quad q}{p \wedge q}\ (\wedge I)$$

In words, if two statements are both true, then so is their conjunction.

$$\frac{p \wedge q}{p}\ (\wedge E) \qquad\qquad \frac{p \wedge q}{q}\ (\wedge E)$$

In words, if a conjunction is true, then so is each conjunct separately. (We regard this as one rule, with two variants. One may regard each variant as a separate rule, if one wishes, but that seems unnecessary.)

There is really not much more to say about these rules.

2.2.2 Disjunction

The first disjunction rule is simple enough:

$$\frac{p}{p \vee q}\ (\vee I) \qquad\qquad \frac{q}{p \vee q}\ (\vee I)$$

If a statement is true, then so is any disjunction which has that statement as a disjunct (first or second). A moment's reflection should convince you that this is essential to what "or" means. (This is admittedly the "inclusive or", for no restriction is being placed on whether both disjuncts are true or not, merely that at least one is true.) Again, we regard this as one rule, with two variants.

The other disjunction rule is somewhat trickier, so let's consider what's involved first. We want a "$\vee$-elimination" rule, that is to say, a rule which will allow us to use a disjunctive formula as a *premise* in a derivation. But how can we conclude something from a premise of the form $A \vee B$? Well, consider what we know if we claim $A \vee B$ is true: then *one* of the two formulas A, B is true, at least, but we don't know which one. If we want to use this information to prove some conclusion (C let us say) then we shall have to be able to prove C from premise A, as well as from premise B (since we don't know which one is true, we have to "cover both possibilities" in effect). This means we shall have to have two derivations in hand, $A, X \vdash C$ and $B, X \vdash C$, in each case possibly with some other premises (represented by the X) as well, and once we have them both, we can conclude $A \vee B, X \vdash C$. Notice that in this new derivation, although we shall still need the extra premises X, we no longer need the premises A and B, since they are replaced by the new premise $A \vee B$.

This is a bit of a load to think about, so think about it carefully. Here is an example, taken from a simple mathematical proof that multiplying two successive whole numbers always gives an even number as the product.

> We start with the observation (fact) that every whole number is either even or odd, and from this we shall conclude that the product of two successive whole numbers is even. We shall break the proof into two "cases": one assuming the first number is even, the other assuming that it is odd. So, our 1st case is this. Premise: suppose the first number is even. Conclusion: the product of that number with its successor is even, since any even number times another number is even. Our 2nd case is this. Premise: suppose the first number is odd. Conclusion: the product of that number with its successor is even, since that successor itself is even, and so again the product is even. So we are done: since every number is even or odd, that means that any product of two successive whole numbers is even.

Notice the key strategy here: we used the disjunctive hypothesis "the first number is even or odd" by breaking the argument into two *cases*, one for each disjunct (that the first number was even, and that the first number was odd). We've seen this strategy before: in many of the knights and knaves problems, we often started "suppose A is a knight ...", followed by "suppose instead that A is a knave ...", in both cases getting the same desired conclusion, which meant that conclusion had to be true always. This amounted to taking the hypothesis "A is either a knight or a knave" and breaking the argument into two cases, one for each possibility.

So this shall be our structure for the $(\vee E)$ rule: to prove some conclusion r from a disjunctive premise $p\vee q$, we shall "break the argument into cases", *i.e.* we shall need to have two subderivations, each with the conclusion r, but one with premise p, the other with premise q (other premises may also be present). In our final conclusion, these premises p, q will be "discharged"—they shall no longer be considered as premises, but will be removed from the premises in the new argument and replaced by the new premise $p \vee q$.

We represent this as follows:

$$\dfrac{p \vee q \qquad \begin{matrix}[p]^1 \\ \vdots \\ r\end{matrix} \quad \begin{matrix}[q]^1 \\ \vdots \\ r\end{matrix}}{r}\ (\vee E)^1$$

The superscripts on the premises, placed in brackets, indicate that those premises have been "discharged" by the application of the rule with the matching index (the number 1 shown here, though any other matching number or symbol could be used). Notice that above the "premise line" we have one premise $p \vee q$ and two *arguments*, namely the two arguments or subderivations $p, \ldots \vdash r$ and $q, \ldots \vdash r$. The vertical dots in our rule just indicate whatever is necessary for those two subderivations to work. The final conclusion is r, underneath the horizontal line. It has all the premises above the line, except for the "discharged" p, q.

Here's another example.

> If you win the lottery, everybody will be after your money, you'll retreat into a shell, isolating yourself from the rest of society, and you'll die a miserable lonely person. But if you don't win the lottery, you'll end up on skid road, poor and grubby; no one will want to be anywhere near you, and you'll die a miserable lonely person. So, win the lottery or not, you'll still die a miserable lonely person.

This is an argument whose premise is a disjunction ("win the lottery or not"), proved by examining each possible case separately, in other words, using $(\vee E)$.

This is a tricky rule, and we shall soon see many examples, which should help you become more familiar with how it works. Try to keep in mind the essential idea: to prove something from a disjunction, it is necessary to break the argument into cases, one for each disjunct. Prove what you want in each case, and you've got it from the disjunction itself.

[OK: take a breath!]

2.2.3 Implication

How do we prove a conditional formula $A \longrightarrow B$? What does it mean to prove such a formula? Well, $A \longrightarrow B$ means that if A is true, then B must also be true; this suggests that proving this would amount to constructing a proof of B (as conclusion) from A (as premise). And this is exactly right. If we have a proof of B from A (with other premises perhaps), then we also have a proof of $A \longrightarrow B$, with the premise A no longer necessary in that proof. (This is another instance of a premise being "discharged", as we had with $(\vee E)$.)

In our usual symbolism, this may be written as follows.

$$\dfrac{\begin{matrix}[p]^1\\ \vdots\\ q\end{matrix}}{p \longrightarrow q}\ (\longrightarrow I)^1$$

where we use the same trick with superscripts to indicate the discharged premise p.

This isn't quite as odd as it may seem; let's consider an example. Here is an argument you might make if you don't want to lend your car to your crazy cousin:

> Suppose you borrow my car. Whenever you have a car, you collect all your friends and go driving. When you're with your friends you always drink too much. When you drink too much you love to show how fast you can drive. When you drive fast, you have accidents. When you have accidents, you wreck cars. So you'll wreck my car. Therefore if you borrow my car, then you'll wreck it.

What you've done here is construct an argument starting from the premise "suppose you borrow my car", and finished with the conclusion "so you'll wreck my car". That justified you in claiming that "if you borrow my car, then you'll wreck it".

The elimination rule for implication is more straightforward.

$$\dfrac{p \quad p \longrightarrow q}{q}\ (\longrightarrow E)$$

This rule is famously known as *modus ponens.* It says that if an implication is true, as well as its antecedent, then its consequent must be true also.

Here is a simple example: If Jones shot Smith intentionally, then Jones is guilty of murder; Jones intended to shoot Smith, and he did shoot Smith. Therefore Jones is guilty of murder.

We can prove this is valid using $(\wedge I)$ (to conclude "Jones shot Smith intentionally" from "Jones intended to shoot Smith" and "he did shoot Smith") and $(\longrightarrow E)$ (to conclude "Jones is guilty of murder" from "if Jones shot Smith intentionally, then Jones is guilty of murder" and "Jones shot Smith intentionally").

2.2.4 Negation

Here's a wonderful little conceptual trick: the formula $\neg p$ is equivalent to the formula $p \longrightarrow \bot$. (Exercise: construct the truth tables for these and compare them.) It's easy enough to see this from the meaning of the material implication $\longrightarrow$: $p \longrightarrow \bot$ is false only if its premise p is true and its conclusion $\bot$ is false (which it always is). Consider what was just said: $p \longrightarrow \bot$ is false only if p is true; so $p \longrightarrow \bot$ is true only if p is false. In other words, $p \longrightarrow \bot$ and p have "opposite" truth values; in other words $p \longrightarrow \bot$ is (equivalent to) $\neg p$.

This means we can construct the derivation rules for $\neg p$ directly from the rules for $\longrightarrow$, just using the special case $p \longrightarrow \bot$. If we do this, we get the following two rules.

$$\dfrac{\begin{array}{c}[p]^1\\ \vdots\\ \bot\end{array}}{\neg p}\ (\neg I)^1 \qquad\qquad \dfrac{p \quad \neg p}{\bot}\ (\neg E)$$

In words, if you can derive a contradiction ($\bot$) from assuming p then you can derive $\neg p$ without that assumption, and you can derive a contradiction from two premises of the form p and $\neg p$. The first rule is often called "proof by contradiction", and the second rule is often called "the law of contradiction".

Remark: Be careful when using proof by contradiction, $(\neg I)$, that you use only exact substitution instances of this rule. As its name should remind you, it *always* introduces a new $\neg$ sign to the statement involved; it cannot *remove* a $\neg$ sign. For example, if you have a derivation of $\bot$ from an assumption $\neg p$, then one may only conclude $\neg\neg p$; you may **not** conclude p:

$$\dfrac{\begin{array}{c}[\neg p]^1\\ \vdots\\ \bot\end{array}}{\neg\neg p}\ (\neg I)^1 \qquad\qquad \dfrac{\begin{array}{c}[\neg p]^1\\ \vdots\\ \bot\end{array}}{p}\ (\text{not correct!})^1$$

To finally conclude p in this situation, you need to use an additional rule, the "law of the excluded middle" ($(\neg\neg E)$ below), which justifies that further conclusion. There are two reasons for care here: first, it is good for your soul to get used to the precision needed to construct correct proofs carefully, and secondly, there are strong philosophical and practical reasons to keep track of what arguments need the $(\neg\neg E)$ rule and which do not. Being careless of the $(\neg I)$ rule can hinder that effort. We shall return to this point later.

2.2.5 Two additional rules

There are two extra rules we need. The first expresses the notion that a false premise allows any conclusion one wishes: this is what I referred to earlier as the "all bets are off", or "anything goes", situation that results from a false context or premise.

$$\dfrac{\bot}{p}\ (\bot E)$$

The second rule expresses the fact that in this logic, there are only two truth possibilities a statement may have (true and false), so that if a statement is not false, it must be true:

$$\dfrac{\neg\neg p}{p}\ (\neg\neg E)$$

This is often called "the law of the excluded middle".

Remark: We will generally be wary of using the rule $(\neg\neg E)$, since it seems to me to be less well justified philosophically than any other of the rules we've considered (it's manifestly obvious that there are many statements that don't easily fit into the "true or false" box, even many purely mathematical statements). In the early twentieth century, a school of logic, known as "intuitionism", formalised this objection, and considered logic without the $(\neg\neg E)$ rule. This logic is very important today, and finds applications in many situations where a more "constructivist" approach is necessary. I will tend to point out derivations where we use the $(\neg\neg E)$ rule, and so try to keep track of what sort of logical conclusions are valid without it, and what sort require it.

2.2.6 Examples

Here are some derivations, using the informal presentation above, with premises separated from conclusions by horizontal lines. We "stack" one derivation on top of another, by proving in the top derivations the premises of the lower ones. This is the basis for a formal presentation of derivations, but one that is somewhat unwieldy, and so one we shall not use extensively.[2]

These derivations in fact correspond to well-known derivation rules.[3] We don't need these rules, since they may be derived as shown, although they can make nice short-cuts in longer derivations. I have indicated their traditional names.

Disjunctive Syllogism:

$$\dfrac{p \lor q \quad [p]^1 \quad \dfrac{\dfrac{[q]^1 \quad \neg q}{\bot}\,(\neg E)}{p}\,(\bot E)}{p}\,(\lor E)^1$$

Modus Tollens:

$$\dfrac{\dfrac{\dfrac{[p]^1 \quad p \longrightarrow q}{q}\,(\longrightarrow E) \quad \neg q}{\bot}\,(\neg E)}{\neg p}\,(\neg I)^1$$

Hypothetical Syllogism:

$$\dfrac{\dfrac{\dfrac{[p]^1 \quad p \longrightarrow q}{q}\,(\longrightarrow E) \quad q \longrightarrow r}{r}\,(\longrightarrow E)}{p \longrightarrow r}\,(\longrightarrow I)^1$$

Constructive Dilemma:

$$\dfrac{p \lor r \quad \dfrac{\dfrac{[p]^1 \quad p \longrightarrow q}{q}\,(\longrightarrow E)}{q \lor s}\,(\lor I) \quad \dfrac{\dfrac{[r]^1 \quad r \longrightarrow s}{s}\,(\longrightarrow E)}{q \lor s}\,(\lor I)}{q \lor s}\,(\lor E)^1$$

Equivalence Rule:

$$\dfrac{\dfrac{\dfrac{[p]^1 \quad \dfrac{\dfrac{p \leftrightarrow q}{(p \longrightarrow q) \land (q \longrightarrow p)}\,(Def)}{p \longrightarrow q}\,(\land E)}{q}\,(\longrightarrow E) \quad q \longrightarrow r}{r}\,(\longrightarrow E)}{p \longrightarrow r}\,(\longrightarrow I)^1$$

Disjunctive Syllogism II:

$$\dfrac{\dfrac{\dfrac{p \quad [q]^1}{p \land q}\,(\land I) \quad \neg(p \land q)}{\bot}\,(\neg E)}{\neg q}\,(\neg I)^1$$

(Note that the Equivalence Rule is the formal version of the statement that in any valid argument deriving r from assumption q, if $p \leftrightarrow q$ you can replace q by p as assumption, and have a valid argument deriving r from assumption p.)

So: we have a set of rules that reflect the basic properties of the logical connectives, but actually putting them together to create more complicated arguments seems somewhat unwieldy. In the

[2] I am being a bit misleading here—this presentation, with the premises separated from conclusions by horizontal lines, and stacking derivations above one another, is not at all informal, nor really unwieldy (though it can take a lot of page space), and indeed, many completely rigorous research papers and books have been written using this notation, including many by myself! But the Fitch-style presentation of the next chapter does seem to be somewhat simpler for our purposes.

[3] These include the four basic schemata of the Stoics, in fact, without the redundant one about exclusive or.

next chapter we shall revisit the derivation rules with a more streamlined presentation, more suited to writing down longer derivations.

2.3 Answers to the exercises

Exercise 2.1.2:
(a): 8 (b): 6 (**not** 3) (c): 4 (d): none (**not** 5) (e): 1 (f): 7 (g): 2

Chapter 3

Formal Proof—Fitch-style Natural Deduction

3.1 The Natural Deduction Rules, revisited

Our task now is to develop a simpler way to write formal derivations, using these derivation rules. We shall adopt the following formalism. Please take care: the point of this section is the formalism itself, with its very fussy attention to detail, and its very particular formation rules. This is one place where "getting the rough idea" just doesn't cut it! You *must* pay attention to the nitty gritty details, and be very precise in how you present things. Getting the details wrong is simply getting it wrong—not much in the way of "feel-good consolation" to be found here!

A derivation will consist of a collection of numbered lines, beginning with the premises of the argument. These will be separated from the rest of the derivation by a horizontal line. We shall also use a vertical line (often called a "spline") at the far left to indicate the "scope" of the derivation (where it starts and ends); the line numbers will be to the left of this vertical line. At some points we may insert a subderivation—that will have its own vertical line to indicate where it starts and ends, "nested" one "level" in (or "one level down") from the main derivation, but the line numbering will continue at the far left, including the lines of the subderivation. This nesting of subderivations may go as deep as you wish: subderivations may themselves contain subderivations. Every line will contain a formula, and if that formula is not a premise, it must contain a notation indicating the justification for the formula. That justification will consist of the derivation rule used, and the line numbers of whatever premises are needed for that derivation rule to be applied.

In the following discussion of the rules, the line numbers are indicated by variables m, n, k; in practice these would just be ordinary numbers, obtained by simply numbering the lines as they are written.

3.1.1 Conjunction

We begin with the "easiest" rules: if each of two facts is true, then so is their conjunct, and if a conjunction is true, so is each of the two facts separately. This simple observation is the basis of the introduction and elimination rules for conjunction. Since we don't care which conjunct comes first, we have two variants of $(\wedge I)$, and since we can conclude either conjunct if the conjunction is true, we have two variants of $(\wedge E)$.

Conjunction introduction ($\wedge I$):

$$\begin{array}{l|ll} \vdots & \vdots & \\ m & p & \\ \vdots & \vdots & \\ n & q & \\ \vdots & \vdots & \\ k & p \wedge q & (\wedge\mathrm{I}),\ m,\ n \end{array} \qquad\qquad \begin{array}{l|ll} \vdots & \vdots & \\ m & q & \\ \vdots & \vdots & \\ n & p & \\ \vdots & \vdots & \\ k & p \wedge q & (\wedge\mathrm{I}),\ m,\ n \end{array}$$

Conjunction elimination ($\wedge E$)

$$\begin{array}{l|ll} \vdots & \vdots & \\ m & p \wedge q & \\ \vdots & \vdots & \\ n & p & (\wedge\mathrm{E}),\ m \end{array} \qquad\qquad \begin{array}{l|ll} \vdots & \vdots & \\ m & p \wedge q & \\ \vdots & \vdots & \\ n & q & (\wedge\mathrm{E}),\ m \end{array}$$

Note that the elimination rules correspond to the tautologies $(p \wedge q) \longrightarrow p$, $(p \wedge q) \longrightarrow q$.

In these rules, all the formulas occur at the same "level", as indicated by the fact that they are all aligned along the same vertical line.

Here are two simple examples. The first simply shows $A \wedge (B \wedge C) \vdash C$, using $(\wedge E)$ twice. The second shows $A \wedge B \vdash B \wedge A$, using both $\wedge$ rules, elimination to break $A \wedge B$ into its constituents, and then introduction to put them back together again, but this time in the reverse order.

$$\begin{array}{l|ll} 1 & \underline{A \wedge (B \wedge C)} & \\ 2 & B \wedge C & (\wedge\mathrm{E}),\ 1 \\ 3 & C & (\wedge\mathrm{E}),\ 2 \end{array} \qquad\qquad \begin{array}{l|ll} 1 & \underline{A \wedge B} & \\ 2 & A & (\wedge\mathrm{E}),\ 1 \\ 3 & B & (\wedge\mathrm{E}),\ 1 \\ 4 & B \wedge A & (\wedge\mathrm{I}),\ 2,\ 3 \end{array}$$

3.1.2 Disjunction

The introduction rule for $\vee$ is simple enough: if a statement is true, then so is any disjunction which contains it as a disjunct. There are two variants of the rule, depending on whether we are looking at the first or second disjunct. These correspond to the tautologies $p \longrightarrow (p \vee q)$ and $q \longrightarrow (p \vee q)$.

Disjunction Introduction ($\vee I$)

$$\begin{array}{l|ll} \vdots & \vdots & \\ m & p & \\ \vdots & \vdots & \\ n & p \vee q & (\vee\mathrm{I}),\ m \end{array} \qquad\qquad \begin{array}{l|ll} \vdots & \vdots & \\ m & q & \\ \vdots & \vdots & \\ n & p \vee q & (\vee\mathrm{I}),\ m \end{array}$$

Here is a simple example; we derive $A \vdash A \wedge (A \vee B)$ by using $(\vee I)$ to derive the necessary disjunction, then $(\wedge I)$ to derive the required conjunction.

$$\begin{array}{l|ll} 1 & \underline{A} & \\ 2 & A \vee B & (\vee\mathrm{I}),\ 1 \\ 3 & A \wedge (A \vee B) & (\wedge\mathrm{I}),\ 1,\ 2 \end{array}$$

The elimination rule is rather trickier, however. This is "proof by cases", where one proves an assertion with a disjunctive premise by examining all of the cases obtained by using one of the disjuncts as the premise instead. We saw lots of arguments of that form when we considered knights and knaves: often in those problems we obtained our solution by just this sort of considering all the cases relevant to the problem. This rule corresponds to the tautology $[(p \longrightarrow r) \wedge (q \longrightarrow r)] \longrightarrow [(p \vee q) \longrightarrow r]$. (Exercise: verify (1) that this is a tautology, and (2) that is really does correspond to $(\vee E)$. In fact, this is an equivalence: $[(p \longrightarrow r) \wedge (q \longrightarrow r)] \leftrightarrow [(p \vee q) \longrightarrow r]$, the reverse direction following from $(\vee I)$.) Here is the rule:

Disjunction Elimination $(\vee E)$

$$\begin{array}{l|l}
\vdots & \vdots \\
\ell & p \vee q \\
\vdots & \vdots \\
m & \quad \vert\, \underline{p} \\
\vdots & \quad \vert\, \vdots \\
n & \quad \vert\, r \\
n+1 & \quad \vert\, \underline{q} \\
\vdots & \quad \vert\, \vdots \\
k & \quad \vert\, r \\
k+1 & r \quad (\vee\text{E}),\ \ell,\ m\text{–}n,\ (n+1)\text{–}k
\end{array}$$

Note that each case is a subderivation: a little derivation in its own right. Each of these may use all the premises of the main derivation one level up, and all of the statements derived from those premises, up till the point where the case subderivation began. However, each of the case subderivations is independent of the other, and you may not use statements proved in one when working on the other.

Each subderivation is marked by its own vertical line, which marks the "scope" of the subderivation, and its own horizontal line, which marks the premise (or hypothesis) for that subderivation. That premise is added to the others in the main derivation, and indeed, in the subderivation, any statements assumed or proven in the main derivation may be used in the subderivation (but not *vice versa*: the only thing proven in the subderivation that may be "lifted" back to the main derivation is the conclusion r which the $(\vee E)$ rule explicitly says you can bring back to the main derivation). The scope (vertical) line indicates the scope of the subderivation's premise: everything written to the right of that line depends on that premise as well as the other premises of the main derivation. But once one leaves the subderivation (once one is no longer to the right of its scope line), its additional premise, and all that was derived from it, is no longer available for further deduction (apart from whatever was brought back to the main derivation by the $(\vee E)$ rule).

Example: We derive $A \vee B \vdash B \vee A$ by eliminating the $\vee$ in the premise, and deriving the conclusion in each of the two cases.

$$\begin{array}{l|ll}
1 & \underline{A \vee B} & \\
2 & \quad \vert\, \underline{A} & \\
3 & \quad \vert\, B \vee A & (\vee\text{I}),\ 2 \\
4 & \quad \vert\, \underline{B} & \\
5 & \quad \vert\, B \vee A & (\vee\text{I}),\ 4 \\
6 & B \vee A & (\vee\text{E}),\ 1,\ 2\text{–}3,\ 4\text{–}5
\end{array}$$

3.1.3 Implication

This is another rule which requires a subderivation: in order to prove an implication, we must prove that the premise really does entail the conclusion. We construct a subderivation with the new, temporary, assumption, and derive the necessary conclusion in that subderivation. This allows us to go back to the main derivation one level up and conclude that we have proved the implication. Here is the rule:

Implication Introduction ($\longrightarrow I$)

$$\begin{array}{l|lll}
\vdots & \vdots & & \\
m & & \multicolumn{1}{|l}{\underline{p}} & \\
\vdots & & \multicolumn{1}{|l}{\vdots} & \\
n & & \multicolumn{1}{|l}{q} & \\
n+1 & \multicolumn{2}{l}{p \longrightarrow q} & (\longrightarrow\text{I}),\ m\text{–}n
\end{array}$$

Here is an example. We derive $A \vdash B \longrightarrow (A \wedge B)$.

$$\begin{array}{r|lll}
1 & \multicolumn{2}{l}{\underline{A}} & \\
2 & & \multicolumn{1}{|l}{\underline{B}} & \\
3 & & \multicolumn{1}{|l}{A \wedge B} & (\wedge\text{I}),\ 1,\ 2 \\
4 & \multicolumn{2}{l}{B \longrightarrow (A \wedge B)} & (\longrightarrow\text{I}),\ 2\text{–}3
\end{array}$$

The elimination rule for implication is a venerable friend to logicians; known as *Modus Ponens*, it is one of the logical laws well known to the ancient Greek philosophers: if both an implication and its premise are true, so must its conclusion be true also. This corresponds to the tautology $[p \wedge (p \longrightarrow q)] \longrightarrow q$. We have two variants, since we don't care what order the two necessary parts occur in. One thing to be careful about, however: we need **both** the implication and its premise, before we can derive the conclusion. **Do not** try to use this rule when you have just one part. For example, merely $p \longrightarrow q$ being true is *not* enough to justify the claim that q is true; p must also be true.

Implication Elimination ($\longrightarrow E$)

$$\begin{array}{l|ll}
\vdots & \vdots & \\
m & p & \\
\vdots & \vdots & \\
n & p \longrightarrow q & \\
\vdots & \vdots & \\
k & q & (\longrightarrow\text{E}),\ m,\ n
\end{array}
\qquad\qquad
\begin{array}{l|ll}
\vdots & \vdots & \\
m & p \longrightarrow q & \\
\vdots & \vdots & \\
n & p & \\
\vdots & \vdots & \\
k & q & (\longrightarrow\text{E}),\ m,\ n
\end{array}$$

Here is an example. We derive $A \longrightarrow B \vdash (B \longrightarrow C) \longrightarrow (A \longrightarrow C)$. Note we have a subsubderivation nested inside a subderivation—in principle, one may have as many levels of nested subderivations as one wishes or needs.

$$\begin{array}{r|llll}
1 & \multicolumn{3}{l}{\underline{A \longrightarrow B}} & \\
2 & & \multicolumn{2}{|l}{\underline{B \longrightarrow C}} & \\
3 & & \multicolumn{1}{|l}{} & \multicolumn{1}{|l}{\underline{A}} & \\
4 & & \multicolumn{1}{|l}{} & \multicolumn{1}{|l}{B} & (\longrightarrow\text{E}),\ 1,\ 3 \\
5 & & \multicolumn{1}{|l}{} & \multicolumn{1}{|l}{C} & (\longrightarrow\text{E}),\ 2,\ 4 \\
6 & & \multicolumn{2}{|l}{A \longrightarrow C} & (\longrightarrow\text{I}),\ 3\text{–}5 \\
7 & \multicolumn{3}{l}{(B \longrightarrow C) \longrightarrow (A \longrightarrow C)} & (\longrightarrow\text{I}),\ 2\text{–}6
\end{array}$$

Remark: We do not have any deduction rules for the biconditional $\leftrightarrow$, preferring to treat $p \leftrightarrow q$ as merely an abbreviation of $(p \longrightarrow q) \wedge (q \longrightarrow p)$. Of course, one could create appropriate rules, but they wouldn't really save us very much effort, and so hardly seem worth the bother.

3.1.4 Negation

The rules for negation parallel the rules for implication, because of the tautological equivalence $\neg p \leftrightarrow (p \longrightarrow \bot)$.

The introduction rule for negation says you can prove a statement false by assuming it is true and then deriving a contradiction (this is a frequent strategy mathematicians use in proving things, and goes by the name "proof by contradiction").

Negation introduction $(\neg I)$

$$
\begin{array}{l|ll}
\vdots & \vdots & \\
\begin{array}{l} m \\ \vdots \\ n \end{array} & \quad\begin{array}{|l} p \\ \hline \vdots \\ \bot \end{array} & \\
n+1 & \neg p & (\neg\text{I}),\ m\text{–}n
\end{array}
$$

The elimination rule simply says that if you've proved both a statement and its negation, then in fact you've arrived at a contradiction. This corresponds to the tautology $(p \wedge \neg p) \longrightarrow \bot$.

Negation elimination $(\neg E)$

$$
\begin{array}{l|ll}
\vdots & \vdots & \\
m & p & \\
\vdots & \vdots & \\
n & \neg p & \\
\vdots & \vdots & \\
k & \bot & (\neg\text{E}),\ m,\ n
\end{array}
\qquad\qquad
\begin{array}{l|ll}
\vdots & \vdots & \\
m & \neg p & \\
\vdots & \vdots & \\
n & p & \\
\vdots & \vdots & \\
k & \bot & (\neg\text{E}),\ m,\ n
\end{array}
$$

Here is an example of a simple derivation using both of these negation rules. We shall derive $A \longrightarrow B \vdash \neg B \longrightarrow \neg A$ by setting ourselves up to arrive at the conclusion $\neg B \longrightarrow \neg A$ *via* the $(\longrightarrow I)$ rule, with a subderivation whose premise is $\neg B$. But to derive $\neg A$, which we must, we use $(\neg I)$, and so set up a subsubderivation with premise A, aiming to derive a contradiction $\bot$. That we manage with $(\longrightarrow E)$ and $(\neg E)$.

$$
\begin{array}{l|ll}
1 & A \longrightarrow B & \\ \hline
\begin{array}{l} 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{array}{|l} \neg B \\ \hline \begin{array}{|l} A \\ \hline B \\ \bot \end{array} \\ \neg A \end{array} & \begin{array}{l} \\ \\ (\longrightarrow\text{E}),\ 1,\ 3 \\ (\neg\text{E}),\ 2,\ 4 \\ (\neg\text{I}),\ 3\text{–}5 \end{array} \\
7 & \neg B \longrightarrow \neg A & (\longrightarrow\text{I}),\ 2\text{–}6
\end{array}
$$

Here is another example. I leave it to you to see why each step is “natural”. $P \longrightarrow Q, P \longrightarrow \neg Q \vdash \neg P$.

$$\begin{array}{rlll}
1 & \vert\ P \longrightarrow Q & & \\
2 & \vert\ \underline{P \longrightarrow \neg Q} & & \\
3 & \vert\ \quad \vert\ \underline{P} & & \\
4 & \vert\ \quad \vert\ Q & & (\longrightarrow\text{E}),\ 1,\ 3 \\
5 & \vert\ \quad \vert\ \neg Q & & (\longrightarrow\text{E}),\ 2,\ 3 \\
6 & \vert\ \quad \vert\ \bot & & (\neg\text{E}),\ 4,\ 5 \\
7 & \vert\ \neg P & & (\neg\text{I}),\ 3\text{–}6
\end{array}$$

We have two other rules which involve negation: the first expresses the property of propositional logic that a contradiction entails anything. We’ve discussed this before (I referred to this as the “anything goes” or “all bets are off” principle, when an assumption in an argument is in fact false); it may still seem strange to you, but if so, you have to work a bit more on getting used to it! This rule corresponds to the tautology $\bot \longrightarrow p$.

Contradiction Elimination ($\bot E$)

$$\begin{array}{rll}
\vdots & \vert\ \vdots & \\
m & \vert\ \bot & \\
\vdots & \vert\ \vdots & \\
n & \vert\ p & (\bot\text{E}),\ m
\end{array}$$

Example: $\neg P \vdash P \longrightarrow Q$

$$\begin{array}{rll}
1 & \vert\ \underline{\neg P} & \\
2 & \vert\ \quad \vert\ \underline{P} & \\
3 & \vert\ \quad \vert\ \bot & (\neg\text{E}),\ 1,\ 2 \\
4 & \vert\ \quad \vert\ Q & (\bot\text{E}),\ 3 \\
5 & \vert\ P \longrightarrow Q & (\longrightarrow\text{I}),\ 2\text{–}4
\end{array}$$

The last negation rule expresses the property of (classical) propositional logic that there are only two truth values, and so if a statement is not false, it must then be true. This corresponds to the tautology $p \leftrightarrow \neg\neg p$.

Double Negation Elimination ($\neg\neg E$)

$$\begin{array}{rll}
\vdots & \vert\ \vdots & \\
m & \vert\ \neg\neg p & \\
\vdots & \vert\ \vdots & \\
n & \vert\ p & (\neg\neg\text{E}),\ m
\end{array}$$

Example: A typical use of ($\neg\neg E$) is in a second form of “proof by contradiction”, where we assume $\neg p$, prove a contradiction $\bot$, and then conclude $\neg\neg p$ (by ($\neg I$)), and hence p, by ($\neg\neg E$). For example, the following derivation shows $p \vee \neg p$ is a tautology (because it shows $\vdash p \vee \neg p$).

```
1 │ │ ¬(p ∨ ¬p)
  │ ├──────────
2 │ │ │ p
  │ │ ├───
3 │ │ │ p ∨ ¬p   (∨I), 2
4 │ │ │ ⊥        (¬E), 1, 3
5 │ │ ¬p         (¬I), 2–4
6 │ │ p ∨ ¬p     (∨I), 5
7 │ │ ⊥          (¬E), 1, 6
8 │ ¬¬(p ∨ ¬p)   (¬I), 1–7
9 │ p ∨ ¬p       (¬¬E), 8
```

3.1.5 Repetition

Finally, we have a "bookkeeping" rule, which says we can repeat any premise or any statement we have proved from the premises later in the derivation. There are some restrictions, which we shall discuss after seeing the rule.

Repetition (R)

```
m │ p
⋮ │ ⋮
⋮ │ ··· │ ⋮
  │     ├───
n │     │ p   (R), m
```

Example: Here is a simple derivation of $p \vdash q \longrightarrow p$.

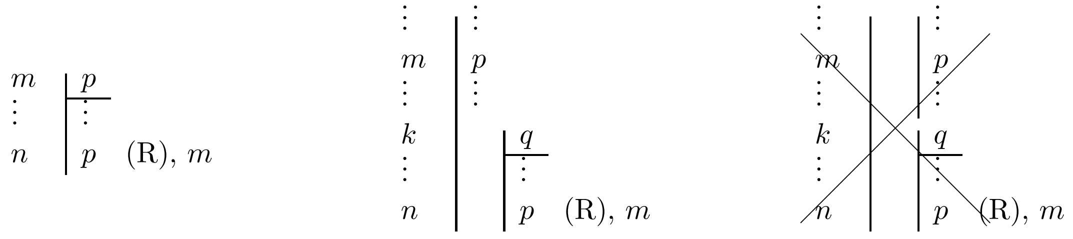
```
1 │ p
  ├───
2 │ │ q
  │ ├───
3 │ │ p      (R), 1
4 │ q ⟶ p   (⟶I), 2–3
```

The Repetition rule needs some care. It is just a bookkeeping rule, and we could actually do without it, although it does help at times to make the derivation clearer. But it must be used only where appropriate. The idea is that once a formula has been proved (or once it is stated as a premise), one may use it at any later stage in the derivation where it is "visible", meaning within the same subderivation, including within any subderivations that appear in that same subderivation. However, instances of a formula within separate (not nested) subderivations are not visible to each other.

Formally, we may define this as follows. Suppose p appears in line m (either as a premise or as a formula already derived); then one may repeat p at line n if $m < n$ and every vertical line (or spline) from line m continues without interruption to line n. This last condition just says that you may repeat p as long as you stay within the same subderivation, but this does not permit repetition between distinct, unconnected subderivations.

It is worth noting that one situation where we may *not* use the repetition rule is in argument by cases, *i.e.* the ($\vee E$) rule: one may not repeat a formula from one case inside the other case, although one may repeat formulas from the main derivation in either case.

For example, in the following derivations, the first two uses of repetition are valid, but the third is **not**, because the first p is in a subderivation unconnected to the subderivation where the repeated p occurs.



3.1.6 Remarks

There are several points to keep in mind when constructing proofs. General strategies will be pointed out as we do some examples, but for now, here are a few things to remember.

- The derivation rules we have developed in this chapter are summarised in Table 3.1.

- You may only use the rules given in Table 3.1. Don't just write down something "because it is obvious"—it is precisely the formal nature of the proofs that gives them their power. They justify the claim that with these few rules, **all** valid arguments may be constructed (at least all such expressible in propositional logic). In other words, these rules are complete for classical propositional logic.

- There is one way one may relax the previous *dictum*: if you *prove* a general entailment, you could then use it as a "derived rule"; a rule not in our system, but one which has already been shown to be valid. This is rather like using defined connectives (like ↔), in that we could always expand the derived rule to its full derivation. Generally I don't suggest you use derived rules; the effort in learning them isn't worth the effort for the very few times you'll find them really useful. In any event, if you do use a derived rule (*e.g.* in a test), you will be required to produce the derivation for the derived rule as well as the derivation in which you use it. There is an example of a derivation using a derived rule in the solutions to Exercise 3.3.2 (#4).

- Notice that each rule serves a very specific purpose. The elimination rules are used when working with a premise, or some intermediate formula you have derived from the premises; these E rules tell you what you can do with a premise, how to break it into its component parts, how to proceed at that point in a derivation. On the other hand, the introduction rules are used when working with a conclusion, or some intermediate conclusion you think will help you get to the end; these I rules help you reach a goal (the conclusion) you are trying to reach, they tell you how you can put component parts together into a compound formula.

  So, the general strategy in making derivations will include two ways to proceed: top-down, working with the premises towards your goal, and bottom-up, starting with the conclusion and seeing what is needed to arrive at that conclusion. In each case the idea will be to examine the formulas (WFFs) you have in front of you (initially just the premises and the conclusion of the argument), for each formula, parse it to see what its main connective is, and then use the introduction or elimination rule for that connective (depending on whether you are working with the formula as a premise or as a conclusion). You will see this in the examples; I will draw attention to this frequently, but you should try to see this principle even when I haven't explicitly discussed it.

$$\begin{array}{l|l}
\vdots & \vdots \\
m & p \\
\vdots & \vdots \\
n & q \\
\vdots & \vdots \\
k & p \wedge q \quad (\wedge\mathrm{I}),\ m,\ n
\end{array}
\qquad
\begin{array}{l|l}
\vdots & \vdots \\
m & q \\
\vdots & \vdots \\
n & p \\
\vdots & \vdots \\
k & p \wedge q \quad (\wedge\mathrm{I}),\ m,\ n
\end{array}$$

Conjunction introduction ($\wedge I$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & p \wedge q \\
\vdots & \vdots \\
n & p \qquad (\wedge\mathrm{E}),\ m
\end{array}
\qquad
\begin{array}{l|l}
\vdots & \vdots \\
m & p \wedge q \\
\vdots & \vdots \\
n & q \qquad (\wedge\mathrm{E}),\ m
\end{array}$$

Conjunction elimination ($\wedge E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & p \\
\vdots & \vdots \\
n & p \vee q \quad (\vee\mathrm{I}),\ m
\end{array}
\qquad
\begin{array}{l|l}
\vdots & \vdots \\
m & q \\
\vdots & \vdots \\
n & p \vee q \quad (\vee\mathrm{I}),\ m
\end{array}$$

Disjunction Introduction ($\vee I$)

$$\begin{array}{l|l}
\vdots & \vdots \\
\ell & p \vee q \\
\vdots & \vdots \\
m & \quad \vert\, \underline{p} \\
\vdots & \quad \vert\, \vdots \\
n & \quad \vert\, r \\
n+1 & \quad \vert\, \underline{q} \\
\vdots & \quad \vert\, \vdots \\
k & \quad \vert\, r \\
k+1 & r \qquad (\vee\mathrm{E}),\ \ell,\ m\text{–}n,\ (n+1)\text{–}k
\end{array}$$

Disjunction Elimination ($\vee E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & \quad \vert\, \underline{p} \\
\vdots & \quad \vert\, \vdots \\
n & \quad \vert\, q \\
n+1 & p \longrightarrow q \quad (\longrightarrow\mathrm{I}),\ m\text{–}n
\end{array}$$

Implication Introduction ($\longrightarrow I$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & p \\
\vdots & \vdots \\
n & p \longrightarrow q \\
\vdots & \vdots \\
k & q \quad (\longrightarrow\mathrm{E}),\ m,\ n
\end{array}
\qquad
\begin{array}{l|l}
\vdots & \vdots \\
m & p \longrightarrow q \\
\vdots & \vdots \\
n & p \\
\vdots & \vdots \\
k & q \quad (\longrightarrow\mathrm{E}),\ m,\ n
\end{array}$$

Implication Elimination ($\longrightarrow E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & \quad \vert\, \underline{p} \\
\vdots & \quad \vert\, \vdots \\
n & \quad \vert\, \bot \\
n+1 & \neg p \qquad (\neg\mathrm{I}),\ m\text{–}n
\end{array}$$

Negation introduction ($\neg I$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & p \\
\vdots & \vdots \\
n & \neg p \\
\vdots & \vdots \\
k & \bot \quad (\neg\mathrm{E}),\ m,\ n
\end{array}
\qquad
\begin{array}{l|l}
\vdots & \vdots \\
m & \neg p \\
\vdots & \vdots \\
n & p \\
\vdots & \vdots \\
k & \bot \quad (\neg\mathrm{E}),\ m,\ n
\end{array}$$

Negation elimination ($\neg E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & \bot \\
\vdots & \vdots \\
n & p \quad (\bot\mathrm{E}),\ m
\end{array}$$

Contradiction Elimination ($\bot E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & \neg\neg p \\
\vdots & \vdots \\
n & p \qquad (\neg\neg\mathrm{E}),\ m
\end{array}$$

Double Negation Elimination ($\neg\neg E$)

$$\begin{array}{l|l}
m & p \\
\vdots & \vdots \\
\vdots & \cdots \vert\, \vdots \\
n & \vert\, \overline{p} \quad (\mathrm{R}),\ m
\end{array}$$

Repetition (R)

Table 3.1: The Fitch-style presentation of the natural deduction rules

Frequently the idea will be to work bottom-up—use the structure of the conclusion to give you hints as to what rules to use. In other circumstances, top-down will work better; experience will help here (the more derivations you try to construct, the better you'll get at it). In many derivations, the structure of the formulas you start with will virtually "force" you to a certain derivation structure; in other examples, you will have some choice how to proceed (choose this or that formula to work on).

- Note that we can only prove tautologies, in effect. This is because all the rules preserve truth. In other words, these rules are consistent. So if we apply a rule to true premises, we'll get a true conclusion. So, any argument constructed from the rules will be valid, and so will correspond to a tautology (just form the implication whose premise is the conjunction of all the premises of the argument, and whose conclusion is the conclusion of the argument).

  So, if you have any reason to doubt that an argument is tautologous, before trying to prove it, it might be a good idea to see if it can be disproved. We shall develop a simple technique for that later, but in principle truth tables could do that for us. (Don't worry—I won't ask you to prove something that cannot be proved! Well, not yet, anyway . . .)

3.2 Examples

These rules are pretty abstract, so we illustrate them with several examples. Remember that the point in each case is to construct a derivation using *only* these rules and the premises stated to arrive at the stated conclusion, in effect constructing a valid argument for that conclusion from the premises.

Our first examples are the same six derived rules we ended the last chapter with. Detailed comments will accompany these derivations—you should refer to the comments while studying the derivations themselves. Of course, the idea is to learn how to create derivations yourself, so try to anticipate the steps as much as you can.

Disjunctive Syllogism:

$p \vee q, \neg q \vdash p$

1	$\mid p \vee q$	
2	$\mid \underline{\neg q}$	
3	$\mid \mid \underline{p}$	
4	$\mid \mid p$	(R), 3
5	$\mid \mid \underline{q}$	
6	$\mid \mid \bot$	(¬E), 2, 5
7	$\mid \mid p$	(⊥E), 6
8	$\mid p$	(∨E), 1, 3–4, 5–7

First, look at the premises ($p \vee q, \neg q$) and the desired conclusion (p): to build a derivation, either we can start top-down, with one of the premises (a $\vee$ and a $\neg$ formula) and use an elimination rule, or we can start bottom-up, with the conclusion (an atomic formula) and use an introduction rule (of course, there is no appropriate introduction rule for an atomic formula, so this isn't a real consideration). In this example this means we should expect to work top-down, considering the rules ($\vee E$), ($\neg E$), and initially *no other rules.* This simplifies things a lot! Let's look at these two rules.

To use ($\neg E$) we need a pair of formulas, one the negation of the other; we do not have that yet, so we can't use ($\neg E$) yet. So we consider using ($\vee E$) with the premise $p \vee q$; this means a proof by cases. Since we have no other obvious options at present, we start our proof with two cases, *i.e.* two subderivations, one with the hypothesis p, the other with hypothesis q. In each case we shall aim for our ultimate conclusion p (we might as well go for broke!).

The first case argument is simple: we have supposed p, so it's easy to conclude p (we just use the repetition rule (R)). The second case (with hypothesis q) is a little more bothersome. But now we recall why we initially didn't use ($\neg E$): to do so, we needed both a formula and its negation.

But now that's exactly what we have: $\neg q$ and q. So we use $(\neg E)$ to get $\bot$. Next we use a trick we'll often find useful: once we get $\bot$, we can derive any formula we like, in particular, whatever formula you're aiming for. Here we want p, so we use $(\bot E)$ to get it.

Finally, we write down the conclusion we may derive from this cases argument. Be sure to justify each step with the rule used, citing the lines that are needed for that rule. For instance, our final conclusion was the result of a cases argument, *i.e.* the $(\vee E)$ rule, which requires a $\vee$ formula (on line 1), and two case subderivations (on lines 3–4, and on lines 5–7).

Notice that with some rules we need only cite one line (because they take only a single premise); (R) and $(\bot E)$ are such rules, and when we justify their use, we only reference one line. Some rules require us to cite two lines, corresponding to the two premises they need; $(\neg E)$ is such a rule, and we referenced the two relevant lines when we used it. Finally, some rules require us to cite not only single lines (corresponding to premises) but also line ranges (corresponding to subderivations); $(\vee E)$ is such a rule, and we were careful to include all that information in the justification.

Modus Tollens:
$p \longrightarrow q, \neg q \vdash \neg p$

$$\begin{array}{r|l l}
1 & p \longrightarrow q & \\
2 & \underline{\neg q} & \\
3 & \quad \vert\ \underline{p} & \\
4 & \quad \vert\ q & (\longrightarrow\text{E}),\ 1,\ 3 \\
5 & \quad \vert\ \bot & (\neg\text{E}),\ 2,\ 4 \\
6 & \neg p & (\neg\text{I}),\ 3\text{–}5
\end{array}$$

Again we look at the premises and conclusion: a $\longrightarrow$ formula and a $\neg$ formula as premises suggest the $(\longrightarrow E)$ and $(\neg E)$ rules; a $\neg$ formula as conclusion suggests the $(\neg I)$ rule. But what do these rules need?

$(\longrightarrow E)$ needs an implication ($p \longrightarrow q$, which we have) and its premise (p, which we don't have), so we cannot use that yet. $(\neg E)$ similarly needs two elements, of which we only have one (what is missing?). So we consider working bottom-up, using $(\neg I)$ to conclude $\neg p$: this means setting up a subderivation with premise p, with the intention of proving $\bot$ in that subderivation.

But once we have added p as a new premise, we are able to use the $(\longrightarrow E)$ rule we were unable to use before. So we do use it to get q. Note we justify this by mentioning not only the rule, but the two line numbers where the necessary data may be found. Next we notice that now we have the necessary data to use $(\neg E)$, which we did not have before. Now the subderivation has accomplished its task, and so we are in a position to finish using the $(\neg I)$ rule we started with. This puts our final conclusion back in the main derivation, and with the necessary justification in place, we're done.

Hypothetical Syllogism:
$p \longrightarrow q, q \longrightarrow r \vdash p \longrightarrow r$

$$\begin{array}{r|l l}
1 & p \longrightarrow q & \\
2 & \underline{q \longrightarrow r} & \\
3 & \quad \vert\ \underline{p} & \\
4 & \quad \vert\ q & (\longrightarrow\text{E}),\ 1,\ 3 \\
5 & \quad \vert\ r & (\longrightarrow\text{E}),\ 2,\ 4 \\
6 & p \longrightarrow r & (\longrightarrow\text{I}),\ 3\text{–}5
\end{array}$$

This is rather similar to the previous derivation, and this is no coincidence: remember that $\neg p$ is equivalent to $p \longrightarrow \bot$, and so the only difference between the two proofs is that we use r instead of $\bot$ (and so change the $\neg$ rules to $\longrightarrow$ rules).

Constructive Dilemma:

$p \longrightarrow q, r \longrightarrow s, p \vee r \vdash q \vee s$

1	$p \longrightarrow q$	
2	$r \longrightarrow s$	
3	$p \vee r$	
4	$\quad p$	
5	$\quad q$	($\longrightarrow$E), 1, 4
6	$\quad q \vee s$	($\vee$I), 5
7	$\quad r$	
8	$\quad s$	($\longrightarrow$E), 2, 7
9	$\quad q \vee s$	($\vee$I), 8
10	$q \vee s$	($\vee$E), 3, 4–6, 7–9

Again, look at the premises (and the corresponding elimination rules) and the conclusion (and the corresponding introduction rule). The $\longrightarrow$ elimination rules need the extra data of the implication-antecedents, which are not available to us (yet!), so we are forced to use the $(\vee E)$ rule, proof by cases. As with [DS], we set up the cases with their temporary premises p and r (since we are applying the rule to the disjunctive premise $p \vee r$).

In each case subderivation we want to derive $q \vee s$ (again, going for our ultimate conclusion at once); there is not much doubt as to what we should do, since we have two implications we can eliminate (use as premises) now, and our conclusion gives us a disjunction to introduce. In other words, we use $(\longrightarrow E)$ (once in each case) and $(\vee I)$ (again, once in each case).

Notice a paradox of proof-construction in these examples: we started conceptually with the $(\vee E)$ elimination rule, but when we look at the finished derivation, that rule seems to be the last one used. This is typical of the bottom-up method of constructing derivations; as you saw in the first example, it is also typical when you start a derivation with $(\vee E)$. Our next example shows that one sometimes really does start with the rule one starts with(!).

Equivalence Rule:

$p \leftrightarrow q, q \longrightarrow r \vdash p \longrightarrow r$

1	$(p \longrightarrow q) \wedge (q \longrightarrow p)$	
2	$q \longrightarrow r$	
3	$p \longrightarrow q$	($\wedge$E), 1
4	$\quad p$	
5	$\quad q$	($\longrightarrow$E), 3, 4
6	$\quad r$	($\longrightarrow$E), 2, 5
7	$p \longrightarrow r$	($\longrightarrow$I), 4–6

This derivation starts with the replacement of the biconditional with its definition (the premise in line 1). Our first step is to use $(\wedge E)$ to get the implication on line 3. After that, we just duplicate the steps used in the derivation of [HS]; notice that lines 2 and 3 just duplicate the premises of [HS].

Disjunctive Syllogism II:

$\neg(p \wedge q), p \vdash \neg q$

1	$\neg(p \wedge q)$	
2	p	
3	$\quad q$	
4	$\quad p \wedge q$	($\wedge$I), 2, 3
5	$\quad \bot$	($\neg$E), 1, 4
6	$\neg q$	($\neg$I), 3–5

In this case, we start with the $(\neg I)$ rule, which more or less finishes the job, since we now have both p and q, and so $p \wedge q$.

A word problem

Here is an argument:

> Unless the Vulcans leave the Federation and join the Romulans, the Klingons will attack the Romulans. If the Klingons attack the Romulans, the Romulans will surrender and join the Klingons to attack the Federation. If the Klingons and Romulans together attack the Federation, the Federation will be destroyed. Therefore, if the Vulcans remain in the Federation, it will be destroyed.

Translate it into a formal argument, and prove it is valid. Use the following abbreviations: V = The Vulcans will leave the Federation; R = The Vulcans will join the Romulans; K = The Klingons will attack the Romulans; S = The Romulans will surrender; J = The Romulans will join the Klingons; A = The Romulans and Klingons will attack the Federation; D = The Federation will be destroyed.

Solution: It's pretty straightforward to translate the English, though you might have been tricked by "unless": but if you think of the truth-functional properties of "p unless q", you will soon realise that this means either q or, if not, then p: $q \vee (\neg q \longrightarrow p)$, which simply means $p \vee q$.[1] You might even think of translating this as $p \vee (\neg p \longrightarrow q)$, but it's easy enough to check that this is also equivalent to $p \vee q$, so you might as well use the simpler form.

So we get this as the translation:

$$(V \wedge R) \vee K, K \longrightarrow ((S \wedge J) \wedge A), A \longrightarrow D \vdash \neg V \longrightarrow D$$

And so we set up the following derivation (with comments).

$$
\begin{array}{rll}
1 & \vert\ (V \wedge R) \vee K & \\
2 & \vert\ K \longrightarrow ((S \wedge J) \wedge A) & \\
3 & \vert\ \underline{A \longrightarrow D} & \\
4 & \vert\ \vert\ \underline{\neg V} & \\
5 & \vert\ \vert\ \vert\ \underline{V \wedge R} & \\
6 & \vert\ \vert\ \vert\ V & (\wedge\mathrm{E}),\ 5 \\
7 & \vert\ \vert\ \vert\ \bot & (\neg\mathrm{E}),\ 4,\ 6 \\
8 & \vert\ \vert\ \vert\ D & (\bot\mathrm{E}),\ 7 \\
9 & \vert\ \vert\ \vert\ \underline{K} & \\
10 & \vert\ \vert\ \vert\ (S \wedge J) \wedge A & (\longrightarrow\mathrm{E}),\ 2,\ 9 \\
11 & \vert\ \vert\ \vert\ A & (\wedge\mathrm{E}),\ 10 \\
12 & \vert\ \vert\ \vert\ D & (\longrightarrow\mathrm{E}),\ 3,\ 11 \\
13 & \vert\ \vert\ D & (\vee\mathrm{E}),\ 1,\ 5\text{–}8,\ 9\text{–}12 \\
14 & \vert\ \neg V \longrightarrow D & (\longrightarrow\mathrm{I}),\ 4\text{–}13
\end{array}
$$

Write out the 3 premises and the conclusion—leave lots of space in between.

We cannot use the premises on lines 2, 3, so we either use $(\vee E)$ ("cases") (from the premise on line 1 which is a disjunction) or $(\longrightarrow I)$ (from the conclusion, which is an implication). In other words, we have a choice here: we can work bottom-up, starting with $(\longrightarrow I)$ or we can work top-down, starting with $(\vee E)$. This is an honest choice: either strategy would work, and we would end up with different derivations depending on which we chose. Here I have opted for the simpler $(\longrightarrow I)$, working bottom-up. So we set up a subderivation with the temporary aim to prove D from premise $\neg V$. (Again, leave lots of space for what comes in between.)

Now we're forced to use cases (with premise on line 1), so we set up the two cases subderivations, lines 5 and 9, again each with D as target. In the first case, we see we can get $\bot$ from V, which gives us anything we want, including D. For the second case, having K "frees up" $(S \wedge J) \wedge A$, and so A, and so finally D, like dominoes! This gets us to the end, finishing off the cases proof and the implication introduction proof.[2]

Study these derivations and the comments carefully—you should often feel a sense of inevitability in these proofs: the structure of the premises and the conclusion really force you to the structure of the derivation. Sometimes you will have a choice (often between working bottom-up or top-down, sometimes you will have a choice of which premise to use first when proceeding top-down). Often you will be led to a derivation just by following the hints produced by the structure of the premises and conclusion. The point is this: when faced with a set of premises and a conclusion to prove,

[1] You need the $(\neg\neg E)$ rule for this.

[2] There's an "almost-animated" version of this explanation on the course web-site, making the step-by-step process clearer. Look under "Assignments and Answers".

there is a definite strategy that should get you started; you don't have to wait for inspiration or a visit from the muses.

I cannot lie to you: sometimes a proof can be very tricky, and these simple steps insufficient to crack it. But I can assure you that you won't meet many like that in this course, and the few you do usually involve the double negation rule ($\neg\neg E$). So don't panic, and try to get the hang of things. You can start with the following exercises!

3.3 Exercises

3.3.1 Some natural deduction problems

Find a Fitch-style natural deduction derivation (proof) for each of the entailments listed below. Remember that the premises are listed above the horizontal line (before the entailment sign $\vdash$ in the compact "in-line" notation), and the conclusion is the last statement of your proof (after the $\vdash$ sign in the compact notation). In the first two questions, I have given both notations. Using the Fitch-style notation, I have written the premises and the conclusion, leaving you to fill in the in-between lines necessary to make a derivation ending with the conclusion. To save space, the rest of the questions are only presented using the "in-line" entailment notation.

You will notice that some entailments have no premises—that is fine, and really just means that there is nothing above the horizontal line which usually indicates premises (and so, indeed, that line isn't necessary anymore). In fact, as you will see by looking at Examples 11 and 12, it's easy enough to "move" premises into the conclusion, using implication, so every entailment could be written without premises.

Entailments marked $\vdash^*$ require the ($\neg\neg E$) rule. This usually means you should aim to contradict the negation of what you want to prove.

1. $A \vdash A \wedge (A \vee B)$

$$\begin{array}{r|l} 1 & A \\ \hline \vdots & \vdots \\ ? & A \wedge (A \vee B) \end{array}$$

Find a proof

2. $A \wedge (A \vee B \rightarrow C) \vdash C \vee D$

$$\begin{array}{r|l} 1 & A \wedge (A \vee B \rightarrow C) \\ \hline \vdots & \vdots \\ ? & C \vee D \end{array}$$

Find a proof

3. $A, A \rightarrow B \vdash A \wedge B$
4. $A \wedge B \rightarrow C, B \rightarrow A, B \vdash C$
5. $A \rightarrow B, B \rightarrow C \vdash A \rightarrow B \wedge C$
6. $A \rightarrow (B \rightarrow C) \vdash B \rightarrow (A \rightarrow C)$
7. $A \rightarrow B \vdash \neg B \rightarrow \neg A$
8. $\vdash^* A \vee \neg A$
9. $(A \rightarrow B) \rightarrow C \vdash B \rightarrow C$
10. $(A \rightarrow B) \rightarrow C \vdash^* A \vee C$
11. $\vdash A \rightarrow \neg\neg A$
12. $\vdash^* \neg\neg A \rightarrow A$
13. $\neg(A \rightarrow B) \vdash \neg B$
14. $A \vee (A \wedge B) \vdash A$
15. $\neg(A \vee B) \vdash \neg A \wedge \neg B$
16. $\vdash \neg(A \vee B) \rightarrow \neg A \wedge \neg B$
17. $\neg A \vee \neg B \vdash \neg(A \wedge B)$
18. $\neg(A \wedge B) \vdash^* \neg A \vee \neg B$
19. $\neg A \vee B \vdash A \rightarrow B$
20. $\neg A \vdash A \rightarrow B$
21. $(A \rightarrow B) \vee (A \rightarrow C) \vdash A \rightarrow B \vee C$
22. $A \rightarrow B \vee C \vdash^* (A \rightarrow B) \vee (A \rightarrow C)$
23. $A \vee B \rightarrow C \vdash (A \rightarrow C) \wedge (B \rightarrow C)$
24. $(A \rightarrow C) \wedge (B \rightarrow C) \vdash A \vee B \rightarrow C$
25. $A \rightarrow B \wedge C \vdash (A \rightarrow B) \wedge (A \rightarrow C)$
26. $(A \rightarrow B) \wedge (A \rightarrow C) \vdash A \rightarrow B \wedge C$

27. $(A \to C) \vee (B \to C) \vdash A \wedge B \to C$

28. $B \wedge C \to A\ ,\ \neg A \to C \vdash^* (C \to B) \to A$

29. $\neg(A \wedge \neg B) \vdash^* A \to B$

30. $A \to B \vdash \neg(A \wedge \neg B)$

31. $A \vee B \vdash \neg B \to (C \to A)$

32. $(B \to A) \to A \vdash^* A \vee B$

33. $(A \to B) \vee C\ ,\ A \to \neg C \vdash (B \to C) \to \neg A$

3.3.2 Some natural deductions from arguments

Translate each of the following into a formal argument, and prove it is valid. (Following the exercises, I have actually done the translations for you—try them yourself first, and then compare your answers with mine. You may use whatever abbreviations you like for the statements, but I have used some pretty obvious ones in my translations.)

1. If life is a carnival, then I'm a clown or a trapeze artist. But life isn't a carnival if there aren't any balloons, and there aren't any balloons if I'm a clown. So, if life is a carnival, then I'm a trapeze artist.

2. Spring has sprung, and the flowers are blooming. If the flowers are blooming, the bees are happy. If the bees are happy but aren't making honey, then spring hasn't sprung. So the bees are making honey.

3. Albert is a Liberal only if Bruce or Carol is. If Bruce is a Liberal, so are Deirdre and Ethel. If Deirdre is a Liberal, then Ethel is a Liberal only if Freda is; but Freda and Albert aren't both Liberals. So Albert is a Liberal only if Carol is.

4. Ladies and gentlemen: either my client has an alibi for this crime, or he is too stupid to have committed it; and anyway he never knew the victim. If he never knew the victim, then clearly he was absent from the crime scene and has no alibi. If he is too stupid to have done the crime, then either he has an alibi or he is innocent. So, either he is innocent, or you are all too incompetent at logic to be jurors.

5. If God exists, he is omnipotent and omniscient; moreover he is benevolent (provided he exists). If God can prevent evil, then if he knows evil exists, he is not benevolent if he doesn't prevent it. If he is omnipotent, he can prevent evil. And if he is omniscient, he knows evil exists if it does exist. Evil does not exist if God prevents it. However, evil does exist. So God does not exist.

6. We say an argument is **inconsistent** if its premises and conclusion, taken together (as a set of premises) allow, as a valid conclusion, the contradiction $\perp$. In symbols, $P_1, P_2, \ldots, P_n \vdash C$ **is inconsistent** if (and only if) $P_1, P_2, \ldots, P_n, C \vdash \perp$ is valid[3]. This is really just saying that the premises and the conclusion together set up a contradiction.
   Show that the following argument is inconsistent.
   Bach is popular only if Beethoven is forgotten. If Bach is unpopular and Beethoven isn't forgotten, then current musical tastes are hopeless. So current musical tastes aren't hopeless, and Beethoven isn't forgotten.

[3] Equivalently, if and only if $P_1, P_2, \ldots, P_n \vdash \neg C$ is valid.

Suggested translations of the word problems

1.

$$\begin{array}{r|l} 1 & L \to C \lor T \\ 2 & \neg B \to \neg L \\ 3 & C \to \neg B \\ \cline{2-2} \vdots & \vdots \\ ? & L \to T \end{array}$$

Find a proof

2.

$$\begin{array}{r|l} 1 & S \land B \\ 2 & B \to H \\ 3 & H \land \neg M \to \neg S \\ \cline{2-2} \vdots & \vdots \\ ? & M \end{array}$$

Find a proof*

3.

$$\begin{array}{r|l} 1 & A \to B \lor C \\ 2 & B \to D \land E \\ 3 & D \to (E \to F) \\ 4 & \neg(F \land A) \\ \cline{2-2} \vdots & \vdots \\ ? & A \to C \end{array}$$

Find a proof

4.

$$\begin{array}{r|l} 1 & (A \lor S) \land \neg K \\ 2 & \neg K \to C \land \neg A \\ 3 & S \to A \lor I \\ \cline{2-2} \vdots & \vdots \\ ? & I \lor J \end{array}$$

Find a proof

5.

$$\begin{array}{r|l} 1 & G \to P \land S \\ 2 & G \to B \\ 3 & C \to (K \to (\neg V \to \neg B)) \\ 4 & P \to C \\ 5 & S \to (E \to K) \\ 6 & V \to \neg E \\ 7 & E \\ \cline{2-2} \vdots & \vdots \\ ? & \neg G \end{array}$$

Find a proof

6.

$$\begin{array}{r|l} 1 & B \to F \\ 2 & \neg B \land \neg F \to H \\ 3 & \neg H \land \neg F \\ \cline{2-2} \vdots & \vdots \\ ? & \bot \end{array}$$

Find a proof

Now try to construct the derivations for these entailments. (The one marked with a $*$ will require the ($\neg\neg E$) rule.)

3.4 Answers to the exercises

Exercise 3.3.1

Note that there are possible variants of these proofs which would be equally correct—ask me if you are unsure of your own attempts.

Entailments requiring the $(\neg\neg E)$ rule are marked $\vdash^*$. What makes these a bit trickier is that by trying to prove $\neg\neg C$ when your desired conclusion is C, you open up a line of attack (*via* the negation introduction rule) that you would not otherwise expect to use. This is the classic mathematical "proof by contradiction": assume what you want is false, derive a contradiction, and conclude that what you want is true.[4] In our system, this requires two rules: $(\neg I)$ and $(\neg\neg E)$. By "warning" you of this, I hoped to make these problems a bit easier than they would otherwise be.

1. $A \vdash A \wedge (A \vee B)$

$$\begin{array}{r|l l}
1 & A & \\ \cline{2-2}
2 & A \vee B & (\vee\mathrm{I}),\ 1 \\
3 & A \wedge (A \vee B) & (\wedge\mathrm{I}),\ 1,\ 2
\end{array}$$

2. $A \wedge (A \vee B \to C) \vdash C \vee D$

$$\begin{array}{r|l l}
1 & A \wedge (A \vee B \to C) & \\ \cline{2-2}
2 & A \vee B \to C & (\wedge\mathrm{E}),\ 1 \\
3 & A & (\wedge\mathrm{E}),\ 1 \\
4 & A \vee B & (\vee\mathrm{I}),\ 3 \\
5 & C & (\to\mathrm{E}),\ 2,\ 4 \\
6 & C \vee D & (\vee\mathrm{I}),\ 5
\end{array}$$

3. $A, A \to B \vdash A \wedge B$

$$\begin{array}{r|l l}
1 & A & \\
2 & A \to B & \\ \cline{2-2}
3 & B & (\to\mathrm{E}),\ 1,\ 2 \\
4 & A \wedge B & (\wedge\mathrm{I}),\ 1,\ 3
\end{array}$$

4. $A \wedge B \to C, B \to A, B \vdash C$

$$\begin{array}{r|l l}
1 & A \wedge B \to C & \\
2 & B \to A & \\
3 & B & \\ \cline{2-2}
4 & A & (\to\mathrm{E}),\ 2,\ 3 \\
5 & A \wedge B & (\wedge\mathrm{I}),\ 3,\ 4 \\
6 & C & (\to\mathrm{E}),\ 1,\ 5
\end{array}$$

5. $A \to B, B \to C \vdash A \to B \wedge C$

$$\begin{array}{r|l|l l}
1 & \multicolumn{2}{l}{A \to B} & \\
2 & \multicolumn{2}{l}{B \to C} & \\ \cline{2-3}
3 & & A & \\ \cline{3-3}
4 & & B & (\to\mathrm{E}),\ 1,\ 3 \\
5 & & C & (\to\mathrm{E}),\ 2,\ 4 \\
6 & & B \wedge C & (\wedge\mathrm{I}),\ 4,\ 5 \\
7 & \multicolumn{2}{l}{A \to B \wedge C} & (\to\mathrm{I}),\ 3\text{–}6
\end{array}$$

6. $A \to (B \to C) \vdash B \to (A \to C)$

$$\begin{array}{r|l|l|l l}
1 & \multicolumn{3}{l}{A \to (B \to C)} & \\
2 & & \multicolumn{2}{|l}{B} & \\ \cline{3-4}
3 & & & A & \\ \cline{4-4}
4 & & & B \to C & (\to\mathrm{E}),\ 1,\ 3 \\
5 & & & C & (\to\mathrm{E}),\ 2,\ 4 \\
6 & & \multicolumn{2}{|l}{A \to C} & (\to\mathrm{I}),\ 3\text{–}5 \\
7 & \multicolumn{3}{l}{B \to (A \to C)} & (\to\mathrm{I}),\ 2\text{–}6
\end{array}$$

[4] Contrast this with the other version of "proof by contradiction", where you assume some premise, derive a contradiction, and conclude that what you assumed is false. This is more obviously valid: it is just the rule $(\neg I)$, and does not require $(\neg\neg E)$. It is intuitionistically valid.

7. $A \to B \vdash \neg B \to \neg A$

1	$\mid\ \underline{A \to B}$	
2	$\mid\ \mid\ \underline{\neg B}$	
3	$\mid\ \mid\ \mid\ \underline{A}$	
4	$\mid\ \mid\ \mid\ B$	$(\to\mathrm{E})$, 1, 3
5	$\mid\ \mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 2, 4
6	$\mid\ \mid\ \neg A$	$(\neg\mathrm{I})$, 3–5
7	$\mid\ \neg B \to \neg A$	$(\to\mathrm{I})$, 2–6

8. $\vdash^{*} A \vee \neg A$

1	$\mid\ \mid\ \underline{\neg(A \vee \neg A)}$	
2	$\mid\ \mid\ \mid\ \underline{A}$	
3	$\mid\ \mid\ \mid\ A \vee \neg A$	$(\vee\mathrm{I})$, 2
4	$\mid\ \mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 1, 3
5	$\mid\ \mid\ \neg A$	$(\neg\mathrm{I})$, 2–4
6	$\mid\ \mid\ A \vee \neg A$	$(\vee\mathrm{I})$, 5
7	$\mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 1, 6
8	$\mid\ \neg\neg(A \vee \neg A)$	$(\neg\mathrm{I})$, 1–7
9	$\mid\ A \vee \neg A$	$(\neg\neg\mathrm{E})$, 8

9. $(A \to B) \to C \vdash B \to C$

1	$\mid\ \underline{(A \to B) \to C}$	
2	$\mid\ \mid\ \underline{B}$	
3	$\mid\ \mid\ \mid\ \underline{A}$	
4	$\mid\ \mid\ \mid\ B$	(R), 2
5	$\mid\ \mid\ A \to B$	$(\to\mathrm{I})$, 3–4
6	$\mid\ \mid\ C$	$(\to\mathrm{E})$, 1, 5
7	$\mid\ B \to C$	$(\to\mathrm{I})$, 2–6

10. $(A \to B) \to C \vdash^{*} A \vee C$

1	$\mid\ \underline{(A \to B) \to C}$	
2	$\mid\ \mid\ \underline{\neg(A \vee C)}$	
3	$\mid\ \mid\ \mid\ \underline{A}$	
4	$\mid\ \mid\ \mid\ A \vee C$	$(\vee\mathrm{I})$, 3
5	$\mid\ \mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 2, 4
6	$\mid\ \mid\ \neg A$	$(\neg\mathrm{I})$, 3–5
7	$\mid\ \mid\ \mid\ \underline{A}$	
8	$\mid\ \mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 6, 7
9	$\mid\ \mid\ \mid\ B$	$(\bot\mathrm{E})$, 8
10	$\mid\ \mid\ A \to B$	$(\to\mathrm{I})$, 7–9
11	$\mid\ \mid\ C$	$(\to\mathrm{E})$, 1, 10
12	$\mid\ \mid\ A \vee C$	$(\vee\mathrm{I})$, 11
13	$\mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 2, 12
14	$\mid\ \neg\neg(A \vee C)$	$(\neg\mathrm{I})$, 2–13
15	$\mid\ A \vee C$	$(\neg\neg\mathrm{E})$, 14

11. $\vdash A \to \neg\neg A$

1	$\mid\ \mid\ \underline{A}$	
2	$\mid\ \mid\ \mid\ \underline{\neg A}$	
3	$\mid\ \mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 1, 2
4	$\mid\ \mid\ \neg\neg A$	$(\neg\mathrm{I})$, 2–3
5	$\mid\ A \to \neg\neg A$	$(\to\mathrm{I})$, 1–4

12. $\vdash^{*} \neg\neg A \to A$

1	$\mid\ \mid\ \underline{\neg\neg A}$	
2	$\mid\ \mid\ A$	$(\neg\neg\mathrm{E})$, 1
3	$\mid\ \neg\neg A \to A$	$(\to\mathrm{I})$, 1–2

13. $\neg(A \to B) \vdash \neg B$

1	$\mid\ \underline{\neg(A \to B)}$	
2	$\mid\ \mid\ \underline{B}$	
3	$\mid\ \mid\ \mid\ \underline{A}$	
4	$\mid\ \mid\ \mid\ B$	(R), 2
5	$\mid\ \mid\ A \to B$	$(\to\mathrm{I})$, 3–4
6	$\mid\ \mid\ \bot$	$(\neg\mathrm{E})$, 1, 5
7	$\mid\ \neg B$	$(\neg\mathrm{I})$, 2–6

14. $A \vee (A \wedge B) \vdash A$

1	$\mid\ \underline{A \vee (A \wedge B)}$	
2	$\mid\ \mid\ \underline{A}$	
3	$\mid\ \mid\ A$	(R), 2
4	$\mid\ \mid\ \underline{A \wedge B}$	
5	$\mid\ \mid\ A$	$(\wedge\mathrm{E})$, 4
6	$\mid\ A$	$(\vee\mathrm{E})$, 1, 2–3, 4–5

15. $\neg(A \vee B) \vdash \neg A \wedge \neg B$

```
1   | ¬(A ∨ B)
    |---------
2   |  | A
    |  |------
3   |  | A ∨ B    (∨I), 2
4   |  | ⊥        (¬E), 1, 3
5   | ¬A          (¬I), 2–4
6   |  | B
    |  |------
7   |  | A ∨ B    (∨I), 6
8   |  | ⊥        (¬E), 1, 7
9   | ¬B          (¬I), 6–8
10  | ¬A ∧ ¬B     (∧I), 5, 9
```

Notice how #16 is almost the same as #15.

16. $\vdash \neg(A \vee B) \rightarrow \neg A \wedge \neg B$

```
1   |  | ¬(A ∨ B)
    |  |---------
2   |  |  | A
    |  |  |------
3   |  |  | A ∨ B            (∨I), 2
4   |  |  | ⊥                (¬E), 1, 3
5   |  | ¬A                  (¬I), 2–4
6   |  |  | B
    |  |  |------
7   |  |  | A ∨ B            (∨I), 6
8   |  |  | ⊥                (¬E), 1, 7
9   |  | ¬B                  (¬I), 6–8
10  |  | ¬A ∧ ¬B             (∧I), 5, 9
11  | ¬(A ∨ B) → ¬A ∧ ¬B     (→I), 1–10
```

17. $\neg A \vee \neg B \vdash \neg(A \wedge B)$

```
1   | ¬A ∨ ¬B
    |---------
2   |  | A ∧ B
    |  |--------
3   |  | ¬A ∨ ¬B     (R), 1
4   |  |  | ¬A
    |  |  |-----
5   |  |  | A        (∧E), 2
6   |  |  | ⊥        (¬E), 4, 5
7   |  |  | ¬B
    |  |  |-----
8   |  |  | B        (∧E), 2
9   |  |  | ⊥        (¬E), 7, 8
10  |  | ⊥           (∨E), 3, 4–6, 7–9
11  | ¬(A ∧ B)       (¬I), 2–10
```

18. $\neg(A \wedge B) \vdash^{*} \neg A \vee \neg B$

```
1   | ¬(A ∧ B)
    |---------
2   |  | ¬(¬A ∨ ¬B)
    |  |------------
3   |  |  | ¬A
    |  |  |--------
4   |  |  | ¬A ∨ ¬B    (∨I), 3
5   |  |  | ⊥          (¬E), 2, 4
6   |  | ¬¬A           (¬I), 3–5
7   |  | A             (¬¬E), 6
8   |  |  | ¬B
    |  |  |--------
9   |  |  | ¬A ∨ ¬B    (∨I), 8
10  |  |  | ⊥          (¬E), 2, 9
11  |  | ¬¬B           (¬I), 8–10
12  |  | B             (¬¬E), 11
13  |  | A ∧ B         (∧I), 7, 12
14  |  | ⊥             (¬E), 1, 13
15  | ¬¬(¬A ∨ ¬B)      (¬I), 2–14
16  | ¬A ∨ ¬B          (¬¬E), 15
```

19. $\neg A \vee B \vdash A \rightarrow B$ This can be done two ways: first introduce the $\rightarrow$ or first eliminate the $\vee$.

```
1   | ¬A ∨ B
    |--------
2   |  | A
    |  |------
3   |  | ¬A ∨ B    (R), 1
4   |  |  | ¬A
    |  |  |----
5   |  |  | ⊥      (¬E), 2, 4
6   |  |  | B      (⊥E), 5
7   |  |  | B
    |  |  |----
8   |  |  | B      (R), 7
9   |  | B         (∨E), 3, 4–6, 7–8
10  | A → B        (→I), 2–9
```

```
1   | ¬A ∨ B
    |--------
2   |  | ¬A
    |  |-----
3   |  |  | A
    |  |  |----
4   |  |  | ⊥      (¬E), 2, 3
5   |  |  | B      (⊥E), 4
6   |  | A → B     (→I), 3–5
7   |  | B
    |  |----
8   |  |  | A
    |  |  |----
9   |  |  | B      (R), 7
10  |  | A → B     (→I), 8–9
11  | A → B        (∨E), 1, 2–6, 7–10
```

20. $\neg A \vdash A \to B$

1	$\mid \underline{\neg A}$	
2	$\mid \mid \underline{A}$	
3	$\mid \mid \bot$	$(\neg\mathrm{E})$, 1, 2
4	$\mid \mid B$	$(\bot\mathrm{E})$, 3
5	$\mid A \to B$	$(\to\mathrm{I})$, 2–4

21. $(A \to B) \vee (A \to C) \vdash A \to B \vee C$

1	$\mid \underline{(A \to B) \vee (A \to C)}$	
2	$\mid \mid \underline{A}$	
3	$\mid \mid (A \to B) \vee (A \to C)$	(R), 1
4	$\mid \mid \mid \underline{A \to B}$	
5	$\mid \mid \mid B$	$(\to\mathrm{E})$, 2, 4
6	$\mid \mid \mid B \vee C$	$(\vee\mathrm{I})$, 5
7	$\mid \mid \mid \underline{A \to C}$	
8	$\mid \mid \mid C$	$(\to\mathrm{E})$, 2, 7
9	$\mid \mid \mid B \vee C$	$(\vee\mathrm{I})$, 8
10	$\mid \mid B \vee C$	$(\vee\mathrm{E})$, 3, 4–6, 7–9
11	$\mid A \to B \vee C$	$(\to\mathrm{I})$, 2–10

22. $A \to B \vee C \vdash^{*} (A \to B) \vee (A \to C)$

1	$\mid \underline{A \to B \vee C}$	
2	$\mid \mid \underline{\neg((A \to B) \vee (A \to C))}$	
3	$\mid \mid \mid \underline{\neg A}$	
4	$\mid \mid \mid \mid \underline{A}$	
5	$\mid \mid \mid \mid \bot$	$(\neg\mathrm{E})$, 3, 4
6	$\mid \mid \mid \mid B$	$(\bot\mathrm{E})$, 5
7	$\mid \mid \mid A \to B$	$(\to\mathrm{I})$, 4–6
8	$\mid \mid \mid (A \to B) \vee (A \to C)$	$(\vee\mathrm{I})$, 7
9	$\mid \mid \mid \bot$	$(\neg\mathrm{E})$, 2, 8
10	$\mid \mid \neg\neg A$	$(\neg\mathrm{I})$, 3–9
11	$\mid \mid A$	$(\neg\neg\mathrm{E})$, 10
12	$\mid \mid B \vee C$	$(\to\mathrm{E})$, 1, 11
13	$\mid \mid \mid \underline{B}$	
14	$\mid \mid \mid \mid \underline{A}$	
15	$\mid \mid \mid \mid B$	(R), 13
16	$\mid \mid \mid A \to B$	$(\to\mathrm{I})$, 14–15
17	$\mid \mid \mid (A \to B) \vee (A \to C)$	$(\vee\mathrm{I})$, 16
18	$\mid \mid \mid \underline{C}$	
19	$\mid \mid \mid \mid \underline{A}$	
20	$\mid \mid \mid \mid C$	(R), 18
21	$\mid \mid \mid A \to C$	$(\to\mathrm{I})$, 19–20
22	$\mid \mid \mid (A \to B) \vee (A \to C)$	$(\vee\mathrm{I})$, 21
23	$\mid \mid (A \to B) \vee (A \to C)$	$(\vee\mathrm{E})$, 12, 13–17, 18–22
24	$\mid \mid \bot$	$(\neg\mathrm{E})$, 2, 23
25	$\mid \neg\neg((A \to B) \vee (A \to C))$	$(\neg\mathrm{I})$, 2–24
26	$\mid (A \to B) \vee (A \to C)$	$(\neg\neg\mathrm{E})$, 25

23. $A \vee B \to C \vdash (A \to C) \wedge (B \to C)$

1	$\mid \underline{A \vee B \to C}$	
2	$\mid \mid \underline{A}$	
3	$\mid \mid A \vee B$	$(\vee\mathrm{I})$, 2
4	$\mid \mid C$	$(\to\mathrm{E})$, 1, 3
5	$\mid A \to C$	$(\to\mathrm{I})$, 2–4
6	$\mid \mid \underline{B}$	
7	$\mid \mid A \vee B$	$(\vee\mathrm{I})$, 6
8	$\mid \mid C$	$(\to\mathrm{E})$, 1, 7
9	$\mid B \to C$	$(\to\mathrm{I})$, 6–8
10	$\mid (A \to C) \wedge (B \to C)$	$(\wedge\mathrm{I})$, 5, 9

24. $(A \to C) \land (B \to C) \vdash A \lor B \to C$

1	$\mid (A \to C) \land (B \to C)$	
2	$\mid A \to C$	($\land$E), 1
3	$\mid B \to C$	($\land$E), 1
4	$\mid \mid A \lor B$	
5	$\mid \mid \mid A$	
6	$\mid \mid \mid C$	($\to$E), 2, 5
7	$\mid \mid \mid B$	
8	$\mid \mid \mid C$	($\to$E), 3, 7
9	$\mid \mid C$	($\lor$E), 4, 5–6, 7–8
10	$\mid A \lor B \to C$	($\to$I), 4–9

25. $A \to B \land C \vdash (A \to B) \land (A \to C)$

1	$\mid A \to B \land C$	
2	$\mid \mid A$	
3	$\mid \mid B \land C$	($\to$E), 1, 2
4	$\mid \mid B$	($\land$E), 3
5	$\mid A \to B$	($\to$I), 2–4
6	$\mid \mid A$	
7	$\mid \mid B \land C$	($\to$E), 1, 6
8	$\mid \mid C$	($\land$E), 7
9	$\mid A \to C$	($\to$I), 6–8
10	$\mid (A \to B) \land (A \to C)$	($\land$I), 5, 9

26. $(A \to B) \land (A \to C) \vdash A \to B \land C$

1	$\mid (A \to B) \land (A \to C)$	
2	$\mid A \to B$	($\land$E), 1
3	$\mid A \to C$	($\land$E), 1
4	$\mid \mid A$	
5	$\mid \mid B$	($\to$E), 2, 4
6	$\mid \mid C$	($\to$E), 3, 4
7	$\mid \mid B \land C$	($\land$I), 5, 6
8	$\mid A \to B \land C$	($\to$I), 4–7

27. $(A \to C) \lor (B \to C) \vdash A \land B \to C$

1	$\mid (A \to C) \lor (B \to C)$	
2	$\mid \mid A \land B$	
3	$\mid \mid A$	($\land$E), 2
4	$\mid \mid B$	($\land$E), 2
5	$\mid \mid \mid A \to C$	
6	$\mid \mid \mid C$	($\to$E), 3, 5
7	$\mid \mid \mid B \to C$	
8	$\mid \mid \mid C$	($\to$E), 4, 7
9	$\mid \mid C$	($\lor$E), 1, 5–6, 7–8
10	$\mid A \land B \to C$	($\to$I), 2–9

28. $B \land C \to A, \neg A \to C \vdash^{*} (C \to B) \to A$

1	$\mid B \land C \to A$	
2	$\mid \neg A \to C$	
3	$\mid \mid C \to B$	
4	$\mid \mid \mid \neg A$	
5	$\mid \mid \mid C$	($\to$E), 2, 4
6	$\mid \mid \mid B$	($\to$E), 3, 5
7	$\mid \mid \mid B \land C$	($\land$I), 5, 6
8	$\mid \mid \mid A$	($\to$E), 1, 7
9	$\mid \mid \mid \bot$	($\neg$E), 4, 8
10	$\mid \mid \neg\neg A$	($\neg$I), 4–9
11	$\mid \mid A$	($\neg\neg$E), 10
12	$\mid (C \to B) \to A$	($\to$I), 3–11

29. $\neg(A \land \neg B) \vdash^{*} A \to B$

1	$\mid \neg(A \land \neg B)$	
2	$\mid \mid A$	
3	$\mid \mid \mid \neg B$	
4	$\mid \mid \mid A \land \neg B$	($\land$I), 2, 3
5	$\mid \mid \mid \bot$	($\neg$E), 1, 4
6	$\mid \mid \neg\neg B$	($\neg$I), 3–5
7	$\mid \mid B$	($\neg\neg$E), 6
8	$\mid A \to B$	($\to$I), 2–7

30. $A \to B \vdash \neg(A \wedge \neg B)$

1	$\vert\ \underline{A \to B}$	
2	$\vert\ \vert\ \underline{A \wedge \neg B}$	
3	$\vert\ \vert\ A$	(∧E), 2
4	$\vert\ \vert\ \neg B$	(∧E), 2
5	$\vert\ \vert\ B$	(→E), 1, 3
6	$\vert\ \vert\ \bot$	(¬E), 4, 5
7	$\vert\ \neg(A \wedge \neg B)$	(¬I), 2–6

31. $A \vee B \vdash \neg B \to (C \to A)$

1	$\vert\ \underline{A \vee B}$	
2	$\vert\ \vert\ \underline{\neg B}$	
3	$\vert\ \vert\ \vert\ \underline{A}$	
4	$\vert\ \vert\ \vert\ \vert\ \underline{C}$	
5	$\vert\ \vert\ \vert\ \vert\ A$	(R), 3
6	$\vert\ \vert\ \vert\ C \to A$	(→I), 4–5
7	$\vert\ \vert\ \vert\ \underline{B}$	
8	$\vert\ \vert\ \vert\ \bot$	(¬E), 2, 7
9	$\vert\ \vert\ \vert\ C \to A$	(⊥E), 8
10	$\vert\ \vert\ C \to A$	(∨E), 1, 3–6, 7–9
11	$\vert\ \neg B \to (C \to A)$	(→I), 2–9

32. $(B \to A) \to A \vdash^* A \vee B$

1	$\vert\ \underline{(B \to A) \to A}$	
2	$\vert\ \vert\ \underline{\neg(A \vee B)}$	
3	$\vert\ \vert\ \vert\ \underline{A}$	
4	$\vert\ \vert\ \vert\ A \vee B$	(∨I), 3
5	$\vert\ \vert\ \vert\ \bot$	(¬E), 2, 4
6	$\vert\ \vert\ \neg A$	(¬I), 3–5
7	$\vert\ \vert\ \vert\ \underline{B}$	
8	$\vert\ \vert\ \vert\ A \vee B$	(∨I), 7
9	$\vert\ \vert\ \vert\ \bot$	(¬E), 2, 8
10	$\vert\ \vert\ \neg B$	(¬I), 7–9
11	$\vert\ \vert\ \vert\ \underline{B}$	
12	$\vert\ \vert\ \vert\ \bot$	(¬E), 10, 11
13	$\vert\ \vert\ \vert\ A$	(⊥E), 12
14	$\vert\ \vert\ B \to A$	(→I), 11–13
15	$\vert\ \vert\ A$	(→E), 1, 14
16	$\vert\ \vert\ \bot$	(¬E), 6, 15
17	$\vert\ \neg\neg(A \vee B)$	(¬I), 2–16
18	$\vert\ A \vee B$	(¬¬E), 17

33. $(A \to B) \vee C\ ,\ A \to \neg C \vdash (B \to C) \to \neg A$

1	$\vert\ (A \to B) \vee C$	
2	$\vert\ \underline{A \to \neg C}$	
3	$\vert\ \vert\ \underline{B \to C}$	
4	$\vert\ \vert\ \vert\ \underline{A}$	
5	$\vert\ \vert\ \vert\ \neg C$	(→E), 2, 4
6	$\vert\ \vert\ \vert\ \vert\ \underline{A \to B}$	
7	$\vert\ \vert\ \vert\ \vert\ B$	(→E), 4, 6
8	$\vert\ \vert\ \vert\ \vert\ C$	(→E), 3, 7
9	$\vert\ \vert\ \vert\ \vert\ \underline{C}$	
10	$\vert\ \vert\ \vert\ \vert\ C$	(R), 9
11	$\vert\ \vert\ \vert\ C$	(∨E), 1, 6–8, 9–10
12	$\vert\ \vert\ \vert\ \bot$	(¬E), 5, 11
13	$\vert\ \vert\ \neg A$	(¬I), 4–12
14	$\vert\ (B \to C) \to \neg A$	(→I), 3–13

Exercise 3.3.2

1.

1	$L \to C \lor T$	
2	$\neg B \to \neg L$	
3	$\underline{C \to \neg B}$	
4	$\quad \underline{L}$	
5	$\quad C \lor T$	(→E), 1, 4
6	$\quad\quad \underline{C}$	
7	$\quad\quad \neg B$	(→E), 3, 6
8	$\quad\quad \neg L$	(→E), 2, 7
9	$\quad\quad \bot$	(¬E), 4, 8
10	$\quad\quad T$	(⊥E), 9
11	$\quad\quad \underline{T}$	
12	$\quad\quad T$	(R), 11
13	$\quad T$	(∨E), 5, 6–10, 11–12
14	$L \to T$	(→I), 4–13

2.

1	$S \land B$	
2	$B \to H$	
3	$\underline{H \land \neg M \to \neg S}$	
4	S	(∧E), 1
5	B	(∧E), 1
6	H	(→E), 2, 5
7	$\quad \underline{\neg M}$	
8	$\quad H \land \neg M$	(∧I), 6, 7
9	$\quad \neg S$	(→E), 3, 8
10	$\quad \bot$	(¬E), 4, 9
11	$\neg\neg M$	(¬I), 7–10
12	M	(¬¬E), 11

3.

1	$A \to B \lor C$	
2	$B \to D \land E$	
3	$D \to (E \to F)$	
4	$\underline{\neg(F \land A)}$	
5	$\quad \underline{A}$	
6	$\quad B \lor C$	(→E), 1, 5
7	$\quad\quad \underline{B}$	
8	$\quad\quad D \land E$	(→E), 2, 7
9	$\quad\quad D$	(∧E), 8
10	$\quad\quad E$	(∧E), 8
11	$\quad\quad E \to F$	(→E), 3, 9
12	$\quad\quad F$	(→E), 10, 11
13	$\quad\quad F \land A$	(∧I), 5, 12
14	$\quad\quad \bot$	(¬E), 4, 13
15	$\quad\quad C$	(⊥E), 14
16	$\quad\quad \underline{C}$	
17	$\quad\quad C$	(R), 16
18	$\quad C$	(∨E), 6, 7–15, 16–17
19	$A \to C$	(→I), 5–18

4.

1	$(A \lor S) \land \neg K$	
2	$\neg K \to C \land \neg A$	
3	$\underline{S \to A \lor I}$	
4	$\neg K$	(∧E), 1
5	$C \land \neg A$	(→E), 2, 4
6	$\neg A$	(∧E), 5
7	$A \lor S$	(∧E), 1
8	$\quad \underline{A}$	
9	$\quad \bot$	(¬E), 6, 8
10	$\quad I \lor J$	(⊥E), 9
11	$\quad \underline{S}$	
12	$\quad A \lor I$	(→E), 3, 11
13	$\quad\quad \underline{A}$	
14	$\quad\quad \bot$	(¬E), 6, 13
15	$\quad\quad I \lor J$	(⊥E), 14
16	$\quad\quad \underline{I}$	
17	$\quad\quad I \lor J$	(∨I), 16
18	$\quad I \lor J$	(∨E), 12, 13–15, 16–17
19	$I \lor J$	(∨E), 7, 8–10, 11–18

4. Let's do number 4 again, this time with the derived rule DS:
   $p \vee q, \neg p \vdash q$ (or $p \vee q, \neg q \vdash p$)

```
 1 | (A ∨ S) ∧ ¬K
 2 | ¬K → C ∧ ¬A
 3 | S → A ∨ I
   |------------
 4 | ¬K              (∧E), 1
 5 | C ∧ ¬A          (→E), 2, 4
 6 | ¬A              (∧E), 5
 7 | A ∨ S           (∧E), 1
 8 | S               (DS), 6, 7
 9 | A ∨ I           (→E), 3, 8
10 | I               (DS), 6, 9
11 | I ∨ J           (∨I), 10
```

This "trick" with the derived rule DS can also be used to make numbers 1–3 a bit simpler too, though the savings aren't as great as they are for number 4.

5.

```
 1 | G → P ∧ S
 2 | G → B
 3 | C → (K → (¬V → ¬B))
 4 | P → C
 5 | S → (E → K)
 6 | V → ¬E
 7 | E
   |--------------------
 8 |   | G
   |   |----
 9 |   | P ∧ S                 (→E), 1, 8
10 |   | P                     (∧E), 9
11 |   | S                     (∧E), 9
12 |   | C                     (→E), 4, 10
13 |   | E → K                 (→E), 5, 11
14 |   | K                     (→E), 7, 13
15 |   | K → (¬V → ¬B)         (→E), 3, 12
16 |   | ¬V → ¬B               (→E), 14, 15
17 |   | B                     (→E), 2, 8
18 |   |   | V
   |   |   |----
19 |   |   | ¬E                (→E), 6, 18
20 |   |   | ⊥                 (¬E), 7, 19
21 |   | ¬V                    (¬I), 18–20
22 |   | ¬B                    (→E), 16, 21
23 |   | ⊥                     (¬E), 17, 22
24 | ¬G                        (¬I), 8–23
```

6.

```
 1 | B → F
 2 | ¬B ∧ ¬F → H
 3 | ¬H ∧ ¬F
   |------------
 4 | ¬H              (∧E), 3
 5 | ¬F              (∧E), 3
 6 |   | B
   |   |----
 7 |   | F           (→E), 1, 6
 8 |   | ⊥           (¬E), 5, 7
 9 | ¬B              (¬I), 6–8
10 | ¬B ∧ ¬F         (∧I), 5, 9
11 | H               (→E), 2, 10
12 | ⊥               (¬E), 4, 11
```

Chapter 4

Analytic Tableaux

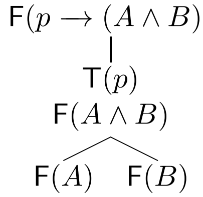
Definition: A signed formula is an expression $\mathsf{T}(A)$ or an expression $\mathsf{F}(A)$, where A is a WFF. (We may drop the parentheses if this doesn't obscure clarity, as in $\mathsf{T}A$.)

4.1 Basic Rules of Knights and Knaves

If we interpret $\mathsf{T}(A)$ as meaning "A is true", and $\mathsf{F}(A)$ as "A is false", then the following are true (and well-known from our experiences on the Island of Knights and Knaves! - take a look again at the table in section 1.5.2):

- If $\mathsf{T}(\neg A)$, then $\mathsf{F}(A)$.
- If $\mathsf{F}(\neg A)$, then $\mathsf{T}(A)$.
- If $\mathsf{T}(A \wedge B)$, then $\mathsf{T}(A)$ and $\mathsf{T}(B)$.
- If $\mathsf{F}(A \wedge B)$, then $\mathsf{F}(A)$ or $\mathsf{F}(B)$.
- If $\mathsf{T}(A \vee B)$, then $\mathsf{T}(A)$ or $\mathsf{T}(B)$.
- If $\mathsf{F}(A \vee B)$, then $\mathsf{F}(A)$ and $\mathsf{F}(B)$.
- If $\mathsf{T}(A \longrightarrow B)$, then $\mathsf{F}(A)$ or $\mathsf{T}(B)$.
- If $\mathsf{F}(A \longrightarrow B)$, then $\mathsf{T}(A)$ and $\mathsf{F}(B)$.

We can use these facts to quickly check the truth values of the components of a WFF, once we know its truth value. For instance, if $p \longrightarrow (A \wedge B)$ is false, then p must be true and one of A, B must be false. We can diagram this in the following way:



where we have used signed formulas to indicate the truth value of the formula, and where we have indicated a choice of two possible truth values by a "fork" in the diagram. We shall call such a diagram a “tree” (sometimes called “Australian trees”, since they branch downwards!). The tree represents the “decomposition” of the original signed formula, and each path in the tree gives a possible set of truth values of the component parts of the initial formula with its specified truth value. In this example, we have two paths, which give the two possible truth value specifications (namely $\mathsf{T}(p), \mathsf{F}(A)$ and $\mathsf{T}(p), \mathsf{F}(B)$) which result in $\mathsf{F}(p \longrightarrow (A \wedge B))$. Any other truth value specification will make $p \longrightarrow (A \wedge B)$ true.

Actually, we should be a bit careful here: we said we had two truth specifications, but in each one, the truth value of one of the atoms was unspecified, and so could have either value. So those two apparent specifications really give three specifications: first, $\mathsf{T}(p), \mathsf{F}(A)$ and either $\mathsf{T}(B)$ or $\mathsf{F}(B)$; and second, $\mathsf{T}(p), \mathsf{F}(B)$ and either $\mathsf{T}(A)$ or $\mathsf{F}(A)$. Looking at that, you will realise this is the three specifications $\mathsf{T}(p), \mathsf{F}(A), \mathsf{T}(B)$ and $\mathsf{T}(p), \mathsf{F}(A), \mathsf{F}(B)$ and $\mathsf{T}(p), \mathsf{T}(A), \mathsf{F}(B)$.

When you decompose a formula, if you find a path which has two oppositely signed instances of the same formula (for example $\mathsf{T}(p)$ and $\mathsf{F}(p)$), then we shall say that path is *closed*, and we shall not add any further formulas or branches of our diagram to that path. In a sense, such closed paths correspond to impossible or contradictory truth specifications (you cannot make a formula be both true and false at the same time).

We may use this simple observation to give a quite powerful method (the “method of analytic tableaux”, or simply the “tableau method”) for determining whether a particular argument is valid or not. The same method can also determine if a particular formula is a tautology or not, and if not, it can determine specifications of truth values for the atomic components which render the formula false (and similarly for an invalid argument). Likewise the method can determine if a formula is satisfiable (*i.e.* not a contradiction), and what truth value specifications satisfy it (*i.e.* make it true). In short, the method does what truth tables do, but usually a lot more efficiently.

There is a reason for this: truth tables start from the values of the atoms and build the values of the compound formula, which is a good way to get all possible values. But tableaux start with a desired truth value of the compound formula, and work back to figure out what values the atoms must have to give it. In effect, tableaux just build the relevant part of the truth table, ignoring the bits you don’t want. Generally this is a lot faster.

4.2 Tableaux rules

A tableau for a signed formula, or for a set of signed formulas, is constructed quite simply. We start with a list of the signed formulas. Then pick one and, using the observations above, decompose that signed formula: when the decomposition gives two definite truth values, list both, and when the decomposition gives a choice of two truth values, create a fork, “splitting” (or “branching”) the tree, as in the example above. Mark the signed formula you picked (so you don’t pick it again), and then choose another signed formula to do the same decomposition operation. The decomposition you perform must be done at the bottom of each path containing the signed formula you are decomposing. Continue in this way until you have decomposed every signed formula in your tree, including the ones you have created in building the tree. Closed paths represent impossible, contradictory assignments of truth values: if every path is closed, then the original specification represented by the signed formula(s) is contradictory and cannot be obtained by any assignment of truth values to its atoms. If any paths are open, then they give truth value assignments which do realize the original specification.

So, we arrive at the following method:

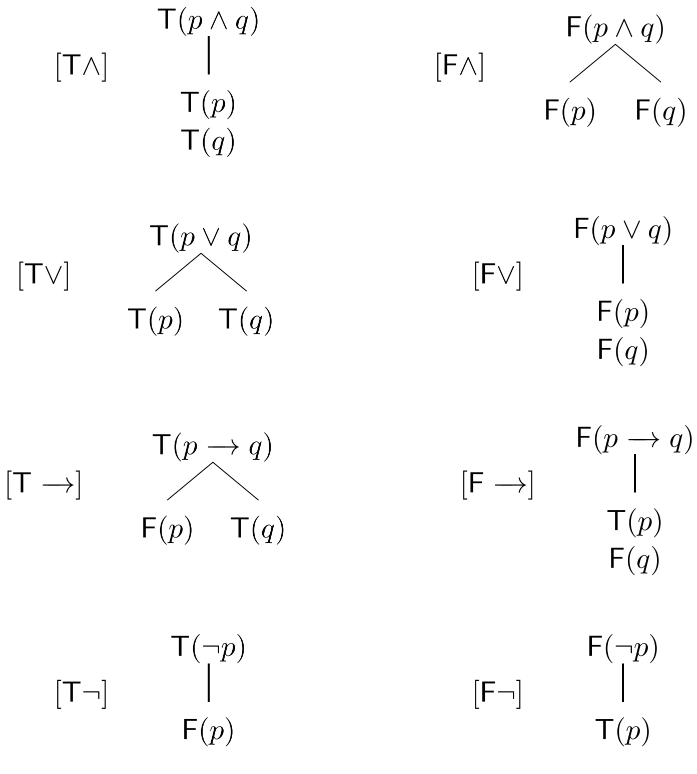
To determine if an argument is valid: Take the premises of the formula, sign them each with a T; take the conclusion of the argument and sign it with an F. Then develop the tableau for this: if all paths close, the argument is valid. On the other hand, if any paths remain open, the argument must be invalid, and you have found a truth value specification invalidating it. (The strategy here is to try to invalidate the argument, by specifying true premises and a false conclusion; if all paths close, you have arrived at a contradiction, so you failed to invalidate the argument, so it must be valid.)

To determine if a formula is a tautology: Here the idea is essentially the same: start with the formula signed F. Develop the tableau, and if all paths close, the formula is a tautology. On the other hand, if any path remains open, you have found a truth value specification which makes the formula false (so showing it is not a tautology).

To determine if a formula is satisfiable: Start with the formula signed T. Develop the tableau: any open path gives a truth value specification satisfying the formula; if all paths close, the formula is not satisfiable.

A summary of the tableaux rules

Here are the basic graphical steps used in building tableaux. Be sure you understand how they reflect the basic observations above, the "basic rules of knights and knaves".
Remark: In practice, we sometimes omit the vertical lines (edges), only keeping the slanted ones where the tree splits or branches. In the first example, I shall not do that, but don't worry if you sometimes find missing vertical edges.



4.3 Examples

Example 1

Show that the following WFF is a tautology:

$$(p \vee (q \wedge r)) \longrightarrow ((p \vee q) \wedge (p \vee r))$$

Also show that the following entailment is valid:

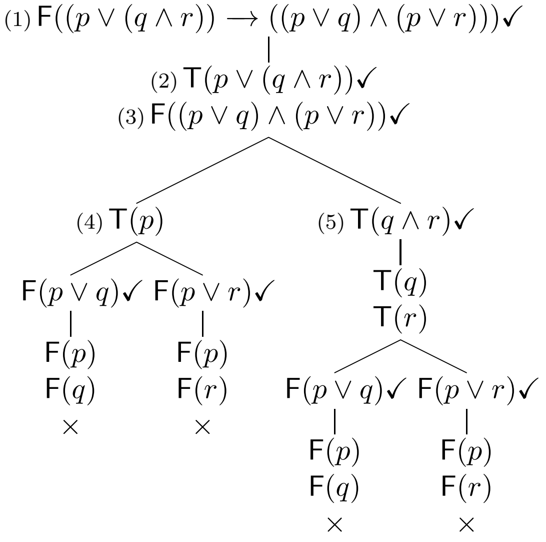
$$(p \vee (q \wedge r)) \vdash ((p \vee q) \wedge (p \vee r))$$

The first point to make is that these two problems use essentially the same tableau; they are equivalent, and that equivalence is reflected in the fact that the same tableau solves them both.

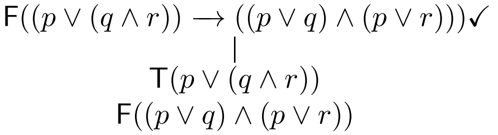
Let's start with the first question:

> show that $(p \vee (q \wedge r)) \longrightarrow ((p \vee q) \wedge (p \vee r))$ is a tautology.

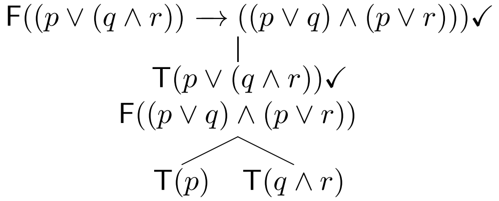
Here is the tableau for this; I shall comment on the steps below. One point to mention is that the little line numbers (1,2,3,4,5) are not part of the tableau, but are just there to make it easier for me to comment on the lines you see there.


(1) $\mathsf{F}((p \vee (q \wedge r)) \longrightarrow ((p \vee q) \wedge (p \vee r)))\checkmark$
(2) $\mathsf{T}(p \vee (q \wedge r))\checkmark$
(3) $\mathsf{F}((p \vee q) \wedge (p \vee r))\checkmark$
(4) $\mathsf{T}(p)$
(5) $\mathsf{T}(q \wedge r)\checkmark$
$\mathsf{F}(p \vee q)\checkmark$
$\mathsf{F}(p \vee r)\checkmark$
$\mathsf{T}(q)$
$\mathsf{T}(r)$
$\mathsf{F}(p)$
$\mathsf{F}(q)$
$\times$
$\mathsf{F}(p)$
$\mathsf{F}(r)$
$\times$
$\mathsf{F}(p \vee q)\checkmark$
$\mathsf{F}(p \vee r)\checkmark$
$\mathsf{F}(p)$
$\mathsf{F}(q)$
$\times$
$\mathsf{F}(p)$
$\mathsf{F}(r)$
$\times$

So: we start with line 1, which is the original formula signed F (remember the strategy is to *deny* the formula, and then try to arrive at a contradiction by having all paths close down, *i.e.* be contradictory). Line 1 is an implication, and according to the rule ($\mathsf{F} \longrightarrow$), a negatively signed implication breaks down into two (non-branching) lines, the premise of the implication signed T and its conclusion signed F. These are the *direct consequences* of line 1, and we write them on lines 2,3. We then check off line 1 (we say it has been "used"), since we've finished decomposing it (we don't want to either miss a decomposition, nor do one twice, so marking the ones you do as you do them helps you keep track of your work). At this point, the tree would look like this:


$\mathsf{F}((p \vee (q \wedge r)) \longrightarrow ((p \vee q) \wedge (p \vee r)))\checkmark$
$\mathsf{T}(p \vee (q \wedge r))$
$\mathsf{F}((p \vee q) \wedge (p \vee r))$

Now we need to decompose another formula (our intent is to work this way till there are no undecomposed formulas left, only atomic ones). We may choose any one we wish (lines 2 and 3 are available to us); we choose to decompose line 2. This is a T signed disjunction, and those branch (consider the ($\mathsf{T}\vee$) rule) into two T signed possibilities, one for each disjunct. By the way: line 3 is an F signed conjunction, which also branches or splits, so even if we'd chosen to decompose line 3 instead of line 2, we'd have had splitting anyway. You may see the splitting of line 2 in the fork leading to lines 4 and 5. Now the tree looks like this:


$\mathsf{F}((p \vee (q \wedge r)) \to ((p \vee q) \wedge (p \vee r)))\checkmark$
$\mathsf{T}(p \vee (q \wedge r))\checkmark$
$\mathsf{F}((p \vee q) \wedge (p \vee r))$
$\mathsf{T}(p)$
$\mathsf{T}(q \wedge r)$

We continue in this way. Line 5 is very simple to decompose, so we do it right away, giving the $\mathsf{T}(q), \mathsf{T}(r)$ you see under line 5. At this point we have to go back to decompose line 3, which we've not done yet (we could have done it earlier, the tableau method works regardless of what order you decompose things, so you should choose an order that makes your job simpler). Line 3 is a negative conjunction, which splits into two negatively signed conjuncts. We must do that splitting at the end of every path containing line 3—this means that splitting occurs in two places, once below line 4, and once below line 5, as you can see. That finally gives us four instances of F signed disjunctions, which do not split, but instead give us the direct consequences you see in the tableau. Look at each path: you will find in each a pair of contradictory atoms. For instance, in the left-most path, there is an $\mathsf{F}(p)$ and a $\mathsf{T}(p)$. So each path is closed (indicated by the × placed at the bottom of each path—in general, mark a path closed, with a ×, as soon as you see it has a contradictory pair of formulas $\mathsf{T}(X), \mathsf{F}(X)$). Hence the tree is closed, meaning that the original formula is a tautology. (We failed to find any way of making the formula false.)

Now let's consider the second problem: show that the following entailment is valid:

$$(p \vee (q \wedge r)) \vdash ((p \vee q) \wedge (p \vee r))$$

We would start this by setting up a tableau with the premise signed T and the conclusion signed F; again, this is because we are trying to disprove our goal (hoping to fail in the attempt!), so we are trying to show the argument is invalid, meaning that it is possible to have the premise true but the conclusion false.

But doing that just gives us lines 2 and 3 of the tableau above, and so continuing would just build the same tableau without line 1. So we again get a closed tableau, and so the entailment is valid.

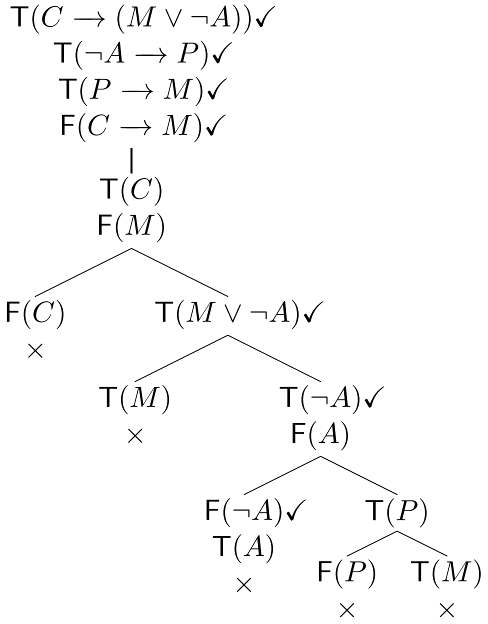
So essentially the same tableau solves both problems.

More examples

Here are tableaux to show each of the following is a valid argument. It's good practice to construct correct derivations for these as well. I'll leave that as an exercise.

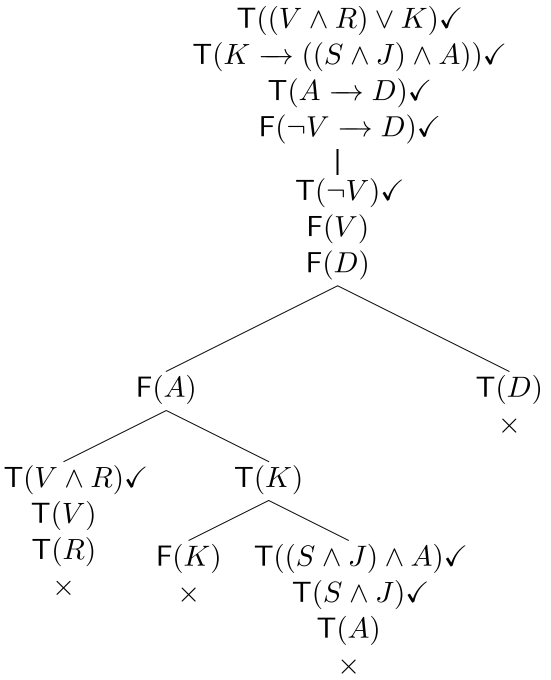
1.
$$\begin{array}{r|l}
1 & C \to (M \vee \neg A) \\
2 & \neg A \to P \\
3 & P \to M \\
\hline
? & C \to M
\end{array}$$

Comments: It's generally good strategy to perform all non-branching decompositions before you do branching ones, since this tends to keep the size of the tree smaller. So we decompose the last initial signed formula, the conclusion $\mathsf{F}(C \to M)$, first. The other three all branch, so our next step is to choose one of them to decompose; I chose the first, mainly because I saw it would give me a $\mathsf{F}(C)$, which would close a path. Decomposing line 3 would also have accomplished this, with a $\mathsf{T}(M)$. Generally you should try to close paths as quickly as you can, again since this will help keep the size of the tree down. Continue this way, and you soon have all compound formulas decomposed into their atoms, and then you can verify that all the paths have been closed. Notice one small trick: whenever you have $\mathsf{T}(\neg X)$ or $\mathsf{F}(\neg X)$, you should immediately remove the $\neg$, switching the sign. (Always do the easy stuff first!)



2.
$$\begin{array}{r|l}
1 & (V \wedge R) \vee K \\
2 & K \to ((S \wedge J) \wedge A) \\
3 & A \to D \\
\hline
? & \neg V \to D
\end{array}$$

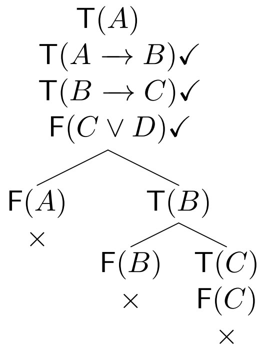
Comments: Ah, this is our old Klingon and Romulan problem ... Well, we know it's valid, as we have a derivation; let's see how easily we get a tableau confirming that. I'll let you study this one—my main strategy was to do the non-branching signed formulas first, starting with the simplest ones (leaving that horrible second premise to last!).


$\mathsf{T}((V \wedge R) \vee K)$✓
$\mathsf{T}(K \to ((S \wedge J) \wedge A))$✓
$\mathsf{T}(A \to D)$✓
$\mathsf{F}(\neg V \to D)$✓
$\mathsf{T}(\neg V)$✓
$\mathsf{F}(V)$
$\mathsf{F}(D)$
$\mathsf{F}(A)$
$\mathsf{T}(D)$
×
$\mathsf{T}(V \wedge R)$✓
$\mathsf{T}(V)$
$\mathsf{T}(R)$
×
$\mathsf{T}(K)$
$\mathsf{F}(K)$
×
$\mathsf{T}((S \wedge J) \wedge A)$✓
$\mathsf{T}(S \wedge J)$✓
$\mathsf{T}(A)$
×

3.

1	A
2	$A \to B$
3	$B \to C$
?	$C \lor D$

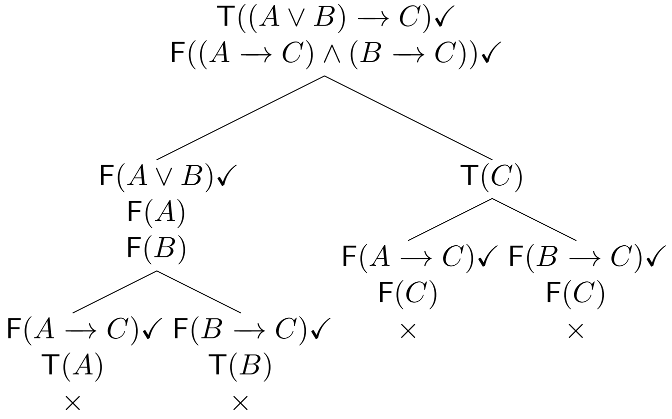
Not a lot to say about this one!



4.

1	$(A \lor B) \to C$
?	$(A \to C) \land (B \to C)$

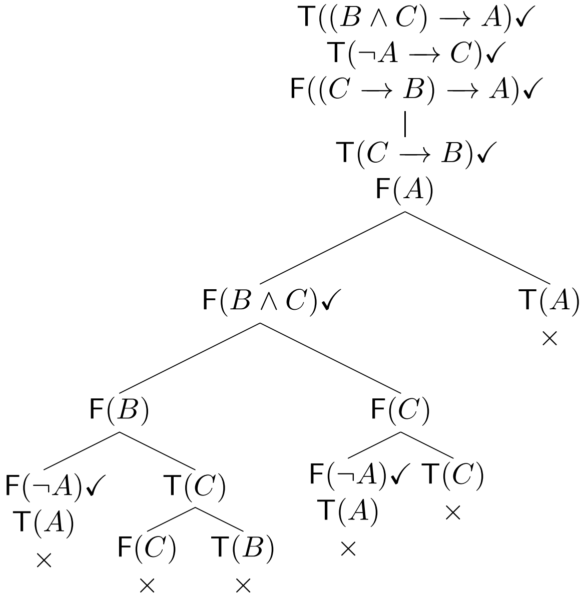
Comments: Both the initial entries split; I chose to split the first (it seemed a bit simpler I guess!). Since the $\mathsf{F}(A \lor B)$ entry decomposes directly (no split), do that right away. Then decompose the other initial entry, at both path ends in its way. This produces two more branching paths, but these quickly decompose to close all paths. Note that I haven't bothered to finish decompositions as soon as I saw the path close. Again, this is merely to finish the job faster.



5.

1	$(B \land C) \to A$
2	$\neg A \to C$
?	$(C \to B) \to A$

Comments: Again, not a lot to say. Do the non-splitting ones first, and simple splitting ones before more complicated ones (which is why I split the $\mathsf{F}(B \land C)$). But there are other ways you could do this tree, without significant differences.



4.3.1 Exercises

Use the method of tableaux to solve the following problems.

1. Show that the following formula is a tautology: $(p \vee q) \longrightarrow (q \vee p)$.
2. Show that the following formula are equivalent: $A \longrightarrow (B \vee C)$ and $(A \longrightarrow B) \vee (A \longrightarrow C)$. (Hint: show that each entails the other; this means constructing two tableaux. You could do this in just one huge tableau, but I think it's probably simpler to use two smaller ones.)
3. Show that each of the following entailments is valid:
   (a) $\neg A \longrightarrow \neg B \vdash B \longrightarrow \neg(A \longrightarrow \neg B)$.
   (b) $(A \longrightarrow B) \vee C, A \longrightarrow \neg C \vdash (B \longrightarrow C) \longrightarrow \neg A$
4. Show that the following formula is not a tautology; find an assignment of truth values which makes it false: $(p \vee q) \longrightarrow (p \wedge q)$. However, show that $(p \wedge q) \longrightarrow (p \vee q)$ is a tautology.
5. Here is a list of useful equivalences; verify that each is a tautology.
   (a) Commutativity:
   i. $(p \wedge q) \leftrightarrow (q \wedge p)$ ii. $(p \vee q) \leftrightarrow (q \vee p)$
   (b) Associativity:
   i. $((p \wedge q) \wedge r) \leftrightarrow (p \wedge (q \wedge r))$ ii. $((p \vee q) \vee r) \leftrightarrow (p \vee (q \vee r))$
   (c) Distributivity:
   i. $((p \wedge q) \vee r) \leftrightarrow (p \vee r) \wedge (q \vee r))$ ii. $((p \vee q) \wedge r) \leftrightarrow (p \wedge r) \vee (q \wedge r))$
   (d) De Morgan Laws:
   i. $\neg(p \wedge q) \leftrightarrow (\neg p \vee \neg q)$ ii. $\neg(p \vee q) \leftrightarrow (\neg p \wedge \neg q)$
   (e) Others:
   i. $(p \longrightarrow q) \leftrightarrow (\neg p \vee q)$ ii. $\neg(p \longrightarrow q) \leftrightarrow (p \wedge \neg q)$
6. Here is an argument: *If there is a blizzard, the highway will be in poor condition. If the highway is in poor condition, I will miss class unless I leave home early. Indeed, there has been a blizzard. Therefore I must leave home early to avoid missing class.* Translate this and show it is valid. (As a contrast, you might also like to construct a derivation for this argument.)
7. (The "Tardy Bus Problem")

   - If Bill takes the bus, then Bill misses his appointment if the bus is late.
   - Bill shouldn't go home, if (a) he misses his appointment and (b) he feels sad.
   - If Bill doesn't get the job, then (a) he feels sad, and (b) he should go home.

   Is it valid to conclude that if Bill doesn't miss his appointment, then (a) he shouldn't go home, and (b) he doesn't get the job?

   Show that this conclusion is *not* a valid consequence of the premises, and provide some truth assignments to the individual statements that show this.

4.3.2 More Exercises

Construct tableaux to show the following are valid; for extra practice, also construct derivations for each.

1. $A \vee B \rightarrow C \ \vdash \ (A \rightarrow C) \wedge (B \rightarrow C)$
2. $B \wedge C \rightarrow A \ , \ \neg A \rightarrow C \ , \ C \rightarrow B \vdash^{*} A$
3. $A \rightarrow B \ , \ (C \vee B) \wedge \neg B \ , \ C \rightarrow D \ \vdash \ A \vee D$
4. $A \rightarrow C \vee D \ , \ \neg B \rightarrow \neg A \ , \ C \rightarrow \neg B \ \vdash \ A \rightarrow D$
5. $(A \rightarrow B) \vee C \ , \ A \rightarrow \neg C \ , \ B \rightarrow C \ \vdash \ \neg A$
6. $(\neg A \vee B) \wedge C \ , \ \neg B \vee \neg C \ \vdash \ \neg A$
7. $P \rightarrow Q \ , \ R \rightarrow S \ , \ P \vee R \ \vdash \ Q \vee S$
8. $(P \rightarrow Q) \rightarrow P \vdash^{*} P$ (Peirce's Law)

Translate the following to appropriate symbols (use the initial for each name), and construct tableau and derivation as above.

1. Either Andy or Betty will run for President. If Andy runs for president, then Carol will be sad. Betty won't run for president if Dave isn't coming home. So, either Carol will be sad, or Dave is coming home.*
2. If God exists, he is omnipotent and omniscient; moreover he is benevolent (provided he exists). If God can prevent evil, then if he knows evil exists, he is not benevolent if he doesn't prevent it. If he is omnipotent, he can prevent evil. And if he is omniscient, he knows evil exists if it does exist. Evil does not exist if God prevents it. However, evil does exist. So God does not exist.

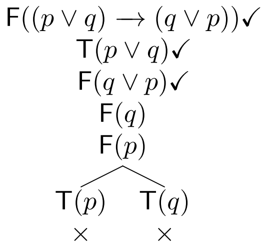
Construct tableaux to show the following are not valid; in each case, give an assignment of truth values to the variables which illustrates the invalidity of the argument.

1. $\neg(P \vee Q) \ , \ P \vee R \ , \ S \rightarrow P \vee U \ \vdash \ \neg S \wedge (Q \vee U)$
2. $A \rightarrow (B \rightarrow C) \ , \ C \wedge D \rightarrow \neg E \ , \ \neg F \rightarrow D \wedge E \ \vdash \ \neg C \rightarrow \neg E \vee \neg F$

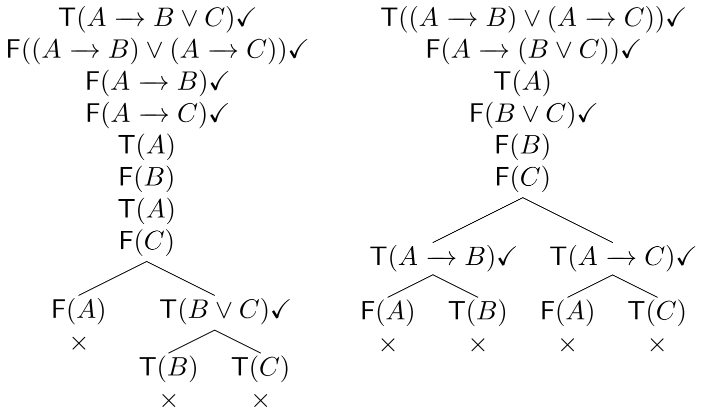
4.4 Answers to the exercises

Exercise 4.3.1

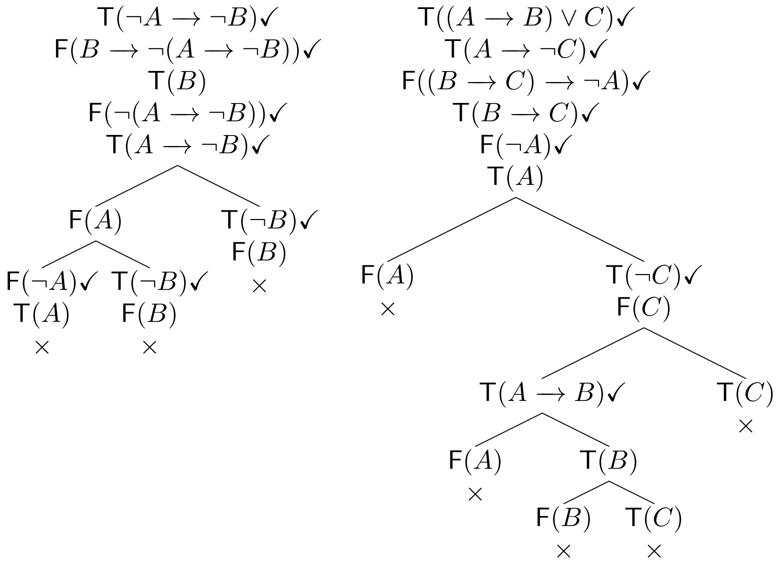
1.



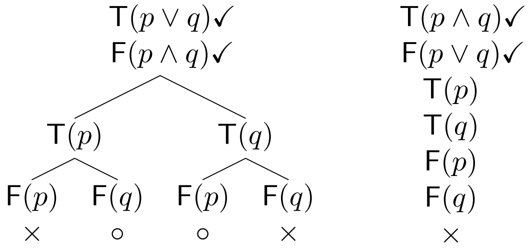
2.


$\mathsf{T}(A \rightarrow B \vee C)$✓
$\mathsf{F}((A \rightarrow B) \vee (A \rightarrow C))$✓
$\mathsf{F}(A \rightarrow B)$✓
$\mathsf{F}(A \rightarrow C)$✓
$\mathsf{T}(A)$
$\mathsf{F}(B)$
$\mathsf{T}(A)$
$\mathsf{F}(C)$
$\mathsf{F}(A)$
×
$\mathsf{T}(B \vee C)$✓
$\mathsf{T}(B)$
×
$\mathsf{T}(C)$
×
$\mathsf{T}((A \rightarrow B) \vee (A \rightarrow C))$✓
$\mathsf{F}(A \rightarrow (B \vee C))$✓
$\mathsf{T}(A)$
$\mathsf{F}(B \vee C)$✓
$\mathsf{F}(B)$
$\mathsf{F}(C)$
$\mathsf{T}(A \rightarrow B)$✓
$\mathsf{T}(A \rightarrow C)$✓
$\mathsf{F}(A)$
×
$\mathsf{T}(B)$
×
$\mathsf{F}(A)$
×
$\mathsf{T}(C)$
×

3.

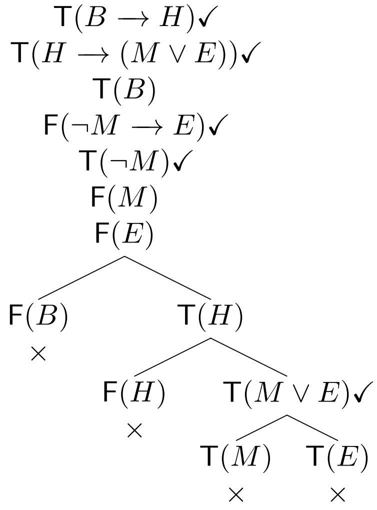

$\mathsf{T}(\neg A \rightarrow \neg B)$✓
$\mathsf{F}(B \rightarrow \neg(A \rightarrow \neg B))$✓
$\mathsf{T}(B)$
$\mathsf{F}(\neg(A \rightarrow \neg B))$✓
$\mathsf{T}(A \rightarrow \neg B)$✓
$\mathsf{F}(A)$
$\mathsf{F}(\neg A)$✓
$\mathsf{T}(A)$
×
$\mathsf{T}(\neg B)$✓
$\mathsf{F}(B)$
×
$\mathsf{T}(\neg B)$✓
$\mathsf{F}(B)$
×
$\mathsf{T}((A \rightarrow B) \vee C)$✓
$\mathsf{T}(A \rightarrow \neg C)$✓
$\mathsf{F}((B \rightarrow C) \rightarrow \neg A)$✓
$\mathsf{T}(B \rightarrow C)$✓
$\mathsf{F}(\neg A)$✓
$\mathsf{T}(A)$
$\mathsf{F}(A)$
×
$\mathsf{T}(\neg C)$✓
$\mathsf{F}(C)$
$\mathsf{T}(A \rightarrow B)$✓
$\mathsf{F}(A)$
×
$\mathsf{T}(B)$
$\mathsf{F}(B)$
×
$\mathsf{T}(C)$
×
$\mathsf{T}(C)$
×

4.

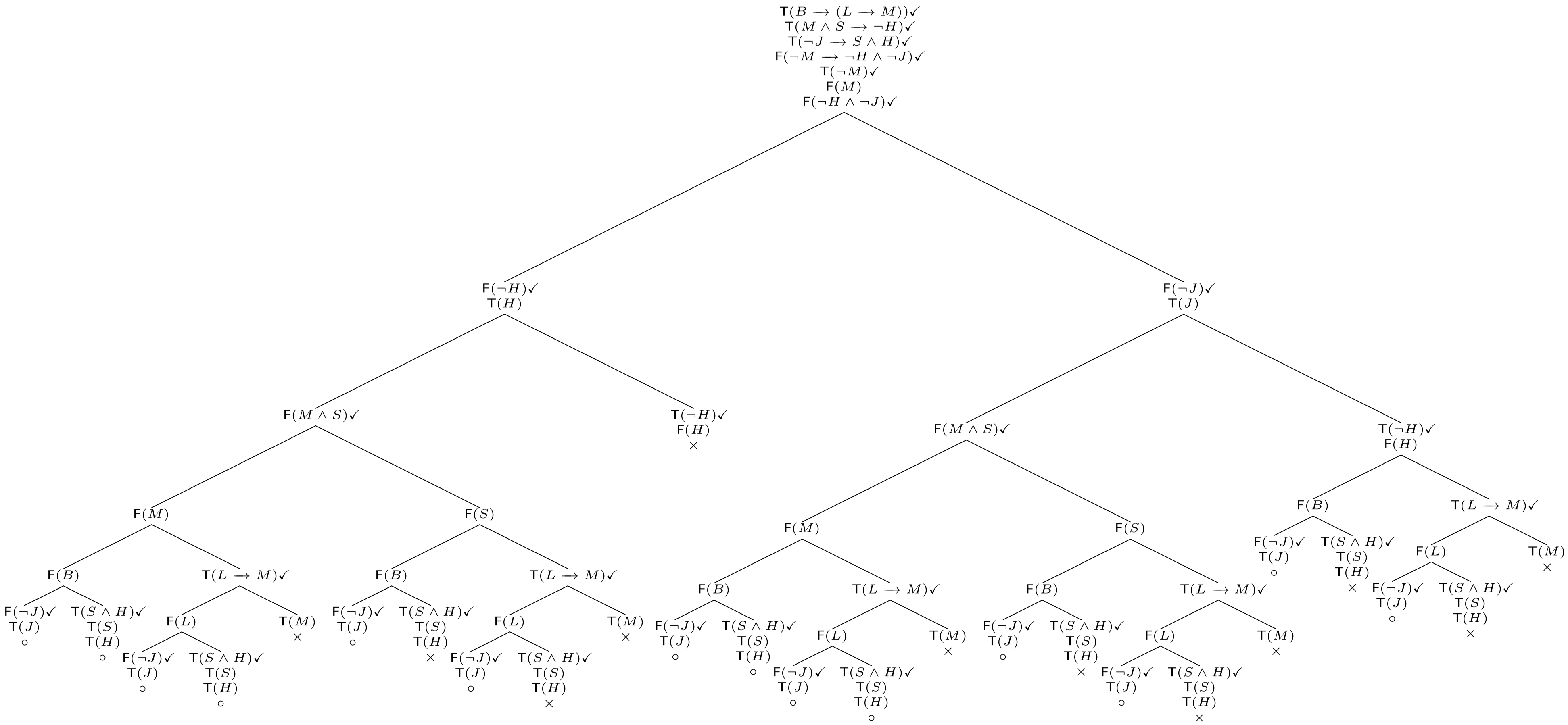


Truth value specification: either $p = \top, q = \bot$ or $p = \bot, q = \top$, read off the two open paths, indicated with a ○.

6.



I'll let you do #5—none are difficult, exercise 1 illustrated the idea, and one aim was to remind you of these basic equivalences! #6: Above, right. #7: Next page

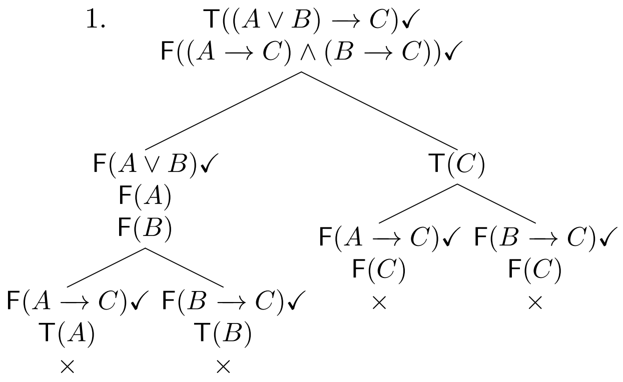


This is so full of open paths! You may choose any one to get a truth-value specification as required. For instance, from the right most open path, we get $H = \bot, J = \top, L = \bot, M = \bot$ (and B, S are arbitrary).

Exercise 4.3.2

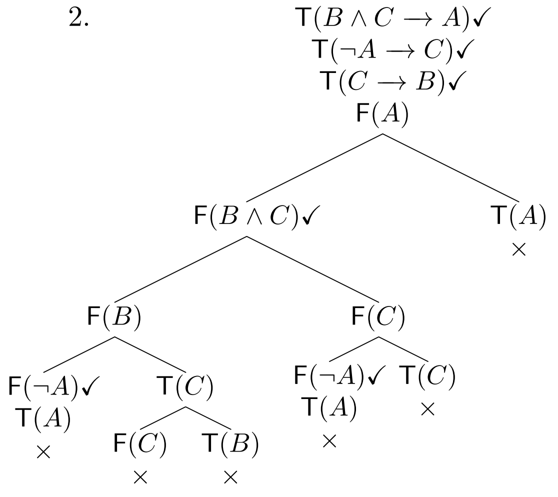
The valid ones:

1.



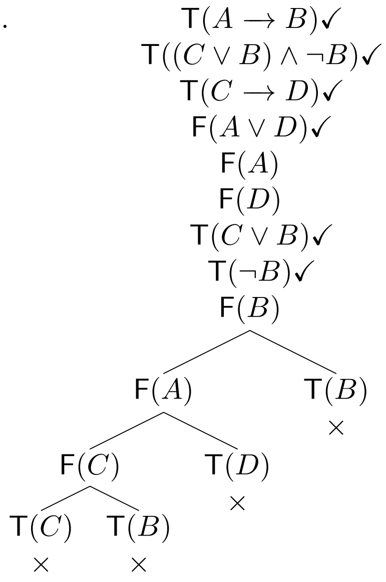
1	$(A \vee B) \to C$	
2	A	
3	$A \vee B$	(∨I), 2
4	C	(→E), 1, 3
5	$A \to C$	(→I), 2–4
6	B	
7	$A \vee B$	(∨I), 6
8	C	(→E), 1, 7
9	$B \to C$	(→I), 6–8
10	$(A \to C) \wedge (B \to C)$	(∧I), 5, 9

2.



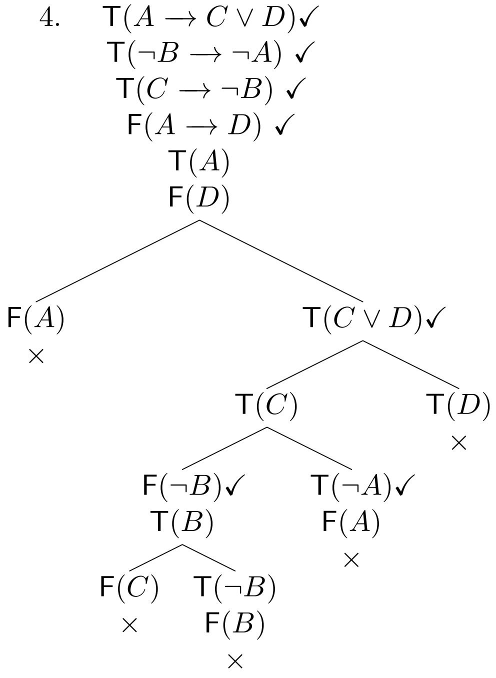
1	$B \wedge C \to A$	
2	$\neg A \to C$	
3	$C \to B$	
4	$\neg A$	
5	C	(→E), 2, 4
6	B	(→E), 3, 5
7	$B \wedge C$	(∧I), 5, 6
8	A	(→E), 1, 7
9	$\bot$	(¬E), 4, 8
10	$\neg\neg A$	(¬I), 4–9
11	A	(¬¬E), 10

3.



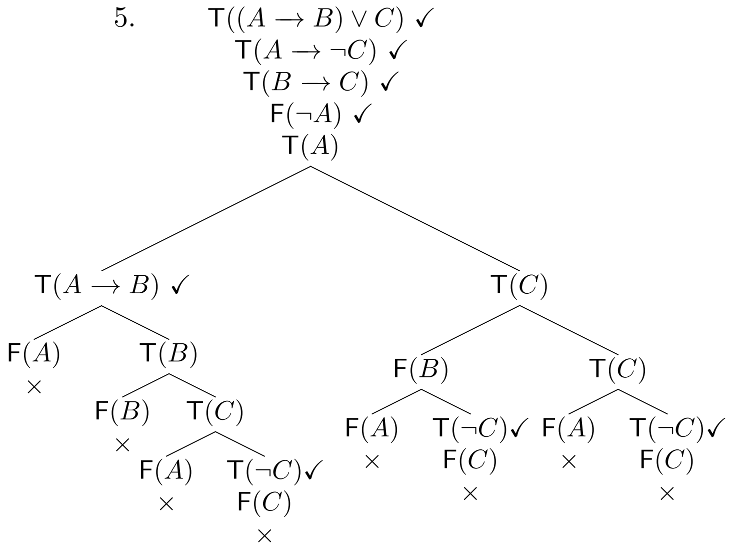
$$\begin{array}{rlll}
1 & \mid & A \to B & \\
2 & \mid & (C \lor B) \land \neg B & \\
3 & \mid & \underline{C \to D} & \\
4 & \mid & C \lor B & (\land\mathrm{E}),\ 2 \\
5 & \mid & \neg B & (\land\mathrm{E}),\ 2 \\
6 & \mid & \quad \mid \underline{C} & \\
7 & \mid & \quad \mid D & (\to\mathrm{E}),\ 3,\ 6 \\
8 & \mid & \quad \mid A \lor D & (\lor\mathrm{I}),\ 7 \\
9 & \mid & \quad \mid \underline{B} & \\
10 & \mid & \quad \mid \bot & (\neg\mathrm{E}),\ 5,\ 9 \\
11 & \mid & \quad \mid A \lor D & (\bot\mathrm{E}),\ 10 \\
12 & \mid & A \lor D & (\lor\mathrm{E}),\ 4,\ 6\text{–}8,\ 9\text{–}11
\end{array}$$

4.



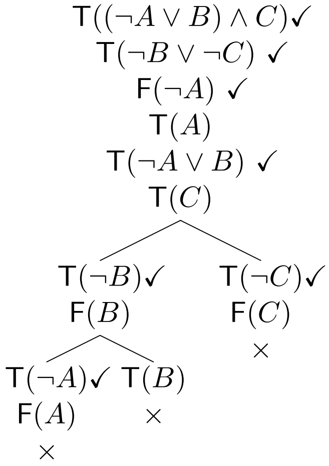
$$\begin{array}{rlll}
1 & \mid & A \to C \lor D & \\
2 & \mid & \neg B \to \neg A & \\
3 & \mid & \underline{C \to \neg B} & \\
4 & \mid & \quad \mid \underline{A} & \\
5 & \mid & \quad \mid C \lor D & (\to\mathrm{E}),\ 1,\ 4 \\
6 & \mid & \quad \mid \quad \mid \underline{C} & \\
7 & \mid & \quad \mid \quad \mid \neg B & (\to\mathrm{E}),\ 3,\ 6 \\
8 & \mid & \quad \mid \quad \mid \neg A & (\to\mathrm{E}),\ 2,\ 7 \\
9 & \mid & \quad \mid \quad \mid \bot & (\neg\mathrm{E}),\ 4,\ 8 \\
10 & \mid & \quad \mid \quad \mid D & (\bot\mathrm{E}),\ 9 \\
11 & \mid & \quad \mid \quad \mid \underline{D} & \\
12 & \mid & \quad \mid \quad \mid D & (\mathrm{R}),\ 11 \\
13 & \mid & \quad \mid D & (\lor\mathrm{E}),\ 5,\ 6\text{–}10,\ 11\text{–}12 \\
14 & \mid & A \to D & (\to\mathrm{I}),\ 4\text{–}13
\end{array}$$

5.



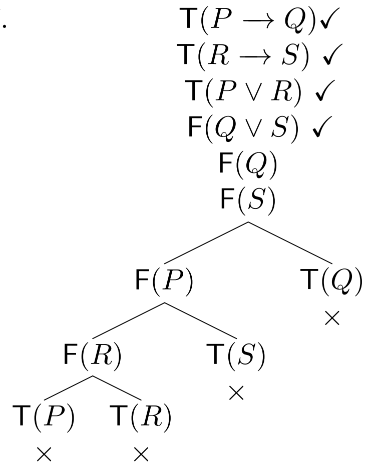
1	│ $(A \to B) \lor C$	
2	│ $A \to \neg C$	
3	│ $B \to C$	
4	│ │ A	
5	│ │ $\neg C$	($\to$E), 2, 4
6	│ │ │ $A \to B$	
7	│ │ │ B	($\to$E), 4, 6
8	│ │ │ C	($\to$E), 3, 7
9	│ │ │ $\perp$	($\neg$E), 5, 8
10	│ │ │ C	
11	│ │ │ $\perp$	($\neg$E), 5, 10
12	│ │ $\perp$	($\lor$E), 1, 6–9, 10–11
13	│ $\neg A$	($\neg$I), 4–12

6.



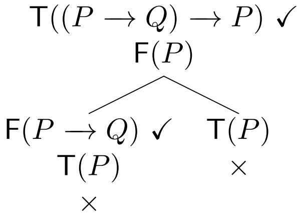
1	│ $(\neg A \lor B) \land C$	
2	│ $\neg B \lor \neg C$	
3	│ │ A	
4	│ │ $\neg A \lor B$	($\land$E), 1
5	│ │ C	($\land$E), 1
6	│ │ │ $\neg A$	
7	│ │ │ $\perp$	($\neg$E), 3, 6
8	│ │ │ B	
9	│ │ │ │ $\neg B$	
10	│ │ │ │ $\perp$	($\neg$E), 8, 9
11	│ │ │ │ $\neg C$	
12	│ │ │ │ $\perp$	($\neg$E), 5, 11
13	│ │ │ $\perp$	($\lor$E), 2, 9–10, 11–12
14	│ │ $\perp$	($\lor$E), 4, 6–7, 8–13
15	│ $\neg A$	($\neg$I), 3–14

7.



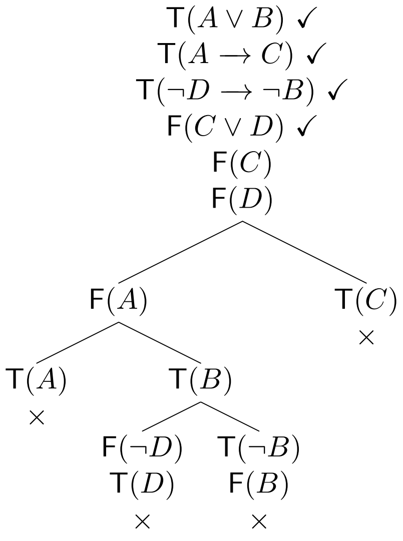
1	$\mid P \to Q$	
2	$\mid R \to S$	
3	$\mid P \lor R$	
4	$\mid \mid P$	
5	$\mid \mid Q$	($\to$E), 1, 4
6	$\mid \mid Q \lor S$	($\lor$I), 5
7	$\mid \mid R$	
8	$\mid \mid S$	($\to$E), 2, 7
9	$\mid \mid Q \lor S$	($\lor$I), 8
10	$\mid Q \lor S$	($\lor$E), 3, 4–6, 7–9

8.



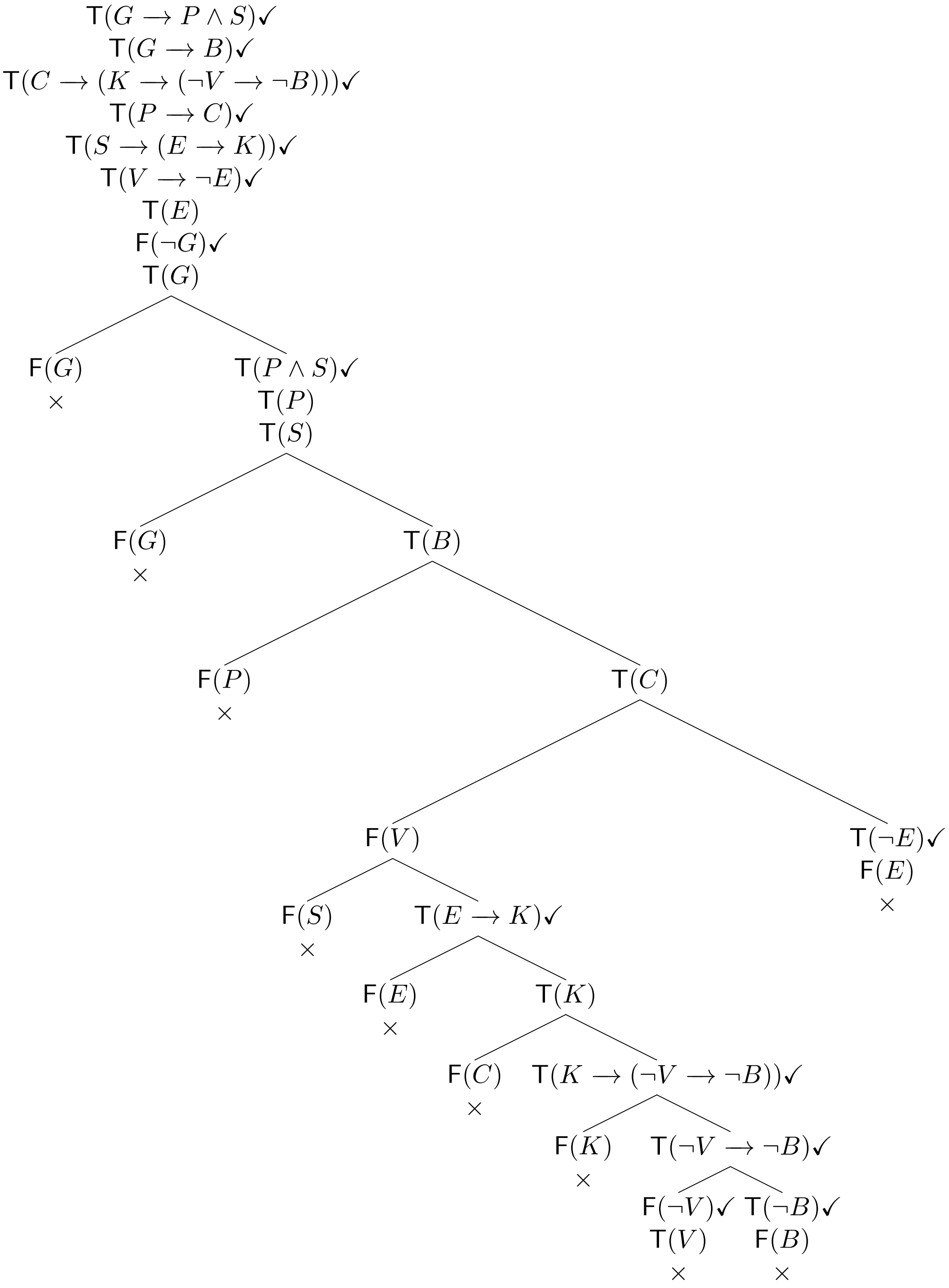
1	$\mid (P \to Q) \to P$	
2	$\mid \mid \neg P$	
3	$\mid \mid \mid P$	
4	$\mid \mid \mid \bot$	($\neg$E), 2, 3
5	$\mid \mid \mid Q$	($\bot$E), 4
6	$\mid \mid P \to Q$	($\to$I), 3–5
7	$\mid \mid P$	($\to$E), 1, 6
8	$\mid \mid \bot$	($\neg$E), 2, 7
9	$\mid \neg\neg P$	($\neg$I), 2–8
10	$\mid P$	($\neg\neg$E), 9

1. The “word problems”



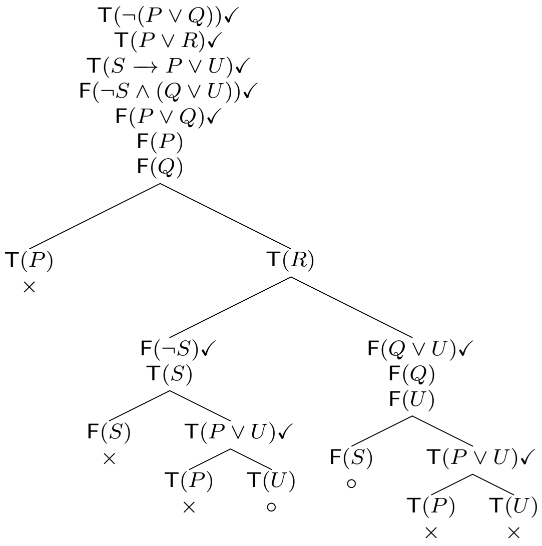
1	$\mid A \lor B$	
2	$\mid A \to C$	
3	$\mid \neg D \to \neg B$	
4	$\mid \mid A$	
5	$\mid \mid C$	($\to$E), 2, 4
6	$\mid \mid C \lor D$	($\lor$I), 5
7	$\mid \mid B$	
8	$\mid \mid \mid \neg D$	
9	$\mid \mid \mid \neg B$	($\to$E), 3, 8
10	$\mid \mid \mid \bot$	($\neg$E), 7, 9
11	$\mid \mid \neg\neg D$	($\neg$I), 8–10
12	$\mid \mid D$	($\neg\neg$E), 11
13	$\mid \mid C \lor D$	($\lor$I), 12
14	$\mid C \lor D$	($\lor$E), 1, 4–6, 7–13

2. (The derivation was done in Chapter 3: Exercise 3.3.2 Q5)


$\mathsf{T}(G \rightarrow P \wedge S)\checkmark$
$\mathsf{T}(G \rightarrow B)\checkmark$
$\mathsf{T}(C \rightarrow (K \rightarrow (\neg V \rightarrow \neg B)))\checkmark$
$\mathsf{T}(P \rightarrow C)\checkmark$
$\mathsf{T}(S \rightarrow (E \rightarrow K))\checkmark$
$\mathsf{T}(V \rightarrow \neg E)\checkmark$
$\mathsf{T}(E)$
$\mathsf{F}(\neg G)\checkmark$
$\mathsf{T}(G)$
$\mathsf{F}(G)$
$\times$
$\mathsf{T}(P \wedge S)\checkmark$
$\mathsf{T}(P)$
$\mathsf{T}(S)$
$\mathsf{F}(G)$
$\times$
$\mathsf{T}(B)$
$\mathsf{F}(P)$
$\times$
$\mathsf{T}(C)$
$\mathsf{F}(V)$
$\mathsf{T}(\neg E)\checkmark$
$\mathsf{F}(E)$
$\times$
$\mathsf{F}(S)$
$\times$
$\mathsf{T}(E \rightarrow K)\checkmark$
$\mathsf{F}(E)$
$\times$
$\mathsf{T}(K)$
$\mathsf{F}(C)$
$\times$
$\mathsf{T}(K \rightarrow (\neg V \rightarrow \neg B))\checkmark$
$\mathsf{F}(K)$
$\times$
$\mathsf{T}(\neg V \rightarrow \neg B)\checkmark$
$\mathsf{F}(\neg V)\checkmark$
$\mathsf{T}(V)$
$\times$
$\mathsf{T}(\neg B)\checkmark$
$\mathsf{F}(B)$
$\times$

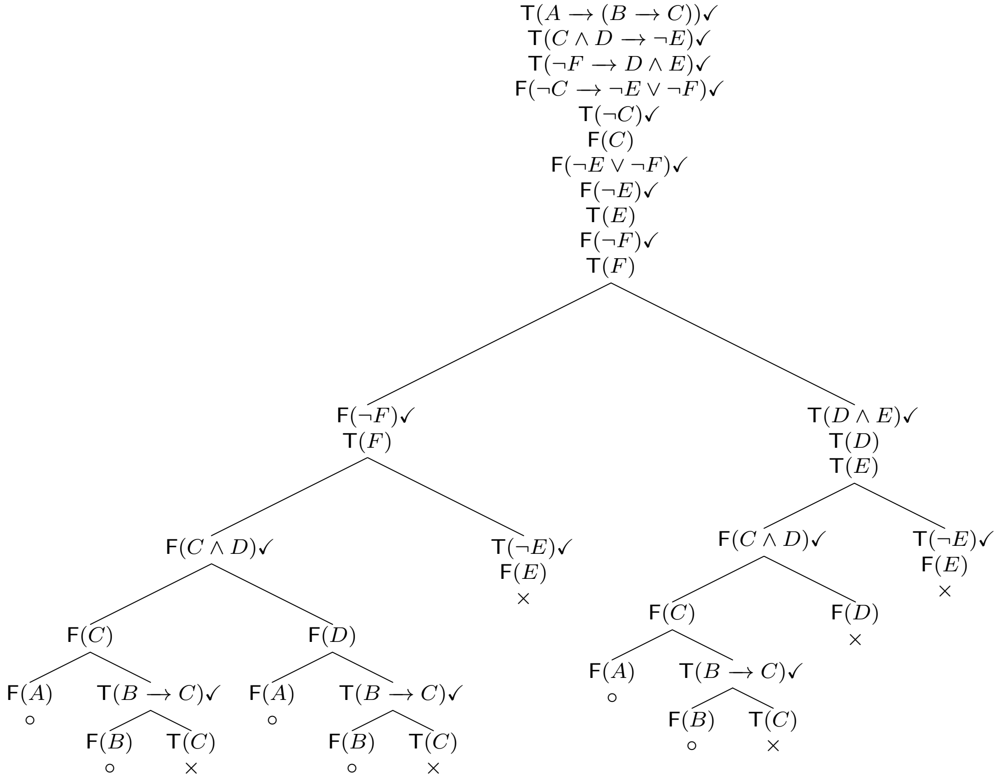
The invalid ones:

1.



So: $P = Q = \bot$, $R = S = U = \top$
or $P = Q = S = U = \bot$, $R = \top$

2.



So: Many possibilities, essentially amounting to
A or $B = \bot$, $C = \bot$, $E = F = \top$, and possibly $D = \top$.

Interlude I

Some other logics

This section may be considered "optional". Its main point is to flesh out to a small extent my comment in Chapter 1 that there were many different logics, constructed for many purposes, but logics where mathematical rigour and clarity still determine the structure of the logic, and where techniques such as we have seen with classical propositional logic are still useful. The three examples below illustrate three possible ways one could alter the classical propositional logic we've been studying. First, one could drop some derivation rules from our set to get a logic that cannot construct all the derivations possible in classical propositional logic. Or one could add additional sentence constructors. Or one could alter the deep structure underlying our derivation rules.

I.1 Intuitionistic Logic

The first logic I will consider here is intuitionistic logic. Early in the twentieth century, dissatisfaction with one aspect of classical propositional logic found eloquent expression in the writings of several mathematicians, Brouwer and Poincaré in particular. The point was this: there are many contexts, even within mathematics (and certainly outside mathematics) where the idea that 'if a statement is not false, then it must be true' seemed ... well, fishy! (though I suppose it would sound more "scholarly" were I to say "suspect"). In other words, one may reasonably doubt that $p \vee \neg p$ is a tautology, in the sense that there are statements for which it seems not obviously true. Here is an example: consider the decimal expansion of the number π: $3.1415926535\ldots$. Define a new number δ in the following manner: start by duplicating the decimal expansion of π, so that the first few decimals of δ are 3.1415926535, but make the following crucial distinction: if the decimal expansion of π contains a sequence of 5 million (or more) consecutive 1s, then change the first 5 million 1s (and no others) to 0s. Now, here's the problem: at present, no one knows whether or not the decimal expansion of π does contain a sequence of 5 million 1s, so at present we don't know if δ equals π, or if it is strictly smaller. Although we have a definite algorithm for calculating δ, we don't know how big it is, really. And so it's an article of faith to assert that "either δ is smaller than π or it isn't". This is related to another philosophical observation: if one asserts "$p \vee q$", one should know which of p or q it is that is true. That is not so with classical propositional logic. So one of the intentions of intuitionistic logic was that one should be more "constructive" in one's assertions. If $\vdash p \vee q$ is valid, then one should have either $\vdash p$ or $\vdash q$. (Notice this is not true of classical propositional logic: one has $\vdash p \vee \neg p$ without either $\vdash p$ or $\vdash \neg p$.)

It turns out that a formal logic which captures these ideas may be obtained simply from our natural deduction presentation of classical propositional logic (and also from the predicate logic of Chapter 5) merely by dropping one rule: eliminate the $(\neg\neg E)$ rule, and we have intuitionistic logic.

To be sure, we do lose many tautologies and many valid argument-forms, but we retain many as well. (This is why I indicated with an $*$ all those derivations which used this rule—the ones without that $*$ are still valid intuitionistically.)

Some things do become more complicated in the intuitionistic setting. We can no longer use truth tables or analytic tableau (since they built the double negation rule in from the start). Versions of these which comply with the intuitionistic view are more complicated than is worth describing here. But the payoff is that in some contexts, derivations that are intuitionistically valid have constructive content which is lacking in classical derivations (which use the double negation rule), constructive content that makes them useful in contexts such as theoretical computer science. I cannot go into full detail here, but the following interpretation of intuitionistic propositional logic might make this claim at least plausible. To describe this interpretation, I need some simple ideas from set theory—you may delay reading the next paragraphs until after we have studied sets in Chapter 6, or you might just remember enough from high school to make sense of this.

The idea is to think of formulas as sets (you may even think of identifying a formula with the set of its proofs, in a sense—I will use that language, but in fact one may be more abstract and not bother identifying just what these sets contain). $\perp$ may be thought of as an empty set (it has no proofs), and $\top$ has one element (since we think of $\top$ as an obviously true statement, it needs only one obviously correct proof(!)). A conjunction $A \wedge B$ may be thought of as all ordered pairs $\langle a, b\rangle$, where a is a proof of A, and b is a proof of B. A disjunction $A \vee B$ may be thought of as the set of pairs $\langle i, x\rangle$, where i is either 0 or 1, and if $i = 0$ then x must be a proof of A, and if $i = 1$, then x must be a proof of B. (Notice that an element of the "set" $A \vee B$ tells you which of A or B it comes from, and so this is what represents the notion that if $A \vee B$ is true, one should be able to determine which disjunct is responsible for that.)

An implication $A \longrightarrow B$ may be thought of as the set of all functions which send proofs of A to proofs of B (so a typical element of the set representing $A \longrightarrow B$ would be a function f, whose domain is the set representing A, and for an a in that set—thought of as a proof of A—$f(a)$ is an element of the set representing B, thought of as a proof of B). This also tells us how to interpret $\neg A$, since we may regard it as $A \longrightarrow \perp$: it is the set of all functions which take proofs of A to proofs of $\perp$—but there are no such proofs of $\perp$, so $\neg A$ must be empty unless A itself is empty, in which case $\neg A$ would have exactly one element (the empty function between the empty set and itself, in a sense, the function which does nothing to nothing). This does fit the basic idea: there should not be any proofs of $\neg A$ if there are proofs of A, but if A has no proof, then $\neg A$ may have a proof (and in fact just one).

What does this mean for $\neg\neg A$? If A were not empty (so there are proofs of A), then $\neg A$ would be empty, and in that case $\neg\neg A$ would be a single element set. There is clearly a map (in fact, only one) from A to $\neg\neg A$ in this case. Suppose A were empty, however (so there are no proofs of A). In this case, $\neg A$ has a single element, so is not empty, and then $\neg\neg A$ would have to be empty, so again, we have a map (just one) from A to $\neg\neg A$. In each case there is only one such map, so there is no doubt what it must be. (We shall make this more precise in a moment.)

But now consider the reverse direction. Is there in general any reason to believe there should always be maps from $\neg\neg A$ to A; any reason to regard $\neg\neg A \longrightarrow A$ as always "inhabited" by a proof? The answer is "no", because without knowing whether A is inhabited or not, you don't know what function might take you from a proof of $\neg\neg A$ to a proof of A. If A is inhabited, there are generally many such functions to choose between (one for each element of A), and so there is no general description we can use to specify a general map from $\neg\neg A$ to A.

We can specify the function taking you from proofs of A to proofs of $\neg\neg A$ in a somewhat abstract, but completely general, manner; it is a bit of a tongue-twister, so let's go slowly. A proof of $\neg\neg A$ is a function which takes proofs of $\neg A$ to proofs of $\perp$, so what we really want is a function which takes us from a proof of A together with a proof of $\neg A$ to a proof of $\bot$. But a proof of $\neg A$ is itself a function taking proofs of A to proofs of $\bot$, so all we need to do is apply that function to the proof of A we started with, and we get that illusive proof of $\bot$. (This is a very curious process, since there are no proofs of $\bot$(!), but it really does work, and so this justifies the intuitionist view that $A \longrightarrow \neg\neg A$ is tautologous, but not the converse $\neg\neg A \longrightarrow A$.)

By the way, this is a special case (when $B = \bot$) of the fact that there is always a map from proofs of A to proofs of $(A \longrightarrow B) \longrightarrow B$, given essentially by "evaluation": given an element of A and a map from A to B, we get the desired element of B by applying the map to the element of A.

The idea that propositions in intuitionistic logic behave like sets has been found useful in several contexts. One is in the study of computer programs: one might treat a program (or better still, the specification of a program) as an entailment between "propositions" which in fact describe datatypes. A valid entailment, or better, a derivation, would then be a program that correctly carries out the intended specification. One may then use the techniques of logic to debug programs as you write them. This idea has been at the core of a lot of research activity in theoretical computer science, and has informed the development of programming languages and the analysis of programming paradigms.

I.2 Modal Logic

Modal logic is a general term used to describe a large family of logics which intend to analyse and describe such notions as "possibly", "necessarily", and others. We shall briefly describe two variants which have been found useful, either philosophically or mathematically.

In both cases, our language will be the same: to the usual connectives of classical propositional logic, we add two operators: $\Box$ and $\Diamond$. The intended meaning of $\Box p$ is "necessarily p", and the intended meaning of $\Diamond p$ is "possibly p". In addition, these are dual in both our examples: $\Diamond p$ is logically equivalent to (may even be defined as) $\neg\Box\neg p$.

So, for instance, if p represented the statement "football is fun", then $\Box p$ would mean "football is necessarily fun" and $\Diamond p$ would mean "football is possibly fun" or "it's possible that football is fun". Note that this could also be expressed by "it's not necessarily the case that football is not fun".

I.2.1 $S4$

One very common modal logic is generally known as $S4$. It adds to the usual rules of classical propositional logic the following rules. First, we have four "axioms": these are statement-forms which are intended to be tautological, in that they should be true, regardless of the truth values of their constituent atoms. They could be regarded as derivation rules with no premises, only the formulas as conclusions.

$$\begin{array}{rcl} \Box p & \longrightarrow & p \\ \Box p & \longrightarrow & \Box\Box p \\ \top & \longrightarrow & \Box\top \\ \Box p \wedge \Box q & \longrightarrow & \Box(p \wedge q) \end{array}$$

Furthermore, we add the derivation rule

$$\frac{p \longrightarrow q}{\Box p \longrightarrow \Box q}$$

You might like to interpret these rules and see if you think they capture the meaning of "necessarily". (One reason there are so many variants of modal logic—and there are many variants indeed—is that there are many disagreements as to just what properties really do characterize "necessarily".) For example, the first axiom says that if some statement is necessarily true, then it is in fact true. Furthermore, you might notice that these rules are a bit stronger than they might appear. For instance, the second axiom isn't merely an implication, but in view of the first axiom, it really establishes an equivalence $\Box p \leftrightarrow \Box\Box p$. A similar remark might be made of the third and fourth axioms. In addition, there is a derived rule:

$$\frac{p}{\Box p}$$

There is a subtlety here worth mentioning. In $S4$ it is not true that $p \longrightarrow \Box p$ is tautologous, but even if p doesn't tautologically imply $\Box p$, it is true that if you can prove p without premises, then you can prove $\Box p$ (again, without premises). (This is not true if there are other premises involved.)

By the way, if you wonder what $S4$ has to say about "possibly", remember that the usual rules of propositional logic will generate properties of $\Diamond$, since it is dual to $\Box$. So, for instance, the first axiom above will force the implication $p \longrightarrow \Diamond p$.

I.2.2 G

One variant modal logic is of interest to us particularly because it captures the notion of "provable". In this system, sometimes known as G (after Gödel, one of the great twentieth century logicians), we add the following two axioms and the following derivation rule to the usual rules of classical propositional logic.

$$\begin{aligned} \Box(p \longrightarrow q) \quad &\longrightarrow \quad (\Box p \longrightarrow \Box q) \\ \Box(\Box p \longrightarrow p) \quad &\longrightarrow \quad \Box p \end{aligned}$$

$$\frac{p}{\Box p}$$

Before discussing the meaning of $\Box$ in system G, it's worth pointing out that this system is different from $S4$. For instance, in G you cannot prove $\Box p \longrightarrow p$, nor is any statement of the form $\Diamond p$ provable in G. But if one added $\Box p \longrightarrow p$ and $\Box p \longrightarrow \Box\Box p$ as additional axioms to G, then the resulting system would just be (equivalent to) $S4$.

So, just what is the intention behind system G? One could think of $\Box p$ as "necessarily p", but more insight into the system may be obtained by thinking of $\Box p$ as meaning "p is provable". So, for instance, the first axiom above would be interpreted as saying that if you can prove $p \longrightarrow q$ and p, then you can prove q as well (which is true). This interpretation of system G is related to what are known as Gödel's incompleteness theorems, which we shall explore at the end of the course.

I.3 Substructural Logic

Our last family of variant logics attempt to address, among other things, the seeming paradox that one can have an implication $p \longrightarrow q$ be true without any connection between p and q, as in "if you are a purple unicorn, then pigs can fly". Other "issues" such logics address involve sequentiality and preservation of resources, in the following sense. In classical propositional logic, $A \wedge B$ is equivalent to $B \wedge A$, but there are contexts where that seems a bit unlikely: consider for example "John flew to Toronto and he had breakfast" compared to "John had breakfast and he flew to

Toronto". There are many uses for a logic in which "A and B" is not equivalent to "B and A" (we shall see one application when exploring sentence-generation in linguistics, at the end of the course). Furthermore, in classical propositional logic, one may re-use premises as often as necessary (this is the essence of the repetition rule), but again, one may imagine a scenario where this isn't likely. Consider for instance the fact (in classical propositional logic) that if $A \vdash B$ and $A \vdash C$, then also $A \vdash B \wedge C$. Imagine now that A represents the statement "I have $1", B represents the statement "I can buy a chocolate bar", and C represents "I can buy a pack of chips". With this interpretation, that fact of propositional logic seems questionable: "if I have $1 then I can buy a chocolate bar" and "if I have $1 then I can buy a pack of chips" may well be true and yet "if I have $1, then I can buy a chocolate bar and I can buy a pack of chips" may nonetheless be false (it is almost certainly false, in fact!). Classical propositional logic is not intended to handle such matters of limited resources, but one can easily imagine that it might be useful to have a logic which can; in fact such logics are proving very useful in computer science, as well as in quantum physics.

We will see in full detail a simple such logic in the section on sentence-generation in linguistics, so for now, I shall just point out how one might arrive at such a logic. The basic idea is to alter the fundamental context in which our derivation rules are situated. There are two basic assumptions our rules operated with: one was that it didn't matter what order the premises were given, that as long as the required premises came before the suitable conclusions, all would be well (in fact, we explicitly had some variant rules to guarantee this). That would have to be discarded. The order of premises would become important in any logic that attempts to differentiate between statements such as $A \wedge B \longrightarrow C$ and $B \wedge A \longrightarrow C$.

Next, we allowed premises to be used repeatedly in our rules—the repetition rule made this explicit, but we used this principle even without explicitly using the repetition rule. Again, that would have to be abandoned as a general principle: a resource-sensitive logic would have to be sure that every premise was used exactly once, not more, not less. (Actually, unused premises are less problematic than repeatedly used premises, and one could relax things in that direction.) So, if one needed to use a premise twice, one would have to list it twice. In our example with the money and the candy, we would be happy with the symbolic summary that $A \vdash B$ and $A \vdash C$ implies $A \wedge A \vdash B \wedge C$. "If I have $1 and $1, then I can buy a chocolate bar and a pack of chips". But in this logic, we would not expect to have a valid derivation of $A \vdash A \wedge A$. (After all, money doesn't grow on trees . . .)

Chapter 5

Formal Proof—Predicate Logic

5.1 Limitations of Propositional Logic

All men are mortal.
Socrates is a man.
Therefore Socrates is mortal.

This is the standard example of a valid deductive argument, but you cannot prove its validity in propositional logic. "All men are mortal" is a simple statement. Suppose we symbolize it with a propositional constant H. "Socrates is a man" is another simple statement, S, and "Socrates is mortal" is another, M. The argument then looks like:

$$H, S \vdash M$$

which clearly is an invalid argument. Its form is

$$p, q \vdash r$$

and it's easy enough to think of a substitution instance of r which is false when substitution instances of both p and q are true. ("$1 + 1 = 2$", "John Abbott College is a CEGEP" $\vdash$ "Bart Simpson is Prime Minister of Canada". Hardly a valid argument!) By the methods of propositional logic, the argument is invalid.

Of course, this is silly. The problem is that these sentences are not simple statements, but instead involve several component parts. "All men are mortal" considers "men" and "mortal", and makes a connection between them. It has the form "All As are B" (where A represents "man" and B represents "mortal"). Similarly for the other two statements.

Predicate logic lets us translate and construct derivations for arguments whose validity depends on the components of simple statements. This chapter introduces the symbolism of predicate logic and illustrates the kind of derivation rules it uses.

5.1.1 New argument forms

The form of the Socrates argument is:

All As are/have B.	(Premise)
C is an A.	(Premise)
C is/has B.	(Conclusion)

Other arguments that have the same form are:

All cats are furry.
Dusty is a cat.
Therefore Dusty is furry.

and

All blitzgedorffs have plurak zingers.
Gnafftzku is a blitzgedorff.
Therefore Gnafftzku has plurak zingers.

It doesn't matter what you replace the letters with, as long as A is replaced with a *kind of thing*, C with a *thing of that kind*, and B with a *property that such things have* (and you adjust for grammatical correctness). In every argument of this form, whenever the first two sentences make true statements, the third sentence will make a true statement.

The two premise-statements need not be true (not all cats need be furry). It doesn't even matter whether they are meaningful (what is a blitzgedorff? plurak? a zinger?) for us to be able to say that the argument is valid. In every argument of that form, if the premises are true, the conclusion cannot be false.

By contrast, the argument:

All Norwegians are human.
All Europeans are human.
Therefore all Norwegians are Europeans.

is *invalid* (even though its premises and conclusion are all true). It is invalid because it is an argument of the form:

All As are B.	(Premise)
All Cs are B.	(Premise)
Therefore all Cs are A.	(Conclusion)

If we replace A with "woman" and B with "human" (making appropriate grammatical adjustments), and replace C with "men", we get:

All women are human.
All men are human.
Therefore all men are women.

The premises of this argument are true; its conclusion is false. But the argument *has the same form* as the Norwegians argument, which shows that the Norwegians argument is invalid.

Note that we have not (indeed, cannot) use our propositional truth tables to show invalidity, since this analysis doesn't fit into the framework of propositional logic. We have used another technique, the method of *counterexample*: find another argument that has the same form and that has true premises and a false conclusion. But this is not a suitable method for showing an argument is valid (one could never be sure one had considered all possible counterexamples), so in this chapter we shall develop a suitable extension of the Fitch-style natural deduction to construct derivations of valid arguments. We shall not spend much time on the extension of analytic tableau to include predicate logic, but a brief introduction may be found at the end of the chapter.

5.2 Predicates

In propositional logic, the statement expressed by “Socrates was mortal” would be symbolized by a statement symbol S, “Socrates was bald” by a different symbol B. We lose the information that both statements are about the same subject. “Gandhi was bald” might be symbolized as G. “Socrates was bald and Gandhi was bald” would be $B \wedge G$, while “Socrates was mortal and Gandhi was bald” would be $S \wedge G$. The first two statements have something in common (they are both about baldness). The second pair does not. The difference is lost in the symbolism.

The “All men are mortal” argument shows that such subtleties are important for some kinds of inference. In that argument the second premise and the conclusion are both about the same *entity* (Socrates). The first and second premises are both about things that have the same *property* (being human). The link between the premise-statements is the property of humanity that all men and Socrates have in common.

“Socrates is a man” states that a *thing or entity or being* (Socrates) has a *property* (humanness).[1] The statement is true if and only if that thing actually has that property.

Predicate logic extends the formalism and methods of propositional logic so that the logical relations between subjects and predicates can be considered.

There are new WFF-rules. We symbolize a simple (non-compound) statement using two different kinds of symbols: (1) one kind of symbol to point to the *thing* that has a property, and (2) another kind of symbol that represents the *property* the thing is asserted to have. A symbol that points to a thing is called a **term**. A symbol representing a property is a **predicate** symbol. So, we assume we have a *language* (a set of basic symbols) containing a set of term constants (symbols for some of the entities which we wish to consider), and a set of predicate symbols, each of which is associated with a natural number $(0, 1, 2, 3, 4, \ldots)$ called the “arity” of the predicate symbol. (We shall explain the meaning and role of “arity” below.)

In most mathematical contexts it is also very useful to have “function symbols” of various arities, which represent functions which assign entities to entities. For example, if we were considering our entities to be numbers, we might want to have a symbol s which represented the function “+1”, so that if n was a term (representing some number), then $s(n)$ would be the term representing that number plus 1 (“the next number”, in an obvious sense). Similarly, one might want a symbol M representing the multiplication function, so $M(m, n)$ would be the term representing the product of the numbers represented by the terms m, n. Generally we won’t consider function symbols very much in this chapter, but we shall see them again when we consider Gödel’s theorem at the end of the course.

If P is a predicate symbol of arity n, and if $t_1, t_2, \ldots, t_n$ is a list of n terms, then $P(t_1, t_2, \ldots, t_n)$ is the predicate P applied to the terms $t_1, t_2, \ldots, t_n$. For example, “Socrates is mortal” becomes $M(\mathsf{s})$ where M is the symbol for the predicate “being-mortal” and s is the term or symbol for the entity called “Socrates”. “Socrates is bald” is $B(\mathsf{s})$. “Gandhi is bald” is $B(\mathsf{g})$. These two statements ($B(\mathsf{s})$ and $B(\mathsf{g})$) ascribe the same property (the property of “baldness”, represented by the symbol B) to two different things (represented by the terms s and g).

Likewise, if F is a function symbol of arity n, and if $t_1, t_2, \ldots, t_n$ is a list of n terms, then $F(t_1, t_2, \ldots, t_n)$ is the term obtained when the “function” F is applied to the terms $t_1, t_2, \ldots, t_n$, as with our example above of multiplication.

A predicate alone is a sort of pattern for a possible sentence. The predicate B in the above example stands for something like “... is bald”. The ellipsis indicates that something is missing. Clearly, “... is bald” is not a sentence and does not make a statement. It only makes a statement

[1] We shall assume that “man” in this context really means “human”—this is not a gender issue!

when we provide the missing something: a term. The truth of the resulting statement depends both on what subject the term names *and* on what property the predicate ascribes to that subject.

Because "… is bald" takes just one term as its argument (as a substitution for the "…"), we call B a unary or 1-ary predicate. But predicates may take more than one term to make them complete; such predicates are often called "relations". (It's typical of mathematicians that they then turn around and also call unary predicates "1-ary relations", but we'll not do that generally.)

By the way, a 0-ary predicate is just a propositional statement, as we've been dealing with for the past several chapters. In a similar manner, a 0-ary function symbol is just an entity constant, *i.e.* a term.

5.2.1 Predicates and relations

The sentence "Socrates is shorter than Plato" resembles "Socrates is bald" in being about both a property and entities. We analyzed the second sentence as ascribing a property (the property of being bald) to the thing called Socrates. Socrates is the subject and "… is bald" is the predicate. But it seems perverse to consider "… is shorter than Plato" as a logical predicate in "Socrates is shorter than Plato". Plato is an entity just as much as Socrates is. Plato should also be represented by a term and treated as an entity which could be replaced by another entity. We would say Socrates and Plato are *logical subjects* of the statement "Socrates is shorter than Plato". So, it makes sense in this context to represent "shorter" as a two-place ("binary") predicate, T for example (for "tiny"?), which takes two arguments: $T(\mathsf{s}, \mathsf{p})$ represents "Socrates is shorter than Plato".

"Montreal is north of Burlington" is a similar sentence, which could be represented by $N(\mathsf{m}, \mathsf{b})$, where N represents the predicate "… is north of …". We say such predicates are binary, or 2-ary, because they need two logical subjects. n-ary predicates are just predicates which require n logical subjects, where n is some number. Such predicates express relations between their subjects. For example, "Montreal is between Kingston and Quebec City" has three logical subjects (terms standing for distinct entities). It should be symbolized by something like $B(\mathsf{m}, \mathsf{k}, \mathsf{q})$. It does not assert a property of Montreal. It expresses a *relation* between three subjects. It uses a 3-ary or ternary (three-place) predicate symbol to express this relation. The predicate B is "… is between … and … ". Similarly, "I love Lucy" would be $L(\mathsf{i}, \mathsf{lucy})$[2] ("… love(s) …" is a binary predicate). "Arnie loves himself" would be $L(\mathsf{a}, \mathsf{a})$. Since "Arnie" and "himself" are two ways of referring to the same entity, we would use the same symbol a for both. A relation can be a relation of a thing to itself.

One binary relation gets special treatment in predicate logic. We use the symbol "=" to indicate a relation between two terms that name the same thing. A statement like "Lewis Carroll was Charles Dodgson" (the writer of *Alice in Wonderland* was the same person as the Oxford logician) might be symbolized as $\ell = \mathsf{c}$.

The ambiguity of the verb "to be" has caused lots of problems in philosophy. In statements like "Socrates was bald" the verb is *the "is" of predication.* It predicates the property of baldness to Socrates. In the Lewis Carroll statement, the verb is called *the "is" of identity.* In logic we use different symbols to avoid this dangerous ambiguity. So, we write $B(\mathsf{s})$ to represent "Socrates is bald" rather than $B = \mathsf{s}$. The last statement is often said "not to type correctly", meaning that when we write "$X = Y$", both X and Y should be "the same type of thing"; "Socrates" is a human, "bald" is a property of humans: these are not at all the same type of thing.

[2] Notice here that we treat "lucy" as a single symbol—even though it looks as if it is made up of several symbols! A "symbol" is not quite the same thing as a letter of the alphabet, and may *seem* to be compound, even when it is not. Usually the context makes this clear.

For example, we might have terms $\mathsf{c}, \ell, \mathsf{a}$ representing the entities Charles Dodgson, Lewis Carroll, and *Alice in Wonderland.* We might even have 'ℓ' representing the name 'Lewis Carroll' (as opposed to the person ℓ by that name). We might also have several predicates $P(x)$ ("x is a person"), $S(x, y)$ ("x is a pseudonym of y"), $T(x)$ ("x is the title of a book"), $A(x, y)$ ("x is the author of y"). Then we could "say" $P(\mathsf{c})$, $S(`\ell\text{'}, \mathsf{c})$, $T(\mathsf{a})$, $A(\ell, \mathsf{a})$, $A(\mathsf{c}, \mathsf{a})$, and maybe even $\ell = \mathsf{c}$.[3] (There is a branch of logic, called "type theory", which studies this idea of giving things "types", and restricting the logic to respect the notion of typing.)

Since $H(\mathsf{s})$ and $N(\mathsf{m}, \mathsf{b})$ and $\mathsf{c} = \mathsf{d}$ symbolize statements, we can use the connectives of propositional logic to construct new statements. Thus, we can say $H(\mathsf{s}) \wedge N(\mathsf{m}, \mathsf{b})$ (Socrates is human and Montreal is north of Burlington), and $\neg N(\mathsf{b}, \mathsf{m})$ (Burlington is not north of Montreal) and so on. "You're not funny!" would be $\neg F(\mathsf{y})$.

5.3 Quantifiers: the tale of ∀belard and ∃loise

We need still more tools before we can do the "all men are mortal" derivation. How should we symbolize "all men are mortal"?

This statement is equivalent to the conjunction "If a is a man then a is mortal and if b is a man then b is mortal and ... ". We could symbolize this as:

$$(H(\mathsf{a}) \rightarrow M(\mathsf{a})) \wedge (H(\mathsf{b}) \rightarrow M(\mathsf{b})) \wedge (H(\mathsf{c}) \rightarrow M(\mathsf{c})) \wedge \ldots$$

but the statement would be hugely long, with one conditional for every thing in our universe (the ellipsis is not meaningful in our symbolism).

"In our universe" brings up the notion of a **universe of discourse**. The universe of discourse consists of everything that could be a term in our statements. It relates to the idea of "relevance" in ordinary discourse. When you visit someone who has just redecorated her house and she says, "What do you think?" there is an implicit context where your remarks will be understood to be relevant to the redecoration. That is, if you say "Everything is boring" it is understood that "everything" includes the colours, the fabric patterns, the layout, the furniture styles, *etc.* "Everything" would not be taken to include your new car or the movie you saw last night. In predicate logic, the universe of discourse is everything that exists (*i.e.*, everything that could be referred to by a term) in any statements in a particular piece of discourse (a particular argument, a book about logic, *etc.*).

If we restrict our universe of discourse to Socrates, Plato, Aristotle and Hypatia, represented by s, p, a, and y, then "All humans are mortal" becomes

$$(H(\mathsf{s}) \rightarrow M(\mathsf{s})) \wedge (H(\mathsf{p}) \rightarrow M(\mathsf{p})) \wedge (H(\mathsf{a}) \rightarrow M(\mathsf{a})) \wedge (H(\mathsf{y}) \rightarrow M(\mathsf{y}))$$

Most discourse involves a wider range of entities. We need another solution.

When a mathematician wants to say something about a whole lot of numbers, she doesn't name every particular number to which the statement applies. She uses a variable to represent numbers. To say, "Any number greater than three is greater than two", she could say "If $x > 3$ then $x > 2$", or "$x > 3 \rightarrow x > 2$". The mathematician's universe of discourse could be specified to include only numbers, so we'd know that x has to stand for a number. In English, pronouns perform a similar function. "He" can be used to represent any male person (or cat or whatever). "He" works as a

[3] But notice that we could not have 'ℓ'$= \mathsf{c}$, nor $S(\ell, \mathsf{c})$, as these would confuse a name with an entity bearing that name; these type wrongly.

term as long as either (1) we know what person (or cat, *etc.*) it points to, or (2) we know that it doesn't matter what particular person, cat, *etc.* it points to.

a, s, p and so on are particular terms. They are like constants, like the particular numbers 3 and 2 in the arithmetic expressions above. They are not variables. We must add variables (for terms) to our setting, in order to handle unspecified terms.

$H(x)$ symbolizes "… is human". $x > 3$ works like "… is greater than three". In each case the ellipsis needs to be filled in with something that identifies just what thing(s) the predicate or relation applies to. $x > 3$ is not true when x is 1 or 2 or 3. $H(x)$ is not true when x is Dusty (my cat). We have to say what thing the variable stands for.

We used statement forms to describe general rules for operating on any statement. Expressions like $H(x)$ and $N(x, y)$ and $x > 3$ are also statement forms. They are not true or false until the variable terms x and y are given values. Statement forms like $H(x)$ and $N(x, y)$ and $x > 3$ are "propositional functions". A propositional function is the form of a predicate-logic statement.

What things does the predicate H apply to? Between what things does the relation N hold? For this course, these questions are answered by the notion of the universe of discourse: it is just the collection of things our predicates apply to.[4]

Now that we have a notion of variable term, we can express the notions of "every" and "some" in the following way, using "quantifiers".

The **Universal quantifier** is the symbol $\forall$; a universally quantified formula is the symbol $\forall$ plus a variable term (as $\forall x$) placed before a propositional function. $\forall x$ is read "for all x … " or "for any x … ". It says that the propositional function is true when x is replaced by any term that points to any logical subject in the universe of discourse. Thus "all men are mortal" would be symbolized $\forall x(H(x) \longrightarrow M(x))$. It says "for all x, if x is human then x is mortal" or "take anything in the universe of discourse: if that thing is human then that thing is mortal".

The **Existential quantifier** is the symbol $\exists$; an existentially quantified formula is the symbol $\exists$ plus a variable term (as $\exists x$) placed before a propositional function. $\exists x$ is read "there is at least one x such that … ", or (geek-speak!) "there exists at least one x so that …". It says that the propositional function is true when x is replaced by at least one of the terms that point to subjects in the universe of discourse. Thus $\exists x B(x)$ says "there is at least one x such that x is bald" or "at least one thing in the universe of discourse is bald". This is equivalent to a disjunction like $B(\mathsf{s}) \vee B(\mathsf{p}) \vee B(\mathsf{a}) \vee B(\mathsf{y})$ in the limited universe of discourse of Socrates, Plato, Aristotle and Hypatia. It is equivalent to an indefinitely long disjunction in a larger universe.

When a quantifier is applied to a statement containing a variable, that variable "disappears" as far as the meaning of the statement goes. Consider the difference between $H(x)$ and $\exists x H(x)$. The first says of (some unspecified entity called x) that x is human, whereas the second says "something is human". The x is in a sense merely a "place-holder", and could be replaced by any other variable: $\exists y H(y)$ is equivalent to $\exists x H(x)$ for any variables x, y. We say that the variable x has become "bound" by the quantifier; variables that are not bound by a quantifier are called "free" (in a sense they retain their individuality!). We make this precise in the following definition.

We assume our formal language includes a collection of variables, for example $x, y, z, w, \ldots,$ $x', y', z', \ldots, x'', \ldots$, a collection of (entity or term) constants, and a collection of predicate symbols (each with its associated arity). We define "terms" as being either variables or constants, and formulas (WFFs) as follows.

[4] In a more subtle setting, we may have several different universes of discourse, and the arity of a predicate symbol would specify which universes each of its arguments ranged over. For instance, we might say *SN* is a predicate of arity $\mathsf{P} \times \mathsf{N}$, with the intention that $SN(p, n)$ represented the statement "p has student number n", restricting p to be a *person* and n to be a *number*. This is essentially what "type theory" deals with, as referred to above.

> An *atomic formula* is an expression of the form $P(t_1, t_2, \ldots, t_n)$, where P is an n-ary predicate symbol and $t_1, t_2, \ldots, t_n$ are terms.
> A *formula* is either: an atomic formula, or
> an expression $\neg\varphi$, where φ is a formula, or
> an expression $\varphi \wedge \psi$, where φ and ψ are formulas, or
> an expression $\varphi \vee \psi$, where φ and ψ are formulas, or
> an expression $\varphi \longrightarrow \psi$, where φ and ψ are formulas, or
> an expression $\exists x\varphi$, where x is a variable and φ is a formula, or
> an expression $\forall x\varphi$, where x is a variable and φ is a formula.
>
> In the expressions $\exists x\varphi$, $\forall x\varphi$, we say the variable x is *bound* by the quantifier. We say $\exists x\varphi$ is the *scope* of the $\exists$ quantifier in $\exists x\varphi$, and similarly $\forall x\varphi$ is the scope of $\forall$ in $\forall x\varphi$.

We must be careful about parentheses in these expressions: for example, whenever φ, ψ are compound formulas, they should be enclosed in parentheses before forming the new compound formulas above. So, $\forall xP(x)$ is well-formed if $P(x)$ is atomic, but we should write $\forall x(A(x) \vee B(x))$, where we enclose the compound formula $A(x) \vee B(x)$ in parentheses before prefixing it with the quantifier.

That formula also illustrates the meaning of the scope of a quantifier and shows the need for suitable parentheses: in the expression $\forall x(A(x) \vee B(x))$, the scope is the entire formula, the x refers to the same entity in both $A(x)$ and $B(x)$. This WFF says " everything has either property A or property B".

But in $(\forall xA(x)) \vee B(x)$ the scope of $\forall$ is only $\forall xA(x)$; the x in $B(x)$ is not related to (and need not reference the same entity as) the x in $A(x)$. This is an open formula, it has one free variable (the x in $B(x)$), unlike the first formula, which had no free variables. This WFF says "either everything has property A, or x has property B".[5]

Note also that this indicates the need to be careful not to accidentally forget to include parentheses, since they indicate the scope. Our convention on parentheses implies that $\forall xA(x) \vee B(x)$ should be interpreted as $(\forall xA(x)) \vee B(x)$; if that is not the intended meaning, one *must* use parentheses to make the real meaning clear. This example also illustrates another point: it's good practice to use different variables for free variables and for bound variables, in order to make the formula easier to read. So $(\forall xA(x)) \vee B(x)$ is less likely to cause confusion if it is written $(\forall yA(y)) \vee B(x)$; remember that changing a bound variable has no effect on the meaning of a WFF.

We distinguish between WFFs that are statements (often called "sentences" or "closed formulas"), and those that are not (*i.e.* propositional functions, which sometimes we call "open formulas"). In this technical sense **sentences** are WFFs without free variables, in other words, those WFFs in which all variables have been bound by a quantifier. Sentences are those WFFs for which it makes sense to ask if they are true or not. For example, $H(x)$ is not a sentence (because x is free), and without knowing just what x was, it's not really meaningful to ask if this sentence is true. But $\forall xH(x)$ is a sentence. $\forall xH(x)$ would be read as "for any x, x is human", which means "everything is human". This is a statement of which it makes sense to ask 'is this true?"; whether or not it is true would depend on just what one's universe of discourse is. If the universe of discourse consists of everyone registered for this course, then the answer is (probably!) "yes, it is true", but if it refers to all living creatures in my house, the answer might be "no, it is false, since

[5] You might try to wriggle out of this scope-problem by thinking "x could refer to anything, so 'x has property B' really just means 'everything has property B' ". Well, you cannot get off that easily! For you still have a different statement. For instance, if we were talking about people, "everyone is male or female" doesn't mean the same thing as "everyone is male or everyone is female". You really do need to pay attention to the scope of a free variable within a quantified formula.

my cats are not human". The mathematical statement "there is at least one x, so that x is greater than three and x is less than six", or "something is between three and six", would be symbolized $\exists x(x > 3 \wedge x < 6)$. Depending on the universe of discourse, this is probably a true statement, but one cannot meaningfully say that about the open formula $x > 3 \wedge x < 6$. That is really a predicate, stating a property of the variable x.

5.3.1 Translation

Four classical "all" and "some" statement-types are symbolized as:

Form	Symbolized
All A is/are B.	$\forall x(Ax \rightarrow Bx)$
No A is/are B.	$\forall x(Ax \rightarrow \neg Bx)$
Some A is/are B.	$\exists x(Ax \wedge Bx)$
Some A is/are not B.	$\exists x(Ax \wedge \neg Bx)$

Do not translate "all whales are mammals" as $\forall x(W(x) \wedge M(x))$. It is $\forall x(W(x) \rightarrow M(x))$. The first (a universally quantified conjunction) says that everything is a whale and a mammal. That is wrong; it is just not what "all whales are mammals" means. The second ("take anything you like, if it's a whale then it's a mammal") is right. Translating "Some mathematicians are women" as $\exists x(M(x) \rightarrow W(x))$ would also be a mistake. This says that there is something such that, if it's a mathematician then it's a woman. But such a thing could be the shoe on my right foot!—that is not what was meant. What we meant to say is $\exists x(M(x) \wedge W(x))$ (there is something that is both a mathematician and a woman).

Multiple Quantifiers

Statements like "someone loves someone" require more than one quantifier. We translate it as $\exists x \exists y L(x, y)$.

How about "someone doesn't love anyone"? It's $\exists x(\neg \exists y L(x, y))$ which says there is something (x) such that it is not the case that there is something (y) such that x loves y. x and y don't have to refer to two distinct things. "Someone doesn't love anyone" includes that the person doesn't love himself. Another equally good translation would have been $\exists x \forall y \neg L(x, y)$ (there is someone (x) such that, no matter whom you pick (call that person y), x does not love y). (Exercise: What if you want to exclude self-hatred, and say "someone doesn't love anyone but himself"? Hint: you will need to use the equality predicate.)

The universe of discourse in these examples is people. If the universe of discourse included other kinds of stuff and I wanted my statements to refer just to people loving people, I would have had to use $P(x)$ (x is a person). "Someone doesn't love anyone" would then be $\exists x(P(x) \wedge \forall y(P(y) \rightarrow \neg L(x, y)))$ ("some person does not love any person").

Examples

Example: "The only good test is one that some students will fail".

$$\forall x((T(x) \wedge G(x)) \rightarrow \exists y(S(y) \wedge F(x, y)))$$

If anything (x) is a good test ("is a test and is good") then something (y) is a student and y will fail x.

Example: "Any test that every student fails is a bad test".

$$\forall x(T(x) \wedge (\forall y(S(y) \rightarrow F(x, y))) \rightarrow \neg G(x))$$

If anything (x) is a test and everything that is a student fails x, then x is not good (bad). Another way we could symbolize this is

$$\forall x(T(x) \longrightarrow (\forall y(S(y) \longrightarrow F(x,y)) \longrightarrow \neg G(x)))$$

If anything (x) is a test then if every student fails x then x is bad. These are equivalent because (exercise!) $(p \wedge q) \longrightarrow r$ is equivalent to $p \longrightarrow (q \longrightarrow r)$.

Example: "I'll go first". This means something like "I will go and, if anything goes and that thing is not me, then that thing goes after me". Using $G(\mathsf{i})$ for "I will go" and $A(x,y)$ for "x goes after y" I get $G(\mathsf{i}) \wedge \forall x((G(x) \wedge \neg(x = \mathsf{i})) \longrightarrow A(x,\mathsf{i}))$.

Try the following exercises, and if you need more help with translation, check out the *Alberta Notes*: I've provided a link to the appropriate section on the course webpage.

5.3.2 Translation exercise

Translate these statements into the symbolism of predicate logic. Specify what each of your predicates means, as "$P(x) = x$ is a politician; $C(x) = x$ is a crook", *etc.*

1. All politicians are crooks.
2. Some crooks are not politicians.
3. Some numbers are even and some are odd.
4. No non-scientists are able to repair flush toilets.
5. Some Unitarians believe in a deity.
6. Not all males are male chauvinists.
7. Some people don't love everybody.
8. Nobody knows everybody.
9. Nobody knows anybody.
10. A platypus is a mammal.
11. A number that can only be divided evenly by itself and 1 is a prime number.
12. Barbers shave all and only those who are not barbers.[6]
13. He jests at scars who never felt a wound.[7]
14. A lawyer who pleads his own case has a fool for a client.
15. Whosoever sheddeth man's blood, by man shall his blood be shed.[8]
16. Every Christian obeys all the commandments.
17. No psychiatrist can help anyone who doesn't want to be helped.
18. The first cut is the deepest.

[6] The equivalence connective may be helpful for the notion of "all and only".

[7] (Shakespeare) Include wounds and scars as *things* in your universe of discourse. "John did not feel a wound" might be symbolized as $\exists x(W(x) \wedge \neg F(\mathsf{j},x))$.

[8] Genesis 9:6.

5.4 Derivation Rules for Quantifiers

We retain all the derivation rules for propositional logic. In addition, we need special predicate-logic derivation rules to get rid of quantifiers or add quantifiers. These can be a little technical, due to the need to be careful with variables (especially when they are bound or unbound by the addition or the removal of quantifiers), but in essence they should seem very familiar. The thing to keep in mind is that $\forall$ behaves somewhat like a huge (maybe even infinitary) conjunction, and $\exists$ like a huge disjunction, as we saw when the quantifiers were introduced. So not too surprisingly, the rules for $\forall$ look a bit like the rules for $\wedge$, and the rules for $\exists$ look a bit like the rules for $\vee$. Try to see the resemblance/analogy, and it should help you remember these rules.

5.4.1 Universal quantifier rules

Suppose $\forall x P(x)$ is true: then for any entity represented by any term t, $P(t)$ is also true. For instance, if we have as a premise $\forall x(H(x) \longrightarrow M(x))$ (for any x, if x is human then x is mortal), then we can infer $H(\mathsf{s}) \longrightarrow M(\mathsf{s})$ (if Socrates is human, then Socrates is mortal). In general, from the statement that something is true of any arbitrarily selected thing x, it follows that it is true of some particular instance (thing) t.

This gives us the following elimination rule for $\forall$.

$$\begin{array}{r|ll}
\vdots & \vdots & \\
m & \forall x P(x) & \\
\vdots & \vdots & \\
n & P(t) & \quad (\forall\text{E}),\ m
\end{array}$$

where t is any term. Here is an example:

$$\begin{array}{r|ll}
1 & \forall x(H(x) \longrightarrow M(x)) & \\
2 & \underline{H(\mathsf{s})} & \\
3 & H(\mathsf{s}) \longrightarrow M(\mathsf{s}) & \quad (\forall\text{E}),\ 1 \\
4 & M(\mathsf{s}) & \quad (\longrightarrow\text{E}),\ 2,\ 3
\end{array}$$

which proves the classic "All men are mortal. Socrates is a man. Therefore Socrates is mortal". On line (3) we eliminate the quantifier by "instantiating" the variable x in the generalization (line (1)) by the particular instance: Socrates. The derivation ends with the $(\longrightarrow E)$ rule, familiar from propositional logic.

This rule should remind you of $(\wedge E)$, in that we are starting from a premise asserting many facts (for that is just what $\forall x P(x)$ does assert, *viz* $P(x)$ for every entity x in the universe of discourse), and then concluding one of those facts, *viz* $P(t)$. Remember that in effect, t is one of the many possible values x represents.

Be careful about the form of the rule: we must replace *all* occurrences of the variable x with the term t, and we must keep the *entire* predicate P when we do so; we cannot just keep part of the predicate. To use rule $(\forall E)$, the universal quantifier must be the first thing on the line, and the whole line must be in its scope. We remove the quantifier from the start of the line and replace *every* instance of the variable that was bound by that quantifier with the same particular term. For example, we can not use rule $(\forall E)$ on the statement $\forall x(H(x) \longrightarrow M(x)) \wedge H(\mathsf{s})$, because the universal quantifier has only part of the line in its scope. In this case we would first have to use $(\wedge E)$ to get the quantified expression on a line of its own, and then use $(\forall E)$.

For instance, the following are **not** valid:

$$\begin{array}{r|ll} 1 & \forall x(H(x) \to M(x)) & \\ \hline 2 & M(\mathsf{d}) & (\forall\mathrm{E}),\ 1 \end{array} \qquad \begin{array}{r|ll} 1 & \forall x(H(x) \to M(x)) & \\ \hline 2 & H(\mathsf{s}) \to M(\mathsf{d}) & (\forall\mathrm{E}),\ 1 \end{array}$$

In the first example, we have used the true premise that all men are mortal to derive "Donny (my pet rock) is mortal". But Donny (being a rock) is not mortal, so the inference is invalid. The problem is that all of the predicate $H(x) \to M(x)$ must be used in the substitution, not just $M(x)$.

In the second example, we used the true premise that all men are mortal to derive "If Socrates is a man then Donny (my pet rock) is mortal". Donny is still not mortal, but Socrates is/was a man, so the conditional on line (2) is false even if the premise (1) is true. That shows that the inference is invalid.

The introduction rule for ∀ is technically a bit trickier, as it involves a new construct in our derivations. The idea is simple enough: if we can prove $P(u)$ is true for every entity u in the universe of discourse, then we've proven $\forall x P(x)$ is true. The problem is how to write down and prove $P(u)$ for every u in the universe—in principle this would seem potentially to need an infinite number of statements and proofs. We get around this by introducing a "new" variable (for example u) into the context, a variable which does not appear anywhere else in the derivation.[9] Then we suppose that we have a subderivation without additional premises which uses that new variable, and proves it has the property represented by P, *i.e.* the subderivation should prove $P(u)$. If we can do that, in other words, if we can prove $P(u)$ for *any* u whatsoever, with no special assumptions or knowledge about u itself, then we can say we've proven "$P(u)$ for any u", or in other words, we've proven $\forall u P(u)$, and so equivalently $\forall x P(x)$. This rule is written this way:

$$\begin{array}{l|l|ll} \vdots & \vdots & & \\ m & u & \vdots & \\ \vdots & & \vdots & \\ n & & P(u) & \\ n+1 & \multicolumn{2}{|l}{\forall x P(x)} & (\forall\mathrm{I}),\ m\text{–}n \end{array}$$

where u is a "new" variable (one that doesn't appear elsewhere in the derivation, outside the specified subderivation). We "decorate" the subderivation with the u to indicate that it is new, and that its scope is only as shown, *i.e.* that it appears only in the indicated subderivation.

The idea is that this subderivation represents the potentially infinite number of derivations of the potentially infinite number of statements $P(u)$, and so in this way, this rule is analogous to $(\wedge I)$. We prove one "general" case instead of lots of specific cases.

Here is an example:

$$\begin{array}{r|l|ll} 1 & \multicolumn{2}{|l}{\forall x(P(x) \to Q(x))} & \\ 2 & \multicolumn{2}{|l}{\forall x P(x)} & \\ \hline 3 & u & P(u) \to Q(u) & (\forall\mathrm{E}),\ 1 \\ 4 & & P(u) & (\forall\mathrm{E}),\ 2 \\ 5 & & Q(u) & (\to\mathrm{E}),\ 3,\ 4 \\ 6 & \multicolumn{2}{|l}{\forall x Q(x)} & (\forall\mathrm{I}),\ 3\text{–}5 \end{array}$$

[9] Actually, we could be a little more liberal, but making precise the technical conditions on just where and how such a variable might appear elsewhere harmlessly is more trouble than it is worth, so for simplicity's sake, we shall just require that the variable be entirely new.

Note that we proved $Q(u)$ for an arbitrary variable u: since it was new, we could have used *any* other variable and got the same result, so we are (morally!) justified in claiming that we've proven $\forall x Q(x)$. (That is the intent of the $(\forall I)$ rule.)

Remark: we have essentially just shown that $\forall x(P(x) \to Q(x)) \vdash \forall x P(x) \to \forall x Q(x)$. You might be tempted to think that in fact $\forall x(P(x) \to Q(x))$ is equivalent to $\forall x P(x) \to \forall x Q(x)$, but this is not in fact true. Find an interpretation for P and Q that convinces you there is no entailment the other way, in other words, that $\forall x P(x) \to \forall x Q(x) \vdash \forall x(P(x) \to Q(x))$ is *not* valid.[10]

5.4.2 Existential quantifier rules

From the statement that Socrates is bald, we can validly infer that someone is bald. Given a statement that ascribes some property or relation to some particular thing (represented by a particular term), we can infer the statement that results by replacing zero or more instances of that term with a variable and putting the whole resulting expression within the scope of an existential quantifier using that same variable. So, from $B(\mathsf{s})$ we can infer $\exists x B(x)$. This is the basis for the introduction rule for the existential quantifier. It may be written thus:

$$\begin{array}{r|ll} \vdots & \vdots & \\ m & P(t) & \\ \vdots & \vdots & \\ n & \exists x P(x) & (\exists\text{I}),\, m \end{array}$$

The reverse inference does not work. Knowing that somebody was bald does not permit us to infer that John Lennon in particular was bald. We shall soon have a more subtle rule for existential elimination.

Here is an example. This example exposes a tiny subtle point about our quantifier rules, however—see if you can spot it.[11]

$$\begin{array}{r|ll} 1 & \forall x P(x) & \\ \cline{2-2} 2 & P(t) & (\forall\text{E}),\, 1 \\ 3 & \exists x P(x) & (\exists\text{I}),\, 2 \end{array}$$

In the $(\exists I)$ rule, please note that you do **not** have to replace all (or even any) occurrences of the term. From the premise $H(\mathsf{r}) \wedge (B(\mathsf{r}) \wedge S(\mathsf{r}))$ (Little Robin is handsome and brave and strong) we could validly infer $\exists x(H(x) \wedge (B(x) \wedge S(x)))$ ("Somebody is handsome, brave and strong"). It would also be valid to infer $\exists x(H(x) \wedge (B(\mathsf{r}) \wedge S(\mathsf{r})))$ ("Somebody is handsome and Little Robin is brave and strong"). It would not be correct to derive $\exists x H(x) \wedge (B(\mathsf{r}) \wedge S(\mathsf{r}))$, however, because $(\exists I)$ cannot be used on part (just the $H(\mathsf{r})$ part, in this example) of a line. The existential quantifier has to be put in front of the whole line, and parentheses may have to be added to ensure that the whole line

[10] This might be easier to read if we used different bound variables wherever possible: we want to show that $\forall x P(x) \to \forall y Q(y) \vdash \forall z(P(z) \to Q(z))$ is not valid. For example, take $P(x)$ to be "x is disgusting" and $Q(x)$ to be "x is disgusted". (It's possible disgusting people disgust other disgusting people, but not themselves.)

[11] The subtle point? This entailment should only be valid if our universe of discourse has some entities; otherwise the premise is vacuously true, regardless of what P is (because there are no x to verify the condition), and the conclusion is obviously false (since it says something exists, which isn't true if the universe of discourse is empty). How does this get reflected in the derivation above? Well, that derivation only works if the language you are using has at least one term. Since we are supposing we have variables, we always have them as terms, but then that does raise the following philosophical issue. Suppose the universe of discourse is empty, and so has no entities. Then we don't want this derivation to be correct. That would mean there really should be no terms, including variables, making this entailment invalid, as required. To keep things simple, we shall ignore this issue further and shall assume our universe of discourse always has some entities, and is never empty.

is within the scope of the quantifier. By the way, if we really wanted to infer $\exists xH(x)\wedge(B(\mathsf{r})\wedge S(\mathsf{r}))$ from the premise above, we could first use $(\wedge E)$ to get $H(\mathsf{r})$ on a line by itself, use $(\exists I)$ to get $\exists xH(x)$, then use $(\wedge E)$ again to get $B(\mathsf{r})\wedge S(\mathsf{r})$ from the premise, and then use $(\wedge I)$ to get the desired result.

Finally we turn to the elimination rule for the existential quantifier. Our analogy will be with the $(\vee E)$ rule, proof by cases. We shall have a version of "proof by cases" for $\exists$ as well. Let's see how this might work. Suppose $\exists xP(x)$ is true. We want to use this as a premise to derive some formula C. In the case of disjunction, we had to have subderivations for each possibility represented by the disjunctive premise; here that would amount to subderivations for each possibility $P(u)$, where u could be any entity in our universe of discourse: a potentially infinite collection of subderivations. To avoid that, we use the "new variable" trick again: we construct a subderivation with premise $P(u)$ and conclusion C, where u is a new variable, one that appears nowhere else but in this subderivation. Then since we know nothing whatsoever about u, we have in effect proved C from merely knowing that P held for something $(P(u))$. Since u is "new", it could be anything, and so this subderivation in effect represents all the cases implicit in the premise $\exists xP(x)$. So we are justified in concluding C from $\exists xP(x)$. Here is what the rule looks like

$$\begin{array}{l|lll}
\vdots & \vdots & & \\
k & \exists xP(x) & & \\
\vdots & \vdots & & \\
m & u & \multicolumn{1}{|l}{P(u)} & \\ \cline{3-3}
\vdots & & \multicolumn{1}{|l}{\vdots} & \\
n & & \multicolumn{1}{|l}{C} & \\
n+1 & C & & (\exists\text{E}),\ k,\ m\text{--}n
\end{array}$$

where u is a "new" variable (one that doesn't appear elsewhere in the derivation, outside the specified subderivation). We "decorate" the subderivation with the u to indicate that it is new, and that its scope is only as shown.

Here is an example:

$$\begin{array}{r|lll}
1 & \multicolumn{2}{l}{\exists x\forall yP(x,y)} & \\ \cline{2-3}
2 & u & \multicolumn{1}{|l}{\forall yP(u,y)} & \\ \cline{3-3}
3 & & \multicolumn{1}{|l}{P(u,u)} & (\forall\text{E}),\ 2 \\
4 & & \multicolumn{1}{|l}{\exists xP(x,x)} & (\exists\text{I}),\ 3 \\
5 & \multicolumn{2}{l}{\exists xP(x,x)} & (\exists\text{E}),\ 1,\ 2\text{--}4
\end{array}$$

5.4.3 Other rules

Since bound variables don't have any effect on the truth of a formula, we may change bound variables whenever we want. In fact, this can be proven; for example we have a (derived) rule as follows.

$$\begin{array}{l|ll}
\vdots & \vdots & \\
m & \exists xP(x) & \\
\vdots & \vdots & \\
n & \exists yP(y) & \text{Change of bound variables, } m
\end{array}$$

This (and the corresponding rule for $\forall$) is an exercise at the end of this chapter.

As with propositional logic, you may use derived rules, as long as you explicitly state them and prove them. I'll point out some examples later. Generally, this isn't something you should worry

$$\begin{array}{l|ll}
\vdots & \vdots & \\
m & u & \multicolumn{1}{|l}{\vdots} \\
\vdots & & \multicolumn{1}{|l}{\vdots} \\
n & & \multicolumn{1}{|l}{P(u)} \\
n+1 & \multicolumn{2}{l}{\forall x P(x) \qquad (\forall \mathrm{I}),\ m\text{–}n}
\end{array}$$

Universal Introduction ($\forall I$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & \forall x P(x) \\
\vdots & \vdots \\
n & P(t) \qquad (\forall \mathrm{E}),\ m
\end{array}$$

Universal Elimination ($\forall E$)

$$\begin{array}{l|l}
\vdots & \vdots \\
m & P(t) \\
\vdots & \vdots \\
n & \exists x P(x) \quad (\exists \mathrm{I}),\ m
\end{array}$$

Existential Introduction ($\exists I$)

$$\begin{array}{l|ll}
\vdots & \vdots & \\
k & \exists x P(x) & \\
\vdots & \vdots & \\
m & u & \multicolumn{1}{|l}{P(u)} \\
\cline{3-3}
\vdots & & \multicolumn{1}{|l}{\vdots} \\
n & & \multicolumn{1}{|l}{C} \\
n+1 & \multicolumn{2}{l}{C \qquad\qquad (\exists \mathrm{E}),\ k,\ m\text{–}n}
\end{array}$$

Existential Elimination ($\exists E$)

where u is a new variable in ($\forall I$) and ($\exists E$).

Table 5.1: Natural deduction rules for quantifiers

too much about; if a derived rule is really useful, I'll mention it. Otherwise you should expect to use only the four introduction and elimination rules and the rules for propositional logic.

The quantifier rules are summarized in Table 5.1.

5.5 Examples

5.5.1 Example

Let's warm up with a simple example: "Ptah is an Egyptian god. Ptah is the father of all Egyptian gods. Therefore Ptah is his own father". This translates as follows.

$$G(\mathsf{p}), \forall x(G(x) \to F(\mathsf{p}, x)) \vdash F(\mathsf{p}, \mathsf{p})$$

Here is a derivation. The idea behind this one is that we want $F(\mathsf{p}, \mathsf{p})$, which seems related to what's inside the $\forall$ sentence. So we strip away ("eliminate") that quantifier *via* ($\forall E$) which gives us an implication, which we can also eliminate to get what we want. The only trick is to choose the right term to use in the ($\forall E$) rule, but since the only term (entity) mentioned is Ptah, it's a good guess that that's the one to use!

$$\begin{array}{r|ll}
1 & G(\mathsf{p}) & \\
2 & \forall x(G(x) \to F(\mathsf{p}, x)) & \\
\cline{2-2}
3 & G(\mathsf{p}) \to F(\mathsf{p}, \mathsf{p}) & (\forall \mathrm{E}),\ 2 \\
4 & F(\mathsf{p}, \mathsf{p}) & (\to \mathrm{E}),\ 1,\ 3
\end{array}$$

5.5.2 Example

Here is a lovely example of an argument whose validity is not obvious, but which can be shown to be valid by a derivation. The argument is

> All the world loves a lover.
> Bob does not love Jane.
> Therefore Jane does not love herself.

Does that conclusion follow from those premises?

The trickiest bit in the derivation is translating the first premise. I took "All the world loves a lover" to be equivalent to "Everybody loves everybody who loves somebody". That leads to "Take any x: if there is some y that x loves, then everybody loves x". That leads in turn to "Take any x, if there is some y that x loves, then whatever z you pick, z loves x". The other two are easy enough, so we get this entailment representing the argument:

$$\forall x(\exists y L(x,y) \to \forall z L(z,x)), \neg L(\mathsf{b},\mathsf{j}) \vdash \neg L(\mathsf{j},\mathsf{j})$$

Here is a derivation: see if you can see the strategy first, before reading my comments below.

$$\begin{array}{lll}
1 & \vert\ \forall x(\exists y L(x,y) \to \forall z L(z,x)) & \\
2 & \vert\ \underline{\neg L(\mathsf{b},\mathsf{j})} & \\
3 & \vert\quad \vert\ \underline{L(\mathsf{j},\mathsf{j})} & \\
4 & \vert\quad \vert\ \exists y L(\mathsf{j},y) \to \forall z L(z,\mathsf{j}) & (\forall\text{E}),\ 1 \\
5 & \vert\quad \vert\ \exists y L(\mathsf{j},y) & (\exists\text{I}),\ 3 \\
6 & \vert\quad \vert\ \forall z L(z,\mathsf{j}) & (\to\text{E}),\ 4,\ 5 \\
7 & \vert\quad \vert\ L(\mathsf{b},\mathsf{j}) & (\forall\text{E}),\ 6 \\
8 & \vert\quad \vert\ \bot & (\neg\text{E}),\ 2,\ 7 \\
9 & \vert\ \neg L(\mathsf{j},\mathsf{j}) & (\neg\text{I}),\ 3\text{–}8
\end{array}$$

So the conclusion is a valid consequence of the premises. Consider the strategy used in constructing this derivation: We want to show $\neg L(\mathsf{j},\mathsf{j})$, so we expect to use the $(\neg I)$ rule, which tells us to introduce a subderivation with premise $L(\mathsf{j},\mathsf{j})$, which we do on line 3. Line 1 is a universal generalization. The scope of $\forall x$ includes the whole statement. Since the statement on line 1 is true for any x, it must be true when x is Jane, and that's what line 4 says. Line 5 uses $(\exists I)$ on line 3 (if "Jane loves Jane" is true, then there is at least one person or thing that Jane loves, *i.e.* Jane is a lover). From $(\to E)$ (modus ponens) we can conclude line 6 (if Jane is a lover, then everybody loves her), giving us a universally quantified statement on which we can use rule $(\forall E)$ (since everybody loves her, so does Bob), to get a contradiction (line 8) (since Bob doesn't love her). That finishes our use of $(\neg I)$, as we hoped: we may conclude that Jane does not love Jane, since if we supposed she did love herself we'd get a contradiction.

Remark: There is a possible alternate translation of the interpretation of the sentence "All the world loves a lover", which although equivalent, would make the derivation slightly different, as well as an alternate inequivalent translation of this sentence, which would make the argument invalid! Such are the consequences of the ambiguity of natural language ...

An alternate equivalent translation might be $\forall z \forall x(\exists y L(x,y) \to L(z,x))$. With this translation of "All the world loves a lover", the derivation representing the argument might look like this:

$$
\begin{array}{r|l|l}
1 & \forall z\forall x(\exists yL(x,y)\to L(z,x)) & \\
2 & \neg L(\mathsf{b},\mathsf{j}) & \\
\hline
3 & \quad \vert\ \underline{L(\mathsf{j},\mathsf{j})} & \\
4 & \quad \vert\ \forall x(\exists yL(x,y)\to L(\mathsf{b},x)) & (\forall\mathrm{E}),\ 1 \\
5 & \quad \vert\ \exists yL(\mathsf{j},y)\to L(\mathsf{b},\mathsf{j}) & (\forall\mathrm{E}),\ 4 \\
6 & \quad \vert\ \exists yL(\mathsf{j},y) & (\exists\mathrm{I}),\ 3 \\
7 & \quad \vert\ L(\mathsf{b},\mathsf{j}) & (\to\mathrm{E}),\ 5,\ 6 \\
8 & \quad \vert\ \bot & (\neg\mathrm{E}),\ 2,\ 7 \\
9 & \neg L(\mathsf{j},\mathsf{j}) & (\neg\mathrm{E}),\ 3\text{–}8
\end{array}
$$

This is equivalent to our first translation, since

$$\forall z\forall x(\exists yL(x,y)\to L(z,x))\equiv\forall x(\exists yL(x,y)\to\forall zL(z,x))$$

To verify this, prove (exercise!) that in general

$$\forall z\forall x(P(x)\to Q(z,x))\equiv\forall x(P(x)\to\forall zQ(z,x))$$

so the previous equivalence is just the case where $P(x)$ is $\exists yL(x,y)$ and $Q=L$. (The answer may be found in the "Answers to the exercises".)

But there is another translation of "All the world loves a lover", which is **not** equivalent—it simply does not mean the same thing as the translations above (this is because "All the world loves a lover" has two possible meanings in English). In our translations above, we imagined that if x was a lover, then everybody loved x, but one might have meant something more modest: that everybody loves some x who is a lover, but not necessarily the same x for everybody, in other words, everybody loves a lover, but not necessarily all lovers. Some folks might love one lover, and others might love another. That might translate thus: $\forall x\exists y(L(x,y)\wedge\exists zL(y,z))$. And with this WFF, the argument is no longer valid, simply because Bob might love someone other than Jane (and not love Jane), and so be a lover, and Jane might love someone, even herself, and so be a lover, without any contradiction. Isn't ambiguous natural language lovely?

5.5.3 Example

"If there are no unicorns, then if Benji is a unicorn then Benji is orange". This could be done two ways: we could take "There are no unicorns" and "Benji is a unicorn" as premises, and have the conclusion be simply "Benji is orange". Or we could treat the whole sentence as one single unit, a conclusion without any premises. (There's another way: can you see it?) We'll illustrate the second view, a conclusion without premises. That gives us the entailment

$$\vdash\neg\exists xU(x)\to(U(\mathsf{b})\to O(\mathsf{b}))$$

which may be proved as follows. (Note that having two implications suggests we use $(\to I)$ twice, which means having doubly nested subderivations, each with its own premise. The inner subderivation is essentially what one would produce if one handled this as a simple conclusion with two premises.)

$$\begin{array}{rll}
1 & \mid\;\mid \underline{\neg\exists xU(x)} & \\
2 & \mid\;\mid\;\mid \underline{U(\mathsf{b})} & \\
3 & \mid\;\mid\;\mid \exists xU(x) & (\exists\mathrm{I}),\ 2 \\
4 & \mid\;\mid\;\mid \bot & (\neg\mathrm{E}),\ 1,\ 3 \\
5 & \mid\;\mid\;\mid O(\mathsf{b}) & (\bot\mathrm{E}),\ 4 \\
6 & \mid\;\mid U(\mathsf{b}) \to O(\mathsf{b}) & (\to\mathrm{I}),\ 2\text{–}5 \\
7 & \mid \neg\exists xU(x) \to (U(\mathsf{b}) \to O(\mathsf{b})) & (\to\mathrm{I}),\ 1\text{–}6
\end{array}$$

Note the use of contradiction to produce what we wanted at line 5. The whole point of this argument is that by starting with an assumption that nothing has a particular property, any further conclusions about things with that property must be valid, however silly they might seem (because they are conclusions about non-existent entities). That's the way material implication works, as we've seen before.

5.5.4 Example

Here is a very important entailment (for reasons we'll discuss in a moment):

$$\forall x(\neg P(x)) \vdash \neg\exists xP(x)$$

$$\begin{array}{rll}
1 & \mid \underline{\forall x(\neg P(x))} & \\
2 & \mid\;\mid \underline{\exists xP(x)} & \\
3 & \mid\;\mid\; u \mid \underline{P(u)} & \\
4 & \mid\;\mid\;\quad \mid \neg P(u) & (\forall\mathrm{E}),\ 1 \\
5 & \mid\;\mid\;\quad \mid \bot & (\neg\mathrm{E}),\ 3,\ 4 \\
6 & \mid\;\mid \bot & (\exists\mathrm{E}),\ 2,\ 3\text{–}5 \\
7 & \mid \neg\exists xP(x) & (\neg\mathrm{I}),\ 2\text{–}6
\end{array}$$

This is one half of an equivalence, in fact. Here is the other direction.

$$\neg\exists xP(x) \vdash \forall x(\neg P(x))$$

$$\begin{array}{rll}
1 & \mid \underline{\neg\exists xP(x)} & \\
2 & \mid\; u \mid\;\mid \underline{P(u)} & \\
3 & \mid\;\quad \mid\;\mid \exists xP(x) & (\exists\mathrm{I}),\ 2 \\
4 & \mid\;\quad \mid\;\mid \bot & (\neg\mathrm{E}),\ 1,\ 3 \\
5 & \mid\;\quad \mid \neg P(u) & (\neg\mathrm{I}),\ 2\text{–}4 \\
6 & \mid \forall x(\neg P(x)) & (\forall\mathrm{I}),\ 2\text{–}5
\end{array}$$

We see here that negation and the quantifiers work together much the same way negation works with conjunction and disjunction: it "flips" between them. Explicitly, here is the equivalence above, and a simple variant (which you should try to prove yourself!)

$$\begin{array}{rcl}
\neg\exists xP(x) & \leftrightarrow & \forall x(\neg P(x)) \\
\exists x(\neg P(x)) & \leftrightarrow & \neg\forall xP(x)
\end{array}$$

(The second equivalence requires the double negation rule $(\neg\neg E)$.)

You may use this equivalence when working with quantified formulas, should you think it helps. Just reference it as "Equivalence", giving the line number of the original quantified statement you are replacing with an equivalent one. Here is an example:

$$\neg\exists xF(x) \vdash F(\mathsf{a}) \to G(\mathsf{a})$$

$$\begin{array}{r|l l}
1 & \neg\exists x F(x) & \\ \cline{2-2}
2 & \forall x(\neg F(x)) & \text{Equivalence, 1} \\
3 & \quad \begin{array}{|l} F(\mathsf{a}) \\ \hline \neg F(\mathsf{a}) \\ \bot \\ G(\mathsf{a}) \end{array} & \begin{array}{l} \\ (\forall \mathrm{E}),\ 2 \\ (\neg \mathrm{E}),\ 3,\ 4 \\ (\bot \mathrm{E}),\ 5 \end{array} \\
7 & F(\mathsf{a}) \rightarrow G(\mathsf{a}) & (\rightarrow \mathrm{I}),\ 3\text{–}6
\end{array}$$

(Exercise: Construct a derivation without using the equivalence. It must be possible, as you could always just "build in" the proof of the half-equivalence needed here. But see if you can find a shorter way. Here's a hint: the direct derivation is actually one line shorter! Another hint: "Benji"!)

5.5.5 Remark

There is a famous[12] paradox of quantification, at least in classical logic, called "The drinkers' paradox". The following statement is a tautology "there's someone in the bar with the property that if they're drinking, then everyone's drinking", $\vdash \exists x(D(x) \rightarrow \forall y D(y))$. This may easily be seen to be valid, since $\exists x(D(x) \rightarrow \forall y D(y))$ is equivalent to $\neg\forall y D(y) \vee \forall y D(y)$, using tautologies we've seen already: $\exists x(D(x) \rightarrow \forall y D(y) \equiv \exists x(\neg D(x) \vee \forall y D(y)) \equiv \exists x(\neg D(x)) \vee \forall y D(y) \equiv \neg\forall x D(x) \vee \forall y D(y)$. This requires the $(\neg\neg E)$ rule; in fact it cannot be proven in intuitionist logic, without $(\neg\neg E)$. So maybe only intuitionists are sober . . .[13]

5.6 Exercises

Construct derivations for each of the following entailments. (As before, a starred exercise uses the $(\neg\neg E)$ rule.) In questions 25, 26 and 27, where I've explicitly indicated "not the converse", find a model or situation which shows the converse is not valid.

1. $P(\mathsf{a}), Q(\mathsf{a}) \vdash \exists x(P(x) \wedge Q(x))$
2. $\forall x(P(x) \rightarrow Q(x)), P(\mathsf{a}) \vdash Q(\mathsf{a})$
3. $\exists x\forall y A(x,y) \vdash \neg\forall x\exists y\neg A(x,y)$
4. $\forall x(R(x) \rightarrow B(x)), \neg B(\mathsf{a}) \vdash \neg R(\mathsf{a})$
5. $R(\mathsf{a}), \forall x(\neg G(x) \rightarrow \neg R(x)), M(\mathsf{b}) \vdash^* \exists x G(x) \wedge \exists x M(x)$
6. $\forall x((R(x) \wedge A(x)) \rightarrow T(x)), A(\mathsf{b}), R(\mathsf{b}) \vdash \exists x T(x)$
7. $\forall x(P(x) \wedge Q(x)) \vdash \forall x P(x) \wedge \forall x Q(x)$
8. $\forall x P(x) \wedge \forall x Q(x) \vdash \forall x(P(x) \wedge Q(x))$
9. $\forall x(\exists y P(y) \rightarrow Q(x)) \vdash \exists y P(y) \rightarrow \forall x Q(x)$ (where there is no free variable x in $\exists y P(y)$)
10. $\exists y\forall x A(x,y), \forall x\forall y(A(x,y) \rightarrow B(x,y)) \vdash \forall x\exists y B(x,y)$
11. $\forall x\exists y A(x,y), \forall x\forall y(A(x,y) \rightarrow A(y,x)), \forall x\forall y\forall z(A(x,y) \wedge A(y,z) \rightarrow A(x,z)) \vdash \forall x A(x,x)$
12. $\exists x(P(x) \rightarrow \forall y Q(y)) \vdash \forall x P(x) \rightarrow \forall x Q(x)$
13. $\forall x(A(x) \rightarrow B(x)) \vdash \exists x A(x) \rightarrow \exists x B(x)$
14. $P \rightarrow \forall x Q(x) \vdash \forall x(P \rightarrow Q(x))$
15. $\forall x R(x) \vee \forall x S(x) \vdash \forall x(R(x) \vee S(x))$

[12] I think this was first made famous by Raymond Smullyan (in *What is the Name of this Book*).

[13] But there is a classically equivalent statement which *is* provable intuitionistically—so they're only moderately sober.

16. $\forall x(P(x) \rightarrow Q) \vdash \exists x P(x) \rightarrow Q$

17. $\forall x(P(x) \rightarrow Q(x)), \forall x(Q(x) \rightarrow R(x)) \vdash \forall x(P(x) \rightarrow R(x))$

18. $\forall x \forall y P(x, y) \vdash P(\mathsf{a}, \mathsf{a})$ 19. $\forall x \forall y P(x, y) \vdash \forall x P(x, x)$

20. $\forall x(P(x) \wedge Q(x) \rightarrow R(x)), Q(\mathsf{a}) \wedge \forall z P(z) \vdash P(\mathsf{a}) \wedge R(\mathsf{a})$

21. $\exists x P(x) \vdash \forall x Q(x) \rightarrow \exists x(P(x) \wedge Q(x))$

22. $\forall x \forall y(R(x, y) \rightarrow (P(x) \wedge \neg P(y))), \exists x \exists y(R(x, y) \wedge R(y, x)) \vdash \exists x(P(x) \wedge \neg P(x))$
(and hence Q23:)

23. $\forall x \forall y(R(x, y) \rightarrow (P(x) \wedge \neg P(y))), \exists x \exists y(R(x, y) \wedge R(y, x)) \vdash \bot$

24. $\exists z R(z, z), \exists y \forall x S(y, x) \vdash \exists y \exists z(S(z, y) \rightarrow R(y, y))$

25. $\exists x(P(x) \wedge Q(x)) \vdash \exists x P(x) \wedge \exists x Q(x)$ (but not the converse!)

26. $\forall x P(x) \vee \forall x Q(x) \vdash \forall x(P(x) \vee Q(x))$ (but not the converse!)

27. $\exists x \forall y R(x, y) \vdash \forall y \exists x R(x, y)$ (but not the converse!)

28. $\exists x \exists y R(x, y) \vdash \exists y \exists x R(x, y)$ and $\exists y \exists x R(x, y) \vdash \exists x \exists y R(x, y)$

29. $\forall x \forall y R(x, y) \vdash \forall y \forall x R(x, y)$ and $\forall y \forall x R(x, y) \vdash \forall x \forall y R(x, y)$

30. $\neg \exists x(P(x) \wedge Q(x)) \vdash \forall x(P(x) \rightarrow \neg Q(x))$

31. $\forall x(P(x) \rightarrow \neg Q(x)) \vdash \neg \exists x(P(x) \wedge Q(x))$

32. $\exists x(P(x) \vee Q(x)) \vdash \exists x P(x) \vee \exists x Q(x)$

33. $\exists x P(x) \vee \exists x Q(x) \vdash \exists x(P(x) \vee Q(x))$

34. $\forall x(P(x) \wedge Q(x)) \vdash \forall x P(x) \wedge \forall x Q(x)$ 35. $\forall x P(x) \wedge \forall x Q(x) \vdash \forall x(P(x) \wedge Q(x))$

36. $\forall x(P(x) \rightarrow Q(x)), \exists x(P(x) \wedge R(x)) \vdash \exists x(Q(x) \wedge R(x))$

37. $\forall x(P(x) \vee Q(x)), \exists x \neg P(x) \vdash \exists x Q(x)$

5.6.1 Word problem exercises

For each of the following, translate the argument and construct a derivation for it.

1. Generous people are happy; Albie is intelligent, but not happy. Everyone is either generous or they're not very free with their money. Hence, someone is intelligent but not very free with their money.

2. Groucho isn't a member of any club that is willing to have him as a member. Any club that isn't willing to have Groucho as a member doesn't. Therefore Groucho isn't a member of any club.

3. Bruce is charismatic. Bruce will retire to Australia only if everyone is satisfied. Everyone is happy if they are satisfied. Everyone will retire to Australia if someone is charismatic. Hence everyone is happy.

4. France is a country bigger than Luxembourg; some country is bigger than France. If something is bigger than a second thing, and the second is bigger than a third, then the first is bigger than the third. So, some country is bigger than either France or Luxembourg (*i.e.* bigger than France *and* bigger than Luxembourg—this is a case where in English usage or really means $\wedge$—do you see why?—which just goes to show how much more confusing everyday language is than logic!).

5. Only people who are neither wealthy nor famous are logicians. Anybody who doesn't need to ask the price of anything is wealthy. So logicians need to ask the price of something. (*)

6. Everybody loves somebody. Hence nobody doesn't love anybody.

7. Here's an almost biblical example:
   If the first be greater than the second, then the second cannot be greater than the first. God is that which is greater than all things. That which **is** is greater than that which **is not**. Something exists.[14] Therefore, God exists. (*)

   (If you find this too confusing, look up the translation in the solutions, and then try to construct the derivation. Don't just give up!)

8. Some students like Roger; all teachers like any student; Roger is a teacher. Therefore there is someone who both likes and is liked by Roger.

9. All horses are animals; therefore all heads of horses are heads of animals

10. Some teachers like all students; no teacher likes any jerk. Therefore no students are jerks.

11. Show that the following argument is inconsistent (meaning that these statements, taken as premises, allow a contradiction $\bot$ to be derived).

    People who steal are breaking the law. People who download songs *via* BitTorrent are stealing. Some people do download songs *via* BitTorrent. People who download songs *via* BitTorrent aren't breaking the law.

    (It's interesting to note that without the third assumption ("Some people do download songs *via* BitTorrent"), this argument is not inconsistent—can you see why?)

[14] Maybe you do, because you think??

Appendix: Tableau for Predicate Logic

In Chapter 4 we considered analytic tableau for propositional logic, as a technical method for determining if an entailment is valid, invalid, or satisfiable. Is there an extension of tableau for predicate logic? The answer is yes, and we shall briefly describe what that is, with a few examples to illustrate such tableau.

We need four new tableau rules for the quantifiers. We leave it to the reader to verify that these do indeed capture the "meaning" of the quantifiers, as embodied by the classical derivation rules. (As with propositional logic, the $(\neg\neg E)$ rule is embedded in these rules, so the quantifiers are dual to each other.)

$$[\mathsf{T}\forall] \quad \begin{array}{c} \mathsf{T}(\forall x A(x)) \\ | \\ \mathsf{T}(A(t)) \end{array} \qquad\qquad [\mathsf{F}\forall] \quad \begin{array}{c} \mathsf{F}(\forall x A(x)) \\ | \\ \mathsf{F}(A(u)) \end{array}$$

$$[\mathsf{T}\exists] \quad \begin{array}{c} \mathsf{T}(\exists x A(x)) \\ | \\ \mathsf{T}(A(u)) \end{array} \qquad\qquad [\mathsf{F}\exists] \quad \begin{array}{c} \mathsf{F}(\exists x A(x)) \\ | \\ \mathsf{F}(A(t)) \end{array}$$

where t is any term, and u is a new variable.

For example, here is a tableau showing that $\forall x(H(x) \to M(x)), H(\mathsf{s}) \vdash M(\mathsf{s})$ is valid (our old Socratic friend).

$$\begin{array}{c} \mathsf{T}(\forall x(H(x) \to M(x))) \checkmark \\ \mathsf{T}(H(\mathsf{s})) \\ \mathsf{F}(M(\mathsf{s})) \\ \mathsf{T}(H(\mathsf{s}) \to M(\mathsf{s})) \checkmark \\ \diagup \qquad \diagdown \\ \begin{array}{cc} \mathsf{F}(H(\mathsf{s})) & \mathsf{T}(M(\mathsf{s})) \\ \times & \times \end{array} \end{array}$$

Notice in the fourth line that we've used the $(\mathsf{T}\forall)$ rule, replacing the x in the first line with s.

Another example: we can show that $\exists x \forall y R(x,y) \vdash \forall y \exists x R(x,y)$ is valid with this tableau. (This is just Exercise 5.6, #27.)

$$\begin{array}{c} \mathsf{T}(\exists x \forall y R(x,y)) \checkmark \\ \mathsf{F}(\forall y \exists x R(x,y)) \checkmark \\ \mathsf{T}(\forall y R(u,y)) \checkmark \\ \mathsf{F}(\exists x R(x,v)) \checkmark \\ \mathsf{T}(R(u,v)) \\ \mathsf{F}(R(u,v)) \\ \times \end{array}$$

This simple tableau uses all four quantifier rules, with u, v being first introduced as new variables, then introduced as instances of an arbitrary term. Check to be sure you understand how the rules have been used.

And more, we can show that the converse $\forall y \exists x R(x, y) \vdash \exists x \forall y R(x, y)$ is *not* valid:

$$\begin{gathered}
\mathsf{T}(\forall y \exists x R(x,y)) \checkmark \\
\mathsf{F}(\exists x \forall y R(x,y)) \checkmark \\
\mathsf{T}(\exists x R(x,t)) \checkmark \\
\mathsf{T}(R(u,t)) \\
\mathsf{F}(\forall y R(u,y)) \checkmark \\
\mathsf{F}(R(u,v)) \\
\circ
\end{gathered}$$

This tableau (with one open path) shows that as long as the predicate R allows for some object u to have two objects t, v so that one pair satisfies R and the other does not (as shown above), then the reverse entailment will indeed be invalid. An example of such a R might be $R(x, y)$ means x is y's parent. Surely a single u might well be the parent of one individual t, but not of another v. And indeed, the argument "If everyone has a parent, then there is one individual who is everyone's parent" is surely invalid, though "If there is an individual who is everyone's parent, then everyone has a parent" is valid.

And finally, a slightly more fun example, showing that the following statement is not satisfiable:

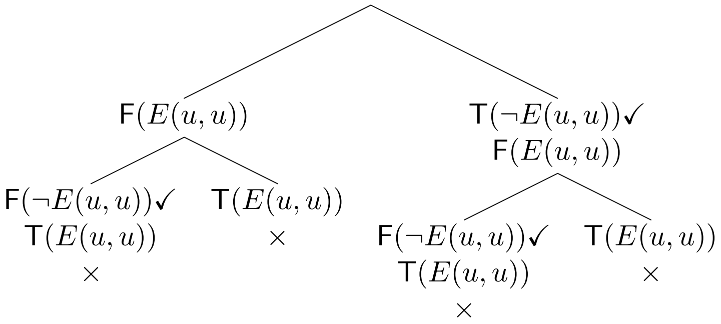
> There is a set whose members are exactly those sets which are not members of themselves.

We translate this using the predicates $S(x) =$"x is a set", and $E(y, x) =$"y is an element of x", so the sentence above becomes

$$\exists x[S(x) \wedge \forall y(E(y,x) \longleftrightarrow \neg E(y,y))]$$

Here is a tableau that shows any attempt to make this true (T) ends in contradiction (*i.e.* a closed path). (The contradiction is already evident by the fifth line; the rest is merely following the rules of tableau to close the paths.)

$$\begin{gathered}
\mathsf{T}(\exists x[S(x) \wedge \forall y(E(y,x) \longleftrightarrow \neg E(y,y))])\checkmark \\
\mathsf{T}(S(u) \wedge \forall y(E(y,u) \longleftrightarrow \neg E(y,y)))\checkmark \\
\mathsf{T}(S(u)) \\
\mathsf{T}(\forall y(E(y,u) \longleftrightarrow \neg E(y,y)))\checkmark \\
\mathsf{T}(E(u,u) \longleftrightarrow \neg E(u,u))\checkmark \\
\mathsf{T}(E(u,u) \longrightarrow \neg E(u,u))\checkmark \\
\mathsf{T}(\neg E(u,u) \longrightarrow E(u,u))\checkmark
\end{gathered}$$



Tableaux Exercises

1. Prove the following entailments with tableaux:
   (a) $\forall x(P(x) \to Q) \vdash \exists x P(x) \to Q$ (b) $\forall x \exists y P(x,y) \vdash \neg \exists x \forall y \neg P(x,y)$

2. Try the following one yourself: translate the following argument and construct a tableau to show that it is valid (though, if you succeed, not sound!). Build a derivation as well.

   > If anyone can solve this problem, then any mathematician can solve it. Bob is a mathematician, but he cannot solve it. Therefore, nobody can solve it.

   Use $S(x) =$“x can solve this problem”, $M(x) =$“x is a mathematician”, and $\mathsf{b} =$“Bob”.

3. Finally, show that change of bound variables can be derived from the other four quantifier rules: $\exists x P(x) \vdash \exists y P(y)$ and $\forall x P(x) \vdash \forall y P(y)$. Construct derivations and tableaux, for these (very simple!) entailments.

Appendix: Deduction rules for equality

Earlier (section 5.2.1) we considered (very briefly) having an equality predicate $s = t$, which is essential for the mathematical use of logic; one might ask if equality can be integrated into formal logic in the same way the logical connectives and quantifiers are handled, with introduction and elimination rules. The answer is simply “yes”. In this appendix, we'll see how this may be done; the rather nice fact is that the introduction and elimination rules for equality are basic and very familiar properties of equality—in the best sense, its addition retains the “natural” nature of natural deduction. The other usual (“standard”) properties of equality follow from these, as we'll see from some simple exercises. Some details will be left to the reader.

So we start with the elimination rule for equality (or rather one version of it—there is a “permuted variation of this rule, as we've seen with the $(\wedge I)$ rule, for example):

$$\begin{array}{l|ll}
\vdots & \vdots & \\
m & t = s & \\
\vdots & \vdots & \\
n & P(t) & \\
\vdots & \vdots & \\
k & P(s) & (=\mathrm{E}),\, m,\, n
\end{array}$$

(and the permuted version, where the premises may be given in the opposite order).

This rule is often known as the “Substitution” rule (meaning that one may substitute a term with an equal term in any WFF).[15]

A subtle point: As with the $(\exists I)$ rule, we have some flexibility with the $(= E)$ rule in terms of substitutions. If $t = s$ and $P(t)$, then when we use $(= E)$ to conclude $P(s)$, we do not actually have to replace *all* instances of t by s. For example (a silly example, for sure(!), but one which will reappear in the exercises), if $P(t)$ were $t = t$, then with the additional hypothesis $t = s$, we could use $(= E)$ to correctly give us the conclusion $s = t$, by substituting s for t in the first occurrence

[15] Note that the $(= E)$ rule may not hold for non-truth-functional statements. For example, “Jack knows that $1 + 1 = 2$” may be true, and yet “Jack knows that $1 + 1 = \sum_{i=0}^{\infty} 2^{-i}$” may be false, even though $\sum_{i=0}^{\infty} 2^{-i} = 2$ is in fact true (Jack may not know this!).

of t in $P(t)$, *i.e.* in $t = t$. We could also get the conclusion $s = s$ by substituting s for t in both occurrences.

The equality introduction rule is an axiom (*i.e.* a rule with no premises):

$$\begin{array}{c|l} \vdots & \vdots \\ k & t = t \quad (=\mathrm{I}) \end{array}$$

for any term t. This axiom is often known as "Reflexivity".

From these simple rules, the usual properties of equality follow. We illustrate this with the following exercises, which begin with some simple derived rules.

Equality Exercises

1. Show that the Substitution rule $(= E)$ is easily generalized:

$$\begin{array}{l|l} \vdots & \vdots \\ n_1 & t_1 = s_1 \\ n_2 & t_2 = s_2 \\ \vdots & \vdots \\ n_k & t_k = s_k \\ \vdots & \vdots \\ m & P(t_1, t_2, \ldots, t_k) \\ \vdots & \vdots \\ \ell & P(s_1, s_2, \ldots, s_k) \quad (=\mathrm{E}),\ n_1,\ .\ .\ ,\ n_k,\ m \end{array}$$

   (and permuted versions)

2. Show these two other properties of equality are derivable:

   First, symmetry:

$$\begin{array}{c|l} \vdots & \vdots \\ n & t = s \\ \vdots & \vdots \\ k & s = t \quad (\mathrm{Sym}),\ n \end{array}$$

   And also transitivity:

$$\begin{array}{l|l} \vdots & \vdots \\ n & t_1 = t_2 \\ n+1 & t_2 = t_3 \\ \vdots & \vdots \\ m & t_1 = t_3 \quad (\mathrm{Trans}),\ n,\ n+1 \end{array}$$

   (hint: you'll need to use symmetry to derive transitivity)

   Some other derivation exercises:

3. (A very simple one!:) Show $\vdash \exists x(x = t)$ for any term t (corresponding to an object in the universe of discourse). Obviously, this assumes we have excluded the "empty" universe of discourse from consideration; this is for the same reason we mentioned in the discussion (footnote) of the validity of $\forall x P(x) \vdash \exists x P(x)$, which also is not valid for an empty universe of discourse.

4. Show these WFFs are equivalent: $\exists x(x = t \wedge P(x))$ and $P(t)$.

5. Show these WFFs are equivalent: $\forall x(x = t \longrightarrow P(x))$ and $P(t)$.

   In other words, show $\forall x(x = t \longrightarrow P(x)) \vdash P(t)$ and $P(t) \vdash \forall x(x = t \longrightarrow P(x))$.

6. Similarly, show these WFFs are equivalent: $\forall x \forall y(x = y \longrightarrow P(x, y))$ and $\forall x P(x, x)$.

7. Next, show these entailments are equivalent: $P(t(x)) \vdash Q(x)$ and $P(y) \vdash \forall x(t(x) = y \longrightarrow Q(x))$

   In other words show that if $\pi_1(x)$ is a derivation $P(t(x)) \vdash Q(x)$ then there is a derivation $\pi_2(y)$ of $P(y) \vdash \forall x(t(x) = y \longrightarrow Q(x))$, and conversely, if $\pi_2(y)$ is a derivation $P(y) \vdash \forall x(t(x) = y \longrightarrow Q(x))$ then there is a derivation $\pi_1(x)$ of $P(t(x)) \vdash Q(x)$. (Note that the term t has a free variable x; it is or is constructed from a function symbol. Also, note that the derivations themselves depend on free variables, as indicated by the notation. Actually, there is even more structure here, but we shall leave that unexplored in this text.)

8. Similarly, show these entailments are equivalent: $P(x) \vdash Q(t(x))$ and $\exists x(t(x) = y \wedge P(x)) \vdash Q(y)$.

   Again, this means given a derivation of one sequent, one can construct a derivation of the other. These derivations depend on the free variables that appear in the entailments, as in the previous exercise.

9. Show that the following two WFFs are equivalent and that they have the (same) meaning, namely that the universe discourse has exactly one element with property P:

   $\exists x(P(x) \wedge \forall y(P(y) \longrightarrow x = y))$ and $\exists x P(x) \wedge \forall x \forall y(P(x) \wedge P(y) \longrightarrow x = y)$

Now try some translation problems: translate the following into suitable symbols and show the arguments are valid.

1. All logicians are crazy. Bob is a logician. Bob is Professor Frankenstein, so therefore Prof Frankenstein is crazy.

2. No logician is sensible. Bob is a logician. Professor Frankenstein is sensible, so therefore Prof Frankenstein is not Bob.

3. Only Bob and Harry are at work late tonight; they are both doing logic homework. Therefore everyone working late tonight is doing logic homework. (“Only” is a bit tricky—there are several ways it could be translated; I’ve given one in the answers. Feel free to explore some other possibilities, but be sure that you render “only” with both parts of its meaning: that Bob and Harry both do work late, and that no one else does.)

4. There is at most one logician at John Abbott. Bob is a logician. Harry is not Bob. Therefore Harry is not a logician.

5.7 Answers to the exercises

Exercise 5.3.2

Variations are also possible—ask me if you are not sure about your own answers. (I'll leave you to guess my notation!)

1. $\forall x(P(x) \to C(x))$
2. $\exists x(C(x) \land \neg P(x))$
3. $\exists x(N(x) \land E(x)) \land \exists x(N(x) \land O(x))$
4. $\forall x(RFT(x) \to S(x))$
5. $\exists x(U(x) \land BD(x))$
6. $\exists x(M(x) \land \neg MC(x))$
7. $\exists x \neg \forall y L(x,y)$
8. $\neg \exists x \forall y K(x,y)$
9. $\neg \exists x \exists y K(x,y)$
10. $\forall x(P(x) \to M(x))$
11. $\forall x(DE(x) \to P(x))$
12. $\forall x(B(x) \to \forall y(S(x,y) \equiv \neg B(y)))$
13. $\forall x(\forall y(W(y) \to \neg F(x,y)) \to JS(x))$
14. $\forall x(L(x) \land PHOC(x) \to FFC(x))$
15. $\forall x(\exists y\, SB(x,y) \to \exists y\, SB(y,x))$
16. $\forall x \forall y(Chr(x) \land Comm(y) \to Obey(x,y))$
17. $\forall x \forall y(P(x) \land \neg W(y) \to \neg H(x,y))$
18. $\forall x(C(x) \land F(x) \to D(x))$

Example 5.5.2 (the story of Bob and Jane)

The equivalence: $\forall z \forall x(P(x) \to Q(z,x)) \equiv \forall x(P(x) \to \forall z Q(z,x))$

1		$\forall z \forall x(P(x) \to Q(z,x))$	
2	u	$P(u)$	
3	v	$\forall x(P(x) \to Q(v,x))$	($\forall$E), 1
4		$P(u) \to Q(v,u)$	($\forall$E), 3
5		$Q(v,u)$	($\to$E), 2, 4
6		$\forall z Q(z,u)$	($\forall$I), 3–5
7		$P(u) \to \forall z Q(z,u)$	($\to$E), 2–6
8		$\forall x(P(x) \to \forall z Q(z,x))$	($\forall$I), 2–7

1		$\forall x(P(x) \to \forall z Q(z,x))$	
2	v u	$P(u)$	
3		$P(u) \to \forall z Q(z,u)$	($\forall$E), 1
4		$\forall z Q(z,u)$	($\to$E), 2, 3
5		$Q(v,u)$	($\forall$E), 4
6		$P(u) \to Q(v,u))$	($\to$I), 2–5
7		$\forall x(P(x) \to Q(v,x))$	($\forall$I), 2–6
8		$\forall z \forall x(P(x) \to Q(z,x))$	($\forall$I), 2–7

Exercise 5.6

1. $P(\mathsf{a}), Q(\mathsf{a}) \vdash \exists x(P(x) \land Q(x))$

1	$P(\mathsf{a})$	
2	$Q(\mathsf{a})$	
3	$P(\mathsf{a}) \land Q(\mathsf{a})$	($\land$I), 1, 2
4	$\exists x(P(x) \land Q(x))$	($\exists$I), 3

2. $\forall x(P(x) \to Q(x)), P(\mathsf{a}) \vdash Q(\mathsf{a})$

1	$\forall x(P(x) \to Q(x))$	
2	$P(\mathsf{a})$	
3	$P(\mathsf{a}) \to Q(\mathsf{a})$	($\forall$E), 1
4	$Q(\mathsf{a})$	($\to$E), 2, 3

3. $\exists x\forall yA(x,y) \vdash \neg\forall x\exists y\neg A(x,y)$

$$\begin{array}{rll}
1 & \mid\ \underline{\exists x\forall yA(x,y)} & \\
2 & \mid\ \mid\ \underline{\forall x\exists y\neg A(x,y)} & \\
3 & \mid\ \mid\ u\ \mid\ \underline{\forall yA(u,y)} & \\
4 & \mid\ \mid\ \quad\ \mid\ \exists y\neg A(u,y) & (\forall\mathrm{E}),\ 2 \\
5 & \mid\ \mid\ \quad\ \mid\ v\ \mid\ \underline{\neg A(u,v)} & \\
6 & \mid\ \mid\ \quad\ \mid\ \quad\ \mid\ A(u,v) & (\forall\mathrm{E}),\ 3 \\
7 & \mid\ \mid\ \quad\ \mid\ \quad\ \mid\ \bot & (\neg\mathrm{E}),\ 5,\ 6 \\
8 & \mid\ \mid\ \quad\ \mid\ \bot & (\exists\mathrm{E}),\ 4,\ 5\text{–}7 \\
9 & \mid\ \mid\ \bot & (\exists\mathrm{E}),\ 1,\ 3\text{–}8 \\
10 & \mid\ \neg\forall x\exists y\neg A(x,y) & (\neg\mathrm{I}),\ 2\text{–}9
\end{array}$$

4. $\forall x(R(x) \to B(x)), \neg B(\mathsf{a}) \vdash \neg R(\mathsf{a})$

$$\begin{array}{rll}
1 & \mid\ \forall x(R(x) \to B(x)) & \\
2 & \mid\ \underline{\neg B(\mathsf{a})} & \\
3 & \mid\ \mid\ \underline{R(\mathsf{a})} & \\
4 & \mid\ \mid\ R(\mathsf{a}) \to B(\mathsf{a}) & (\forall\mathrm{E}),\ 1 \\
5 & \mid\ \mid\ B(\mathsf{a}) & (\to\mathrm{E}),\ 3,\ 4 \\
6 & \mid\ \mid\ \bot & (\neg\mathrm{E}),\ 2,\ 5 \\
7 & \mid\ \neg R(\mathsf{a}) & (\neg\mathrm{I}),\ 3\text{–}6
\end{array}$$

5. $R(\mathsf{a}), \forall x(\neg G(x) \to \neg R(x)), M(\mathsf{b}) \vdash \exists xG(x) \wedge \exists xM(x)$

$$\begin{array}{rll}
1 & \mid\ R(\mathsf{a}) & \\
2 & \mid\ \forall x(\neg G(x) \to \neg R(x)) & \\
3 & \mid\ \underline{M(\mathsf{b})} & \\
4 & \mid\ \neg G(\mathsf{a}) \to \neg R(\mathsf{a}) & (\forall\mathrm{E}),\ 2 \\
5 & \mid\ \mid\ \underline{\neg G(\mathsf{a})} & \\
6 & \mid\ \mid\ \neg R(\mathsf{a}) & (\to\mathrm{E}),\ 4,\ 5 \\
7 & \mid\ \mid\ \bot & (\neg\mathrm{E}),\ 1,\ 6 \\
8 & \mid\ \neg\neg G(\mathsf{a}) & (\neg\mathrm{I}),\ 5\text{–}7 \\
9 & \mid\ G(\mathsf{a}) & (\neg\neg\mathrm{E}),\ 8 \\
10 & \mid\ \exists xG(x) & (\exists\mathrm{I}),\ 9 \\
11 & \mid\ \exists xM(x) & (\exists\mathrm{I}),\ 3 \\
12 & \mid\ \exists xG(x) \wedge \exists xM(x) & (\wedge\mathrm{I}),\ 10,\ 11
\end{array}$$

6. $\forall x((R(x) \wedge A(x)) \to T(x)), A(\mathsf{b}), R(\mathsf{b}) \vdash \exists xT(x)$

$$\begin{array}{rll}
1 & \mid\ \forall x((R(x) \wedge A(x)) \to T(x)) & \\
2 & \mid\ A(\mathsf{b}) & \\
3 & \mid\ \underline{R(\mathsf{b})} & \\
4 & \mid\ R(\mathsf{b}) \wedge A(\mathsf{b}) \to T(\mathsf{b}) & (\forall\mathrm{E}),\ 1 \\
5 & \mid\ R(\mathsf{b}) \wedge A(\mathsf{b}) & (\wedge\mathrm{I}),\ 2,\ 3 \\
6 & \mid\ T(\mathsf{b}) & (\to\mathrm{E}),\ 4,\ 5 \\
7 & \mid\ \exists xT(x) & (\exists\mathrm{I}),\ 6
\end{array}$$

7. $\forall x(P(x) \wedge Q(x)) \vdash \forall xP(x) \wedge \forall xQ(x)$

$$\begin{array}{rll}
1 & \mid\ \underline{\forall x(P(x) \wedge Q(x))} & \\
2 & \mid\ u\ \mid\ P(u) \wedge Q(u) & (\forall\mathrm{E}),\ 1 \\
3 & \mid\ \quad\ \mid\ P(u) & (\wedge\mathrm{E}),\ 2 \\
4 & \mid\ \forall xP(x) & (\forall\mathrm{I}),\ 2\text{–}3 \\
5 & \mid\ v\ \mid\ P(v) \wedge Q(v) & (\forall\mathrm{E}),\ 1 \\
6 & \mid\ \quad\ \mid\ Q(v) & (\wedge\mathrm{E}),\ 5 \\
7 & \mid\ \forall xQ(x) & (\forall\mathrm{I}),\ 5\text{–}6 \\
8 & \mid\ \forall xP(x) \wedge \forall xQ(x) & (\wedge\mathrm{I}),\ 4,\ 7
\end{array}$$

8. $\forall xP(x) \wedge \forall xQ(x) \vdash \forall x(P(x) \wedge Q(x))$

$$\begin{array}{rll}
1 & \mid\ \underline{\forall xP(x) \wedge \forall xQ(x)} & \\
2 & \mid\ \forall xP(x) & (\wedge\mathrm{E}),\ 1 \\
3 & \mid\ \forall xQ(x) & (\wedge\mathrm{E}),\ 1 \\
4 & \mid\ u\ \mid\ P(u) & (\forall\mathrm{E}),\ 2 \\
5 & \mid\ \quad\ \mid\ Q(u) & (\forall\mathrm{E}),\ 3 \\
6 & \mid\ \quad\ \mid\ P(u) \wedge Q(u) & (\wedge\mathrm{I}),\ 4,\ 5 \\
7 & \mid\ \forall x(P(x) \wedge Q(x)) & (\forall\mathrm{I}),\ 4\text{–}6
\end{array}$$

9. $\forall x(\exists yP(y) \to Q(x)) \vdash \exists yP(y) \to \forall xQ(x)$ (where there is no free variable x in $\exists yP(y)$)

1	$\mid \forall x(\exists yP(y) \to Q(x))$	
2	$\mid \quad \mid \exists yP(y)$	
3	$\mid \quad \mid u \mid \exists yP(y) \to Q(u)$	($\forall$E), 1
4	$\mid \quad \mid \quad \mid Q(u)$	($\to$E), 2, 3
5	$\mid \quad \mid \forall xQ(x)$	($\forall$I), 3–4
6	$\mid \exists yP(y) \to \forall xQ(x)$	($\to$I), 2–5

10. $\exists y\forall xA(x,y), \forall x\forall y(A(x,y) \to B(x,y)) \vdash \forall x\exists yB(x,y)$

1	$\mid \exists y\forall xA(x,y)$	
2	$\mid \forall x\forall y(A(x,y) \to B(x,y))$	
3	$\mid u \mid v \mid \forall xA(x,v)$	
4	$\mid \quad \mid \quad \mid A(u,v)$	($\forall$E), 3
5	$\mid \quad \mid \quad \mid A(u,v) \to B(u,v)$	($\forall$E), 2
6	$\mid \quad \mid \quad \mid B(u,v)$	($\to$E), 4, 5
7	$\mid \quad \mid \quad \mid \exists yB(u,y)$	($\exists$I), 6
8	$\mid \quad \mid \exists yB(u,y)$	($\exists$E), 1, 3–7
9	$\mid \forall x\exists yB(x,y)$	($\forall$I), 3–8

11. $\forall x\exists yA(x,y), \forall x\forall y(A(x,y) \to A(y,x)), \forall x\forall y\forall z(A(x,y) \land A(y,z) \to A(x,z)) \vdash \forall xA(x,x)$

1	$\mid \forall x\exists yA(x,y)$	
2	$\mid \forall x\forall y(A(x,y) \to A(y,x))$	
3	$\mid \forall x\forall y\forall z(A(x,y) \land A(y,z) \to A(x,z))$	
4	$\mid u \mid \exists yA(u,y)$	($\forall$E), 1
5	$\mid \quad \mid v \mid A(u,v)$	
6	$\mid \quad \mid \quad \mid \forall y(A(u,y) \to A(y,u))$	($\forall$E), 2
7	$\mid \quad \mid \quad \mid A(u,v) \to A(v,u)$	($\forall$E), 6
8	$\mid \quad \mid \quad \mid A(v,u)$	($\to$E), 5, 7
9	$\mid \quad \mid \quad \mid A(u,v) \land A(v,u)$	($\land$I), 5, 8
10	$\mid \quad \mid \quad \mid \forall y\forall z(A(u,y) \land A(y,z) \to A(u,z))$	($\forall$E), 3
11	$\mid \quad \mid \quad \mid \forall z(A(u,v) \land A(v,z) \to A(u,z))$	($\forall$E), 10
12	$\mid \quad \mid \quad \mid A(u,v) \land A(v,u) \to A(u,u)$	($\forall$E), 11
13	$\mid \quad \mid \quad \mid A(u,u)$	($\to$E), 9, 12
14	$\mid \quad \mid A(u,u)$	($\exists$E), 4, 5–13
15	$\mid \forall xA(x,x)$	($\forall$I), 4–14

12. $\exists x(P(x) \to \forall yQ(y)) \vdash \forall xP(x) \to \forall xQ(x)$

$$\begin{array}{rll}
1 & \vert\ \underline{\exists x(P(x) \to \forall yQ(y))} & \\
2 & \vert\quad \vert\ \underline{\forall xP(x)} & \\
3 & \vert\quad \vert\ u\ \vert\ \underline{P(u) \to \forall yQ(y)} & \\
4 & \vert\quad \vert\quad\ \ \vert\ P(u) & (\forall\text{E}),\ 2 \\
5 & \vert\quad \vert\quad\ \ \vert\ \forall yQ(y) & (\to\text{E}),\ 3,\ 4 \\
6 & \vert\quad \vert\quad\ \ \vert\ \forall xQ(x)\ \text{(Change of bound variables), 5} & \\
7 & \vert\quad \vert\ \forall xQ(x) & (\exists\text{E}),\ 1,\ 3\text{–}6 \\
8 & \vert\ \forall xP(x) \to \forall xQ(x) & (\to\text{I}),\ 2\text{–}7
\end{array}$$

13. $\forall x(A(x) \to B(x)) \vdash \exists xA(x) \to \exists xB(x)$

$$\begin{array}{rll}
1 & \vert\ \underline{\forall x(A(x) \to B(x))} & \\
2 & \vert\quad \vert\ \underline{\exists xA(x)} & \\
3 & \vert\quad \vert\ u\ \vert\ \underline{A(u)} & \\
4 & \vert\quad \vert\quad\ \ \vert\ A(u) \to B(u) & (\forall\text{E}),\ 1 \\
5 & \vert\quad \vert\quad\ \ \vert\ B(u) & (\to\text{E}),\ 3,\ 4 \\
6 & \vert\quad \vert\quad\ \ \vert\ \exists xB(x) & (\exists\text{I}),\ 5 \\
7 & \vert\quad \vert\ \exists xB(x) & (\exists\text{E}),\ 2,\ 3\text{–}6 \\
8 & \vert\ \exists xA(x) \to \exists xB(x) & (\to\text{I}),\ 2\text{–}7
\end{array}$$

14. $P \to \forall xQ(x) \vdash \forall x(P \to Q(x))$

$$\begin{array}{rll}
1 & \vert\ \underline{P \to \forall xQ(x)} & \\
2 & \vert\ u\ \vert\quad \vert\ \underline{P} & \\
3 & \vert\quad\ \ \vert\quad \vert\ \forall xQ(x) & (\to\text{E}),\ 1,\ 2 \\
4 & \vert\quad\ \ \vert\quad \vert\ Q(u) & (\forall\text{E}),\ 3 \\
5 & \vert\quad\ \ \vert\ P \to Q(u) & (\to\text{I}),\ 2\text{–}4 \\
6 & \vert\ \forall x(P \to Q(x)) & (\forall\text{I}),\ 2\text{–}5
\end{array}$$

15. $\forall xR(x) \lor \forall xS(x) \vdash \forall x(R(x) \lor S(x))$

$$\begin{array}{rll}
1 & \vert\ \underline{\forall xR(x) \lor \forall xS(x)} & \\
2 & \vert\ u\ \vert\quad \vert\ \underline{\forall xR(x)} & \\
3 & \vert\quad\ \ \vert\quad \vert\ R(u) & (\forall\text{E}),\ 2 \\
4 & \vert\quad\ \ \vert\quad \vert\ R(u) \lor S(u) & (\lor\text{I}),\ 3 \\
5 & \vert\quad\ \ \vert\quad \vert\ \underline{\forall xS(x)} & \\
6 & \vert\quad\ \ \vert\quad \vert\ S(u) & (\forall\text{E}),\ 5 \\
7 & \vert\quad\ \ \vert\quad \vert\ R(u) \lor S(u) & (\lor\text{I}),\ 6 \\
8 & \vert\quad\ \ \vert\ R(u) \lor S(u) & (\lor\text{E}),\ 1,\ 2\text{–}4,\ 5\text{–}7 \\
9 & \vert\ \forall x(R(x) \lor S(x)) & (\forall\text{I}),\ 2\text{–}8
\end{array}$$

16. $\forall x(P(x) \to Q) \vdash \exists xP(x) \to Q$

$$\begin{array}{rll}
1 & \vert\ \underline{\forall x(P(x) \to Q)} & \\
2 & \vert\quad \vert\ \underline{\exists xP(x)} & \\
3 & \vert\quad \vert\ u\ \vert\ \underline{P(u)} & \\
4 & \vert\quad \vert\quad\ \ \vert\ P(u) \to Q & (\forall\text{E}),\ 1 \\
5 & \vert\quad \vert\quad\ \ \vert\ Q & (\to\text{E}),\ 3,\ 4 \\
6 & \vert\quad \vert\ Q & (\exists\text{E}),\ 2,\ 3\text{–}5 \\
7 & \vert\ \exists xP(x) \to Q & (\to\text{E}),\ 2\text{–}6
\end{array}$$

17. $\forall x(P(x){\to}Q(x)), \forall x(Q(x){\to}R(x)) \vdash \forall x(P(x){\to}R(x))$

$$\begin{array}{rll}
1 & \vert\ \forall x(P(x) \to Q(x)) & \\
2 & \vert\ \underline{\forall x(Q(x) \to R(x))} & \\
3 & \vert\ u\ \vert\quad \vert\ \underline{P(u)} & \\
4 & \vert\quad\ \ \vert\quad \vert\ P(u) \to Q(u) & (\forall\text{E}),\ 2 \\
5 & \vert\quad\ \ \vert\quad \vert\ Q(u) \to R(u) & (\forall\text{E}),\ 2 \\
6 & \vert\quad\ \ \vert\quad \vert\ Q(u) & (\to\text{E}),\ 3,\ 4 \\
7 & \vert\quad\ \ \vert\quad \vert\ R(u) & (\to\text{E}),\ 5,\ 6 \\
8 & \vert\quad\ \ \vert\ P(u) \to R(u) & (\to\text{I}),\ 3\text{–}7 \\
9 & \vert\ \forall x(P(x) \to R(x)) & (\forall\text{I}),\ 3\text{–}8
\end{array}$$

18. $\forall x\forall yP(x, y) \vdash P(\mathsf{a}, \mathsf{a})$

$$\begin{array}{rll}
1 & \vert\ \underline{\forall x\forall yP(x, y)} & \\
2 & \vert\ \forall yP(\mathsf{a}, y) & (\forall\text{E}),\ 1 \\
3 & \vert\ P(\mathsf{a}, \mathsf{a}) & (\forall\text{E}),\ 2
\end{array}$$

19. $\forall x\forall yP(x, y) \vdash \forall xP(x, x)$

$$\begin{array}{rll}
1 & \vert\ \underline{\forall x\forall yP(x, y)} & \\
2 & \vert\ u\ \vert\ \forall yP(u, y) & (\forall\text{E}),\ 1 \\
3 & \vert\quad\ \ \vert\ P(u, u) & (\forall\text{E}),\ 2 \\
4 & \vert\ \forall xP(x, x) & (\forall\text{I}),\ 2\text{–}3
\end{array}$$

20. $\forall x(P(x) \land Q(x) \to R(x)), Q(\mathsf{a}) \land \forall z P(z) \vdash P(\mathsf{a}) \land R(\mathsf{a})$

$$\begin{array}{rll}
1 & \vert\ \forall x(P(x) \land Q(x) \to R(x)) & \\
2 & \vert\ \underline{Q(\mathsf{a}) \land \forall z P(z)} & \\
3 & \vert\ Q(\mathsf{a}) & (\land\mathrm{E}),\ 2 \\
4 & \vert\ \forall z P(z) & (\land\mathrm{E}),\ 2 \\
5 & \vert\ P(\mathsf{a}) & (\forall\mathrm{E}),\ 4 \\
6 & \vert\ P(\mathsf{a}) \land Q(\mathsf{a}) & (\land\mathrm{I}),\ 3,\ 5 \\
7 & \vert\ P(\mathsf{a}) \land Q(\mathsf{a}) \to R(\mathsf{a}) & (\forall\mathrm{E}),\ 1 \\
8 & \vert\ R(\mathsf{a}) & (\to\mathrm{E}),\ 6,\ 7 \\
9 & \vert\ P(\mathsf{a}) \land R(\mathsf{a}) & (\land\mathrm{I}),\ 5,\ 8
\end{array}$$

21. $\exists x P(x) \vdash \forall x Q(x) \to \exists x(P(x) \land Q(x))$

$$\begin{array}{rll}
1 & \vert\ \underline{\exists x P(x)} & \\
2 & \vert\ \quad \vert\ \underline{\forall x Q(x)} & \\
3 & \vert\ \quad \vert\ u\ \vert\ \underline{P(u)} & \\
4 & \vert\ \quad \vert\ \quad \vert\ Q(u) & (\forall\mathrm{E}),\ 2 \\
5 & \vert\ \quad \vert\ \quad \vert\ P(u) \land Q(u) & (\land\mathrm{I}),\ 3,\ 4 \\
6 & \vert\ \quad \vert\ \quad \vert\ \exists x(P(x) \land Q(x)) & (\exists\mathrm{I}),\ 5 \\
7 & \vert\ \quad \vert\ \exists x(P(x) \land Q(x)) & (\exists\mathrm{E}),\ 1,\ 3\text{–}6 \\
8 & \vert\ \forall x Q(x) \to \exists x(P(x) \land Q(x)) & (\to\mathrm{I}),\ 2\text{–}7
\end{array}$$

22. $\forall x \forall y(R(x,y) \to (P(x) \land \neg P(y))), \exists x \exists y(R(x,y) \land R(y,x)) \vdash \exists x(P(x) \land \neg P(x))$

Notice in Q22, in lines 7,8 the ($\forall$E) rule is used twice each.

$$\begin{array}{rll}
1 & \vert\ \forall x \forall y(R(x,y) \to (P(x) \land \neg P(y))) & \\
2 & \vert\ \underline{\exists x \exists y(R(x,y) \land R(y,x))} & \\
3 & \vert\ u\ \vert\ \underline{\exists y(R(u,y) \land R(y,u))} & \\
4 & \vert\ \quad \vert\ v\ \vert\ \underline{R(u,v) \land R(v,u)} & \\
5 & \vert\ \quad \vert\ \quad \vert\ R(u,v) & (\land\mathrm{E}),\ 4 \\
6 & \vert\ \quad \vert\ \quad \vert\ R(v,u) & (\land\mathrm{E}),\ 4 \\
7 & \vert\ \quad \vert\ \quad \vert\ R(u,v) \to P(u) \land \neg P(v) & (\forall\mathrm{E}),\ 1 \\
8 & \vert\ \quad \vert\ \quad \vert\ R(v,u) \to P(v) \land \neg P(u) & (\forall\mathrm{E}),\ 1 \\
9 & \vert\ \quad \vert\ \quad \vert\ P(u) \land \neg P(v) & (\to\mathrm{E}),\ 5,\ 7 \\
10 & \vert\ \quad \vert\ \quad \vert\ P(v) \land \neg P(u) & (\to\mathrm{E}),\ 6,\ 8 \\
11 & \vert\ \quad \vert\ \quad \vert\ P(u) & (\land\mathrm{E}),\ 9 \\
12 & \vert\ \quad \vert\ \quad \vert\ \neg P(u) & (\land\mathrm{E}),\ 10 \\
13 & \vert\ \quad \vert\ \quad \vert\ P(u) \land \neg P(u)) & (\land\mathrm{I}),\ 11,\ 12 \\
14 & \vert\ \quad \vert\ \quad \vert\ \exists x(P(x) \land \neg P(x)) & (\exists\mathrm{I}),\ 13 \\
15 & \vert\ \quad \vert\ \exists x(P(x) \land \neg P(x)) & (\exists\mathrm{E}),\ 3,\ 4\text{–}14 \\
16 & \vert\ \exists x(P(x) \land \neg P(x)) & (\exists\mathrm{E}),\ 2,\ 3\text{–}15
\end{array}$$

23. The simplest way to do Q23 is to use the derivation in Q22, but change line 13 to $\perp$ (*via* the ($\neg$E) rule), and then line 14 can be dropped, and lines 15 and 16 also become $\perp$. But notice that the conclusion is "obvious", since "saying" that there is an x so that $P(x)$ and $\neg P(x)$ are both true is "obviously" a contradiction.

24. $\exists z R(z,z), \exists y \forall x S(y,x) \vdash \exists y \exists z (S(z,y) \rightarrow R(y,y))$

1	$\mid\ \exists z R(z,z)$	
2	$\mid\ \exists y \forall x S(y,x)$	
3	$\mid\ u\ \mid\ R(u,u)$	
4	$\mid\ \quad \mid\ \mid\ S(u,u)$	
5	$\mid\ \quad \mid\ \mid\ R(u,u)$	(R), 3
6	$\mid\ \quad \mid\ S(u,u) \rightarrow R(u,u)$	($\rightarrow$I), 4–5
7	$\mid\ \quad \mid\ \exists z(S(z,u) \rightarrow R(u,u))$	($\exists$I), 6
8	$\mid\ \quad \mid\ \exists y \exists z(S(z,y) \rightarrow R(y,y))$	($\exists$I), 7
9	$\mid\ \exists y \exists z(S(z,y) \rightarrow R(y,y))$	($\exists$E), 1, 2–8

25. $\exists x(P(x) \wedge Q(x)) \vdash \exists x P(x) \wedge \exists x Q(x)$

1	$\mid\ \exists x(P(x) \wedge Q(x))$	
2	$\mid\ u\ \mid\ P(u) \wedge Q(u)$	
3	$\mid\ \quad \mid\ P(u)$	($\wedge$E), 2
4	$\mid\ \quad \mid\ \exists x P(x)$	($\exists$I), 3
5	$\mid\ \quad \mid\ Q(u)$	($\wedge$E), 2
6	$\mid\ \quad \mid\ \exists x Q(x)$	($\exists$I), 5
7	$\mid\ \quad \mid\ \exists x P(x) \wedge \exists x Q(x)$	($\wedge$I), 4, 6
8	$\mid\ \exists x P(x) \wedge \exists x Q(x)$	($\exists$I), 1, 2–7

26. $\forall x P(x) \vee \forall x Q(x) \vdash \forall x(P(x) \vee Q(x))$

1	$\mid\ \forall x P(x) \vee \forall x Q(x)$	
2	$\mid\ u\ \mid\ \quad \mid\ \forall x P(x)$	
3	$\mid\ \quad \mid\ \quad \mid\ P(u)$	($\forall$E), 2
4	$\mid\ \quad \mid\ \quad \mid\ P(u) \vee Q(u)$	($\vee$I), 3
5	$\mid\ \quad \mid\ \quad \mid\ \forall x Q(x)$	
6	$\mid\ \quad \mid\ \quad \mid\ Q(u)$	($\forall$E), 5
7	$\mid\ \quad \mid\ \quad \mid\ P(u) \vee Q(u)$	($\vee$I), 6
8	$\mid\ \quad \mid\ P(u) \vee Q(u)$	($\vee$E), 1, 2–4, 5–7
9	$\mid\ \forall x(P(x) \vee Q(x))$	($\forall$I), 2–8

27. $\exists x \forall y R(x,y) \vdash \forall y \exists x R(x,y)$

1	$\mid\ \exists x \forall y R(x,y)$	
2	$\mid\ u\ \mid\ v\ \mid\ \forall y R(v,y)$	
3	$\mid\ \quad \mid\ \quad \mid\ R(v,u)$	($\forall$E), 2
4	$\mid\ \quad \mid\ \quad \mid\ \exists x R(x,u)$	($\exists$I), 3
5	$\mid\ \quad \mid\ \exists x R(x,u)$	($\exists$E), 1, 2–4
6	$\mid\ \forall y \exists x R(x,y)$	($\forall$I), 2–5

The two problems in Q28 and in Q29 are done similarly: just change the roles of Qx and Qy in each case, for the appropriate quantifier: "$Q = \exists$" in Q28, and "$Q = \forall$" in Q29. Here's half for each.

28. $\exists x \exists y R(x,y) \vdash \exists y \exists x R(x,y)$

1	$\mid\ \exists x \exists y R(x,y)$	
2	$\mid\ u\ \mid\ \exists y R(u,y)$	
3	$\mid\ \quad \mid\ v\ \mid\ R(u,v)$	
4	$\mid\ \quad \mid\ \quad \mid\ \exists x R(x,v)$	($\exists$I), 3
5	$\mid\ \quad \mid\ \quad \mid\ \exists y \exists x R(x,y)$	($\exists$I), 4
6	$\mid\ \quad \mid\ \exists y \exists x R(x,y)$	($\exists$E), 1
7	$\mid\ \exists y \exists x R(x,y)$	($\exists$E), 1

29. $\forall x \forall y R(x,y) \vdash \forall y \forall x R(y,x)$

1	$\mid\ \forall x \forall y R(x,y)$	
2	$\mid\ u\ \mid\ v\ \mid\ \forall y R(u,y)$	($\forall$E), 1
3	$\mid\ \quad \mid\ \quad \mid\ R(u,v)$	($\forall$E), 1
4	$\mid\ \quad \mid\ \forall x R(u,x)$	($\forall$I), 2–3
5	$\mid\ \forall y \forall x R(y,x)$	($\forall$I), 2–4

30. $\neg\exists x(P(x) \wedge Q(x)) \vdash \forall x(P(x) \rightarrow \neg Q(x))$

1	$\mid \underline{\neg\exists x(P(x) \wedge Q(x))}$	
2	$\mid u \mid \mid \underline{P(u)}$	
3	$\mid \quad \mid \mid \underline{Q(u)}$	
4	$\mid \quad \mid \mid \mid P(u) \wedge Q(u)$	$(\wedge\mathrm{I})$, 2, 3
5	$\mid \quad \mid \mid \mid \exists x(P(x) \wedge Q(x))$	$(\exists\mathrm{I})$, 4
6	$\mid \quad \mid \mid \mid \bot$	$(\neg\mathrm{E})$, 1, 5
7	$\mid \quad \mid \mid \neg Q(u)$	$(\neg\mathrm{I})$, 3–6
8	$\mid \quad \mid P(u) \rightarrow \neg Q(u)$	$(\rightarrow\mathrm{I})$, 2–7
9	$\mid \forall x(P(x) \rightarrow \neg Q(x))$	$(\forall\mathrm{I})$, 2–8

31. $\forall x(P(x) \rightarrow \neg Q(x)) \vdash \neg\exists x(P(x) \wedge Q(x))$

1	$\mid \underline{\forall x(P(x) \rightarrow \neg Q(x))}$	
2	$\mid \mid \underline{\exists x(P(x) \wedge Q(x))}$	
3	$\mid \mid u \mid \underline{P(u) \wedge Q(u)}$	
4	$\mid \mid \quad \mid P(u)$	$(\wedge\mathrm{E})$, 3
5	$\mid \mid \quad \mid Q(u)$	$(\wedge\mathrm{E})$, 3
6	$\mid \mid \quad \mid P(u) \rightarrow \neg Q(u)$	$(\forall\mathrm{E})$, 1
7	$\mid \mid \quad \mid \neg Q(u)$	$(\rightarrow\mathrm{E})$, 4, 6
8	$\mid \mid \quad \mid \bot$	$(\neg\mathrm{E})$, 5, 7
9	$\mid \mid \bot$	$(\exists\mathrm{E})$, 2, 3–8
10	$\mid \neg\exists x(P(x) \wedge Q(x))$	$(\neg\mathrm{I})$, 2–9

32. $\exists x(P(x) \vee Q(x)) \vdash \exists xP(x) \vee \exists xQ(x)$

1	$\mid \underline{\exists x(P(x) \vee Q(x))}$	
2	$\mid u \mid \underline{P(u) \vee Q(u)}$	
3	$\mid \quad \mid \mid \underline{P(u)}$	
4	$\mid \quad \mid \mid \exists xP(x)$	$(\exists\mathrm{I})$, 3
5	$\mid \quad \mid \mid \exists xP(x) \vee \exists xQ(x)$	$(\vee\mathrm{I})$, 4
6	$\mid \quad \mid \mid \underline{Q(u)}$	
7	$\mid \quad \mid \mid \exists xQ(x)$	$(\exists\mathrm{I})$, 6
8	$\mid \quad \mid \mid \exists xP(x) \vee \exists xQ(x)$	$(\vee\mathrm{I})$, 7
9	$\mid \quad \mid \exists xP(x) \vee \exists xQ(x)$	$(\vee\mathrm{E})$, 2, 3–5, 6–8
10	$\mid \exists xP(x) \vee \exists xQ(x)$	$(\exists\mathrm{I})$, 1, 2–9

33. $\exists xP(x) \vee \exists xQ(x) \vdash \exists x(P(x) \vee Q(x))$

1	$\mid \underline{\exists xP(x) \vee \exists xQ(x)}$	
2	$\mid \mid \underline{\exists xP(x)}$	
3	$\mid \mid u \mid \underline{P(u)}$	
4	$\mid \mid \quad \mid P(u) \vee Q(u)$	$(\vee\mathrm{I})$, 3
5	$\mid \mid \quad \mid \exists x(P(x) \vee Q(x))$	$(\exists\mathrm{I})$, 4
6	$\mid \mid \exists x(P(x) \vee Q(x))$	$(\exists\mathrm{E})$, 2, 3–5
7	$\mid \mid \underline{\exists xQ(x)}$	
8	$\mid \mid v \mid \underline{Q(v)}$	
9	$\mid \mid \quad \mid P(v) \vee Q(v)$	$(\vee\mathrm{I})$, 8
10	$\mid \mid \quad \mid \exists x(P(x) \vee Q(x))$	$(\exists\mathrm{I})$, 9
11	$\mid \mid \exists x(P(x) \vee Q(x))$	$(\exists\mathrm{E})$, 7, 8–10
12	$\mid \exists x(P(x) \vee Q(x))$	$(\exists\mathrm{E})$, 1, 2–6, 7–11

34. $\forall x(P(x) \wedge Q(x)) \vdash \forall xP(x) \wedge \forall xQ(x)$

1	$\mid \underline{\forall x(P(x) \wedge Q(x))}$	
2	$\mid u \mid P(u) \wedge Q(u)$	$(\forall\mathrm{E})$, 1
3	$\mid \quad \mid P(u)$	$(\wedge\mathrm{E})$, 2
4	$\mid \forall xP(x)$	$(\forall\mathrm{I})$, 2–3
5	$\mid v \mid P(v) \wedge Q(v)$	$(\forall\mathrm{E})$, 1
6	$\mid \quad \mid Q(v)$	$(\wedge\mathrm{E})$, 5
7	$\mid \forall xQ(x)$	$(\forall\mathrm{I})$, 5–6
8	$\mid \forall xP(x) \wedge \forall xQ(x)$	$(\wedge\mathrm{I})$, 4, 8

35. $\forall xP(x) \wedge \forall xQ(x) \vdash \forall x(P(x) \wedge Q(x))$

1	$\mid \underline{\forall xP(x) \wedge \forall xQ(x)}$	
2	$\mid u \mid \forall xP(x)$	$(\wedge\mathrm{E})$, 1
3	$\mid \quad \mid P(u)$	$(\forall\mathrm{E})$, 2
4	$\mid \quad \mid \forall xQ(x)$	$(\wedge\mathrm{E})$, 1
5	$\mid \quad \mid Q(u)$	$(\forall\mathrm{E})$, 4
6	$\mid \quad \mid P(u) \wedge Q(u)$	$(\wedge\mathrm{I})$, 3, 5
7	$\mid \forall x(P(x) \wedge Q(x))$	$(\forall\mathrm{I})$, 2–6

36. $\forall x(P(x){\rightarrow}Q(x)), \exists x(P(x) \land R(x)) \vdash \exists x(Q(x) \land R(x))$

1		$\forall x(P(x) \rightarrow Q(x))$	
2		$\exists x(P(x) \land R(x))$	
3	u	$\quad P(u) \land R(u)$	
4		$\quad P(u)$	$(\land\text{E})$, 3
5		$\quad R(u)$	$(\land\text{E})$, 3
6		$\quad P(u) \rightarrow Q(u)$	$(\forall\text{E})$, 1
7		$\quad Q(u)$	$(\rightarrow\text{E})$, 4, 6
8		$\quad Q(u) \land R(u)$	$(\land\text{I})$, 5, 7
9		$\quad \exists x(Q(x) \land R(x))$	$(\exists\text{I})$, 8
10		$\exists x(Q(x) \land R(x))$	$(\exists\text{E})$, 2, 3–9

37. $\forall x(P(x){\lor}Q(x)), \exists x \neg P(x) \vdash \exists x Q(x)$

1		$\forall x(P(x) \lor Q(x))$	
2		$\exists x \neg P(x)$	
3	u	$\quad \neg P(u)$	
4		$\quad P(u) \lor Q(u)$	$(\forall\text{E})$, 1
5		$\quad\quad P(u)$	
6		$\quad\quad \bot$	$(\neg\text{E})$, 3, 5
7		$\quad\quad Q(u)$	$(\bot\text{E})$, 6
8		$\quad\quad Q(u)$	
9		$\quad\quad Q(u)$	(R), 8
10		$\quad Q(u)$	$(\lor\text{E})$, 4, 5–7, 8–9
11		$\quad \exists x Q(x)$	$(\exists\text{I})$, 10
12		$\exists x Q(x)$	$(\exists\text{E})$, 2, 3–11

Some remarks:

Q24 Notice that we don't actually need the second premise—(basically because any proposition of the form $P \rightarrow \top$ is equivalent to $\top$—in this case, knowing that there is a z so that $R(z, z)$ means that any statement $\exists y(Q \rightarrow R(y, y))$ will be "true").

As for the questions (25, 26, & 27) whose converses are invalid, here are some suggestions:

Q25 Notice that the two x's need not be the same in $\exists x P(x) \land \exists x Q(x)$, but they must be the same in $\exists x(P(x) \land Q(x)$, so just think of a situation where they might be different. For example, just because you know there is a blond person and a blue-eyed person in the class, does not guarantee that there is a blond&blue-eyed person there.

Q26 This is similar: for example, it may well be true that everyone in the class is either male or female, but that does not guarantee that everyone is male or that everyone is female.

Q27 This says if there is an x so that $R(x, y)$ holds for all y, then for any y, there is an x (the same x in fact) that satisfies $R(x, y)$, which is "obvious". But the converse says that if for any y there is an x satisfying $R(x, y)$, then one x will work for all the y. This is not always true: you might have different x's for different y's.

For example, consider the positive integers (*i.e.*, the positive whole numbers $1, 2, 3, 4, 5, \ldots$), and let R be "is not smaller than" (*i.e.* "is equal or greater than"): $R(x, y) \equiv$ "$x \geq y$". Then $\forall y \exists x R(x, y)$ is true: it simply says that for any integer, there's a number not smaller than it (*e.g.* the given number itself, or any bigger one). But that need not imply that there is a single "biggest" number, *i.e.* an x so that for every number y, $x \geq y$, for, in fact, there is no "biggest" integer. So $\exists x \forall y R(x, y)$ is false. Similar examples could deal with any situation where there's no single "extreme" instance ("biggest", "smallest", "richest", and so on), even though "locally" there is.

Q27–29 Note that a consequence of exercises 27–29 is that you can change the order of quantifiers if they are the same type, but not if they are different types. "$\exists x \exists y \equiv \exists y \exists x$", "$\forall x \forall y \equiv \forall y \forall x$", but "$\exists x \forall y$" is not equivalent to "$\forall y \exists x$".

(This is "obvious" in retrospect, if you regard $\exists$ as "being like $\lor$", and $\forall$ as a "being like $\land$", since we know that you can bracket $\lor$'s anyway you want, and likewise $\land$'s, but mixing $\lor$'s and $\land$'s becomes sensitive to brackets. And from this perspective, exercises 25 and 26 fit into the same "story".)

Exercise 5.6.1

In some exercises I've footnoted quantifier rules to help you see what they are doing in each case.

1.

$$
\begin{array}{r|l l}
1 & \forall x(G(x) \longrightarrow H(x)) & \\
2 & I(\mathsf{a}) \land \neg H(\mathsf{a}) & \\
3 & \underline{\forall x(G(x) \lor \neg F(x))} & \\
4 & I(\mathsf{a}) & (\land\text{E}), 2 \\
5 & \neg H(\mathsf{a}) & (\land\text{E}), 2 \\
{}^{a}6 & G(\mathsf{a}) \longrightarrow H(\mathsf{a}) & (\forall\text{E}), 1 \\
{}^{b}7 & G(\mathsf{a}) \lor \neg F(\mathsf{a}) & (\forall\text{E}), 3 \\
8 & \quad | \; \underline{G(\mathsf{a})} & \\
9 & \quad | \; H(\mathsf{a}) & (\longrightarrow\text{E}), 6, 8 \\
10 & \quad | \; \bot & (\neg\text{E}), 5, 9 \\
11 & \quad | \; I(\mathsf{a}) \land \neg F(\mathsf{a}) & (\bot\text{E}), 10 \\
12 & \quad | \; \underline{\neg F(\mathsf{a})} & \\
13 & \quad | \; I(\mathsf{a}) \land \neg F(\mathsf{a}) & (\land\text{I}), 4, 12 \\
14 & I(\mathsf{a}) \land \neg F(\mathsf{a}) & (\lor\text{E}), 7, 8\text{–}11, 12\text{–}13 \\
{}^{c}15 & \exists x(I(x) \land \neg F(x)) & (\exists\text{I}), 14
\end{array}
$$

[a] (a replacing x)

[b] (a replacing x)

[c] (taking a as the required x)

2.

1	$\forall x(W(x,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},x))$	
2	$\forall x(\neg W(x,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},x))$	
[a][b]3	$u \mid W(u,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},u)$	(∀E), 1
[c]4	$\quad \mid \neg W(u,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},u)$	(∀E), 2
5	$\quad \mid \quad \mid M(\mathsf{g},u)$	
6	$\quad \mid \quad \mid \quad \mid W(u,\mathsf{g})$	
7	$\quad \mid \quad \mid \quad \mid \neg M(\mathsf{g},u)$	(→E), 3, 6
8	$\quad \mid \quad \mid \quad \mid \bot$	(¬E), 5, 7
9	$\quad \mid \quad \mid \neg W(u,\mathsf{g})$	(¬I), 6–8
10	$\quad \mid \quad \mid \quad \mid \neg W(u,\mathsf{g})$	
11	$\quad \mid \quad \mid \quad \mid \neg M(\mathsf{g},u)$	(→E), 4, 10
12	$\quad \mid \quad \mid \quad \mid \bot$	(¬E), 5, 11
13	$\quad \mid \quad \mid \neg\neg W(u,\mathsf{g})$	(¬I), 10–12
14	$\quad \mid \quad \mid \bot$	(¬E), 9, 13
15	$\quad \mid \neg M(\mathsf{g},u)$	(¬I), 5–14
[d]16	$\forall x(\neg M(\mathsf{g},x))$	(∀I), 3–15

[a] (we're going to try to prove the conclusion $\neg M(\mathsf{g},u)$ for any arbitrary, *i.e.* "fresh", u)

[b] (g replacing x)

[c] (g replacing x)

[d] (since we got $\neg M(\mathsf{g},u)$ for any u, we got $\neg M(\mathsf{g},x)$ for all x).

An alternate derivation using the tautology $\vdash A \vee \neg A$ may be found at the end of this Answers section.

3.

1	$C(\mathsf{b})$	
2	$R(\mathsf{b}) \longrightarrow \forall x S(x)$	
3	$\forall x(S(x) \longrightarrow H(x))$	
4	$\exists x C(x) \longrightarrow \forall y R(y)$	
[a]5	$\exists x C(x)$	(∃I), 1
6	$\forall y R(y)$	(→E), 4, 5
[b]7	$R(\mathsf{b})$	(∀E), 6
8	$\forall x S(x)$	(→E), 2, 7
[c][d]9	$u \mid S(u) \longrightarrow H(u)$	(∀E), 3
[e]10	$\quad \mid S(u)$	(∀E), 8
11	$\quad \mid H(u)$	(→E), 9, 10
[f]12	$\forall x H(x)$	(∀I), 9–11

[a] (taking b as the required x)

[b] (b replacing y)

[c] (aiming to prove the conclusion for any u)

[d] (u replacing x)

[e] (ditto)

[f] (proving $H(u)$ for any u justifies $\forall x H(x)$)

4.

1	$C(\mathsf{f}) \wedge B(\mathsf{f},\mathsf{l})$	
2	$\exists x(C(x) \wedge B(x,\mathsf{f}))$	
3	$\forall x \forall y \forall z[(B(x,y) \wedge B(y,z)) \longrightarrow B(x,z)]$	
[a]4	$u \mid \underline{C(u) \wedge B(u,\mathsf{f})}$	
5	$\mid B(u,\mathsf{f})$	$(\wedge$E), 4
6	$\mid B(\mathsf{f},\mathsf{l})$	$(\wedge$E), 1
7	$\mid B(u,\mathsf{f}) \wedge B(\mathsf{f},\mathsf{l})$	$(\wedge$I), 5, 6
[b]8	$\mid \forall y \forall z[(B(u,y) \wedge B(y,z)) \longrightarrow B(u,z)]$	$(\forall$E), 3
[c]9	$\mid \forall z[(B(u,\mathsf{f}) \wedge B(\mathsf{f},z)) \longrightarrow B(u,z)]$	$(\forall$E), 8
[d]10	$\mid B(u,\mathsf{f}) \wedge B(\mathsf{f},\mathsf{l})) \longrightarrow B(u,\mathsf{l})$	$(\forall$E), 9
11	$\mid B(u,\mathsf{l})$	$(\longrightarrow$E), 7, 10
12	$\mid B(u,\mathsf{f}) \wedge B(u,\mathsf{l})$	$(\wedge$I), 5, 11
13	$\mid C(u)$	$(\wedge$E), 4
14	$\mid C(u) \wedge (B(u,F) \wedge B(u,\mathsf{l}))$	$(\wedge$I), 12, 13
[e]15	$\mid \exists z[C(z) \wedge (B(z,\mathsf{f}) \wedge B(z,\mathsf{l}))]$	$(\exists$I), 14
[f]16	$\exists z[C(z) \wedge (B(z,\mathsf{f}) \wedge B(z,\mathsf{l}))]$	$(\exists$E), 2, 4–15

[a] (taking an arbitrary "fresh" u for x, as per line 2)
[b] (u replacing x)
[c] (f replacing y)
[d] (l replacing z)
[e] (taking u as the required z)
[f] (if 4-15 works for u, it'll work for whatever x satisfies line 2)

5.

1	$\forall x(L(x) \longrightarrow [P(x) \wedge \neg W(x) \wedge \neg F(x)])$	
2	$\forall x([P(x) \wedge \neg \exists y N(x,y)] \longrightarrow W(x))$	
[a]3	$b \mid\ \mid \underline{L(b)}$	
4	$\mid\ \mid\ \mid \underline{\neg \exists y N(b,y)}$	
[b]5	$\mid\ \mid\ \mid L(b) \longrightarrow P(b) \wedge \neg W(b) \wedge \neg F(b)$	$(\forall$E), 1
6	$\mid\ \mid\ \mid P(b) \wedge \neg W(b) \wedge \neg F(b)$	$(\longrightarrow$E), 3, 5
7	$\mid\ \mid\ \mid P(b)$	$(\wedge$E), 6
8	$\mid\ \mid\ \mid \neg W(b)$	$(\wedge$E), 6
9	$\mid\ \mid\ \mid P(b) \wedge \neg \exists y N(b,y)$	$(\wedge$I), 4, 7
[c]10	$\mid\ \mid\ \mid P(b) \wedge \neg \exists y N(b,y) \longrightarrow W(b)$	$(\forall$E), 2
11	$\mid\ \mid\ \mid W(b)$	$(\longrightarrow$E), 9, 10
12	$\mid\ \mid\ \mid \bot$	$(\neg$E), 8, 11
13	$\mid\ \mid \neg\neg \exists y N(b,y)$	$(\neg$I), 4–12
14	$\mid\ \mid \exists y N(b,y)$	$(\neg\neg$E), 13
15	$\mid L(b) \longrightarrow \exists y N(b,y)$	$(\longrightarrow$I), 3–14
[d]16	$\forall x(L(x) \longrightarrow \exists y N(x,y))$	$(\forall$I), 3–15

[a] (We're going to prove the conclusion $L(b) \longrightarrow \exists y N(b,y)$ for a "fresh" variable b.)
[b] (b replacing x)
[c] (b replacing x)
[d] (Since we got $L(b) \longrightarrow \exists y N(b,y)$ for any b, we have $L(x) \longrightarrow \exists y N(x,y)$ for all x.)

6.

$$\begin{array}{lll}
1 & \mid \underline{\forall x \exists y L(x,y)} & \\
2 & \mid\ \ \mid \underline{\exists x \forall y \neg L(x,y)} & \\
3 & \mid\ \ \mid\ u\ \mid \underline{\forall y \neg L(u,y)} & \\
\text{[a]}4 & \mid\ \ \mid\ \ \ \ \mid \exists y L(u,y) & (\forall\text{E}),\ 1 \\
5 & \mid\ \ \mid\ \ \ \ \mid\ v\ \mid \underline{L(u,v)} & \\
\text{[b]}6 & \mid\ \ \mid\ \ \ \ \mid\ \ \ \ \mid \neg L(u,v) & (\forall\text{E}),\ 3 \\
7 & \mid\ \ \mid\ \ \ \ \mid\ \ \ \ \mid \bot & (\neg\text{E}),\ 5,\ 6 \\
\text{[c]}8 & \mid\ \ \mid\ \ \ \ \mid \bot & (\exists\text{E}),\ 4,\ 5\text{–}7 \\
\text{[d]}9 & \mid\ \ \mid \bot & (\exists\text{E}),\ 2,\ 3\text{–}8 \\
10 & \mid \neg\exists x \forall y \neg L(x,y) & (\neg\text{I}),\ 2\text{–}9
\end{array}$$

[a] (u replacing x)

[b] (v replacing y)

[c] (if 5-7 works for arbitrary v, it will work for any y satisfying line 4)

[d] (if 3-8 works for arbitrary u, it will work for any x satisfying line 2)

7.

$$\begin{array}{lll}
1 & \mid \forall x \forall y[G(x,y) \longrightarrow \neg G(y,x)] & \\
2 & \mid \forall y G(\mathsf{d},y) & \\
3 & \mid \forall x \forall y[E(x) \wedge \neg E(y) \longrightarrow G(x,y)] & \\
4 & \mid \underline{\exists x E(x)} & \\
5 & \mid\ \ \mid \underline{\neg E(\mathsf{d})} & \\
6 & \mid\ \ \mid\ u\ \mid \underline{E(u)} & \\
7 & \mid\ \ \mid\ \ \ \ \mid E(u) \wedge \neg E(\mathsf{d}) & (\wedge\text{I}),\ 5,\ 6 \\
\text{[a]}8 & \mid\ \ \mid\ \ \ \ \mid E(u) \wedge \neg E(\mathsf{d}) \longrightarrow G(u,\mathsf{d}) & (\forall\text{E}),\ 3 \\
9 & \mid\ \ \mid\ \ \ \ \mid G(u,\mathsf{d}) & (\longrightarrow\text{E}),\ 7,\ 8 \\
\text{[b]}10 & \mid\ \ \mid\ \ \ \ \mid G(u,\mathsf{d}) \longrightarrow \neg G(\mathsf{d},u) & (\forall\text{E}),\ 1 \\
11 & \mid\ \ \mid\ \ \ \ \mid \neg G(\mathsf{d},u) & (\longrightarrow\text{E}),\ 9,\ 10 \\
\text{[c]}12 & \mid\ \ \mid\ \ \ \ \mid G(\mathsf{d},u) & (\forall\text{E}),\ 2 \\
13 & \mid\ \ \mid\ \ \ \ \mid \bot & (\neg\text{E}),\ 11,\ 12 \\
\text{[d]}14 & \mid\ \ \mid \bot & (\exists\text{E}),\ 4,\ 6\text{–}13 \\
15 & \mid \neg\neg E(\mathsf{d}) & (\neg\text{I}),\ 5\text{–}14 \\
16 & \mid E(\mathsf{d}) & (\neg\neg\text{E}),\ 15
\end{array}$$

[a] (actually we use (∀E) twice, replacing x with u and y with d)

[b] (again we use (∀E) twice, with the same substitutions)

[c] (this time replace y with u)

[d] (if 6-13 works for an arbitrary u, it will work for anything that satisfies line 4)

8.

1	$\exists x(S(x) \wedge A(x, \mathsf{r}))$	
2	$\forall x \forall y(T(x) \wedge S(y) \rightarrow A(x, y))$	
3	$T(\mathsf{r})$	
4	$u \mid S(u) \wedge A(u, \mathsf{r})$	
5	$\quad \mid \forall y(T(\mathsf{r}) \wedge S(y) \rightarrow A(\mathsf{r}, y))$	(∀E), 2
6	$\quad \mid T(\mathsf{r}) \wedge S(u) \rightarrow A(\mathsf{r}, u)$	(∀E), 5
7	$\quad \mid S(u)$	(∧E), 4
8	$\quad \mid T(\mathsf{r}) \wedge S(u)$	(∧I), 3, 7
9	$\quad \mid A(\mathsf{r}, u)$	(→E), 6, 8
10	$\quad \mid A(u, \mathsf{r})$	(∧E), 4
11	$\quad \mid A(u, \mathsf{r}) \wedge A(\mathsf{r}, u)$	(∧I), 9, 10
12	$\quad \mid \exists x(A(x, \mathsf{r}) \wedge A(\mathsf{r}, x))$	(∃I), 11
13	$\exists x(A(x, \mathsf{r}) \wedge A(\mathsf{r}, x))$	(∃E), 1

9.

1	$\forall y(H(y) \rightarrow A(y))$	
2	$u \mid \quad \mid \exists y(T(u, y) \wedge H(y))$	
3	$\quad \mid \quad \mid v \mid T(u, v) \wedge H(v)$	
4	$\quad \mid \quad \mid \quad \mid T(u, v)$	(∧E), 3
5	$\quad \mid \quad \mid \quad \mid H(v)$	(∧E), 3
6	$\quad \mid \quad \mid \quad \mid H(v) \rightarrow A(v)$	(∀E), 1
7	$\quad \mid \quad \mid \quad \mid A(v)$	(→E), 5, 6
8	$\quad \mid \quad \mid \quad \mid T(u, v) \wedge A(v)$	(∧I), 4, 7
9	$\quad \mid \quad \mid \quad \mid \exists y(T(u, y) \wedge A(y))$	(∃I), 8
10	$\quad \mid \quad \mid \exists y(T(u, y) \wedge A(y))$	(∃E), 2
11	$\quad \mid \exists y(T(u, y) \wedge H(y)) \rightarrow \exists y(T(u, y) \wedge A(y))$	(→I), 2–10
12	$\forall x(\exists y(T(x, y) \wedge H(y)) \rightarrow \exists y(T(x, y) \wedge A(y)))$	(∀I), 2–11

10.

1	$\mid\ \exists x(T(x)\wedge\forall y(S(y)\rightarrow A(x,y)))$	
2	$\mid\ \forall x\forall y(T(x)\wedge J(y)\rightarrow\neg A(x,y))$	
3	$\mid\ \mid\ \exists y(S(y)\wedge J(y))$	
4	$\mid\ \mid\ u\ \mid\ S(u)\wedge J(u)$	
5	$\mid\ \mid\ \mid\ v\ \mid\ T(v)\wedge\forall y(S(y)\rightarrow A(v,y))$	
6	$\mid\ \mid\ \mid\ \mid\ \forall y(T(v)\wedge J(y)\rightarrow\neg A(v,y))$	($\forall$E), 2
7	$\mid\ \mid\ \mid\ \mid\ T(v)\wedge J(u)\rightarrow\neg A(v,u)$	($\forall$E), 6
8	$\mid\ \mid\ \mid\ \mid\ T(v)$	($\wedge$E), 5
9	$\mid\ \mid\ \mid\ \mid\ J(u)$	($\wedge$E), 4
10	$\mid\ \mid\ \mid\ \mid\ T(v)\wedge J(u)$	($\wedge$I), 8, 9
11	$\mid\ \mid\ \mid\ \mid\ \neg A(v,u)$	($\rightarrow$E), 7, 10
12	$\mid\ \mid\ \mid\ \mid\ \forall y(S(y)\rightarrow A(v,y))$	($\wedge$E), 5
13	$\mid\ \mid\ \mid\ \mid\ S(u)\rightarrow A(v,u)$	($\forall$E), 12
14	$\mid\ \mid\ \mid\ \mid\ S(u)$	($\wedge$E), 4
15	$\mid\ \mid\ \mid\ \mid\ A(v,u)$	($\rightarrow$E), 13, 14
16	$\mid\ \mid\ \mid\ \mid\ \bot$	($\neg$E), 11, 15
17	$\mid\ \mid\ \mid\ \bot$	($\exists$E), 1, 5–16
18	$\mid\ \mid\ \bot$	($\exists$E), 3, 4–17
19	$\mid\ \neg\exists y(S(y)\wedge J(y))$	($\neg$I), 3–18

11.

1	$\mid\ \forall x(S(x)\rightarrow L(x))$	
2	$\mid\ \forall x(B(x)\rightarrow S(x))$	
3	$\mid\ \exists xB(x)$	
4	$\mid\ \forall x(B(x)\rightarrow\neg L(x))$	
[a]5	$\mid\ u\ \mid\ B(u)$	
[b]6	$\mid\ \mid\ S(u)\rightarrow L(u)$	($\forall$E), 1
[c]7	$\mid\ \mid\ B(u)\rightarrow S(u)$	($\forall$E), 2
[d]8	$\mid\ \mid\ B(u)\rightarrow\neg L(u)$	($\forall$E), 4
9	$\mid\ \mid\ S(u)$	($\rightarrow$E), 5, 7
10	$\mid\ \mid\ L(u)$	($\rightarrow$E), 6, 9
11	$\mid\ \mid\ \neg L(u)$	($\rightarrow$E), 5, 8
12	$\mid\ \mid\ \bot$	($\neg$E), 10, 11
[e]13	$\mid\ \bot$	($\exists$E), 3, 5–12

[a] (taking an arbitrary "fresh" u for x, as per line 3)

[b] (u replacing x)

[c] (u replacing x)

[d] (u replacing x)

[e] (if 5-12 works for u, it'll work for whatever x satisfies line 3)

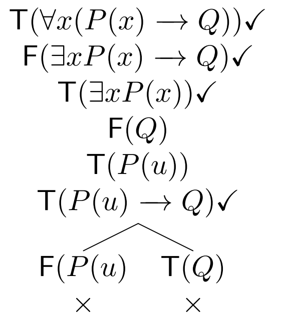
Remark

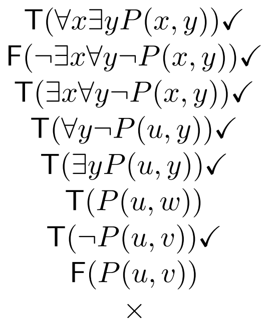
There is an alternate derivation for Q2 (the "Groucho" problem) using the tautology $\vdash A \vee \neg A$ (this is #8 in Exercise 3.3.1). However, as this tautology is equivalent to the $(\neg\neg E)$ rule, which we saw isn't actually necessary in this case, philosophically this alternate derivation isn't as "nice" as the original one (even if it is perhaps simpler(?)).

1	$\forall x(W(x,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},x))$	
2	$\forall x(\neg W(x,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},x))$	
3	$u \mid W(u,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},u)$	$(\forall$E), 1
4	$\mid \neg W(u,\mathsf{g}) \longrightarrow \neg M(\mathsf{g},u)$	$(\forall$E), 2
5	$\mid W(u,\mathsf{g}) \vee \neg W(u,\mathsf{g})$	(Tautology)
6	$\mid \quad \mid W(u,\mathsf{g})$	
7	$\mid \quad \mid \neg M(\mathsf{g},u)$	$(\longrightarrow$E), 3, 6
8	$\mid \quad \mid \neg W(u,\mathsf{g})$	
9	$\mid \quad \mid \neg M(\mathsf{g},u)$	$(\longrightarrow$E), 4, 8
10	$\mid \neg M(\mathsf{g},u)$	$(\vee$E), 5, 6–7, 8–9
11	$\forall x(\neg M(\mathsf{g},x))$	$(\forall$I), 3–10

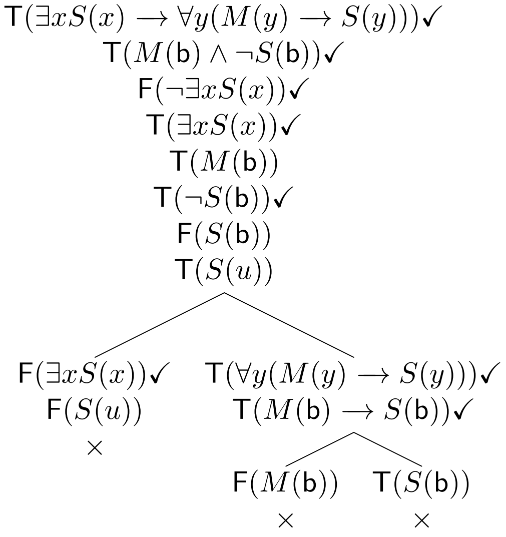
Solutions to the Tableaux Appendix exercises

1.


$\mathsf{T}(\forall x(P(x) \to Q))$✓
$\mathsf{F}(\exists x P(x) \to Q)$✓
$\mathsf{T}(\exists x P(x))$✓
$\mathsf{F}(Q)$
$\mathsf{T}(P(u))$
$\mathsf{T}(P(u) \to Q)$✓
$\mathsf{F}(P(u)$
$\mathsf{T}(Q)$
×
×


$\mathsf{T}(\forall x \exists y P(x,y))$✓
$\mathsf{F}(\neg\exists x \forall y \neg P(x,y))$✓
$\mathsf{T}(\exists x \forall y \neg P(x,y))$✓
$\mathsf{T}(\forall y \neg P(u,y))$✓
$\mathsf{T}(\exists y P(u,y))$✓
$\mathsf{T}(P(u,w))$
$\mathsf{T}(\neg P(u,v))$✓
$\mathsf{F}(P(u,v))$
×

2. To show $\exists x S(x) \to \forall y(M(y) \to S(y)), M(\mathsf{b}) \wedge \neg S(\mathsf{b}) \vdash \neg\exists x S(x)$ is valid:


$\mathsf{T}(\exists x S(x) \to \forall y(M(y) \to S(y)))$✓
$\mathsf{T}(M(\mathsf{b}) \wedge \neg S(\mathsf{b}))$✓
$\mathsf{F}(\neg\exists x S(x))$✓
$\mathsf{T}(\exists x S(x))$✓
$\mathsf{T}(M(\mathsf{b}))$
$\mathsf{T}(\neg S(\mathsf{b}))$✓
$\mathsf{F}(S(\mathsf{b}))$
$\mathsf{T}(S(u))$
$\mathsf{F}(\exists x S(x))$✓
$\mathsf{F}(S(u))$
×
$\mathsf{T}(\forall y(M(y) \to S(y)))$✓
$\mathsf{T}(M(\mathsf{b}) \to S(\mathsf{b}))$✓
$\mathsf{F}(M(\mathsf{b}))$
$\mathsf{T}(S(\mathsf{b}))$
×
×

As a derivation:

1	$\exists x S(x) \to \forall y(M(y) \to S(y))$	
2	$M(\mathsf{b}) \wedge \neg S(\mathsf{b})$	
3	$M(\mathsf{b})$	(∧E), 2
4	$\neg S(\mathsf{b})$	(∧E), 2
5	$\mid\ \exists x S(x)$	
6	$\mid\ \forall y(M(y) \to S(y))$	(→E), 1, 5
7	$\mid\ M(\mathsf{b}) \to S(\mathsf{b})$	(∀E), 6
8	$\mid\ S(\mathsf{b})$	(→E), 3, 7
9	$\mid\ \bot$	(¬E), 4, 8
10	$\neg\exists x S(x)$	(¬I), 5–9

3. Finally, $\exists x P(x) \vdash \exists y P(y)$, $\forall x P(x) \vdash \forall y P(y)$:

$\mathsf{T}(\exists x P(x))$ ✓
$\mathsf{F}(\exists y P(y))$ ✓
$\mathsf{T}(P(u))$
$\mathsf{F}(P(u))$
×

1	$\exists x P(x)$	
2	$u \mid P(u)$	
3	$\quad\mid\ \exists y P(y)$	(∃I), 2
4	$\exists y P(y)$	(∃E), 2–3

$\mathsf{T}(\forall x P(x))$ ✓
$\mathsf{F}(\forall y P(y))$ ✓
$\mathsf{F}(P(u))$
$\mathsf{T}(P(u))$
×

1	$\forall x P(x)$	
2	$u \mid P(u)$	(∀E), 1
3	$\forall y P(y)$	(∀I), 2–2

Solutions to the Equality Appendix exercises

1.

$$\begin{array}{l|ll}
\vdots & \vdots & \\
n_1 & t_1 = s_1 & \\
n_2 & t_2 = s_2 & \\
\vdots & \vdots & \\
n_k & t_k = s_k & \\
m & P(t_1, t_2, \ldots, t_k) & \\
\vdots & \vdots & \\
\ell_1 & P(s_1, t_2, \ldots, t_k) & (=\text{E}),\ n_1,\ m \\
\ell_2 & P(s_1, s_2, \ldots, t_k) & (=\text{E}),\ n_2,\ \ell_1 \\
\vdots & \vdots & \\
\ell_k & P(s_1, s_2, \ldots, s_{k-1}, t_k) & (=\text{E}),\ \ell_1,\ \ell_2,\ .\ .\ . \\
\ell & P(s_1, s_2, \ldots, s_k) & (=\text{E}),\ \ell_1,\ \ell_2,\ .\ .\ ,\ \ell_k
\end{array}$$

2. Symmetry:

We use $(= E)$ with $P(t_1, t_2)$ as $t_1 = t_2$; we start with $P(t, t)$, then derive $P(s, t)$ from $t = s$ and $t = t$:

$$\begin{array}{l|ll}
\vdots & \vdots & \\
n & t = s & \\
n+1 & t = t & (=\text{I}) \\
\vdots & \vdots & \\
k & s = t & (=\text{E}),\ n,\ n+1
\end{array}$$

And transitivity:

$$\begin{array}{l|ll}
\vdots & \vdots & \\
n & t_1 = t_2 & \\
n+1 & t_2 = t_3 & \\
\vdots & \vdots & \\
m & t_1 = t_3 & (=\text{E}),\ n,\ n+1
\end{array}$$

3.

$$\begin{array}{l|ll}
1 & t = t & (=\text{I}) \\
2 & \exists x(x = t) & (\exists\text{I}),\ 1
\end{array}$$

4.

$$\begin{array}{l|l}
1 & \underline{\exists x(x = t \wedge P(x)} \\
2 & u \quad \begin{array}{|l} \underline{u = t \wedge P(u)} \\ u = t \\ P(u) \\ P(t) \end{array} \\
3 & \\
4 & \\
5 & \\
6 & P(t)
\end{array}
\qquad
\begin{array}{l}
\\ \\ (\wedge\text{E}),\ 2 \\ (\wedge\text{E}),\ 2 \\ (=\text{E}),\ 3,\ 4 \\ (\exists\text{E}),\ 1,\ 2\text{–}5
\end{array}$$

$$\begin{array}{l|ll}
1 & \underline{P(t)} & \\
2 & t = t & (=\text{I}) \\
3 & t = t \wedge P(t) & (\wedge\text{I}),\ 1,\ 2 \\
4 & \exists x(x = t \wedge P(x) & (\exists\text{I}),\ 3
\end{array}$$

5.

$$\begin{array}{rll}
1 & \mid\ \underline{\forall x(x = t \to P(x))} & \\
2 & \mid\ t = t \to P(t) & (\forall\mathrm{E}),\ 1 \\
3 & \mid\ t = t & (=\mathrm{I}) \\
4 & \mid\ P(t) & (\to\mathrm{E}),\ 2,\ 3
\end{array}$$

$$\begin{array}{rll}
1 & \mid\ \underline{P(t)} & \\
2 & \mid\ u \mid\ \mid\ \underline{u = t} & \\
3 & \mid\ \ \ \mid\ \mid\ P(u) & (=\mathrm{E}),\ 1,\ 2 \\
4 & \mid\ \ \ \mid\ u = t \to P(u) & (\to\mathrm{I}),\ 2\text{–}3 \\
5 & \mid\ \forall x(x = t \to P(x)) & (\forall\mathrm{I}),\ 2\text{–}4
\end{array}$$

6.

$$\begin{array}{rll}
1 & \mid\ \underline{\forall x \forall y(x = y \to P(x, y))} & \\
2 & \mid\ u \mid\ \forall y(u = y \to P(u, y) & (\forall\mathrm{E}),\ 1 \\
3 & \mid\ \ \ \mid\ u = u \to P(u, y) & (\forall\mathrm{E}),\ 2 \\
4 & \mid\ \ \ \mid\ u = u & (=\mathrm{I}) \\
5 & \mid\ \ \ \mid\ P(u, u) & (\to\mathrm{E}),\ 3,\ 4 \\
6 & \mid\ \forall x P(x, x) & (\forall\mathrm{I}),\ 2\text{–}5
\end{array}$$

$$\begin{array}{rll}
1 & \mid\ \underline{\forall x P(x, x)} & \\
2 & \mid\ u \mid\ v \mid\ \mid\ \underline{u = v} & \\
3 & \mid\ \ \ \mid\ \ \ \mid\ \mid\ P(u, u) & (\forall\mathrm{I}),\ 1 \\
4 & \mid\ \ \ \mid\ \ \ \mid\ \mid\ P(u, v) & (=\mathrm{E},\ 2,\ 3 \\
5 & \mid\ \ \ \mid\ \ \ \mid\ u = v \to P(u, v) & (\to\mathrm{I}),\ 2\text{–}4 \\
6 & \mid\ \ \ \mid\ \forall y(u = y \to P(u, y)) & (\forall\mathrm{I}),\ 2\text{–}5 \\
7 & \mid\ \forall x \forall y(x = y \to P(x, y)) & (\forall\mathrm{I}),\ 2\text{–}6
\end{array}$$

7.

$$\begin{array}{rll}
1 & \mid\ \underline{P(y)} & \\
2 & \mid\ u \mid\ \mid\ \underline{t(u) = y} & \\
3 & \mid\ \ \ \mid\ \mid\ P(t(u)) & (=\mathrm{E}),\ 1,\ 2 \\
4 & \mid\ \ \ \mid\ \mid\ Q(u) & \pi_1(u) \\
5 & \mid\ \ \ \mid\ t(u) = y \to Q(u) & (\to\mathrm{I}),\ 2\text{–}4 \\
6 & \mid\ \forall x(t(x) = y \to Q(x) & (\forall\mathrm{I}),\ 2\text{–}5
\end{array}$$

$$\begin{array}{rll}
1 & \mid\ \underline{P(t(u))} & \\
2 & \mid\ \forall x(t(x) = t(x) \to Q(x)) & \pi_2(t(x)) \\
3 & \mid\ t(x) = t(x) \to Q(x) & (\forall\mathrm{E}),\ 2 \\
4 & \mid\ t(x) = t(x) & (=\mathrm{I}) \\
5 & \mid\ Q(x) & (\to\mathrm{E}),\ 3,\ 4
\end{array}$$

8.

$$\begin{array}{rll}
1 & \mid\ \underline{\exists x(t(x) = y \wedge P(x)} & \\
2 & \mid\ u \mid\ \underline{t(u) = y \wedge P(u)} & \\
3 & \mid\ \ \ \mid\ P(u) & (\wedge\mathrm{E}),\ 2 \\
4 & \mid\ \ \ \mid\ Q(t(u) & \pi_1(u) \\
5 & \mid\ \ \ \mid\ t(u) = y & (\wedge\mathrm{E}),\ 2 \\
6 & \mid\ \ \ \mid\ Q(y) & (=\mathrm{E}),\ 4,\ 5 \\
7 & \mid\ Q(y) & (\exists\mathrm{E}),\ 1,\ 2\text{–}6
\end{array}$$

$$\begin{array}{rll}
1 & \mid\ \underline{P(x)} & \\
2 & \mid\ t(x) = t(x) & (=\mathrm{I}) \\
3 & \mid\ t(x) = t(x) \wedge P(x) & (\wedge\mathrm{I}),\ 1,\ 2 \\
4 & \mid\ \exists x(t(x) = t(x) \wedge P(x) & (\exists\mathrm{I}),\ 3 \\
5 & \mid\ Q(t(x)) & \pi_2(t(x))
\end{array}$$

9.

1	$\exists x(P(x) \land \forall y(P(y) \to x = y))$	
2	u $P(u) \land \forall y(P(y) \to u = y)$	
3	v w $P(v) \land P(w)$	
4	$\forall y(P(y) \to u = y)$	$(\land\text{E})$, 2
5	$P(v)$	$(\land\text{E})$, 3
6	$P(w)$	$(\land\text{E})$, 3
7	$P(v) \to u = v$	$(\forall\text{E})$, 4
8	$P(w) \to u = w$	$(\forall\text{E})$, 4
9	$u = v$	$(\to\text{E})$, 5, 7
10	$u = w$	$(\to\text{E})$, 6, 8
11	$v = u$	(Sym), 10
12	$v = w$	(Trans), 11, 10
13	$P(v) \land P(w) \to v = w$	$(\to\text{I})$, 3–12
14	$\forall y(P(v) \land P(y) \to v = y)$	$(\forall\text{I})$, 3–13
15	$\forall x \forall y(P(x) \land P(y) \to x = y)$	$(\forall\text{I})$, 3–14
16	$\forall x \forall y(P(x) \land P(y) \to x = y)$	$(\exists\text{E})$, 1, 2–15
17	r $P(r) \land \forall y(P(y) \to r = y)$	
18	$P(r)$	$(\land\text{E})$, 17
19	$\exists x P(x)$	$(\exists\text{I})$, 18
20	$\exists x P(x)$	$(\exists\text{E})$, 1, 17–19
21	$\exists x P(x) \land \forall x \forall y(P(x) \land P(y) \to x = y)$	$(\land\text{I})$, 16, 20

1	$\exists x P(x) \land \forall x \forall y(P(x) \land P(y) \to x = y)$	
2	$\exists x P(x)$	$(\land\text{I})$, 1
3	$\forall x \forall y(P(x) \land P(y) \to x = y)$	$(\land\text{I})$, 1
4	u $P(u)$	
5	u $P(v)$	
6	$P(u) \land P(v)$	$(\land\text{I})$, 4, 5
7	$P(u) \land P(v) \to u = v$	$(\forall\text{E})$, 3
8	$u = v$	$(\to\text{E})$, 6, 7
9	$P(v) \to u = v$	$(\to\text{I})$, 5–8
10	$\forall y(P(y) \to u = y)$	$(\forall\text{I})$, 5–9
11	$P(u) \land \forall y(P(y) \to u = y)$	$(\land\text{I})$, 4, 10
12	$\exists x(P(x) \land \forall y(P(y) \to x = y))$	$(\exists\text{I})$, 11
13	$\exists x(P(x) \land \forall y(P(y) \to x = y))$	$(\exists\text{E})$, 3, 4–12

(Note: $(\forall E)$ is used twice in line 7.)

The translation problems (with “obvious” symbols):

1.

1	$\forall x(L(x) \to C(x))$	
2	$L(\mathsf{b})$	
3	$\mathsf{b} = \mathsf{f}$	
4	$L(\mathsf{b}) \to C(\mathsf{b})$	$(\forall\mathrm{E})$, 1
5	$C(\mathsf{b})$	$(\to\mathrm{E})$, 2, 4
6	$C(\mathsf{f})$	$(=\mathrm{E})$, 3, 5

2.

1	$\forall x(L(x) \to \neg S(x))$	
2	$L(\mathsf{b})$	
3	$S(\mathsf{f})$	
4	$\mid\ \mathsf{f} = \mathsf{b}$	
5	$\mid\ S(\mathsf{b})$	$(=\mathrm{E})$, 3, 4
6	$\mid\ L(\mathsf{b}) \to \neg S(\mathsf{b})$	$(\forall\mathrm{E})$, 1
7	$\mid\ \neg S(\mathsf{b})$	$(\to\mathrm{E})$, 2, 6
8	$\mid\ \bot$	$(\neg\mathrm{E})$, 2, 6
9	$\neg(\mathsf{f} = \mathsf{b})$	$(\neg\mathrm{I})$, 4–9

3.

1	$W(\mathsf{b})$	
2	$W(\mathsf{h})$	
3	$\forall x(W(x) \to x = \mathsf{b} \vee x = \mathsf{h})$	
4	$L(\mathsf{b})$	
5	$L(\mathsf{h})$	
6	$u \mid\ \mid\ W(u)$	
7	$\mid\ \mid\ W(u) \to u = \mathsf{b} \vee u = \mathsf{h}$	$(\forall\mathrm{E})$, 3
8	$\mid\ \mid\ u = \mathsf{b} \vee u = \mathsf{h}$	$(\to\mathrm{E})$, 6, 7
9	$\mid\ \mid\ \mid\ u = \mathsf{b}$	
10	$\mid\ \mid\ \mid\ L(u)$	$(=\mathrm{E})$, 4, 9
11	$\mid\ \mid\ \mid\ u = \mathsf{h}$	
12	$\mid\ \mid\ \mid\ L(u)$	$(=\mathrm{E})$, 5, 11
13	$\mid\ \mid\ L(u)$	$(\vee\mathrm{E})$, 8, 9–10, 11–12
14	$\mid\ W(u) \to L(u)$	$(\to\mathrm{E})$, 6–13
15	$\forall x(W(x) \to L(x))$	$(\forall\mathrm{I})$, 6–14

4.

$$\begin{array}{r|lll}
1 & \forall x \forall y(L(x) \wedge L(y) \longrightarrow x = y) & & \\
2 & L(\mathsf{b}) & & \\
3 & \neg \mathsf{h} = \mathsf{b} & & \\
\hline
4 & \quad \vert \; L(\mathsf{h}) & & \\
5 & \quad \vert \; L(\mathsf{h}) \wedge L(\mathsf{b}) \longrightarrow \mathsf{h} = \mathsf{b} & & (\forall \mathrm{E}),\ 1 \\
6 & \quad \vert \; L(\mathsf{h}) \wedge L(\mathsf{b}) & & (\wedge \mathrm{I}),\ 2,\ 4 \\
7 & \quad \vert \; \mathsf{h} = \mathsf{b} & & (\longrightarrow \mathrm{E}),\ 5,\ 6 \\
8 & \quad \vert \; \bot & & (\neg \mathrm{E}),\ 3,\ 7 \\
9 & \neg L(\mathsf{h}) & & (\neg \mathrm{I}),\ 4\text{–}8
\end{array}$$

(Note: $(\forall E)$ is used twice in line 5.)

A final remark about equality:

One nice technical point is that having equality as a predicate, we can now define functions in terms of predicates, in the following way. Suppose we want to have a symbol s representing "+1". One could define it in terms of an "addition predicate" $A(m, n, k)$, interpreted as "k is the sum of m and n", as follows. Assuming that our entities are ordinary numbers, with the usual properties (later we shall see how one could construct a "theory" where such properties are required as "axioms"), then we would want the following statements to be true.

$$\forall x \forall y \exists z A(x, y, z)$$
$$\forall x \forall y \forall z \forall z'(A(x, y, z) \wedge A(x, y, z') \quad \longrightarrow \quad z = z')$$

This says that for any pair of numbers m, n, their sum is uniquely determined. We could then introduce a new term constructor s with arity 1, and "define" it by the assertion $A(n, 1, s(n))$, which just says that $s(n) = n + 1$, as intended. With the assertions above, we would know that $s(n)$ has the properties $n + 1$ ought to have.

And next . . . ?

Now that we have a good understanding of what a (formal) proof is, we may relax somewhat, and use these logical principles more informally, which is exactly what mathematicians do when they want to prove things about the mathematical universe. In the remainder of the course, we shall look at several mathematical topics, and shall construct mathematical proofs of various claims and statements. Those proofs will be informal, conversational in fact, but it should always be clear that one could translate these informal proofs into a formal presentation of the sort we have been considering in these past few chapters. In other words, the arguments use informal versions of the derivation rules we have been using for formal derivations.

Chapter 6

Sets and Things

6.1 Sets

Membership in a set is a kind of pattern that is basic to our ability to count and, ultimately, to conceive of numbers and arithmetic operations on numbers.

The concept of "number" depends on the "set" concept. Set theory also includes a lot of propositional and predicate logic that we touched on in the last chapters. Sets are a bridge between logic and number theory.

A set is a collection of things. Counting how many things are in the collection is basic to the concept of number.

When we say "a set is a collection of things", what sort of things do we have in mind: what is a "thing"? We allow "things" to be almost anything you can imagine, not only concrete objects like stones and people, but also abstract entities as well, like colours and qualities ("goodness" is a thing, in this view). There are some minimal properties "things" must have. A set must come equipped with a notion of "equality": it must be possible to say whether or not two things in a set are equal. The things in a set must be the kind of thing about which it makes sense to say how many of them there are (so we can count them). Things must be self-identical (so a thing continues to be that thing for as long as it matters) and distinct (so we can somehow distinguish this thing from that thing).

A set may be represented by a predicate, in the sense that it is the collection of all things which satisfy that predicate. For instance, one might define a set by saying it contains all people with blue eyes (*i.e.* all things having the property that they are people and have blue eyes). Or a set might just be specified by listing the things in the collection.

It is important to be able to distinguish between a collection of things, and the things in that collection. We can consider the properties of the collection independently of the properties of the things. For example, the students in a college class form a collection. One property of everything in that collection is that he/she has a mother. But the class does not have a mother. On the other hand, a property of the college class is its male/female ratio, but a student cannot have a male/female ratio.

So a collection, group or set can be itself a thing with properties. A set is a new kind of (compound) thing: since one set can be distinguished from another, we see that sets are things. So a set could be a collection of sets (a set of sets). (This means there should be a notion of equality for sets—we shall return to this shortly.)

For greater clarity and less ambiguity mathematicians have developed an artificial language of symbols for dealing with sets and things.

Sets and things will be symbolized with variables, similarly to how we handled terms in predicate logic. Suppose we have a collection of Beanie Babies, consisting of Bamboo, Prissy, Purplebeary, two Romeos and Shadow. The two Romeos have to be given distinct names, so we might call them Romeo1 and Romeo2 (or OldRomeo and NewRomeo). We may symbolize these as b (Bamboo), p (Prissy), q (Purplebeary), r (Romeo1), m (Romeo2) and s (Shadow). The set of my Beanie Babies might be symbolized as B. Special sets (like the empty set, the universal set, and certain sets of numbers) have special identifiers (names). When we deal with numbers, the identifiers for individual things are *numerals*—the names of the numbers.

6.1.1 Specifying sets

There are two ways to define a set. The *denotative* definition defines a set by giving its *denotation*. The denotation of a set is a list of all the members of the set. The definition of the set B is $B = \{b, p, q, r, m, s\}$. We list the names of the members inside curly brackets. $T = \{1, 2, 3, \ldots, 10\}$ can represent the set of positive integers less than 11. The ellipsis ($\ldots$) indicates that there are members not specifically listed. We can define a set with infinitely many members as $C = \{a, b, c, \ldots\}$, the ellipsis indicating that the list goes on indefinitely. For example $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ is the set of all non-negative integers, usually called the *natural numbers*. There is no last natural number; this list goes on without stop, the set $\mathbb{N}$ is *infinite*.

To say that a thing is a *member* or an *element* of a set (*i.e.* that it is in the set), we use the symbol $\in$. To say that Bamboo is one of my Beanie Babies we write $b \in B$. To deny that a thing is a member (or element) of a set we can either use the $\neg$ symbol or the special symbol $\notin$. To say that Duster (symbolized as d) is not one of my Beanie Babies, we say $d \notin B$ or $\neg\, d \in B$.[1]

Connotative definitions define a set by giving the *connotation* of the name of the set. The connotative definition of a set is a specification of the rule (the property or predicate) which determines whether or not something is a member of the set. It states a rule that something must satisfy if it is to be included in the set. We could define the set "all-my-Beanies-except-the-Romeos" as $C = \{x | x \in B \wedge (x \neq r \wedge x \neq m)\}$. We read this expression as: C is the set of all x such that x is a member of B and x is not r and x is not m. This way of stating the definition is often called "set-builder notation", because the notation specifies how to build the set. This expression defines the property (predicate) that something must have to be included in the set C. The "x" in the notation is a variable that stands for the identifier of a thing.

Remember that implicitly, every set comes equipped with a notion of equality—we assume that we can always tell whether or not two things are to be regarded as "the same thing", meaning that they are equal.[2] Since sets are things, this means we must also be able to determine if two sets are equal. And here hangs a philosophical tale: there are two well-established notions of equality of sets (corresponding to the two ways of specifying sets), and which one we use has a significant effect on how we regard and handle sets.

Extensional equality of sets is determined by the actual elements of the sets: we say that two sets are equal (extensionally), indicated by $P = Q$, if and only if the two sets contain exactly the same elements. $P = Q$ means that every member of P is also a member of Q, and every member of Q is also a member of P:

$$(P = Q) \leftrightarrow \forall x((x \in P \longrightarrow x \in Q) \wedge (x \in Q \longrightarrow x \in P))$$

[1] $\in$ is a (binary) predicate, and we handle it just as we did other predicates before. We may use the other logical operations, of course, so for example, we might write $b \in B \wedge \neg\, d \in B$, to mean Bamboo, but not Duster, is one of my Beanie Babies.

[2] Formal treatments of set theory often take such an equality relation as part of the specification of a set, so that a set is not merely a collection, but is a collection together with a suitable notion of equality.

(meaning "set P is equal to set Q if and only if for any x, if x is a member of P then x is a member of Q and if x is a member of Q then x is a member of P"). Of course this is equivalent to: $(P = Q) \leftrightarrow \forall x(x \in P \leftrightarrow x \in Q)$ (read as "set P is equal to set Q if and only if for any x, x is a member of P if and only if x is a member of Q"). The definition of equality entails that, for any set A, $A = A$.

Intensional equality of sets is determined by the predicates that define them: two sets are intensionally equal if they are defined by the *same* predicate.

So, here's a bit of 60's trivia for you: there are official Beatles recordings on which exactly 1, 2, 3, or 4 of the Beatles (John, Paul, George, Ringo) performed. (For example, only Paul performed on "Yesterday", only John and Paul on "Ballad of John and Yoko", all but Ringo on "Back in the USSR", and they all performed on "Help!".) So, *extensionally* the following two sets are both equal to $\{1, 2, 3, 4\}$ (and so to each other):

$$\begin{aligned} &\{x|x \text{ is a natural number } \wedge 0 < x < 5\} \\ &= \{x|\exists y(y \text{ is a Beatles recording } \wedge\ x \text{ is the number of Beatles performing on } y)\} \end{aligned}$$

However, *intensionally* these sets are not at all equal. It's merely a historical accident that these sets have the same elements; their connotations are quite distinct.

Unless otherwise stated, **we shall always mean extensional equality by** $A = B$. The only times I shall refer to intensional equality will be to point out how that notion makes for a different set theory.

6.2 Two important sets

When considering sets, there are two extremes: the set that has nothing, and the set that has everything. The empty set is the set that has no members. It is represented either as $\{\,\}$ (empty curly brackets) or as $\emptyset$ (a stylized Greek letter phi).[3] No matter what connotative definition you give, the empty set has the same denotative definition, since it contains exactly the same members (no members at all). So, extensionally speaking there is only one empty set. However, different connotative definitions can all define the empty set. (So intensionally speaking, one would have many different empty sets—this is one sign that intensional set theory is more complicated, or "richer", to be more positive, than extensional set theory.) The set $W = \{x|x \text{ is a woman more than 16 meters tall}\}$ states a different rule from that stated by $L = \{x|x \text{ is a leprechaun}\}$. But W is the empty set, and so is L, because neither set has any members. The predicates "... is a woman more than 16 meters tall" and "... is a leprechaun" have the same denotation but different connotations: they have the same extension (or reference), but different intensions (or senses).

Chapter 5 introduced the concept of a "universe of discourse". A similar notion in set theory is the universal set. The universal set consists of every individual thing in the universe of discourse. Usually the universe of discourse is restricted, either explicitly or implicitly. If we are talking about my Beanie Babies, I would probably not have to specify $x \in B$ when defining the set C. I assume you understand that I'm talking about Beanies when I say "Everything is for sale except the Romeos". The set of Beanies for sale could be defined as $C = \{x|x \neq r \wedge x \neq m\}$ on this assumption. The symbol for the universal set is $\mathbb{U}$. If I want to restrict the universe of discourse to my Beanie Babies, I could say $\mathbb{U} = B$ or $\mathbb{U} = \{b, p, q, r, m, s\}$.

We saw that sets may be members of other sets. In fact, a set can include members that are sets as well as members that are individuals. Suppose I need some money and decide to sell my

[3] Be careful here: do not represent the empty set as $\{\emptyset\}$. That is *not* an empty set, as it has (exactly) one member, namely the empty set $\emptyset$. That is, it is a set with one element, and that element is a set itself, but with no elements.

MIDI synthesizer (i), my computer (j) and my collection of Beanie Babies. The set of things I'm selling can be defined denotatively as $S = \{i, j, B\}$. Romeo1 (r) is not a member of this set. I'm selling the Beanie Babies as a set, not individually. The set is one of the things I'm selling. Even when I define the set as $S = \{i, j, \{b, p, q, r, m, s\}\}$, it is not true that $r \in S$, rather, $r \notin S$. S itself just has three elements (i, j, B), no more, no less. None of those elements is r. One of the elements (B) is a set, and that set has r as an element, but S does not have r as an element. This is an important distinction (between sets *as* things, and sets *of* things), and you should be careful to make it.

6.3 Subsets, Proper Subsets, and the Power Set

Set A is a **subset** of set B, symbolized by $A \subseteq B$, if and only if all the elements of set A are also elements of set B. In symbols:

$$A \subseteq B \leftrightarrow \forall x((x \in A) \rightarrow (x \in B))$$

We write $A \subsetneqq B$, to mean $A \subseteq B \wedge A \neq B$; *i.e.* that A is a subset of B, but is definitely not equal to B.[4] We call such a subset a **proper subset**. Notice that saying $A \subsetneqq B$ is giving more information than $A \subseteq B$, and in fact $A \subseteq B$ is equivalent to $A \subsetneqq B \vee A = B$.

The set of Romeos $\{r, m\}$ is a subset of the set B of my Beanie Babies. If we call the set of Romeos $R = \{r, m\}$, we can say $R \subseteq B$. (We can go further and also say $R \subsetneqq B$, since we have that additional information.)

Now we have a conceptual framework (definitions) that permits us to do some simple proofs in set theory. We shall relax the requirements for proof (*i.e.*, derivation) in these mathematical proofs, compared to our formal logical derivations. So in particular, we shall replace the formal structure of derivations with a more "conversational" style, although it is implicit that such strict structure could be applied if a proof were questioned.

6.3.1 Examples

1. Prove that any set is a subset of itself. That is, prove that $\forall X(X \subseteq X)$ (for any set X, X is a subset of X).

   Proof: Clearly, $\forall X(\forall x((x \in X) \rightarrow (x \in X)))$ (for any set X, every member of X is a member of X), so $\forall X(X \subseteq X)$.

2. Prove $\forall X \forall Y(X = Y \rightarrow X \subseteq Y)$ (for any sets X and Y, if $X = Y$ then $X \subseteq Y$).

   Remark: this is a bit subtle, even though it is easy and obvious(!). For when we say $X = Y$, we are not really saying X and Y are the same thing (and so this example isn't just a repeat of the previous one), but rather we are saying that they have the same elements. In particular, the proof of this property is tied to the fact that we are using extensional equality; this proof would not work if we were using intensional equality.

   Proof: By the definition of set equality, whenever $X = Y$, every member of X is also a member of Y. Therefore $X \subseteq Y$. It is also true that $Y \subseteq X$. In fact:

3. Show that we could have defined set equality as

$$X = Y \leftrightarrow (X \subseteq Y \wedge Y \subseteq X)$$

[4] Other notations are used, most commonly $A \subset B$; but since $A \subset B$ is also used by some writers to mean $A \subseteq B$, we'll avoid this potentially confusing notation.

4. Prove $\forall X(\emptyset \subseteq X)$ (the empty set is a subset of every set).

   Proof: Since the empty set $\emptyset$ has no members, then for every $x \in \mathbb{U}$, $x \in \emptyset$ is false. An implication with a false premise is true. So the implication $(x \in \emptyset) \longrightarrow (x \in Y)$ is true for all possible values of x and for any set Y. Every member of the empty set is a member of any set Y. In other words, the empty set is a subset of every set.

5. Prove $\forall X(X \subseteq \mathbb{U})$.

   Proof: The universal set $\mathbb{U}$ includes everything that could be a member of any set X, so every member of any set X must be a member of $\mathbb{U}$.

Be very clear about the difference between one set being a **subset** of another set and a set being a **member** of another set. $\emptyset$ is a *subset* of the set B of Beanie Babies. But $\emptyset$ is not a Beanie Baby, so it is not a *member* of B. That is, $\emptyset \subseteq B$ is true but $\emptyset \in B$ is false.

When a set A is not a subset of a set B we write $\neg A \subseteq B$ or just $A \not\subseteq B$. Don't confuse this with $A \subsetneqq B$ (which says A *is* a subset of B, just not equal to it).

6.3.2 Power sets

The **power set** of a set A is the set of all the subsets of A. We write $\mathcal{P}(A)$ to mean the power set of the set A.

For example, if $A = \{a, b\}$, what are the subsets of A? We proved above that the empty set $\emptyset$ is a subset of every set, so it's a subset of A. We also showed that every set is a subset of itself, so $\{a, b\}$ is a subset of A. What else? The set $\{a\}$ is a subset, because every one of its members is also a member of A. The same is true of the set $\{b\}$. What about $\{b, a\}$? It is a subset of A, but it is the same set as $\{a, b\}$ (by the definition of set equality, since they have the same elements). So all the subsets of A are: $\emptyset, \{a\}, \{b\}$, and $\{a, b\}$. These are all the members of the power set of A. $\mathcal{P}(A)$ is the set that has these sets as members, so

$$\mathcal{P}(A) = \mathcal{P}(\{a, b\}) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$$

You may prefer to use $\{\,\}$ for the empty set, so you might write $\mathcal{P}(A) = \{\{\,\}, \{a\}, \{b\}, \{a, b\}\}$. Note that every element of $\mathcal{P}(A)$ is a set, so if we write sets denotationally (listing elements explicitly), every element is of the form $\{\ldots\}$.

Some more examples: what is the power set of the one-element set $\{a\}$? It's $\{\{\}, \{a\}\}$. The power set of the empty set is $\{\{\}\}$ (that is, it is the set that has one member, and that one member is the empty set). The only subset of the empty set is the empty set (which is a subset of every set).

Later (Chapter 8) we shall prove the fact that if n is the number of elements of a set A, then $\mathcal{P}(A)$ has 2^n elements. Verify this in the examples above (and in the examples in exercises 1–3 below).

6.3.3 Exercise on subsets and power sets

1. Write out all the subsets of the set $A = \{a, b, c, d\}$. How many are there? Which of them are proper subsets? Write the denotative definition of $\mathcal{P}(A)$.

2. Write out the power sets (a) of a three-element set and (b) of a five-element set.

3. Since a power set is a set, so it too will have a power set. Since the power set is a set of sets, its power set will be a set of sets of sets. Based on the previous exercise, what would you think $\mathcal{P}(\mathcal{P}(A))$ will be when $A = \{a, b, c\}$? How many sets are in the power set of the power set of $\{a, b, c\}$?

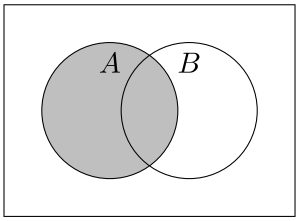
4. Each of the following statements is intended to apply to *all non-empty sets* A and B. Indicate whether each statement is true or false (for all non-empty A, B). (If a statement is only true for some A, B but not for all A, B, then the statement is false.)

   (a) $(A \subseteq B) \rightarrow (A \subsetneqq B)$
   (b) $(A \subsetneqq B) \rightarrow (A \subseteq B)$
   (c) $A \subsetneqq A$
   (d) $A \subseteq A$
   (e) $\emptyset \subseteq A$
   (f) $\emptyset \subsetneqq A$
   (g) $\emptyset \subsetneqq \emptyset$
   (h) $A \subseteq \mathbb{U}$
   (i) $\mathbb{U} \subseteq \emptyset$
   (j) $\emptyset \subsetneqq \mathbb{U}$
   (k) $\emptyset \subseteq \emptyset$
   (l) $A \subsetneqq \mathbb{U}$

6.4 Operations on Sets

In set theory, $=, \subseteq, \in$ and so on are (binary) relations or predicates. The power-set operator $\mathcal{P}$ is a (unary) operator: it operates on a set and produces a set.[5] Other operators on sets that produce sets are *complement*, *union*, *intersection* and *Cartesian product*. The set-complement operation acts on one set (it is a unary operator) and produces a new set. Union, intersection and Cartesian product are binary operators.

6.4.1 Venn diagrams

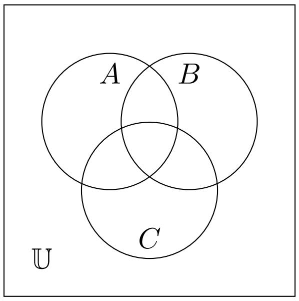


A Venn diagram is often useful to represent sets. We shall use Venn diagrams to illustrate the set operations complement, union and intersection. The basic idea is to make a diagram that shows the universal set as a rectangle with the set or sets (*e.g.* marked A and B in the diagram at left) shown as circles. To show a set, the convention is to shade everything that is in the set and leave the rest un-shaded.[6] The diagram illustrates the set A by shading the region enclosed by A. This is a Venn diagram of the set A. You might like to draw simple Venn diagrams for the set B, for $\mathbb{U}$, and for $\emptyset$.

One important point to make is that the various regions of the rectangle $\mathbb{U}$ may be empty (may have no elements in them), or may be non-empty (may have elements in them). In general we make no assumptions as to whether or not any region (circle or part thereof) is empty or inhabited. So, in the picture at left, there may or may not be things in the lune-shaped region where A and B overlap, or in any of the other regions one might specify.

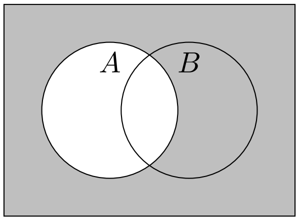
[5] It is actually a function symbol, something we only briefly mentioned in the previous chapter.

[6] This is not universally agreed upon—many authors do exactly the opposite, shading what is *not* in the set, leaving the desired set un-shaded. You should be careful in reading other books *etc.* which have Venn diagrams, and be sure which convention is being used.



You may have as many circles (representing different sets) as you need. For example, to represent three sets in the universe, you would draw three circles. To have the most general configuration possible, they should overlap in all possible ways—a little thought should convince you that this means the diagram must look something like the one at left. Needless to say, as the number of circles increases, this could get a bit complicated!

6.4.2 The complement of a set



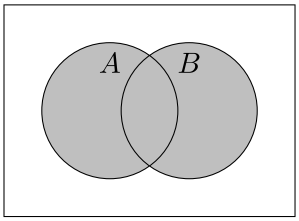
The complement of a set A is another set whose members are everything that is not a member of A. Since the diagram in section 6.4.1 shows the set A, shading everything outside A would diagram the complement of A, as shown in the Venn diagram at left. This gives us the following definition.

The **complement** of a set A is the set of everything that is *not* a member of A—every element in the universal set *except* the things (if any) that are members of A. The complement of A is symbolized as A^{C} (read A-complement). In symbols, we define A^{C} as:

$$A^{\mathsf{C}} = \{x | x \notin A\}.$$

Equivalently: $\forall x((x \in A^{\mathsf{C}}) \leftrightarrow (x \notin A))$.

6.4.3 The union of sets



The **union** of two sets A and B is the set containing all elements that are in either A or B or both, and is symbolized as $A \cup B$.

So we define the set-union as follows:

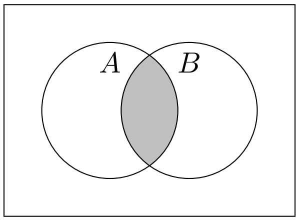
$$A \cup B = \{x | x \in A \vee x \in B\}$$

The diagram at left shows the union $A \cup B$ shaded. Everything that is not in $A \cup B$ is unshaded.

There is a point of English language usage about which you should be careful. Although set union is defined in terms of *inclusive-or*, we often describe a union with "and". For example, we may say something like "All of my colleagues and friends are invited" to mean that the set of invitees consists of the union of the set of my colleagues and the set of my friends, which means everyone who is *either* a colleague *or* a friend or both. Think about this, so you understand what is happening here. In spite of the English usage, the union is the set of elements of one *or* the other set, *not* the set of elements of both. The set of people who are both colleagues and friends is not the union, but instead is the *intersection* (defined next), namely those people who have both properties at once: colleagues who are *also* friends. That's not what is usually meant by the phrase "all of my colleagues and friends": that phrase usually means all your colleagues and also all your friends, including those who are only one or the other. That is the union.

What is going on here? It's actually quite simple: in everyday English "and" has two meanings, roughly corresponding to logical and to arithmetical usage. "It is raining and it is Monday" is the "logical and", meaning that both conditions hold. "The invitees are all my colleagues and friends" is the "arithmetical and", meaning that one adds my colleagues to my friends to get the collection of invitees. There is actually a mathematical reason why the latter is really the dual of the former ("or" *vs* "and", logically speaking), but that would take us way beyond the scope of this course.

6.4.4 The intersection of sets



The **intersection** of two sets A and B is the set containing all elements that are members of both sets, symbolized as $A \cap B$. The definition of set-intersection is:

$$A \cap B = \{x | x \in A \wedge x \in B\}$$

In the Venn diagram of $A \cap B$, at left, only the lune-shaped overlap of the two circles is shaded.

If A is the set of colleagues and B is the set of friends, then the intersection of the two sets is all those people (if there are any) who are both colleagues and friends. Friends who are not also colleagues and colleagues who are not friends are excluded.

If the intersection of two sets is the empty set $\emptyset$ (so the two sets have no elements in common), we say that the sets are *disjoint.*

6.4.5 Set difference

The **difference** of two sets A and B, symbolized as $A \setminus B$, is the set containing all elements that are members of A but *not* members of B. So:

$$A \setminus B = A \cap B^{\mathsf{C}} = \{x | (x \in A) \wedge (x \notin B)\}$$

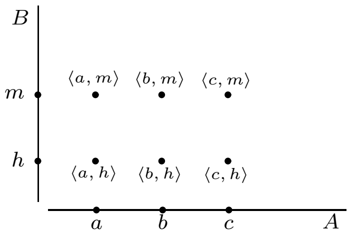
Exercise: Draw the Venn diagram of $A \setminus B$.

6.4.6 Cartesian product

The **Cartesian product** of two sets is a set of *ordered pairs.* An ordered pair has two elements, a first element and a second element, so the order of the elements matters. We indicate an ordered pair using angle-brackets, so that $\langle a, b \rangle$ would be the ordered pair consisting of a and b, in that order. The ordered pair $\langle b, a \rangle$ is a *different* ordered pair. Every ordered pair in the Cartesian product set $A \times B$ has its first element from set A and its second element from set B. So, we define the Cartesian product of set A and set B, denoted by $A \times B$, to be the set of all ordered pairs $\langle a, b \rangle$ such that $a \in A$ and $b \in B$:

$$A \times B = \{\langle a, b \rangle | (a \in A) \wedge (b \in B)\}$$

Example: if $A = \{a, b, c\}$ and $B = \{h, m\}$, then $A \times B = \{\langle a, h \rangle, \langle a, m \rangle, \langle b, h \rangle, \langle b, m \rangle, \langle c, h \rangle, \langle c, m \rangle\}$. The first element in each ordered pair is a member of set A and the second is a member of set B. In this example, $A \times B$ is not the same set as $B \times A$ (exercise: list the elements of $B \times A$ and check the two sets are different) so, using the definition of set equality, $A \times B \neq B \times A$. Prove (for non-empty sets A, B) that $(A \times B = B \times A) \leftrightarrow (A = B)$. This is not true if we allow the empty set, however, as $A \times \emptyset = \emptyset \times A = \emptyset$ for any A; prove this as well.



Generally, Venn diagrams are not a suitable graphic way to represent Cartesian products,[7] but you have already learned a nice way in high school, namely the rectangular axes of a Cartesian coordinate system (the axes could be any sets, not merely the real line, which is how such axes are usually used in high school). The axes for the sets A, B in the previous paragraph contain only three and two points, respectively (though we shall represent them by lines to make the figure more familiar), and the product $A \times B$ has exactly six points, one for each element of the set. This is illustrated at left.

We can also form the Cartesian product of a set with itself, as $A \times A$. In this case, each member of A would be paired with every member of A (including itself). Using set $A = \{a, b, c\}$ from the previous example, we have

$$A \times A = \{\langle a, a\rangle, \langle a, b\rangle, \langle a, c\rangle, \langle b, a\rangle, \langle b, b\rangle, \langle b, c\rangle, \langle c, a\rangle, \langle c, b\rangle, \langle c, c\rangle\}$$

Exercise: Show that if A has n elements, B has m elements (for appropriate numbers n, m), then $A \times B$ will have $n \times m$ elements. Verify this for the examples above, and show it is true for *any* sets A, B.

6.4.7 Exercise on set operators

1. Draw Venn diagrams to illustrate the following sets:
   (a) $(A \cup B) \cup C$ (b) $A \cup (B \cup C)$ (c) $(A \cap B) \cap C$ (d) $A \cap (B \cap C)$
   (e) $A \cap (B \cup C)$ (f) $(A \cap B) \cup (A \cap C)$ (g) $A \cup (B \cap C)$ (h) $(A \cup B) \cap (A \cup C)$
   (i) $A \cup B$ (j) $B \cup A$ (k) $A \cap B$ (l) $B \cap A$
   (m) $A \cap \mathbb{U}$ (n) $A \cup \emptyset$ (o) $A \cup \mathbb{U}$ (p) $A \cap \emptyset$
   (q) $A \cap B^{\mathsf{C}}$ (r) $A^{\mathsf{C}} \cup B$ (s) $\emptyset^{\mathsf{C}}$ (t) $\mathbb{U}^{\mathsf{C}}$

2. Look at the diagrams for 1.(a) and 1.(b), above. Does it make any difference which pair of sets is parenthesized? What about 1.(c) and 1.(d)? Prove (based on the definitions of union and intersection) that it doesn't matter. Notice that in each case the proof essentially uses the same "associative property" of $\vee$ and $\wedge$.[8] In general, operators that have the associative property are said to be associative operators. $\cup$ and $\cap$ are associative operators.

3. Look at the diagrams for 1.(e) and 1.(f), above. Do the same for 1.(g) and 1.(h). This suggests that there are distributive laws for sets, much like the distributive law in arithmetic $a \times (b + c) = (a \times b) + (a \times c)$, and like the distributive rules in propositional logic. We say that union is distributive over intersection and intersection is distributive over union. Prove (based on the definitions) that the union and association operators have these distributive properties. One interesting point to notice here is that each of $\cup$ and $\cap$ distributes over the other, unlike in arithmetic where only one of the two possible distribution properties is true (the distribution $a + (b \times c) = (a + b) \times (a + c)$ is false).

4. Consider the pairs 1.(i), 1.(j) and 1,(k), 1.(l), what can you say about the commutativity of union and intersection? Prove (from the definitions) that the union and intersection operators have the commutative property. Notice in all your proofs that a property of set operators reflects a similar property of logical connectives.

[7] You'd need a 4D picture! Can you see why?

[8] The commutative, associative, and distributive properties for propositional logic were stated and proven to be equivalences in Section 1.3.12 and again in Exercise 4.3.1 #5.

5. What general principles (rules or laws) can you conclude from 1.(m) to 1.(p)? Prove them.

6. What do the pair 1.(q), 1.(r) suggest? These two sets are always complements: prove this from the definitions. Reword the equation using set-difference. And what do the last pair 1.(s), 1.(t) suggest? Prove that these sets are complementary, *i.e.* that $\emptyset^{\mathsf{C}} = \mathbb{U}$ and $\mathbb{U}^{\mathsf{C}} = \emptyset$.

7. Some other equations dealing with complements are:
   (a) $(A \cup B)^{\mathsf{C}} = A^{\mathsf{C}} \cap B^{\mathsf{C}}$ (b) $(A \cap B)^{\mathsf{C}} = A^{\mathsf{C}} \cup B^{\mathsf{C}}$ (c) $(A^{\mathsf{C}})^{\mathsf{C}} = A$

   Prove each of these, and illustrate the equation with Venn diagrams.

8. Can you prove that the Cartesian product operator is commutative? Explain your answer.

9. Given a universe of 26 elements named by the lower-case letters of the alphabet: $\mathbb{U} = \{a, b, c, \ldots, z\}$, a set $A = \{a, l, p, h, b, e, t, s\}$, a second set $B = \{s, e, t, i, n, r, c, o\}$, and $C = \{s, e, t, u, n, i, o\}$, draw a Venn diagram and show the members of the sets by putting the letters into the appropriate areas.

10. Prove that there can be no more than one empty set: if E and N are both empty, then $E = N$. Use only the "official" (extensional) definition of set equality.

11. Some of the following statements are equivalent to $A \subseteq B$, some are not. Identify which statements are which, using Venn diagrams in each case to help justify your answer.
    (a) $A \cap B^{\mathsf{c}} = \emptyset$ (b) $A \setminus B = \emptyset$
    (c) $B^{\mathsf{c}} \subseteq A^{\mathsf{c}}$ (d) $A \cup B = B$
    (e) $A \cap B = A$ (f) $A^{\mathsf{c}} \cap B = \emptyset$

12. (Optional challenge:) Construct a Venn diagram for four sets, showing regions for all 16 subsets (including the empty set). Feel free to try to find a way to represent a suitable Venn diagram for five (or more!) sets.

6.5 Answers to Exercises

Exercises 6.3.3

1. Subsets of $A = \{a, b, c, d\}$: $\emptyset, \{a\}, \{b\}, \{c\}, \{d\},\ \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\},$ $\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\},\ \{a, b, c, d\}$
   (16 subsets). All but the last are proper subsets.
   $\mathcal{P}A = \big\{\emptyset, \{a\}, \{b\}, \{c\}, \{d\}\,,\ \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\},\ \{a, b, c\}, \{a, b, d\},$ $\{a, c, d\}, \{b, c, d\}, \{a, b, c, d\}\big\}$

2. $\mathcal{P}\{a, b, c\} = \big\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\big\}$

   $\mathcal{P}\{a, b, c, d, e\} = \big\{\emptyset, \{a\}, \{b\}, \ldots, \{e\}, \{a, b\}, \{a, c\}, \ldots, \{a, e\}, \{b, a\}, \{b, c\}, \ldots, \{d, e\}\,,$ $\{a, b, c\}, \{a, b, d\}, \ldots, \{c, d, e\}, \{a, b, c, d\}, \ldots, \{b, c, d, e\}, \{a, b, c, d, e\}\big\}$
   I have not written all 32 subsets—I leave it to you to fill in the gaps.

3. $\mathcal{P}(\mathcal{P}(A))$ will have 256 different elements, each being a set of subsets of A, such as the following (see the subsets of A as listed in #2 above—we are now considering sets of these sets):
   $\big\{\emptyset, \{a\}, \{b, c\}\big\}$, $\big\{\{b\}, \{a, b, c\}\big\}$ and $\big\{\{a, b\}, \{a, c\}, \{a, b, c\}\big\}$
   and 253 others … $\mathcal{P}(\mathcal{P}(A))$ will have all these sets as *elements.*

4. (a) F (b) T (c) F (d) T (e) T (f) T (since A is not empty) (g) F (h) T (i) F (j) T (k) T (l) F

Exercises 6.4.7

1. (a – r): see the figure on the next page. (s,t): $\emptyset^{\mathsf{c}} = \mathbb{U}$ and $\mathbb{U}^{\mathsf{c}} = \emptyset$.

2. Clearly these pairs are equal, as shown in the Venn diagrams. To illustrate the proofs, I'll prove (a) = (b).

   > Let $x \in (A \cup B) \cup C$. Then ($x \in A$ or $x \in B$) or $x \in C$. Because $(p \vee q) \vee r \leftrightarrow p \vee (q \vee r)$, we can conclude $x \in A$ or ($x \in B$ or $x \in C$); *i.e.* $x \in A \cup (B \cup C)$. And similarly *vice versa* because of the equivalence. So $(A \cup B) \cup C = A \cup (B \cup C)$.

The other proofs in these exercises are similar. Ask me if you need more details.
Here are hints for the rest of this exercise set:

7. I'll illustrate this with a proof of the first one:

   > Suppose $x \in (A \cup B)^{\mathsf{c}}$: then $x \notin A \cup B$, in other words, $\neg(x \in A \vee x \in B)$. But $\neg(p \vee q) \leftrightarrow (\neg p \wedge \neg q)$, so this is equivalent to $x \notin A \wedge x \notin B$. In other words, $x \in A^{\mathsf{c}} \cap B^{\mathsf{c}}$. (And *vice versa*, because of the equivalence.)

   See the figure for Venn diagrams to illustrate this, and the others.

8. This is discussed in the text; show that $A \times B \neq B \times A$ for the specific sets in the Example in section 6.4.6, for instance.

9. See the figure.

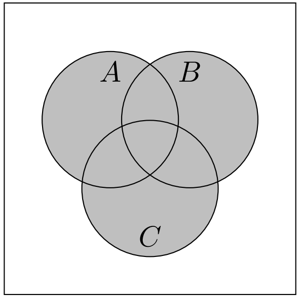
10. For any $x \in E$, $x \in N$ is true (simply because there is no $x \in E$ in the first place). So $E \subseteq N$, and the reverse is similar. So $E = N$.

11. The statements that are equivalent to $A \subseteq B$ are (a) (there is nothing in A 'outside' B), (b) (taking all elements also in B from A leaves nothing behind), (c) (everything not in B is also not in A), (d) (adding elements from A to B doesn't actually add anything to B), (e) (only things from A are in both A and B). (f) is not equivalent (there's nothing in B 'outside' A—that means $B \subseteq A$, not $A \subseteq B$).

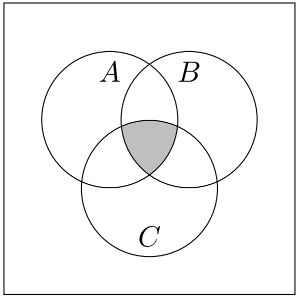
    I'll let you draw Venn diagrams to illustrate this. I've done (a) in the figure.

12. See the figure for one possibility.

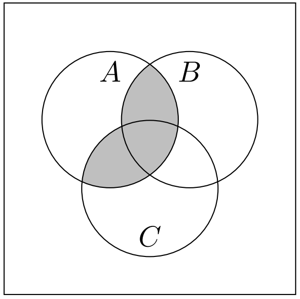
1 (a,b)



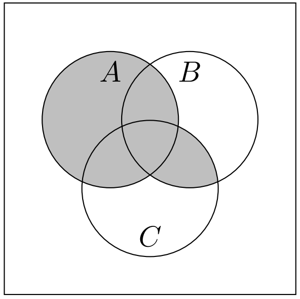
1 (c,d)



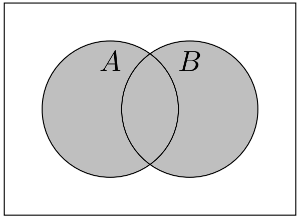
1 (e,f)



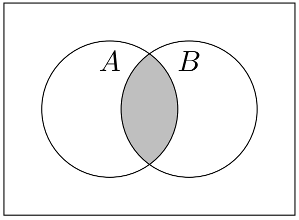
1 (g,h)



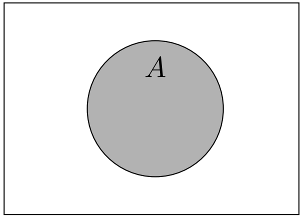
1 (i,j)



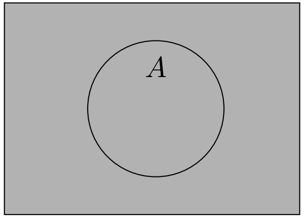
1 (k,l)



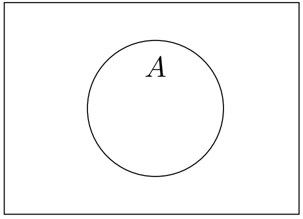
1 (m,n)



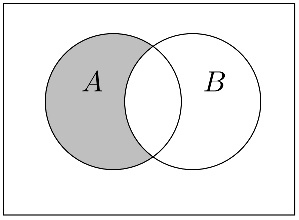
1 (o)



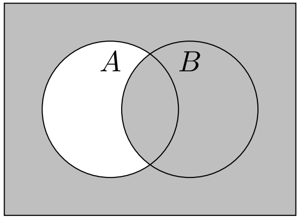
1 (p)



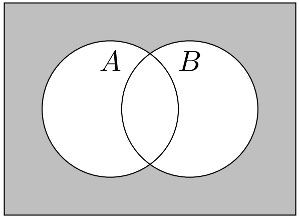
1 (q)



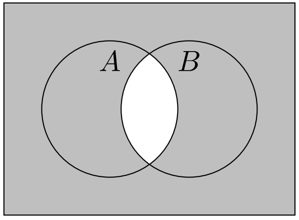
1 (r)



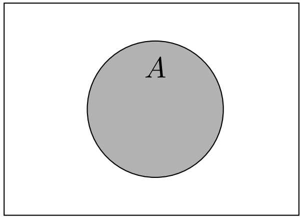
7 (a)



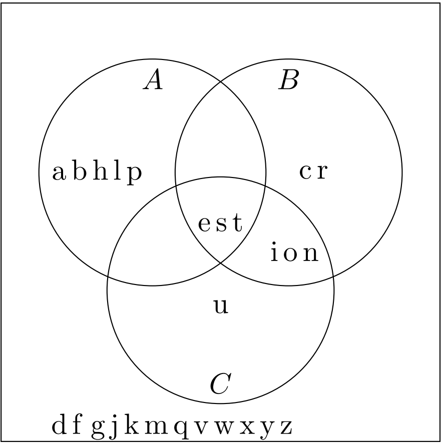
7 (b)



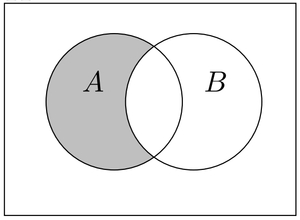
7 (c)



9



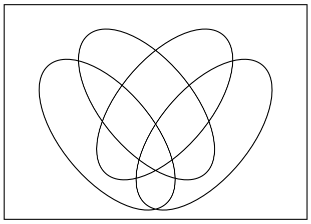
11



If $A \cap B^{\mathsf{c}} = \emptyset$, then the shaded area must be empty—which means that $A \subseteq B$.

12

One possible version of Venn for 4 sets:



Chapter 7

What is a Number?

7.1 Numbers, Numerals, Cardinals, and Ordinals

What is a number? This is not an easy question to answer; think about it a bit to see if you can come up with a satisfactory answer. Numbers are abstract: you cannot touch, draw, or point to a number. (You can draw a numeral; we'll come to that in a moment.) You can hold up five fingers, but you cannot hold up just 'five'. Some abstract notions are tied to some physical reality: for instance "red" corresponds to a wavelength of light. But some are less tied: for instance "goodness" doesn't seem to be linked to a physical phenomenon. Numbers have something of this nature as well, and one way around the problem of defining just what numbers are is similar to a well-known way to define notions like "goodness": "goodness" may be defined to be that property shared by all things that are good (!), and so in a similar fashion (but without the transparent circularity!) we might define "five" as the property shared by all collections (sets) which have exactly five elements.

We shall start this chapter by looking at this idea more carefully, and show that it can be handled in a way to define numbers in a fashion that is well grounded (not being circular in its logic). We shall then look at the history of numbers, in particular, at number systems used by past cultures which have had an influence on our own history. Finally we shall look at some different types of numbers used in mathematics today.

7.1.1 Counting

Let's start with an intuitive idea of what "number" means, and draw the distinction between "number" and "numeral": a **number** is a "quantity of countable things". A **numeral** is a word or symbol that stands for a number. So, when you hold up five fingers, the number five is the quantity of fingers, and the word "five" is the numeral representing that number. The distinction between numbers and numerals is very important. We *invented* numerals; numerals are artefacts of human language, arbitrarily chosen and adopted by convention. We did *not* invent the (natural) numbers. The number of sheep in a valley is some particular number, whether there are people around to count them or not, even if nobody has a name or symbol for that number.

It's worth remarking that counting is a very basic human skill, one that predates the invention of numerals. Prehistoric humans could count quantities even without having much in the way of numeric concepts. In many languages, the largest numerals refer to fairly small numbers, such as two or three or five. To refer to a number that is larger than the one named by their largest numeral, they use a word or phrase that means something like our words "many" or "infinity".

Still, these groups of people can and could count and record large countable quantities.[1]

How do pre-numerate[2] cultures count and record numbers? They use a fundamental principle of set theory, namely that **two sets contain the same number of elements if there is a one-to-one correspondence between the elements of the two sets**. This is what we mean by numerical equality, and the point is that we don't need to know what numbers are to use it.

Imagine a pre-numerate sheep-owner about to send her sheep out with the herd-boy. The owner puts a pebble into a jar as each sheep leaves. When the sheep return, the owner dumps the pebbles into a bowl and puts one back into the jar as each sheep enters the fold. If the owner has extra pebbles in the bowl after all the sheep are in the fold, she has lost sheep. If she runs out of pebbles before all the sheep are in, she has gained sheep. She doesn't need numerals (a way of saying *how many* sheep she has). She cannot answer the question "How many?" in words. She just holds up the jar. The set of pebbles is not a numeral—it is not a *symbol* of the number. It *is* the number. It is a set that has a one-to-one correspondence with the collection of sheep. It answers "How many?" with "This many". Notches carved in sticks or bones, knots in cords, beads on a string, or tally-marks on clay tablets or on paper work the same way.

The point to notice about this is that it is possible to understand what it means for two collections (sets) to have "the same number of things", without having to understand what "number" means. So "same size" would seem to be a more basic concept than "size", and that is what we look at next.

7.1.2 Cardinality and sets

Numbers that answer the "How many?" question are referred to as **cardinal** numbers. The point being made in the previous paragraphs is that it is easier to define what it means for two sets to have the same **cardinality** (or size) than it is to define what cardinal numbers are themselves. From this observation, in the early twentieth century, when mathematical logicians and philosophers tried to define just what is a number, they arrived at the viewpoint that it would be a good strategy to define number in terms of sets, more particularly, in terms of "same-size sets". There were several variations on this theme: we shall briefly look at two. Later in the century, an alternate approach to "foundations of mathematics" (trying to define what maths is about, and in particular, what numbers are) was proposed, which avoids this set-theoretical baggage: I shall also briefly describe that as well. What may one conclude from all this? Well, perhaps mainly that it is possible to give a logical basis for discussing numbers, for to be sure, most practicing mathematicians do not actually bother with such issues, being quite happy to work with numbers regardless of how they are defined.[3]

A naive proposal[4] (essentially what Bertrand Russell and Alfred North Whitehead did in their seminal *Principia Mathematica*) would be to say a cardinal number is a (maximal) set of sets all of which have the same cardinality. So for instance, we could say the number two is the set of all the

[1] You might like to look at *Pi in the Sky: Counting, Thinking, and Being*, by John D. Barrow, Oxford 1992.

[2] "Numeracy" means "ability with or knowledge of numbers". John Allen Paulos (*Innumeracy: Mathematical Illiteracy and its Consequences*, Vintage Books 1988) defines "innumeracy" as "an inability to deal comfortably with the fundamental notions of number and chance". As we speak of "pre-literate" cultures, "pre-numerate" seems a good word for those without a sophisticated vocabulary for dealing with numbers, and "innumerate" is as a kind of correlate to "illiterate".

[3] Actually, I should be a bit more careful: it is *natural* numbers, defined soon, that mathematicians are happy to use without definition; other numbers, such as integers, rationals, reals, . . . , are carefully defined starting with the natural numbers, more or less as we shall do later in this chapter.

[4] There are some technical problems with this naive idea, but they have fairly simple technical resolutions, so we won't worry too much about that here.

sets which have the same size as the set $\{0, 1\}$ (*i.e.* all sets with two elements, but note that we don't need to know what "with two elements" means; we merely need to understand what "which have the same size as $\{0, 1\}$" means). Then, *e.g.*, saying that the set $\{a, b\}$ has two elements would be understood as saying that it belongs to the set we are identifying as the number two.

An alternate view, associated with the logicians Ernst Zermelo and John von Neumann, is to say a cardinal number is a particular representative set, whose size (cardinality) establishes the number. So, for example, the number 0 is defined to be $\emptyset$, the empty set, 1 is the set $\{\emptyset\}$, 2 is the set $\{\emptyset, \{\emptyset\}\}$ (which is also $\{0, 1\}$, which suggests the pattern: each number is the set of previous numbers), and so on. So, according to this definition, to say a set has 2 elements would be to say it has the same cardinality as the set $\{0, 1\}$. What both definitions have in common is that numbers are defined to be particular sets, and to say that a set has a particular "number of elements" is taken to mean that it has the same cardinality as some other set (whose cardinality is pre-determined).

With this view, then, the (philosophical) problem of defining what a number is has been shifted to the problem of what it means for two sets to have the same size or cardinality. We define this as follows:

> Two sets A and B have the same (equal) cardinality if and only if there is a correspondence such that every element of A corresponds to exactly one element of B, in such a way that every element of B corresponds to exactly one element of A. (This is called a one-to-one correspondence.)

We write $\#A = \#B$ to mean A and B have the same cardinality. We often read this as "the cardinality of A equals the cardinality of B", thinking of $\#A$ as "the cardinality of A" or even as "the number of elements of A". But the real notion is not "cardinality" (yet), but "of equal cardinality".

Note that sets may have the same cardinality without being equal; consider for example the sets $A = \{a, m, t, b\}$ and $B = \{0, 1, 2, 3\}$. $\#A = \#B$, but $A \neq B$ (the sets are not equal: A might be a collection of Beanie Babies while B is a set of (natural) numbers).

The idea then is that a cardinal is the cardinality (size) of certain sets.[5] When we say "the cardinality of this set is two", we may understand this as saying "this set has the same cardinality as $\{0, 1\}$". Two (or deux or zwei or 2 or II or whatever numeral we use to refer to it) is the size of any set which is the same size as (can be put into a one-to-one correspondence with) the set of my feet or the set $\{a, b\}$ or the set $\{0, 1\}$. Three (trois, drei, 3, III, or whatever) is the cardinality of any set that is the same size as the set $\{0, 1, 2\}$. And so on. Finite cardinal numbers are called **natural numbers**. The set of natural numbers is an infinite set, denoted $\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$.[6] We use the notation $\#A$ to denote the cardinality of the set A, so for example $\#\{a, e, i, o, u\} = 5$.

What would it mean (in the spirit of the discussion in this section) to say $\#A < \#B$? Think about this a moment: you should be able to convince yourself that this should mean that whenever you try to establish a correspondence between the elements of A and the elements of B, there should be some "left over": every element in A corresponds to exactly one element of B, but there are always (no matter how you do the correspondence) elements in B that are not associated with

[5] Technically, the relation "of equal cardinality" is an equivalence relation, and cardinals are the equivalence classes for this relation (modulo some technical quibbles). In the Zermelo approach, these equivalence classes are represented by canonical representatives. If this idea is unclear, don't worry: all you need to remember is that it is possible to give an unambiguous definition of number in terms of the clear notion of "same size".

[6] Here is another convention about which there is no universal agreement: some people regard 0 as the first natural number (what is the cardinality of $\emptyset$?), some start with 1 (being guilty of anti-$\emptyset$ism). In this course we shall start with 0, but you should be aware that some other books you might look at could use the other convention. This changes very little of importance, however.

elements of A. This last condition means that you cannot associate every element of B with exactly one element of A. This leads us to the following definition:

> $\#A < \#B$ (the cardinality of a set A is less than the cardinality of a set B) if and only if the cardinality of A is equal to the cardinality of a subset of B and the cardinality of B is *not* equal to the cardinality of any subset of A.

What the definition says is that the cardinality of a set A is less than the cardinality of a set B if there is a subset of B that is the same size as A, but there is no subset of A that is the same size as B. The shepherd who lost sheep found that the sheep could be put into a one-to-one correspondence with a subset of the pebbles from the jar, but that there was no way to put all the pebbles from the jar into a one-to-one correspondence with any subset of the sheep (there would always be extra pebbles left over—corresponding to the missing sheep).

We can define $>$ and other order relations (like $\leq, \geq, \not<, \not>, \neq$) in the obvious way.

Remark: For *finite* sets A and B, if $A \subsetneqq B$ (*i.e.* A is a *proper* subset of B) then $\#A < \#B$, *i.e.*, the set A is smaller—has a smaller cardinality—than B. But be careful!! This is only true for *finite* sets, and is definitely *not true* for infinite sets, such as $\mathbb{N}$. We'll discuss this later in this chapter. Cardinalities for *infinite* sets have some *very* surprising properties.

Freeing ourselves from set theory

I said earlier that there was another approach to numbers which avoided the complications of set theory: basically that approach says not to worry about what a number *is*, focus instead on what *properties* it possesses. So one takes (say) the natural numbers to be just some entities, and lists a set of axioms those entities must possess in order for them to have the required properties one expects of the natural numbers. Such axioms we shall see later (the Peano axioms), so I won't go into detail now. But what this view of the basis of number gives us is the freedom not to worry what numbers *are*, instead we can act as if we already know that, and focus on their properties. This is what most of us (mathematicians and non-mathematicians alike) do anyway, and it will be what I shall do from now on in this text (except for now I assume you know the basic properties of numbers, leaving the technicalities for later).

7.1.3 Ordinals and counting

Another aspect of counting gives an order relation on the things counted; this corresponds to the notion of an **ordinal**. As with cardinals, this is a very ancient notion; pre-numerates understand ordinals. A pre-numerate shepherd can know that Dolly was her *first* sheep and Belle was her *second*, or which sheep came home *first* and which came home *second*. These ordinal notions derive from recognition of the order of events in time. Once she had no sheep. Then she had Dolly. Once she had just Dolly, and then she had Dolly and Belle. These statements embody ordinal concepts even before the shepherd has ordinal numerals. "First", "second", and "third" and so on are ordinal numerals. Numerals like "1", "2", and "3" can represent ordinals when they are used to express a place in a series of numbers. They also can represent cardinals (which makes them ambiguous).

The first ordinal number corresponds to the cardinal number of the smallest set: the empty set. We call it 0. The next number could be defined as: 1 is the cardinality of any set X such that $\#X > 0$ and for any other set Y, if $\#Y > 0$ then $\#Y \geq \#X$. Think about this definition carefully. We define "next number" (an ordinal notion) in terms of minimally-greater cardinality. Exercise: define the next cardinal number after 1.

Now, when an ancient shepherd is counting her flock, she can name each sheep (or each pebble in the jar) by using the names of the ordinal numerals in order. So, implicitly (before counting the first sheep) she starts with 0, then, as she counts the first sheep, she uses the name of the next ordinal (1), then the next (2), and so on. When she has matched an ordinal n against the last (n^{th}) sheep, she can report the number of sheep in the flock by stating the last ordinal (n) she used. The last ordinal is the cardinal number of sheep. That's the way we count.

7.2 Transfinite Arithmetic

Discovering the foundations of number in the concept of "set" has many consequences. One of the consequences is that we can speak and write more clearly about infinity. Until Georg Cantor in the 19th century, we were a little like those people whose numerals only go up to three or five. If we were asked how many natural numbers there are, we could only say "infinitely many". Asked how many rational numbers (fractions) there are, we could only say, again, "infinitely many". How many real numbers? Again we would say "infinitely many". No distinction was made between the sizes of these various sets of numbers.

Cantor chose the numeral (symbol) $\aleph_0$ for the cardinality of the set $\mathbb{N}$ (the set of natural numbers). He called the cardinality of an infinite set its "order of infinity". The numeral $\aleph_0$ is aleph (the first letter in the Hebrew alphabet) followed by a subscript zero (0), and it's pronounced "aleph-nought" or "aleph-null" or simply "aleph-zero". Then, by definition, $\#\mathbb{N} = \aleph_0$.

$\aleph_0$ stands for a new kind of number, a transfinite (or infinite) cardinal.

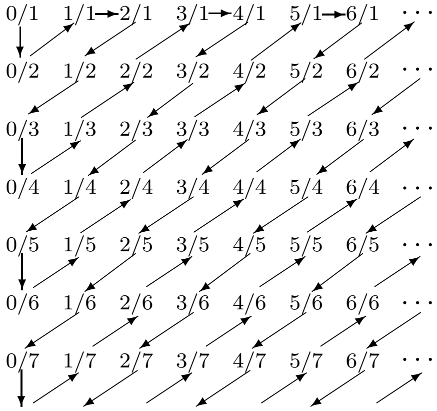
How can we decide whether infinite sets have cardinalities that are less than, the same as, or greater than the cardinality of the set of natural numbers? We go back to the idea that sets of equal size can be mapped one-to-one against each other.

The (infinite) set of *even* natural numbers is the set $\{0, 2, 4, 6, 8, \ldots\}$. There is a simple way to establish a one-to-one correspondence between this set and $\mathbb{N}$, to find a correspondence such that every element of the set of even natural numbers corresponds to exactly one element of $\mathbb{N}$: Just map $0 \leftrightarrow 0, 1 \leftrightarrow 2, 2 \leftrightarrow 4, 3 \leftrightarrow 6$, and so on. It is clear that by this mapping every even number corresponds to one and only one natural number, and every natural number is covered by this correspondence. By the definition of "equal cardinality" these two sets are the same size! But the set of even natural numbers is a *proper* subset of the set of natural numbers. So the whole (the set of natural numbers) is **not** greater than the part (the set of even natural numbers). This seems to violate a basic axiom of arithmetic and geometry. Intuitively, if there are $\aleph_0$ even numbers, there are $2 \cdot \aleph_0$ numbers. It seems that $2 \cdot \aleph_0 = \aleph_0$. You get the same result if you notice that the set of all odd natural numbers can also be mapped one-to-one against the set of natural numbers (each natural number n maps to the odd number $2n + 1$). The cardinality of the set consisting of the union of the two disjoint sets is $\aleph_0 + \aleph_0$. But the union of those two sets is just the set of all natural numbers, whose size is $\aleph_0$. So, **contrary to our intuition**, it follows that $\aleph_0 + \aleph_0 = \aleph_0$. It is possible to show that for any *infinite* cardinal numbers α, β, the sum $\alpha + \beta$ is in fact equal to the larger of the two cardinals. One of the things you must let go of when dealing with infinite cardinalities is your intuition, based on finite numbers. The finite is **not** a good guide in guessing what the infinite is like.

Galileo noticed that the infinite set of all squares of natural numbers can be put into a one-to-one correspondence with the set of natural numbers. Map each natural number against its square, as $0 \leftrightarrow 0, 1 \leftrightarrow 1, 2 \leftrightarrow 4, 3 \leftrightarrow 9, 4 \leftrightarrow 16$, and so on. The set of squares is a proper subset of the set of natural numbers, yet the two sets have equal cardinality.

We defined "less-than" by saying "the cardinality of a set A is less than the cardinality of a set B, $\#A < \#B$, if and only if the cardinality of A is equal to the cardinality of a subset of B and the cardinality of B is not equal to the cardinality of any subset of A". The cardinality of the set of squares is equal to the cardinality of a subset of the set of natural numbers, because the set of squares *is* a subset of the set of natural numbers. Is the cardinality of the set of natural numbers equal to the cardinality of a subset of the set of squares? It is: we have just shown that the cardinality of the set of natural numbers is equal to the cardinality of the set of squares by finding a one-to-one correspondence between them, and the set of squares is a subset of the set of squares. (Any set is a subset of itself.) The sets are "the same size" in the sense that they have equal cardinality.

This seems, and seemed to Cantor's contemporaries, very paradoxical: a set can be the same size as a proper subset? How can the part be the same size as the whole? Well, the answer in one sense is just what we have already seen: the one-to-one mappings show this is exactly what is the case. But we can turn this seeming paradox on its head, and notice that in fact this is a basic *property* of *infinite* sets: it is only infinite sets that can have proper subsets of the same cardinality as the whole set. In fact, this could be a *definition* of infinite set: a set X is infinite if there is a proper subset A with the same cardinality $\#A = \#X$.



There are more surprises for you dealing with infinite sets of numbers. Consider the following fact: the cardinality of the set of all fractions (all "rational numbers", which we shall define more carefully soon, should you not recall them from high school) is also exactly $\aleph_0$. Try to see if you can create a one-to-one association of natural numbers to all fractions. Here's one way: it is possible to arrange all fractions in (infinitely long) rows, stacked one above the other in an infinitely tall and infinitely wide table. Put all the fractions with denominator 1 in the first row ($\frac{0}{1}$,$\frac{1}{1}$,$\frac{2}{1}$,$\frac{3}{1}$,...), all the fractions with denominator 2 in the second row ($\frac{0}{2}$,$\frac{1}{2}$,$\frac{2}{2}$,$\frac{3}{2}$,...), *etc.* This will give you lots of duplicates, but you can easily eliminate them if you want. Then trace a path through this table, starting at the top-left, and working along diagonals right and down. In this way you can list, one after the other, every fraction (skip duplicates if you want), thus giving the one-to-one correspondence you need.[7] Notice that this suggests that $\aleph_0 \cdot \aleph_0 = \aleph_0$.

It is starting to look as if every infinite set has size (cardinality) $\aleph_0$, so that $\aleph_0$ appears to be just another way of saying "infinitely many". It's not. Cantor proved that there are infinite sets whose cardinality is greater than the cardinality of the (infinite) set of natural numbers. In particular, he showed that the cardinality of the set of natural numbers ($\aleph_0$) is less than the cardinality of the power set of the set of natural numbers. The proof is a nice example of subtlety and creativity in mathematics.

Here is a simple variant of this result: we shall show that the set of decimal numbers between 0 and 1 has a strictly greater cardinality than $\aleph_0$, by showing that any enumeration of such decimals (any attempt at a one-to-one correspondence between the decimals and the natural numbers) always omits some decimals. So, consider any enumeration of decimal numbers between 0 and 1: for example, the first decimal in your enumeration might be 0.12345678987654321... , the second

[7] There are many other clever ways to "count" all the fractions. Here is another: collect all the fractions in lowest terms where the sum of numerator and denominator is 1, and arrange them in numerical order; then do the same for those where the sum of numerator and denominator is 2 (omit duplicates), then 3 then 4 .... This gives a list of all fractions, as required. This list would start thus: $\frac{0}{1}, \frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \ldots$.

might be 0.098765432123456789... , and so on. We construct a decimal (let's call it x) which is not in your enumeration as follows: look at the first decimal place in your first decimal number (which is a 1 in our example), and set the first decimal place of x to be something different (*e.g.* add one to the digit, unless the digit is 9, in which case take 8 to be your different digit[8]). So in our example x begins 0.2. For the second decimal place in x, look at the second decimal digit in the second decimal number in your enumeration: in our example, this digit is 9. Again take a different digit (in our example, using the procedure mentioned for the first digit, we would take 8 as our different digit). That is the second digit of x, so x now looks like 0.28. Continue in this way—every digit of x differs in at least one position from every decimal number in your enumeration, so your enumeration cannot contain x. In this way, we can see that **no** enumeration of decimal numbers between 0 and 1 can possibly list *every* decimal number between 0 and 1, and so the cardinality of the set of all decimal numbers between 0 and 1 must be strictly greater than $\aleph_0$. In fact, the cardinality of the set of decimal numbers between 0 and 1 is the same as the cardinality of $\mathcal{P}(\mathbb{N})$, and Cantor's proof in that case is essentially the same. His proof may be modified slightly to show that for every set A, $\#A < \#\mathcal{P}(A)$ (the power set of A always has a strictly greater cardinality than that of A itself). So there are lots (infinitely many) of infinite cardinals.[9]

This proof is a bit subtle, to be sure, but underlying it is a principle (called "diagonalization") that has proved of considerable power in twentieth century mathematical logic (we shall see a similar technique when we consider Gödel's incompleteness theorems, at the end of the course).

Infinite sets whose cardinality is $\aleph_0$ are called "countably infinite sets". Although we could never finish counting all the members, they are "countable" in the sense that we can map them one-to-one against the set of "counting numbers" ($\mathbb{N}$, the natural numbers). Infinite sets whose cardinality is greater than $\aleph_0$ are called uncountably infinite.

"How many natural numbers are there?" "Infinitely many" is vague and ambiguous. "$\aleph_0$ many" is more precise. There are (infinitely many) different "sizes" (orders) of infinity. Not all infinities are the same. A similar story may also be told about infinite ordinals, though not in this text.

7.3 Systems of Numeration

By a system of numeration we mean a set of elementary numerals and a scheme or rule for combining elementary numerals to represent numbers.

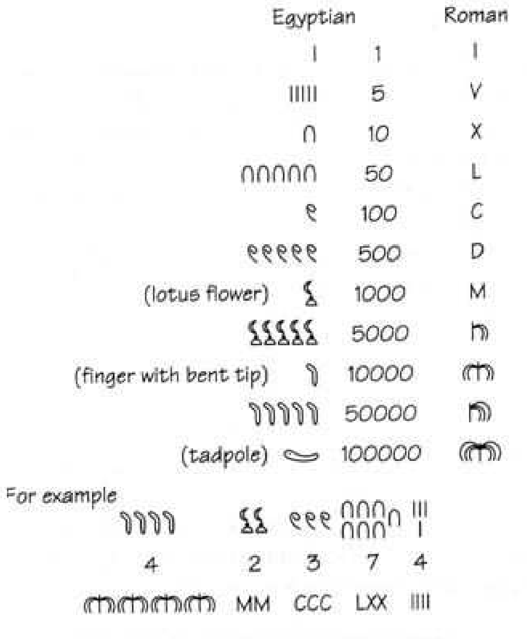
There are all sorts of odd developments in the early history of numerals as they developed beyond the "one, two, many" counting systems. In some cultures, the names for the numbers differed depending on the kind of thing one was counting. In such systems, the numeral for the

[8] This choice avoids the problem that some different decimal numbers are in fact equal, such as $0.49999\ldots = 0.5000\ldots$

[9] This is the tip of a very interesting, if astounding, story. One aspect is the following. The cardinality of $\mathcal{P}(\mathbb{N})$ is usually denoted $2^{\aleph_0}$; we know that $\aleph_0 < 2^{\aleph_0}$, but are there any (infinite) cardinalities in between? One of the significant achievements of twentieth century set theory was to find the (surprising) answer to this question. Cantor had proposed the "Continuum Hypothesis" (CH), his hypothesis that any set of reals (the "continuum", in his terminology) was either finite, or had cardinality $\aleph_0$ or $2^{\aleph_0}$; in other words, that there was no cardinality in between $\aleph_0$ and $2^{\aleph_0}$. In spite of all his efforts, he failed to prove or disprove his Hypothesis. Gödel, in the late 1930's, showed that CH is *consistent* with the usual axioms of set theory, by finding a model in which $2^{\aleph_0}$ was the next infinite cardinal after $\aleph_0$. Later, in the early 1960's, Paul Cohen proved CH *independent* of the usual axioms of set theory, by finding a model in which there *were* other cardinals between $\aleph_0$ and $2^{\aleph_0}$. So no wonder Cantor couldn't prove or disprove his guess—there *is* no such proof or disproof, and in order to settle the matter, one must use principles about sets that go beyond the usual axioms. Since Cohen's proof, many other mathematical questions have been found with this property that they are consistent with but independent of the usual axioms of set theory. The world of the infinite is a very odd place indeed.

number four when applied to sheep was different from the numeral for four when counting pebbles or trees. Traces exist in English, where we say “a pair of shoes”, “a brace of pistols”, “a span of oxen”, and so on.

However, there are two basic patterns we see in various systems: additivity and place-value. These are sometimes “pure”, but frequently are mixed together or with other types of systems.


Egyptian
Roman
1
I
5
V
10
X
50
L
100
C
500
D
(lotus flower)
1000
M
5000
(finger with bent tip)
10000
50000
(tadpole)
100000
For example
4
2
3
7
4
MM
CCC
LXX
IIII

The Egyptian and Roman number systems.

Some of the first systems of numeration were additive systems. In an additive system, the number represented by a particular string of numerals is the sum of the values of the numerals. The purest form of an additive system perhaps is a simple token-based system: there is only one numeric symbol, a stroke perhaps, and numbers are represented by appropriate numbers of strokes. So, *e.g.* three might be |||. Of course, this quickly becomes impracticable, so generally many tokens were used, but the key feature of adding the tokens to derive the number represented by a numeral remained. The most familiar additive system is the Roman system, based on a small set of elementary numerals (such as I, V, X, L, C, and so on), repeating them as necessary to represent larger numbers.[10] The Egyptian system of numeration was similar—you may find a summary of these in the figure to the left.

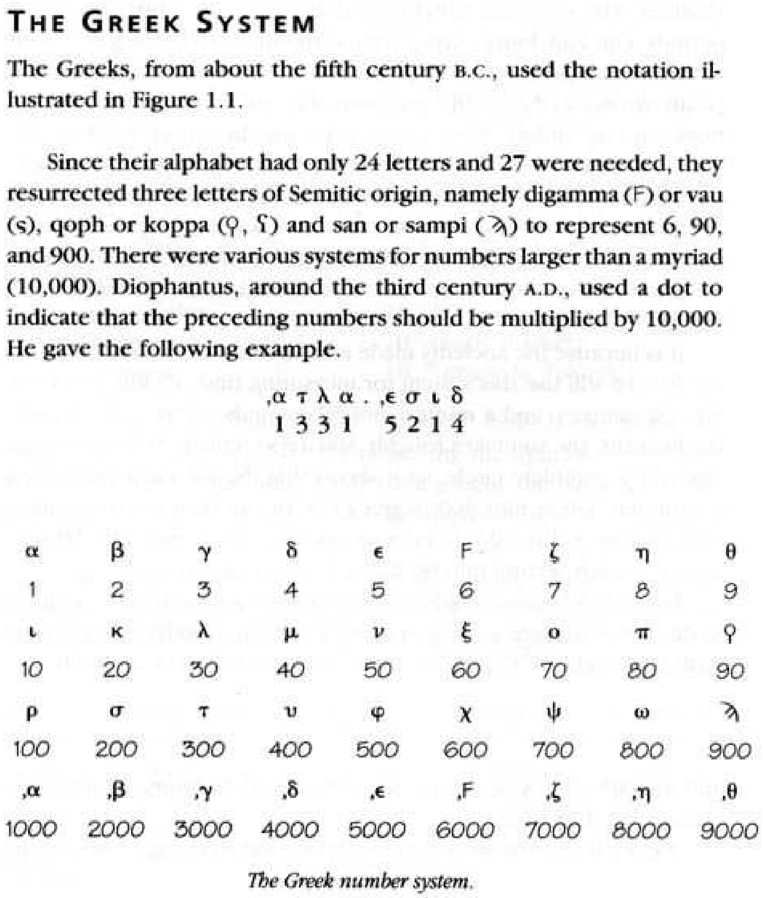
An additive system has the advantage of being simple in concept, but can quickly become unwieldy, especially for representing large numbers. In the Roman system, very large numbers become almost parodies of themselves—consider the spider-like figures used for 100000. But the most serious drawback is in using such numerals for computation. Multiplication and division are so complicated as to be almost impossible—and in fact societies that used Roman numerals generally also used small portable computers (the abacus) for routine calculations. For instance, every shop would have an abacus so as to be able to calculate products like XV times VII (which, by the way, comes to CV—try to figure that out without converting to our numerals!).

Sometimes an additive system used a lot of elementary numerals—such a system is often referred to as a ciphered system. For example, the Ionic Greeks and the Hebrews used a ciphered system. Ciphered systems use a different symbol for each small number. Ancient Greek numerals were the 24 letters from their alphabet plus three letters from the Phoenician alphabet. Nine of these symbols represented the numbers from 1 through 9. Nine different symbols were used for 10, 20, 30, ..., 90, and another nine for 100, 200, 300, ..., 900. For 3000, they used the symbol for 3 (γ, “gamma”) and added a tick-mark, so 3000 was $,\gamma$. The value of a compound numeral was still the sum of the values of the component simple numerals, but there was no repetition of symbols as in a purely additive system.

Well, not exactly: to represent very large numbers, eventually a form of repetition was introduced, by using a dot (or period) to represent multiplication by 10000. With this convention, for example, 5214 would be represented by $,\epsilon\sigma\iota\delta$ (note that comma to get 5000 from the ϵ representing

[10] The medieval trick of using subtraction, as in IV representing 4, being V minus I, or IX being X minus I, or 9, is a late development, added presumably for convenience. You will see the earlier IIII instead of IV on many clocks, although VIIII is rarer, IX being preferred.

5), 1331 by $,\alpha\tau\lambda\alpha$, and these two numbers could be combined by multiplying the second by 10000 and adding, to get 13315214 represented by $,\alpha\tau\lambda\alpha., \epsilon\sigma\iota\delta$.



The Greek System

The Greeks, from about the fifth century B.C., used the notation illustrated in Figure 1.1.

Since their alphabet had only 24 letters and 27 were needed, they resurrected three letters of Semitic origin, namely digamma (F) or vau (ς), qoph or koppa (ϙ, ϟ) and san or sampi (ϡ) to represent 6, 90, and 900. There were various systems for numbers larger than a myriad (10,000). Diophantus, around the third century A.D., used a dot to indicate that the preceding numbers should be multiplied by 10,000. He gave the following example.

,α τ λ α . ,ϵ σ ι δ
1 3 3 1   5 2 1 4

α	β	γ	δ	ϵ	F	ζ	η	θ
1	2	3	4	5	6	7	8	9
ι	κ	λ	μ	ν	ξ	ο	π	ϙ
10	20	30	40	50	60	70	80	90
ρ	σ	τ	υ	φ	χ	ψ	ω	ϡ
100	200	300	400	500	600	700	800	900
,α	,β	,γ	,δ	,ϵ	,F	,ζ	,η	,θ
1000	2000	3000	4000	5000	6000	7000	8000	9000

The Greek number system.

But this system still was very awkward to use, and calculations almost impossible to perform. There was also a high price to pay memorizing all the elementary numerals. But another (unsuspected) problem makes such a system a hindrance in doing mathematics: by using letters to represent numbers, how do you represent variables (as in $3x^2 + 1$)? The Greeks simply didn't have variables, and one wonders if their numeration system didn't have something to do with that. Certainly that made algebra more difficult (though they managed quite well in spite of it all!). In this vein, it's interesting to note that Greek mathematical reasoning, even about number theory, was mainly geometric.

Our own (Hindu-Arabic) numeration system uses a quite different principle: place-value. In a place-value system we use the same symbol 2 to mean two or twenty or two hundred or two thousand, depending on where it occurs. In such a system, addition and multiplication tables are simpler and easier to remember than those in additive systems, like the Greek and Roman systems. Systems like ours require more mathematical insight to construct, but need less memorizing in practice.

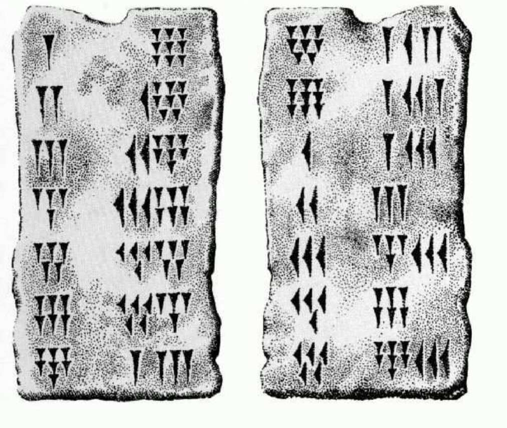
In a place-value system, a compound numeral is organized by position—usually in columns.[11] One column (the right-most column in our system) represents units. The next column represents a multiple of the number used as the *base* of the system. The next represents a multiple of the base times the base, and so on. In our denary (or decimal) system, the base is 10. The Babylonians used a base of 60. Others used other bases.[12]

The Babylonian and Mayan systems used some additive features, to generate the basic numerals (corresponding to our digits) from some simple symbols. Let's take a look at the Babylonian cuneiform system—if only because its rather odd feature of being essentially "base 60" has left a trace in our own culture, where 60 is a fundamental number in time and in measuring angles (60 seconds in a minute, 60 minutes in an hour, 6 times 60 degrees in a circle). The Babylonians had two wedge-shaped symbols: ▾ representing one and ❮ representing ten. (These would be pressed into a clay tablet, using the corner of a piece of wood—not so easy to draw with a pencil, but quite convenient with their writing implement.) They used the additive principle for numbers

[11] Though the Mayans, for example, used a vertical notation, so the different place-values were marked by rows.

[12] A trace of a base 20 system in our own history may be found in terms like "score", as in "four score and seven years ago ...", and "quatre-vingt".

up to 60, so that for example 𒌋𒁹𒁹𒁹𒁹 would represent 14, and 𒌋𒌋𒁹𒁹 would represent 22. But then the remarkable new feature would reveal itself: starting with 60, the Babylonians used place-value. That is, they would start a new column in their number, and values there would represent multiples of 60. So, 𒌋𒌋𒁹𒁹 𒌋𒁹𒁹𒁹𒁹 would represent 22 × 60 + 14 or 1334. This would continue with a new column representing multiples of 60 × 60, and so on. Quite quickly very large numbers indeed could be simply represented with this numeral system. It seems more than likely that the Babylonian expertise in astronomy was related to their development of a system so well suited to recording very large numbers.



In the figure at left are drawings of the front (left) and back (right) of a clay tablet. The cuneiform characters probably represent a simple Babylonian arithmetic lesson (a "9 times" table). From the information above, you should be able to figure out what the symbols mean. Try the symbols in the left column first, starting from the top of the front of the tablet and continuing on the back. Then try the right column.

In Figure 7.1, you can see how the Mayans used a similar mixture of place-value and additivity. They used dots for units and horizontal lines for fives. Four was represented by four dots in a horizontal row. Thirteen was three dots over two lines. Nineteen was four dots over three lines, and was the highest basic numeral (or "digit") they needed: from then on they used a place-value system, much as the Babylonians did. For example, to write the number twenty-two, they would put two dots in one row (not column) and another single dot (representing one twenty) in a row above it.

These place-value systems had two very significant features. We have already seen how easily they can represent large numbers, but also they permit simple algorithms for arithmetic (you probably learnt such algorithms for multiplication and division in primary school!). But there was one small problem, however. Consider a single symbol, such as 𒁹 in cuneiform. This could easily be misinterpreted: for instance, is this 1, or is it 60 (1 × 60 + 0)?

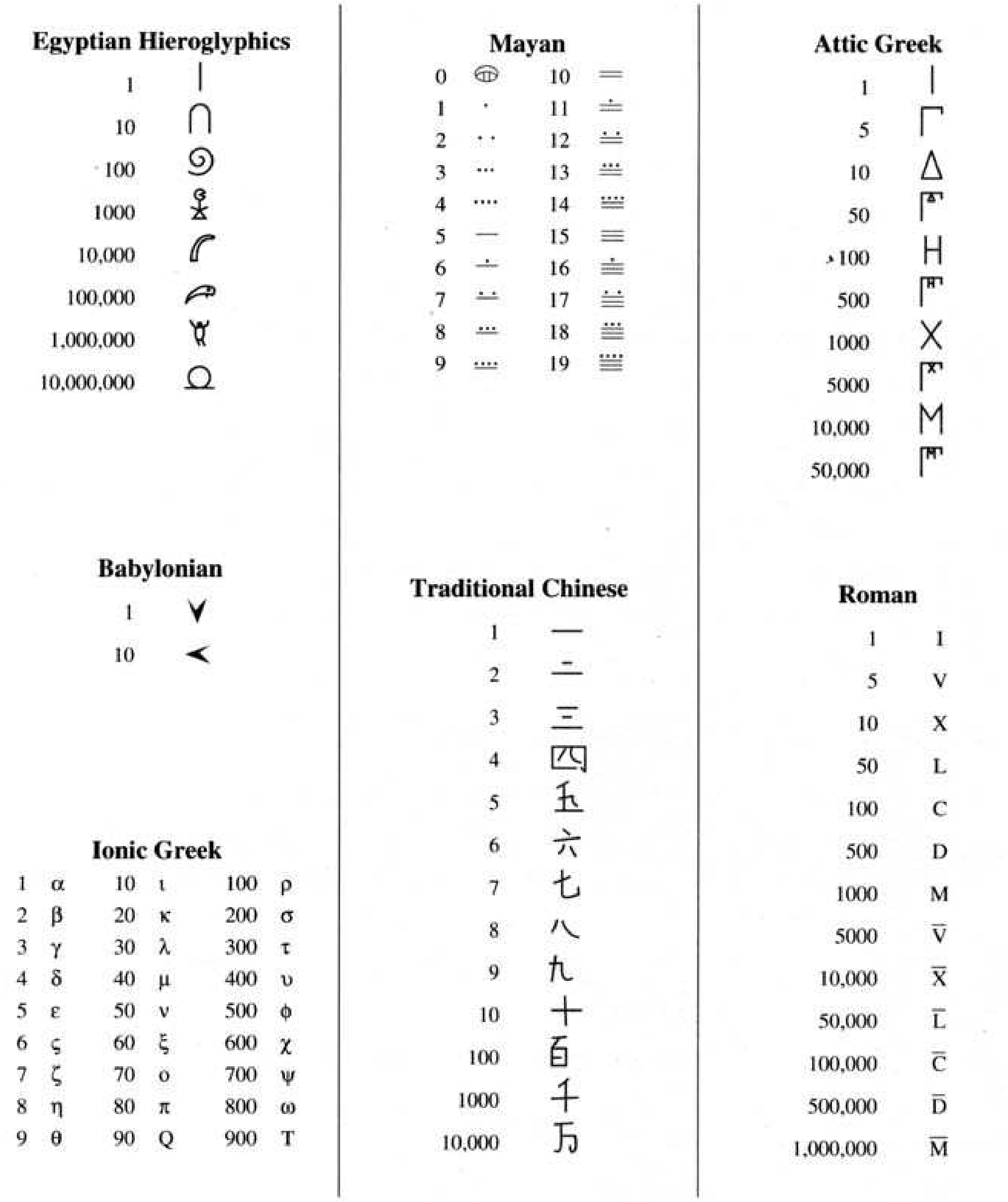
In practice, the Babylonians indicated 60 with a one-wedge in the sixties-column, leaving the ones-column blank. Look at the tablet reproduced above: the cuneiform for 180 (three sixties plus no ones) may be seen in row four in the right column on the back of the tablet. The Mayans did the same sort of thing. To represent 20, they put a single dot in the 20s-row and left the ones-row blank. But this could be ambiguous (is the units-column/row blank, or was the writer just sloppy?).

This wasn't usually a problem; for example, the Babylonians recorded numbers in neat columns, so an empty column usually wasn't hard to spot, but there must have been times when such misinterpretations were made.

In a place-value system, we need elementary numerals for units up to one less than the base. In the decimal system, we need 1, 2, 3, 4, 5, 6, 7, 8, and 9. But the dangerous ambiguity of "blank columns" led to one of the most important inventions in mathematics—the concept of zero. At first, Babylonians just used a special mark to indicate that there was no numeral in a particular column. Eventually the mark came to be seen as a special numeral, and then as a number.

There are now ten digits in the decimal system: the original nine and the new zero. The concept of zero as a number was scandalous to many people. Zero was not seen as the cardinality of a set. To say that it is the cardinality of the empty set seemed ridiculous. If there are no sheep in a field, it did not seem sensible to say there are 0 sheep in the field.[13] No such problem existed with

[13] Perhaps some mathematicians still regard 0 with suspicion, and so start the natural numbers with 1.



Egyptian Hieroglyphics

1	𓏺
10	𓎆
100	𓍢
1000	𓆼
10,000	𓂭
100,000	𓆐
1,000,000	𓁨
10,000,000	𓍶

Mayan

0	𝋠	10	𝋪
1	𝋡	11	𝋫
2	𝋢	12	𝋬
3	𝋣	13	𝋭
4	𝋤	14	𝋮
5	𝋥	15	𝋯
6	𝋦	16	𝋰
7	𝋧	17	𝋱
8	𝋨	18	𝋲
9	𝋩	19	𝋳

Attic Greek

1	Ι
5	Γ
10	Δ
50	𐅄
100	Η
500	𐅅
1000	Χ
5000	𐅆
10,000	Μ
50,000	𐅇

Babylonian

1	𒁹
10	𒌋

Traditional Chinese

1	一
2	二
3	三
4	四
5	五
6	六
7	七
8	八
9	九
10	十
100	百
1000	千
10,000	万

Roman

1	I
5	V
10	X
50	L
100	C
500	D
1000	M
5000	$\overline{\text{V}}$
10,000	$\overline{\text{X}}$
50,000	$\overline{\text{L}}$
100,000	$\overline{\text{C}}$
500,000	$\overline{\text{D}}$
1,000,000	$\overline{\text{M}}$

Ionic Greek

1	α	10	ι	100	ρ
2	β	20	κ	200	σ
3	γ	30	λ	300	τ
4	δ	40	μ	400	υ
5	ε	50	ν	500	φ
6	ς	60	ξ	600	χ
7	ζ	70	ο	700	ψ
8	η	80	π	800	ω
9	θ	90	Q	900	T

Figure 7.1: Various Numeration Systems (variations)

Roman numerals. This was one of the reasons the Hindu-Arabic place-value system was resisted in Europe. Leonardo of Pisa[14] tried to introduce the base-10 place-value system in 1202. In 1299 that system of numeration was outlawed in Florence. It only became widespread after 1479. Eventually, this anti-0ism prejudice disappeared, for the most part, and we're quite comfortable with 0 among our numbers.[15]

7.3.1 Place-value systems

There is some mathematics behind the place-value systems, which we shall look at a bit now.

The numerals representing the natural numbers $0, 1, 2, 3, 4, 5, 6, 7, 8,$ and 9 are the digits of the denary (decimal) system. A place-value system that includes 0 and whose base is n must have n digits (from 0 to $n - 1$).

The numeral for a large decimal number (say 27453) is a string of digits. The location (place) of the digits determines their value, using the operations of addition and multiplication. The numeric value of the right-most digit in the numeral is just the value of that digit. Each other digit has a value that is 10 times as great as it would be if it were one column to the right. In this case, the "3" (the numeral) just means 3 (the number). If the "5" were one column to the right, it would mean 5, so it actually means $10 \cdot 5 = 50$. If the "4" were one column to the right it would mean 40, so it actually means $10 \cdot 40 = 400$. Similarly, the "7" stands for 7000 and the "2" means 20000. So the number 27453 is $3 + (5 \cdot 10) + (4 \cdot 10 \cdot 10) + (7 \cdot 10 \cdot 10 \cdot 10) + (2 \cdot 10 \cdot 10 \cdot 10 \cdot 10)$, or $3 + 50 + 400 + 7000 + 20000$. In general, if the base of a place-value system of numeration is b (where the digits go from 0 to $b - 1$) any number larger than $b - 1$ is represented by a string of digits $\ldots d_5 d_4 d_3 d_2 d_1 d_0$, and the string stands for the number

$$d_0 + d_1 \cdot b + d_2 \cdot b^2 + d_3 \cdot b^3 + d_4 \cdot b^4 + d_5 \cdot b^5 + \ldots$$

Note that this "goes backwards": in the place-value notation of the number, the left-most digits go with the largest powers, and the right-most digits go with the lowest. When you are evaluating a number in a new base, I suggest you start from the right end, as in the expression above.

Binary (or base-2) numerals use a base b of 2. They use numerals for digits from 0 up to one less than the base. That means that the only digits (elementary numerals) in the binary system are 0 and 1. These are the "binary digits", often abbreviated "bits". Any number is represented by a string of zeros and ones $\ldots d_4 d_3 d_2 d_1 d_0$ whose meaning is $d_0 + d_1 \cdot 2 + d_2 \cdot 2^2 + d_3 \cdot 2^3 + d_4 \cdot 2^4 + \ldots$. To translate numbers from any other base to binary, you could try to remember that (in the decimal system) $2^2 = 4$, $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, and so on, and then figure out how many 32's, how many 16's, how many 8's, *etc.* would add up to the number.

For example, to convert 117 to binary, we notice that $2^7 = 128$ won't go into 117 (nor will higher powers of 2). $2^6 = 64$ goes, leaving a remainder of $117 - 64 = 53$. Write down a 1 as the left-most digit of our answer. $2^5 = 32$ goes into 53, leaving a remainder of 21. Write a 1 to the right of our first 1, so 11 are the left-most two digits. 2^4 goes into 21 with a remainder of 5. Write a 1 to the right of our 11, so we have 111 as the left-most three digits. $2^3 = 8$ won't go into 5, so write a 0. $2^2 = 4$ goes, with a remainder of 1, so write another 1. Our first five digits are 11101. $2^1 = 2$ won't go into 1. Write a 0. Write the 1 in the ones column. We're done. The number is 1110101 in binary.

[14] Also known as Leonardo Fibonacci—we'll meet him again.

[15] A most readable account of the "history of zero", including its first use, may be found in Amir Aczel's *Finding Zero*, Palgrave Macmillan 2015. He also successfully argues the case that our numerals originated in India, and later spread to the Arab area, and then to the west. So the "Hindu-Arabic" numerals are properly only "Hindu", or more appropriately, since this is not a religious issue, "Indic numerals".

In order to distinguish numerals written in base b (as opposed to another base), we often use a subscript to indicate the base. So our previous calculation showed that $117_{10} = 1110101_2$ (117 in base 10 equals 1110101 in base 2). Then we adopt the convention that we omit the subscript if the numeral is in base 10.

We shall simplify this algorithm, as illustrated by the following examples. First, we shall spell it out in words, then we shall write this in a convenient manner that makes the calculation somewhat more streamlined.

To convert 117 (base 10) to binary, divide 117 by 2. We get 58 and a remainder of 1. We put that 1 as the right-most digit of our answer. Divide 58 by 2. We get 29 and a remainder of 0. Put that 0 in the next column to the left, giving 01. Divide 29 by 2, giving 14 and remainder of 1. Put the 1 in the next column to the left, giving 101. Divide 14 by 2, giving 7 and a remainder of 0. Put the 0 in the next column to the left, so we have 0101. Divide 7 by 2, giving 3 and a remainder of 1, so put a 1 in the next column to the left, giving 10101. Divide 3 by 2, giving 1 with 1 left over, so a 1 goes into the next column, giving 110101. Divide 1 by 2, giving 0 with 1 left over, so put 1 into the next column. Since we are down to 0, we are finished, so our answer is 1110101_2. To check that this is the right answer, notice that

$$\begin{aligned}
&1 + 0\cdot 2 + 1\cdot 2^2 + 0\cdot 2^3 + 1\cdot 2^4 + 1\cdot 2^5 + 1\cdot 2^6 \\
&\quad= \quad 1 + 0\cdot 2 + 1\cdot 4 + 0\cdot 8 + 1\cdot 16 + 1\cdot 32 + 1\cdot 64 \\
&\quad= \quad 1 + 4 + 16 + 32 + 64 \;=\; 117
\end{aligned}$$

Here's an even easier way to set up and do the conversion. Again, we are converting 117 (base 10) into base 2. We start by setting up a simple division of 2 into 117, which goes 58 times with remainder 1:

$$\begin{array}{r|r|l}
2 & 117 & \\ \hline
 & 58 & 1 \\ \cline{2-3}
\end{array}$$

Next we divide 2 into 58, which goes 29 times with remainder 0: we just write that immediately below the previous division:

$$\begin{array}{r|r|l}
2 & 117 & \\ \hline
 & 58 & 1 \\ \cline{2-3}
 & 29 & 0 \\ \cline{2-3}
\end{array}$$

We continue in this way, till we arrive at 2 into 1, which goes 0, remainder 1. In general, this method consists of repeated divisions, till we get the answer 0 with some remainder.

$$\begin{array}{r|r|l}
2 & 117 & \\ \hline
 & 58 & 1 \\ \cline{2-3}
 & 29 & 0 \\ \cline{2-3}
 & 14 & 1 \\ \cline{2-3}
 & 7 & 0 \\ \cline{2-3}
 & 3 & 1 \\ \cline{2-3}
 & 1 & 1 \\ \cline{2-3}
 & 0 & 1 \\ \cline{2-3}
\multicolumn{1}{r}{} & \multicolumn{1}{r}{} & \uparrow
\end{array}$$

Our final binary numeral is read **from the bottom up** as the remainders we obtained, in this case 1110101_2. (As we already knew!—look at this method and convince yourself it's just the same as we did in the previous paragraph, which is also what we first did. All that's changing is the smoothness of the presentation of the algorithm.)

Example: convert 2763 (base 10) to base 3 (ternary). We use the same method, but using 3 instead of 2.

3	2763	
	921	0
	307	0
	102	1
	34	0
	11	1
	3	2
	1	0
	0	1

Note: Don't stop until you get a 0 as a quotient. Write the ternary equivalent of 2763 by reading the remainders *from the bottom up*: 10210100_3. Check your answer by converting back to base 10:

$$\begin{aligned} & 0+0\cdot 3+1\cdot 3^2+0\cdot 3^3+1\cdot 3^4+2\cdot 3^5+0\cdot 3^6+1\cdot 3^7 \\ = \;& 9+81+486+2187 \;=\; 2763 \end{aligned}$$

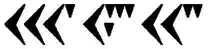
as required.

7.3.2 Exercise on change-of-base

1. Convert the following to bases 2, 3 and 5 (so each number represents three problems!):
   (a) 5239 (b) 128 (c) 357
   (d) 1234 (e) 486 (f) 75

2. Convert the following to base 10:
   (a) 11010011101_2 (b) 1202102_3 (c) 42301_5
   (d) 410_7 (e) 1001101_2 (f) 120210_3
   (g) 43031_5 (h) 56_7 (i) 11010100010_2

3. In many ways, 12 is a better choice of base than 10. You can divide 12 into halves, quarters, thirds, and sixths, whereas 10 only divides into halves and fifths. You can divide 144 (12^2) into halves, thirds, quarters, sixths, eighths, ninths, twelfths, twenty-fourths, thirty-sixths, forty-eighths, and seventy-secondths, but 100 only divides into halves, quarters, fifths, tenths, twentieths, twenty-fifths, and fiftieths. In base 12, we need digits for the numbers that we call 10 and 11 in the decimal system; it is traditional to use letters, so use A and B for these new digits representing the numbers ten and eleven (so your digits go $0,1,2,3,4,5,6,7,8,9,A,B$). Convert the following numbers to base 12, and check your answers.
   (a) 34 (b) 1024 (c) 16
   (d) 17 (e) 298 (f) 1066

4. And some from Babylon:
   (a) Convert  to base 10.
   (b) Convert 112462 to Babylonian numerals.

7.4 Some Common Types of Numbers

Let's return to the question "what is a number?". We've discussed this for the natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, and even seen how different cultures represent these as numerals. Let's start from $\mathbb{N}$ and consider the other numbers you are familiar with, such as $\frac{22}{7}, \pi, 3.1416$. (By the way, those are three different, unequal, numbers—right?—check on your calculator, if you're not sure, though your calculator will only give you an approximation of π.)

These are all *real* numbers, but of different types. $\frac{22}{7}$ and 3.1416 are *rational* numbers, whereas π is *irrational* (it is also *transcendental*). We shall briefly examine the nature of these numbers, from the point of view of starting with $\mathbb{N}$ and then adding new types of numbers to that set, getting larger and larger collections as we do. (To be honest, we shall "fudge" things a bit when we get to the reals, as that is an extension of a more subtle nature than the others. But I hope the flavour of the reals will be a bit clearer, even if the details are partially hidden.) We shall end the chapter with an extension of the reals to include "imaginary" numbers, such as $\sqrt{-1}$.

7.4.1 The integers

We start by adding "negative numbers" to the natural numbers.[16] One way to motivate this might be to ask "what is the solution to the equation $5 + x = 0$, or even $5 + x = 3$?". Notice that we have no problems solving an equation $5 + x = 7$ using the natural numbers: $x = 2$ works just fine. The problem with the first two equations (from our advanced standpoint!) is that the solutions are not values $x \geq 0$, and so are not natural numbers. Formally, what we do is we add **additive inverses** to $\mathbb{N}$.

A brief digression: notice that 0 has a very special property with respect to $+$. For any number n, we have $n + 0 = 0 + n = n$. This property is usually described by saying 0 *is an additive unit* or *an additive identity.* It is a "neutral" element for addition: adding it has no effect on the value of the number to which it is added.

Then we say the **additive inverse** of a number n is a number n' with the property that $n + n' = n' + n = 0$. So adding an additive inverse to a number gives you the additive unit. We can show that if a number has an additive inverse, then it has only one such—additive inverses are unique (if they exist).[17] In the set $\mathbb{N}$, only one element has an additive inverse, namely the number 0 (which is its own inverse). Notice that if n' is the additive inverse of n, then n must also be the additive inverse of n'.

The set of **integers** $\mathbb{Z}$ is obtained from $\mathbb{N}$ by adding additive inverses for every positive natural number. We denote the additive inverse of n by $-n$ in the usual way, so that

$$\mathbb{Z} = \{\ldots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}$$

We can define addition on $\mathbb{Z}$ in a straightforward way.[18] Then it's not hard to see that $\mathbb{Z}$ obeys the

[16] There was a lot of controversy about the idea of negative numbers, just as there was about 0. Many people refused to accept that a "negative number" was a number at all—and probably many folks still only accept them out of habit (or because they don't think clearly about anything). Negative numbers are a big step in the history of thought (as was 0), and it was not universally accepted as an *obvious* step. In fact, each extension of the notion of number had to overcome some resistance, and the terminology such numbers have (*negative, irrational, imaginary*) reflects that resistance. But then, even 1 had its detractors: if "I have a number of cars" means I have more than 1, then can 1 actually be a number?

[17] Here is how: suppose a number n has two additive inverses, n' and n''. Then $n + n' = 0 = n + n''$. Now add n' to each side of these equations: $n' + n + n' = n' + 0 = n' + n + n''$. Since $n' + n = 0$, we get $n' = n' = n''$, or $n' = n''$. In other words, if n has two additive inverses, then in fact they must be equal, and so n really only has one.

[18] Here's how: if m, n are natural numbers with $m \geq n$, we define $(-m) + (-n) = -(m + n)$, $m + (-n) = m - n$ and $(-m) + n = -(m - n)$.

usual commutative and associative laws for addition, and that 0 is still the additive unit. Notice that n must be the additive inverse of $-n$ for any n, and so $-(-n) = n$.

What about multiplication? What is the product of a positive and a negative integer? The product of two negative integers? The key here is that we define $n \cdot (-m) := -(n \cdot m)$ for $n, m \in \mathbb{N}$. This really is a consequence of our wanting to preserve the distributive law (which will now be true). Here is an illustration that shows if the distributive law is true, then $n \cdot (-m) = -(n \cdot m)$ must be true. Start with $m + (-m) = 0$ and multiply both sides by n. On the left side, we get $n(m+(-m)) = nm+n(-m)$. On the right we get $n0 = 0$, so we have $nm+n(-m) = 0$, which shows that $n(-m)$ is the additive inverse of nm, *i.e.* $-(nm) = n(-m)$, as claimed. If you reverse this argument, you will see that taking the definition $n(-m) = -(nm)$ guarantees that the distributive law is preserved, so this definition is equivalent to preserving the distributive law.

From $n(-m) = -(nm)$ (and by commutativity $(-n)m = -(nm)$ also) we get that $(-n)\cdot(-m) = -(n(-m)) = -(-(nm)) = nm$, so the product of two negative numbers is the positive number obtained by multiplying the corresponding positive numbers.

What is the cardinality of the set $\mathbb{Z}$? Since it has "twice as many elements as $\mathbb{N}$", you might expect its cardinality is $\aleph_0$, by analogy with the results we obtained earlier (*e.g.* that the cardinality of the even natural numbers is the same as the cardinality of all natural numbers), and indeed that is so. If we re-write $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, 4, -4, 5, -5, \ldots\}$, you can see how we can find a one-to-one correspondence between $\mathbb{Z}$ and $\mathbb{N}$. Map 0 to 0, map every positive number $n \in \mathbb{Z}$ to $2n - 1$ (so $1 \in \mathbb{Z}$ maps to $1 \in \mathbb{N}$, $2 \in \mathbb{Z}$ maps to $3 \in \mathbb{N}$, *etc.*) and map every negative number $-n \in \mathbb{Z}$ to $2n$, (so $-1 \in \mathbb{Z}$ maps to $2 \in \mathbb{N}$, $-2 \in \mathbb{Z}$ maps to $4 \in \mathbb{N}$, *etc.*).

7.4.2 The rational numbers

In passing from the natural numbers $\mathbb{N}$ to the integers $\mathbb{Z}$, we added additive inverses for all numbers that didn't already have them (only 0 had an additive inverse already, being its own inverse). Now we shall do the same for multiplication.

First notice that 1 has the property of being a *multiplicative unit* or *identity*: for all numbers a, $a \cdot 1 = 1 \cdot a = a$ (1 is a neutral element with respect to multiplication: multiplying by it has no effect on any number).

We say the **multiplicative inverse** of a number a is a number a' with the property that $a \cdot a' = a' \cdot a = 1$. So multiplying a number by its multiplicative inverse gives you the multiplicative unit. As with additive inverses, one can show that if a number has a multiplicative inverse, then it has only one such—multiplicative inverses are unique, if they exist. In the set $\mathbb{Z}$, only two elements have multiplicative inverses: 1 and -1 (each is its own inverse). There is one number that cannot have a multiplicative inverse: that is 0. This is because there can be no number a with the property $a \cdot 0 = 1$ (since $a \cdot 0 = 0$ is true for any number a). Notice (similar to additive inverses) that if a' is the multiplicative inverse of a, then a is the multiplicative inverse of a'.

The multiplicative inverse for a number a is often denoted[19] a^{-1} or (maybe more familiar) $\frac{1}{a}$. Our comments above then would say there can be no number $\frac{1}{0}$, and $\frac{1}{1/a} = a = (a^{-1})^{-1}$.

We mentioned that adding negative numbers could be seen as a way to add solutions to equations like $5 + x = 0$ or $5 + x = 3$. In a similar way, adding multiplicative inverses can be seen as adding solutions to equations like $5x = 1$ or even $5x = 3$. A solution to $5x = 3$ will take the form $x = 3 \cdot \frac{1}{5} = \frac{3}{5}$, a "fraction", so that suggests we extend the integers by including all fractions (like $\frac{3}{5}$) and not just multiplicative inverses (like $\frac{1}{5}$).

[19] This is an *exponent*, about which we'll have something to say soon.

So, we start with the integers, and formally we shall extend them to include fractions in a way that doesn't assume we already know what fractions are. To emphasise the formal nature of this definition, we shall write the "new" numbers just as pairs $\langle m, n\rangle$, which don't have any particular connotation, other than what we give them *via* our definitions. However, to keep a grip on the reality these definitions are meant to capture, you may think of such a pair $\langle m, n\rangle$ as being the fraction $\frac{m}{n}$. With these thoughts in mind, we make the following definition.

> **Definition**: A **rational number**[20] is an ordered pair $\langle m, n\rangle$ of integers, $n \neq 0$. The first member of the ordered pair is called the *numerator* and the second member is called the *denominator*. A rational number $\langle m, n\rangle$ is expressed in the form $\frac{m}{n}$ or m/n where m, n are integers, m is the numerator and n is the denominator, and $n \neq 0$.

Every set must come equipped with a notion of equality; what does it mean for two rational numbers to be equal? It is not necessary for the integers involved to be the same (consider $\frac{1}{2}$ and $\frac{2}{4}$); instead we have this definition:

> Two fractions $\langle m, n\rangle, \langle p, q\rangle$ are **equal** if and only if $mq = np$.

For example, $\langle 6, 8\rangle = \langle 3, 4\rangle$, because $6 \cdot 4 = 8 \cdot 3 = 24$. This is (of course!) just the usual notion of equality of fractions, by "cross-multiplying". Check for yourself that $\langle 1, 2\rangle = \langle 2, 4\rangle$.

We denote the *set of rational numbers* by $\mathbb{Q}$.

The integers are included in the rationals, by associating the rational number $\frac{a}{1}$ with the integer a. We often *identify* the rational $\frac{a}{1}$ with a, treating integers as rational; in other words, we frequently omit the denominator 1. This identification is well behaved with respect to the usual operations of arithmetic (meaning, if you add two fractions corresponding to integers, their sum corresponds to the integer you get by adding the original integers: $\frac{a}{1} + \frac{p}{1} = \frac{a+p}{1}$, and similarly for multiplication).

Addition and multiplication of general rationals are easily defined by the equations

$$\frac{m}{n} + \frac{p}{q} = \frac{mq + np}{nq} \qquad \frac{m}{n} \cdot \frac{p}{q} = \frac{mp}{nq}$$

One can then verify that all the laws of arithmetic still hold for the rationals.

Since there are many pairs which are all equal to each other, we often choose one form as being the simplest form or *reduced form* of the fraction.

> **Definition:** A fraction $\frac{m}{n}$ is said to be in **reduced form** if the greatest common integral divisor of m and n is 1, (*i.e.*, if m and n are relatively prime) and if n is positive.

Two numbers are relatively prime if there is no number that exactly divides them both other than 1 or -1: for example, 6 and 8 are not relatively prime, since 2 exactly divides them both, but 3 and 4 are relatively prime, since no number (other than 1 or -1) divides them both. So $\frac{6}{8}$ is not in reduced form, but $\frac{3}{4}$ is.

We extend the concept of division using rational numbers. The fraction $\frac{a}{b}$ can be thought of as an operation of dividing a by b.[21] We say $\frac{a}{b} = x$ if and only if $a = b \cdot x$. By this rule, $\frac{a}{a} = 1$ for any integer a. We define division of two rational numbers as:

$$\frac{m/n}{p/q} = \frac{mq}{np}$$

[20] "Rational" merely means "having the form of a ratio", so fractions are called "rational numbers". The word "rational" only acquired its secondary meaning ("logical, coherent") later, and thereby hangs a tale which shall have to await another time.

[21] A curious thing is happening here: we are identifying a process (division) with the result of that process (the fraction). $\frac{a}{b}$ is both a verb and a noun!

where, of course, none of n, p, q may $= 0$.

Since $\mathbb{Z}$ has additive inverses, so does $\mathbb{Q}$. Subtraction of two rational numbers is given by $\frac{m}{n} - \frac{p}{q} = \frac{mq-np}{nq}$.

This all gives $\mathbb{Q}$ the structure algebraists call a *field*: it has addition and multiplication, an additive unit 0, a multiplicative unit 1, every element has an additive inverse, every element other than 0 has a multiplicative inverse, and the various commutative, associative, and distributive laws hold. It would seem $\mathbb{Q}$ is an excellent domain for all our usual arithmetical and mathematical calculations. What else could we want? ...

Before we leave $\mathbb{Q}$, we should recall that we have already considered its cardinality: although $\mathbb{Q}$ seems very much larger than $\mathbb{N}$ (at least it has lots more elements), in fact its cardinality is the same: $\aleph_0$. But although the number of elements is the same, $\mathbb{Q}$ has a property neither $\mathbb{N}$ nor $\mathbb{Z}$ has: *density.* This means that between any two unequal rational numbers there is a rational number not equal to either one (take the average, for example), and so, in fact, between any two unequal rationals, there must be an infinity ($\aleph_0$) of rationals. There is no such thing as "the next rational number" (in the usual ordering where, for n, q positive, $\frac{m}{n} < \frac{p}{q}$ if and only if $mq < np$).

7.4.3 Interlude: exponents

We used the notation a^{-1} for $\frac{1}{a}$ before, and you may have wondered at this use of exponents (it might even remind you of something your high school algebra teacher told you about). Why do we use negative exponents in this way? Well, the answer is similar to our exploration of multiplying negative numbers when we were discussing the integers: we want the usual properties of arithmetic to be valid, and that wish forces certain things for us, automatically. Let's look at the use of various exponents from this viewpoint.

We start with the definition that for a positive natural number $n > 0$ and (any) number a, by a^n we mean the product of a with itself n times. So $a^1 = a$, $a^2 = a \cdot a$, $a^3 = a \cdot a \cdot a$, and in general, $a^{n+1} = a^n \cdot a$. One property follows from this definition: if $n, m > 0$ are positive natural numbers, then $a^n \cdot a^m = a^{n+m}$ (just count how many as there are, all multiplied together). So, we'd like this property to remain true when we define a^n for exponents n other than positive natural numbers.

Let's start with 0: what should a^0 mean? Well, we know that $n + 0 = n$, and if our property is to remain true, then $a^n \cdot a^0 = a^{n+0} = a^n$, so a^0 has to be a number which has no effect when multiplying: in other words, a^0 has to $= 1$. So, we may *define* $a^0 = 1$,[22] and now we have preserved our property so that it's true for any natural number n, including 0.

What about negative numbers? What should a^{-n} mean? We know that (for any n), $(-n)+n = 0$, so if we want to preserve our property, we must have $a^{-n} \cdot a^n = a^{(-n)+n} = a^0 = 1$, and this means a^{-n} must be the multiplicative inverse $\frac{1}{a^n}$ of a^n. In the special case $n = 1$ we get $a^{-1} = \frac{1}{a}$. (This only makes sense if $a \neq 0$, since division by 0 is impossible, as we saw before: 0 cannot have a multiplicative inverse.)

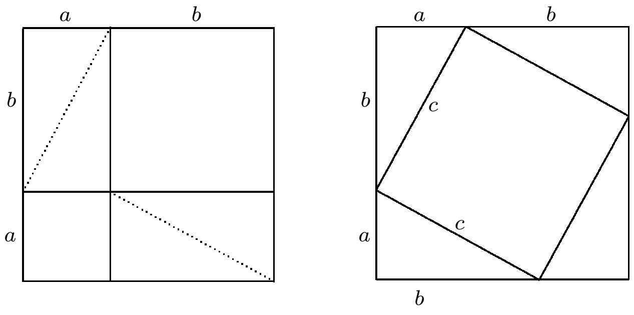
So, those definitions are not arbitrary, they come from a simple wish, namely that the property $a^n \cdot a^m = a^{n+m}$ should remain true for all n, m. We can go even further, by wondering what $a^{\frac{1}{2}}$ might mean (and similarly for other fractions). The same idea gives us the answer. We know that $\frac{1}{2} + \frac{1}{2} = 1$, and so we would want $a^{\frac{1}{2}} \cdot a^{\frac{1}{2}} = a^{\frac{1}{2}+\frac{1}{2}} = a^1 = a$, and so $a^{\frac{1}{2}}$ must be a number b with the property that $b \cdot b = b^2 = a$, in other words, $b = \sqrt{a}$ is the square root of a. So $a^{\frac{1}{2}} = \sqrt{a}$. In a similar manner, we can see that $a^{\frac{1}{3}} = \sqrt[3]{a}$, $a^{\frac{1}{4}} = \sqrt[4]{a}$, and so on: fractional exponents correspond to

[22] There is one case which might seem a bit odd, namely 0^0, since for positive n, $0^n = 0$, and $0^0 = 1$ seems a bit inconsistent with that. However, since 0^x is not even defined for negative x, this isn't quite as odd as it might seem, and we'll keep the definition $0^0 = 1$ for the time being.

appropriate roots. Of course, these numbers (like $\sqrt{a}$) are no longer in the form of simple fractions, so we should look at them a bit closer (in the next section).

7.4.4 Interlude: Pythagoras' Theorem

In high school you may have learned Pythagoras' Theorem, which states that in a right angle triangle, the square on the hypotenuse (the longest side, opposite the right angle) is equal in area to the sum of the areas of the squares on the other two sides: $a^2 + b^2 = c^2$. There are many proofs of this famous theorem (see, for example, `http://www.cut-the-knot.org/pythagoras/`); including several nice geometric, pictorial proofs which show this equation very clearly using clever rearrangements of tiles.



Here is one such, which also illustrates the truth of another famous fact: $(a + b)^2 = a^2+b^2+2ab$, which you might also remember from high school. The point of this illustration is that Pythagoras' Theorem follows just by looking at the figure in the right way. Let's see how. Look first at the square on the left: its sides are each $a + b$, but the way it has been divided, it's clear that its area is also given by two inner squares $(a^2 + b^2)$ *plus* two inner rectangles $(+2ab)$. (In other words, the area of the outer square is $(a + b)^2 = a^2 + b^2 + 2ab$.) Now, look at the square on the right: it is the same square, with sides $a + b$, but it has been differently divided, into four right angle triangles and one square. Each triangle is half of one of the rectangles from the left square (so has area $\frac{1}{2}ab$), and the square inside has its side length given by the hypotenuse of the right angle triangle (so has area c^2), (so this time the area of the outer square may be seen to be $c^2 + 4(\frac{1}{2}ab) = c^2 + 2ab$). Subtract the two rectangles from the left square, subtract the four triangles from the right one, and you have the same result, since you've subtracted the same area from the same larger area. What's left shows that $a^2 + b^2 = c^2$. Or more formally, compare the two squares, and you will see that $a^2 + b^2 + 2ab = c^2 + 2ab$, and hence $a^2 + b^2 = c^2$. So either way we've proven Pythagoras' Theorem.[23]

7.4.5 Irrational numbers

The Greeks had two beliefs about numbers: that numbers were either natural numbers (other than 0) or proportions of natural numbers (meaning essentially rationals), and that numbers corresponded to lengths of line segments. Those beliefs came against a serious contradiction the day one of them proved that there were lengths of line segments which could not be described as proportions of integer lengths,[24] or as we might say it, numbers that were not rational numbers. Here is the simplest such example.

A diagonal line from one corner to the opposite corner of a square, divides the square into two right triangles. If the side of the square is 1 unit long, then by Pythagoras' Theorem, the diagonal

[23] Or rather, we have a good reason to believe it's true, at any rate—whether such a picture actually constitutes a "proof" is a matter for discussion!

[24] There is a story that the person who first discovered this uncomfortable fact was actually executed by his fellows; if so, this is a bad case of shooting the messenger rather than listening to the message. The fact that "irrational" has such a negative connotation still is a hangover from the reaction to the existence of such numbers, however.

must be $\sqrt{2}$ (the square root of 2) units long, meaning that if you squared the length, you would get 2. (Think about this: we want $1^2 + 1^2 = c^2$ where c is the length of that diagonal, but this means $2 = c^2$ and so the diagonal length $c = \sqrt{2}$.) The square root of 2 is a number somewhat less than $\frac{3}{2}$ and more than $\frac{5}{4}$, as you can check by squaring $\frac{3}{2}$ to get $\frac{9}{4}$ and by squaring $\frac{5}{4}$ to get $\frac{25}{16}$ and then checking that one is less than 2, and the other is greater than 2.[25] We can continually subdivide the interval between these two rational numbers into smaller and smaller intervals *ad infinitum*. It would seem that sooner or later we'd come up with the rational number that exactly expresses the square root of two. But such is not the case.

For, we can *prove* that the square root of 2 *cannot* be a rational number. The theorem that the square root of 2 is *ir*rational[26] may be expressed by saying that the side and the diagonal of a square are *incommensurate*: no matter how finely we subdivide a measuring stick, if it can accurately measure the side, it cannot accurately measure the diagonal, and *vice versa*.

Theorem: The square root of two, $\sqrt{2}$, is an irrational number.

Proof: Before we start, we take a moment to point out a rather *special* property of 2: if 2 divides exactly into a product mn, then in fact 2 must either divide exactly into m or divide exactly into n. Think about that for a moment: it says that if a product mn is even, at least one of the factors, m or n (or both) must be even. If this isn't obvious to you, consider the product of two odd numbers ($(2p+1)(2q+1) = 4pq + 2p + 2q + 1$): it must be odd as well. This property is key to the proof, in fact.[27]

The theorem says that there is *no* rational number whose square is two. Since this is a negative conclusion, it suggests that we use proof by contradiction (our old friend, the ($\neg I$) rule). In other words, we shall assume it *is* rational, and derive a contradiction. So, assume that there is a rational number equal to $\sqrt{2}$: *i.e.* there are integers a and b such that

$$2 = \left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}$$

We may also assume that $\frac{a}{b}$ is a fraction in reduced form (*i.e.*, a and b are relatively prime), for if it is not, then we replace a, b with appropriate integers so that the fraction *is* in reduced form.[28] Multiplying both sides of the equation $2 = \frac{a^2}{b^2}$ by b^2, we get $2b^2 = a^2$. But $2b^2$ is an *even* number, so a^2 must also be even. However $a^2 = a \cdot a$ and so a must be even, by our remark above about the special property of 2. Since a is even, there is some number k such that $a = 2k$. Substituting into $2b^2 = a^2$, we get $2b^2 = (2k)^2 = 4k^2$. Dividing both sides by 2 gives $b^2 = 2k^2$. Reasoning as above, b^2 must be even and hence b is even. So we have shown that a and b must have a common factor of 2. But we assumed that a and b are relatively prime. This is our contradiction, and so there is no rational number whose square is two. (QED)[29]

Think about what the irrationality of $\sqrt{2}$ means. It means that there is a number somewhere between (*e.g.*) $\frac{7}{5}$ and $\frac{71}{50}$, and no matter how finely we divide the interval, we'll never find a rational number which is exactly equal to that number.

Similar proofs about roots (square roots, cube roots, fourth roots, *etc.*) of other numbers might lead one to suspect that there are many irrational numbers: are there more irrational than rational

[25] Even "tighter": somewhere between $\frac{7}{5}$ and $\frac{71}{50}$. Find even tighter approximations yourself. Check them by squaring them and verifying that 2 is between the results of squaring your approximations.

[26] **Irrational** merely means "not rational", *i.e.* not a fraction of integers.

[27] Later (next chapter) we shall see that lots of numbers have this property, and so their square roots are not rational either. In fact, if the square root of an integer is not an integer, then it is not rational.

[28] So in particular, a, b are not *both* even.

[29] *Quod erat demonstrandum*: that which was to be proven, or in other words, we're done here.

numbers? In fact, a simple cardinality comparison shows there are *infinitely* many more irrationals than rationals, since the cardinality of the rationals is $\aleph_0$, but the cardinality of the rationals and irrationals together (the real numbers) is strictly greater, being $2^{\aleph_0}$. Since (we mentioned this before) the sum of two cardinals equals the larger one, the cardinality of the rationals and the irrationals together must equal the greater cardinality of the two, and since that total (being the cardinality of the reals) is $2^{\aleph_0}$, the cardinality of the irrationals must also be $2^{\aleph_0}$.

7.4.6 The real numbers

In our discussion of irrational numbers, we rather pulled a fast one—we assumed that there were numbers, such as $\sqrt{2}$, which are not rational. An alternate view might have been that $\sqrt{2}$ actually doesn't exist, and it was that prospect that so bothered the Greek mathematicians. The eventual resolution was that there is indeed a larger set of numbers which includes both rationals and irrationals, which set we now call the **real numbers**, denoted $\mathbb{R}$. The intuition is that this set corresponds to line lengths, so one of the Greek beliefs at least remained. But what does this really mean?

This is not an easy question, and in a sense, it wasn't really dealt with successfully until the nineteenth century. The problem is that the set of reals embodies the notion of "continuity", and getting a good grasp on that proved elusive. Two standard definitions of the reals were arrived at eventually, both essentially infinitary in nature; we shall briefly consider them both.

The first view is that reals amount to partitions of rationals. By a partition of rationals I mean a pair of sets L, U (so we are considering "binary partitions"), with the properties that (1) every element of L is less than any element of U ($x < y$ for all $x \in L, y \in U$), so these are ordered partitions, and (2) together they include all the rationals ($L \cup U = \mathbb{Q}$). Notice that (1) implies L, U are *disjoint*, meaning they have no elements in common ($L \cap U = \emptyset$), and so with (2) we see that L and U "split" $\mathbb{Q}$ in two parts, one "lower" than the other. The idea we have in mind is that such a split defines a "real number" at the point where the two parts meet. That point might already be a rational number: in some partitions, either L will have a greatest element or U will have a least element: these are the partitions corresponding to rational numbers. (In fact, to the rational number that is the greatest element of L or the least element of U—one cannot have both in a partition, since that would mean the same rational would appear in L and U, contradicting our assumption (1), so the rational corresponding to the partition is unambiguously defined in this case).

In other partitions, L will not have a greatest, nor will U have a least, element, and such partitions will correspond to irrationals. The idea here may be illustrated by the example of $\sqrt{2}$: this will correspond to the partition where $L = \{x | x < 0 \vee (x \geq 0 \wedge x^2 < 2)\}$, and U is the complement of L. In this case $\sqrt{2}$ is exactly the number that "squeezes in between" L and U. The appropriate definition of equality of such partitions is reasonably straightforward, as long as you keep the idea in mind that the partition is intended to represent the real number that is "at the place where L meets U".

The second view is that a real number is the sum of an infinite sequence of rational numbers. Not all infinite sequences may be taken; roughly speaking, we want those infinite sums which actually "converge" or give a meaningful approximation to a finite number. It may be helpful if you think of the sum of an infinite sequence in the following simple manner.

Any rational or irrational number can be represented by an integer followed by a decimal point followed by an infinite string of digits. Each digit after the decimal point represents a multiple of a power of the reciprocal of the base (10) of the number system, in other words, a rational number. So for example, 34.1234 means $3 \cdot 10 + 4 + 1 \cdot \frac{1}{10} + 2 \cdot \frac{1}{100} + 3 \cdot \frac{1}{1000} + 4 \cdot \frac{1}{10000}$.

In an infinite decimal fraction there are infinitely many ($\aleph_0$) digits in the string after the decimal point. Any real number can be expressed as an infinite decimal. Integers like 3 can be represented as $3.0000000\ldots$ or even as $2.99999\ldots$ (the ellipsis indicates that the string goes on forever).

When rational numbers are represented as real numbers there is always a string of digits after the decimal point that repeats infinitely. For example, $13/30 = 0.43333\ldots$ and $71/105 = 0.6761904761904\ldots$; in the first of these, the 3 repeats endlessly, in the second, the 761904 repeats endlessly. There may (as in these examples) or may not be an initial string of digits that is not part of the repeating pattern, but eventually one only has the repeating pattern. On the other hand, irrational numbers do *not* contain an infinitely repeating string. There may be some repetitions in the string, but eventually the pattern changes.

The point is that any infinite decimal corresponds to an infinite sum in the same way finite decimals correspond to finite sums. So, for instance, $\pi = 3.1415926\ldots = 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1000} + \frac{5}{10000} + \frac{9}{100000} + \frac{2}{1000000} + \frac{6}{10000000} + \ldots$.

Irrational real numbers may be further divided into two number classes. Those like the square root of two are called *algebraic* irrationals. An algebraic number is any real number x that satisfies some polynomial (algebraic) equation of the form $a_nx^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1x + a_0 = 0$, where all the coefficients a_i are rational numbers. $\sqrt{2}$ satisfies the equation $x^2 - 2 = 0$, and so it is an algebraic irrational. Numbers like π and e (the base of the natural logarithms) and many others are *transcendental* irrationals. There is no algebraic equation of the above form that is satisfied by π or e. Although not a lot of transcendental numbers are known by name, there must be $2^{\aleph_0}$ of them, since there are $2^{\aleph_0}$ real numbers and only $\aleph_0$ algebraic numbers:[30] so "most" numbers are transcendental.

7.4.7 The complex numbers

Our extensions of the natural numbers so far have all been related to extending the sorts of equations we can solve. For instance, integers allowed us to solve equations like $5+x = 3$ and rationals allowed solutions to $5x = 3$; (algebraic) irrationals allowed solutions to equations like $x^2 - 2 = 0$; but what about an equation like $x^2 + 1 = 0$?

Among the real numbers, there is no number x such that $x^2 = -1$. You may remember learning to solve quadratic equations (equations of the form $ax^2 + bx + c = 0$) by substituting a, b, and c into the (quadratic) formula $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$. When $b^2 - 4ac$ was negative, the formula required taking the square root of a negative number. That was impossible as long as our solutions (values of x) had to be real numbers.

It turns out we can extend the set of real numbers in a manner not entirely unlike the process of adding multiplicative inverses to create the rationals, and in this new extension, all such equations will now have solutions (so in particular, taking square roots of negative numbers will become

[30] Why only $\aleph_0$ algebraic numbers? Well, here is a sketch of how we might count them: first, notice that a polynomial equation of degree n (the highest power of x) is determined by $n+1$ rational numbers, the coefficients of the polynomial. Since the cardinality of $\mathbb{Q}$ is $\aleph_0$, the number of polynomial equations of degree n is $\aleph_0^{n+1}$, the product of $\aleph_0$ times itself $n+1$ times. But a finite product of infinite cardinals is the largest in the product, so this product is still just $\aleph_0$. Now, there are a countable number of possible values of n, so to get the total number of polynomial equations (of all degrees n), we just add the number for each degree—but this is $\aleph_0$ many $\aleph_0$s added together, or just $\aleph_0 \cdot \aleph_0 = \aleph_0$: there are only $\aleph_0$ polynomial equations. Now, each of these equations has only has a finite number of solutions; to convince yourself this is so, consider that if a polynomial equation has solutions $x = a, b, c,$ *etc.*, then the polynomial must be equal to $K(x-a)(x-b)(x-c)$, (*etc.*), where K is the coefficient of the highest power, and where "multiple solutions" are repeated appropriately. The idea here is that finding solutions to $p(x) = 0$ (for a polynomial $p(x)$) is equivalent to factoring the polynomial and setting the factors equal to 0. Each factor $(x-a) = 0$ gives a solution $x = a$, and *vice versa*. So, we have $\aleph_0$ many equations, each with a finite number of solutions, so in all we have $\aleph_0 \cdot \aleph_0$ many solutions, or $\aleph_0$ many algebraic numbers.

possible in this extension of the reals). The trick is that we add one new number i (which should be thought of as $\sqrt{-1}$), and then form other new numbers by adding and subtracting. One can multiply i by a real number b, giving the complex number bi, or one can add a real number a to i, giving $a+i$, or one can form a complex number by adding a real number a to the product bi. So $a+bi$ is a general expression we could form using reals and i. Notice that when $a=0$, the form $a+bi$ is equivalent to the form bi. When $b=1$, the form is equivalent to $a+i$. So $a+bi$ is the generic form of any complex number. We'll see soon that this generic form is sufficient to also allow for all the arithmetic operations.

So, just as rational numbers were introduced as ordered pairs of integers, representing fractions, and addition and multiplication were defined for these ordered pairs, we may define **complex numbers** as ordered pairs of *real* numbers $\langle a,b\rangle$. The set of complex numbers is denoted $\mathbb{C}$. Real numbers are a subset of $\mathbb{C}$, the set of complex numbers: the complex number $\langle n,0\rangle$ is identified with the real number n. We think of (and represent) the pair $\langle a,b\rangle$ as $a+bi$, so that i is represented by the pair $\langle 0,1\rangle$. i is called an **imaginary** number.[31] Two complex numbers are equal if and only if they have the same components: $\langle a,b\rangle = \langle c,d\rangle$ if and only if $a=c$ and $b=d$.

Complex numbers are added and multiplied by simple formulas. Those formulas are given below, but it is best to remember a guiding principle here: the idea is that $i^2=-1$ and the ordinary rules of algebra hold for complex numbers. The definitions are taken so that this will be true, and this forces the definitions to be what they are. So for example $(a+bi)+(c+di)=(a+c)+(b+d)i$ by the usual algebra, and so we *define* addition of complex numbers as $\langle a,b\rangle + \langle c,d\rangle = \langle a+c, b+d\rangle$. Likewise $(a+bi)(c+di) = ac+bidi+adi+bic = ac+bdi^2+(ad+bc)i = ac-bd+(ad+bc)i$ (since $i^2=-1$), and so we are forced to the definition that $\langle a,b\rangle\langle c,d\rangle = \langle ac-bd, ad+bc\rangle$. The thing about this definition is that as a result, $\langle 0,1\rangle\langle 0,1\rangle = \langle 0-1, 0+0\rangle = \langle -1,0\rangle$, and identifying $\langle 0,1\rangle$ as i and $\langle -1,0\rangle$ as -1 (remember any $\langle n,0\rangle$ is identified with the real number n), we have $i^2=-1$ as a result of this formal definition. Actually, in $\mathbb{C}$, -1 has two square roots: i and $-i$ (check this: $(-i)(-i) = (-1)(-1)(i)(i) = i^2 = -1$). In fact, in $\mathbb{C}$, every number (real and complex) has two square roots, three cube roots, four fourth roots, *etc.*

This is a big cheat, isn't it?! The reason for the pairs notation is merely to verify that one can define the complex numbers without having to officially assume the existence of a number i with the property that $i^2=-1$, so that even a sceptic can accept the definition; then one turns around and says "well, now you have such a number whose square is -1, represented by $\langle 0,1\rangle$". Once we're convinced, we drop the pretense and just use the $a+bi$ notation, which is much more convenient.

It is possible to add, subtract, multiply and divide complex numbers: they form a field (just as $\mathbb{Q}$ and $\mathbb{R}$ do), with all the usual algebraic properties you know and love from high school. They have one additional property that is very useful, and is usually expressed this way:

> **The Fundamental Theorem of Algebra:** Every algebraic equation of any degree n with real or complex coefficients, $a_nx^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1x + a_0 = 0$ has solutions in the complex numbers, and so has in fact n solutions.

It's worth pausing a moment to consider how remarkable this is: we formed the complex numbers by adding a solution to *one* equation ($x^2+1=0$) to the reals, and ended up with a number system that allows solutions to *all* polynomial equations, of any degree. We got a lot of bang for that buck! By the way, the reason we get n solutions, just from knowing there is one, is by noticing that once

[31] Once again, that negative terminology! Of course i isn't actually "imaginary" in the sense that unicorns are: it is an actual number with practical uses (in the maths of electric currents, for example), and is perfectly well defined—by the trick of the pairs of reals, for example. But initially there was serious resistance to these numbers, and the terminology reflects that. By the way, once you get used to them, complex numbers aren't really "complex" either.

we have one solution $x = a$, we can divide by the polynomial $x - a$ to get a polynomial of degree $n - 1$, which in turn must have a solution, *etc.* Eventually we collect all n solutions to the original equation this way. (Some solutions are "repeated" more than once, to make up the total of n.)

Complex numbers are not just some mathematicians' fantasy. After they were invented they cropped up again and again in physics and engineering. They are also a central feature in fractal graphics (pictures which are some of the most beautiful results of pure mathematics).

7.5 Answers to exercises

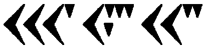
Exercise 7.3.2

1. (a) $= 1010001110111_2 = 2101200 1_3 = 131424_5$ (b) $= 10000000_2 = 11202_3 = 1003_5$ (c) $= 101100101_2 = 111020_3 = 2412_5$ (d) $= 10011010010_2 = 1200201_3 = 14414_5$ (e) $= 111100110_2 = 200000_3 = 3421_5$ (f) $= 1001011_2 = 2210_3 = 300_5$

2. (a) $= 845$ (b) $= 1280$ (c) $= 2826$ (d) $= 203$ (e) $= 77$ (f) $= 426$ (g) $= 2891$ (h) $= 41$ (i) $= 1698$

3. (a) $= 2A$ (b) $= 714$ (c) $= 14$ (d) $= 15$ (e) $= 20A$ (f) $= 74A$

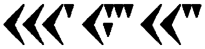
4. Each of these converts to the other. Here are the details:

(a) 

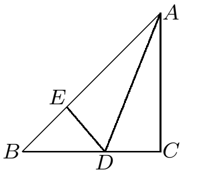
$$\begin{aligned} &= 31 \times 60^2 + 14 \times 60 + 22 \\ &= 31 \times 3600 + 14 \times 60 + 22 \\ &= 111600 + 840 + 22 \quad = \quad 112462. \end{aligned}$$

(b)

60	112462	
	1874	22
	31	14
	0	31
		↑

So, reading the remainders up from the bottom, the answer is 31 14 22, or in Babylonian numerals: .

Appendix:



Here is a rather nice "graphical" proof that $\sqrt{2}$ is irrational, also based on the idea that one cannot have an infinite descending sequence of positive integers. We start with an isosceles right-angled triangle with integer-length sides: ΔABC, as shown. Note that this is only possible if $\sqrt{2}$ is rational (exercise: show why this is true).

"Fold" the triangle so that side AC lies along side AB, creating an angle bisector AD (which meets side BC at point D), and the length $AE = AC$ on side AB. This creates another isosceles right-angled triangle ΔBED, with positive integer-length sides, which is strictly smaller than the one we started with, meaning we can construct an infinite descending sequence of such positive integer-length triangles. This is obviously an impossibility, so $\sqrt{2}$ cannot be rational in the first place.

If you doubt this, first notice that $\angle BED$ is a right angle, that $\angle EDB = \angle EAC = \angle EBD$, and so that $EB = ED = DC$. Moreover, these new lengths are all positive integers as well: for example, $EB = AB - AE = AB - AC$ is an integer, and $BD = BC - DC = BC - EB$ is an integer (since the difference of integers is an integer).

Chapter 8

Number Theory

In this chapter, we'll explore some of the basic structure of the positive natural numbers $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$, culminating with the Fundamental Theorem of Arithmetic, the result that establishes a canonical representation of numbers in terms of "prime numbers", which are the basic building blocks for the system of numbers as a whole. Throughout this chapter, "number" will (unless otherwise stated) always mean "natural number", so we are *not* going to deal with negative numbers, fractional (rational) numbers, *etc.* Essentially all the content of this chapter derives from classical Greek mathematics (*e.g.* most of it may be found in Euclid); in keeping with the spirit of that time, we shall generally ignore 0, though on occasion we'll consider how the situation may be modified to include it.

8.1 Prime Numbers

8.1.1 Division

When considering only the natural numbers, although addition and multiplication are always well defined, subtraction and division are not. Often a division (for example 10/3) is not possible (that was the reason for extending the numbers to include the rationals). Sometimes, however, a division *is* possible (for example $10/5 = 2$).

Where an exact division is possible (b/a in $\mathbb{N}$), we say that a **divides** b, and symbolize this relation as $a|b$. We say a is a **divisor** of b.

> Given two positive numbers a and b, a **divides** b (symbolized $a|b$) if and only if there is some number x such that $ax = b$. In symbols:
>
> $$a|b \leftrightarrow \exists x(ax = b)$$
>
> If $a|b$, then we say that a is a **divisor** of b, and that b is a multiple of a.

We write $a \not| b$ to mean a does not divide b.
Remark: This definition also works for 0, but is not very interesting: if 0 is allowed, then we can say $a|0$ for any a, because $a0 = 0$; moreover, $0 \not| a$ for any $a \neq 0$, since $0x$ can only equal 0. The definition also works perfectly well if we apply it to $\mathbb{Z}$, the set of all integers. In fact, if a positive number a divides a positive number b, then $\pm a \mid \pm b$, and *vice versa*, so by restricting to natural numbers, all we are missing is the $+$ and $-$ signs. So there's nothing important lost by restricting to positive natural numbers.

The definition of divisibility becomes rather trivial, however, if extended to the rationals or the reals, since there everything divides everything (apart from the forbidden division by 0).

Some properties of the | operator are easily proved.

- For any a, $a|a$. (We must find an x so that $ax = a$ in $\mathbb{N}$: just take $x = 1$.)
- $1|a$. (We must find an x so that $1x = a$ in $\mathbb{N}$: just take $x = a$.)

  Note: as a consequence of this and the previous fact, every number has at least two divisors: 1 and the number itself, with one exception (what exception?—well, the one exception is when "1 and the number itself" describes only one number, not two, namely the number 1: 1 has only one divisor).

Notice the structure of proof here: when we want to prove something about $a|b$ we translate the statement into a problem about finding a number expressing that b is a multiple of a, according to the definition of $a|b$, and then see what we can do about finding the multiplier. Here is another example, with a bit more content.

- For any a, b, c, if $a|b$ and $a|c$, then $a|(b + c)$.
  Proof: We must find an x so that $ax = b + c$, given that there is a y so that $ay = b$ and there is a z so that $az = c$, for $x, y, z \in \mathbb{N}$. (Note that $y = b/a$ and $z = c/a$.) We can start with $b + c$ and see what it equals, from this information: $b + c = ay + az = a(y + z)$, so we can take $x = y + z$. So we're done: $a|(b + c)$ does follow from $a|b$ and $a|c$.
- For any a, b, c, if $a|b$ and $a|c$, then $a|(b - c)$.
  See if you can prove this yourself, using the previous proof as a model.

8.1.2 Prime numbers

The "fundamental building blocks" of the positive natural numbers are the **prime numbers**, in the sense that every number can be represented as a product of prime numbers in exactly one way. To prove this statement will be our main goal in this chapter, though we shall take a number of detours on the way. Let's start by defining what *prime numbers* are.

> A **prime number** is any natural number that has *exactly two* distinct divisors, 1 and the number itself.
> A **composite number** is any *positive* natural number that has *more* than two distinct divisors.

So there are four types of natural numbers: 0, 1 (which we could call the *unit*), prime numbers, and composite numbers. 1 is neither prime nor composite; all natural numbers > 1 are either prime or composite. 0 is a rather special case, since although (apart from not being positive!) it seems to fit the definition of a composite number, it is really not composite in the intended sense. It is best to omit 0 from this discussion of primes and composites, and you will see that is just what we do for most of this chapter.

How many primes are there?

Let's look at the distribution of prime numbers. Given a number, it's not a hard matter (in principle!: with large numbers, this could take *a lot* of calculation!) to determine if it's prime or not. All you have to do is start trying to divide smaller numbers into it, and if you never succeed, then the number you are testing is prime. There is a minor shortcut: you need only test numbers up to the square root of the number you are testing, since every potential divisor comes paired with the other multiple making up the number, and if one is greater than the square root, the other must be less. And you only need to test prime numbers, since if any prime doesn't divide some number, neither does any multiple of the prime. For example, suppose we wanted to test 29. We try to divide $2, 3, 5$ into 29; no need to go higher, since the square root of 29 is between 5 and 6, so we can stop at 5, and no need to test 4 since we already know 2 doesn't divide into 29, so 4 cannot either. (Do it: you'll find that none of $2, 3, 5$ do divide exactly into 29, and so 29 is a prime number.) In this manner, you can test numbers, and here is what you'll find.[1] I have marked the primes in boldface, the others are in ordinary text.

1	**2**	**3**	4	**5**	6	**7**	8	9	10	**11**	12	**13**	14	15	16	**17**	18	**19**	20
21	22	**23**	24	25	26	27	28	**29**	30	**31**	32	33	34	35	36	**37**	38	39	40
41	42	**43**	44	45	46	**47**	48	49	50	51	52	**53**	54	55	56	57	58	**59**	60
61	62	63	64	65	66	**67**	68	69	70	**71**	72	**73**	74	75	76	77	78	**79**	80
81	82	**83**	84	85	86	87	88	**89**	90	91	92	93	94	95	96	**97**	98	99	100
101	102	**103**	104	105	106	**107**	108	**109**	110	111	112	**113**	114	115	116	117	118	119	120
121	122	123	124	125	126	**127**	128	129	130	**131**	132	133	134	135	136	**137**	138	**139**	140
141	142	143	144	145	146	147	148	**149**	150	**151**	152	153	154	155	156	**157**	158	159	160
161	162	**163**	164	165	166	167	168	169	170	171	172	**173**	174	175	176	177	178	**179**	180
181	182	183	184	185	186	187	188	189	190	**191**	192	**193**	194	195	196	**197**	198	**199**	200
201	202	203	204	205	206	207	208	209	210	**211**	212	213	214	215	216	217	218	219	220
221	222	**223**	224	225	226	**227**	228	**229**	230	231	232	**233**	234	235	236	237	238	**239**	240
241	242	243	244	245	246	247	248	249	250	**251**	252	253	254	255	256	**257**	258	259	260
261	262	**263**	264	265	266	267	268	**269**	270	**271**	272	273	274	275	276	**277**	278	279	280
281	282	**283**	284	285	286	287	288	289	290	291	292	**293**	294	295	296	297	298	299	300
301	302	303	304	305	306	**307**	308	309	310	**311**	312	**313**	314	315	316	**317**	318	319	320
321	322	323	324	325	326	327	328	329	330	**331**	332	333	334	335	336	**337**	338	339	340
341	342	343	344	345	346	**347**	348	**349**	350	351	352	**353**	354	355	356	357	358	**359**	360
361	362	363	364	365	366	**367**	368	369	370	371	372	**373**	374	375	376	377	378	**379**	380
381	382	383	384	385	386	387	388	**389**	390	391	392	393	394	395	396	**397**	398	399	400
401	402	403	404	405	406	407	408	**409**	410	411	412	413	414	415	416	417	418	**419**	420 . . .

There are some things to notice about this printout: one is that although the primes seem to become a bit "sparse" as the numbers get larger, they never cease. So it's natural to ask (as did the ancient Greeks) if this continues: are there only a finite number of primes, or does the list of primes continue without ever ending. The answer is that they never stop. And the proof of this fact is one of the prettiest little proofs around. We shall present it here, just as it appears in Euclid's *Elements* (well, I have translated it!). First, to set the scene, we need a few remarks, including an important result, the Prime Factorization Theorem:

Lemma: For any prime p and any number n, if $p|n$, then $p \not| \, (n+1)$. (If p divides n, then it does not divide $n+1$.)

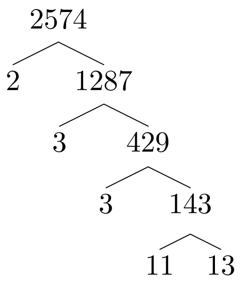
Proof: By our previous remarks about divisibility, if $p|(n+1)$ and $p|n$, then $p|(n+1-n)$, *i.e.* $p|1$. But no prime can divide 1 (only 1 divides 1). (QED)

Theorem: (**Prime factorization**) Any composite number may be represented as a product of primes.

[1] Just for fun, here are the rest of the primes up to 1000: 421, 431, 433, 439, 443, 449, 457, 461, 463, 467, 479, 487, 491, 499, 503, 509, 521, 523, 541, 547, 557, 563, 569, 571, 577, 587, 593, 599, 601, 607, 613, 617, 619, 631, 641, 643, 647, 653, 659, 661, 673, 677, 683, 691, 701, 709, 719, 727, 733, 739, 743, 751, 757, 761, 769, 773, 787, 797, 809, 811, 821, 823, 827, 829, 839, 853, 857, 859, 863, 877, 881, 883, 887, 907, 911, 919, 929, 937, 941, 947, 953, 967, 971, 977, 983, 991, 997.

This is an important property of prime numbers in its own right, deserving of its own name. The proof also establishes an algorithm for producing the prime factorization of any number.

Start with the number n you wish to factor. Try to divide 2 into n: if you succeed, then replace n in this algorithm by the quotient $n/2$ (that will be a natural number), and start over (try to divide 2 into this new number); if you fail, then try to divide the next prime (3) into your number. Continue in this way either (a) until you find a prime that divides exactly into your number, in which case replace your number with the quotient obtained by dividing it by the prime divisor you just found and continue testing with that prime, or (b) until you reach its square root, at which point you have established your number (or the appropriate quotient thereof) is a prime. All the primes which did divide into your number, or into the reduced numbers, make up the prime factorization (notice that some primes may appear multiple times). (QED)



An illustration might make this clearer: we shall find the prime factorization of 2574. The procedure is often written as a tree, as illustrated. Trying to divide 2 into 2574 succeeds, and gives the quotient 1287. Starting over with 1287, we try dividing 2 into it (that fails), then 3 into it: that succeeds, giving a quotient 429. Starting over (we can skip 2 as that failed before), we try dividing 3 (again) into 429: that succeeds again, giving a quotient 143. Again we try 3 into this: it fails. We try successive primes (5, then 7, then 11): 5 and 7 fail, but 11 succeeds, giving a quotient of 13. That is a prime (or, if you prefer, try dividing 11 into 13, which fails, then try the next prime, 13, into 13, which succeeds, giving a quotient 1, and no prime divides into 1). List the primes which did divide into our number and into its reductions: $2, 3, 3, 11, 13$. Their product is 2574 (check!), so the prime factorization is $2 \times 3^2 \times 11 \times 13$.

This proof is actually quite simple, if you focus on the *idea* behind it: given any number > 1, it is either composite or prime. If composite, it has factors (which are smaller than it is). Each of those factors in turn is either composite or prime, and if composite, has factors smaller in turn. Just decompose any composites to get smaller factors; eventually you must terminate with primes (since you cannot count down from any natural number without eventually stopping—there are only a finite number of numbers smaller than any specific number).

Theorem: (**The infinitude of primes**) Given any finite set of primes, there must be some prime number not in that set. In other words, the set of prime numbers cannot be finite, and so is infinite.

Proof: Take any (finite) set of primes (imagine it is of the form $\{p, q, r\}$), and form their product: pqr. Add 1 to that product: $pqr + 1$. Form the prime decomposition of the resulting number. None of the primes in your original set can occur in that prime decomposition, since each of them divides their product (*e.g.* $p|pqr$), and hence cannot divide $pqr + 1$. So the prime decomposition of your resulting number $pqr + 1$ must be made up of primes **not** in your original set. It might be prime itself, or it might be a product of primes, but such primes are not in your set, so either way, you have found primes not in the original set. So that set did not include all the primes. (QED)

Remark: It is worth remarking that this proof is actually constructive: given any finite set of numbers, it gives an algorithm (*via* the algorithm used to prove Prime Factorization) to get primes which do not divide any of the numbers in your set: just multiply your numbers together, add one, and then look at the prime factorization of the result—it must contain primes which don't divide any of the numbers you started with. Some authors give a proof by contradiction of this theorem, missing part of the point of Euclid's proof. I have also kept one rather amusing stylistic device of Euclid's proof, using 3 as a standard representative of "any finite number", by taking the form of the finite collection of primes the theorem talks about to be $\{p, q, r\}$. Of course, any finite collection may be treated in the same manner.

Look again at the table showing the primes distributed among the natural numbers up to 420. The primes seem to be scattered almost randomly throughout the list, though some hints of patterns tantalize us (look at sequences like 7, 23, 37, 53, 67, 83, 97, 113, 127), but patterns are less obvious (or simply do not exist) as we get into higher and higher numbers. Is there a pattern in the sizes of the gaps[2] between successive primes? There is only one pair of primes (2 and 3) that have a gap of 1 between them, because 2 is the only even prime. Some pairs of primes have a gap of only 2, even near the end of our list (347 and 349). Pairs of primes that differ by 2 are called "twin primes". Will there continue to be twin primes even among the very large numbers? Here is a guess at the answer:

Conjecture: There are infinitely many twin primes.

This conjecture was proposed in 1923 by Hardy and Littlewood,[3] but has never been proved or disproved. We don't know. It's an open question. So far …

Look again at the distribution of primes. There are 4 primes among the first 10 numbers (40% of the first 10 numbers are primes). There are 8 among the first 20 numbers (still 40%). There are only 10 primes less than 30 (33%); 12 primes less than 40 (30%); 24 primes less than 100 (24%); 44 primes less than 200 (22%); and so on. Among the first 100 numbers, 24% are prime; 20% of the next 100; 16% of the next 100; 15% of the next 100. The primes thin out as they get larger. Is there a pattern here?

The answer is "yes", and it is one of the major results of number theory: the Prime Number Theorem, which states that the proportion of primes less than n is approximately $1/\ln(n)$ (the reciprocal of the natural logarithm of n). So, roughly $16\frac{1}{2}\%$ of numbers up to 1000 are primes, and the reciprocal of $\ln(1000)$ is roughly $14\frac{1}{2}\%$, which isn't too far off, and this approximation will improve for larger numbers. More precisely, if we denote the number of primes less than or equal to n by $\pi(n)$, the Prime Number Theorem says that the following limit equals 1:

$$\lim_{n\longrightarrow\infty} \frac{\pi(n)/n}{1/\ln(n)} = 1$$

meaning that the value of this fraction is close to 1 for large numbers n, and gets even closer to 1 as n gets larger ("goes to infinity"). This result is at the start of a remarkable journey through a significant part of mathematics—you can find more by looking up books or articles on the Riemann Hypothesis, a famous open problem whose solution is actually worth one million dollars to the person who finally "cracks" it.

We saw that the primes "thin out" as we look at larger numbers. The biggest gap between successive primes in our sample is 14 (the first such gap is from 113 to 127). Is there a limit to how large the gap can be between successive primes? No:

Proposition: There is no limit to the size of the gap between successive primes.

[2] I shall define the *gap* between numbers m, n to be the difference between them: $m - n$ if m is the larger number. So the gap between 5 and 6 is 1, and the gap between 10 and 15 is 5, for example. The number of numbers between m and n is 1 less than the gap between them: there is no number between 5 and 6, and there are 4 numbers between 10 and 15.

[3] Godfrey H. Hardy was a famous British mathematician and cricket fanatic: near the end of his life he wrote a remarkably personal and touching memoir *A Mathematician's Apology*, which I recommend. His longtime collaborator J.E. Littlewood also produced a memoir *A Mathematician's Miscellany* in the form of a collection of academic anecdotes, not as personal, but often amusing.

Proof: We shall display an algorithm which constructs a sequence of as many consecutive composite numbers as we like. The construction involves the concept of the *factorial* function,[4] written $n!$. $n!$ is defined as follows: $0! = 1$, $1! = 1$, $2! = 1 \cdot 2 = 2$, $3! = 1 \cdot 2 \cdot 3 = 6$, and so on, so that $n! = 1 \cdot 2 \cdot 3 \cdots n$. To find the factorial of any natural number > 0, multiply the number by every positive natural number less than or equal to itself. $27! = 1 \cdot 2 \cdot 3 \cdot 4 \cdots 27 = 10888869450418352000000000000$, which is a very large number indeed.

Before we define our algorithm, let's consider an example. Suppose we want to find a string of four consecutive composite numbers. Then we start with 4, and add 1 (to get 5), then form $5! = 120$: the four guaranteed consecutive numbers start with this plus 2, so are $122, 123, 124, 125$, for the following simple reason. By construction, 120 is divisible by $2, 3, 4, 5$ (since we multiplied those to get 120). Therefore $120 + 2$ is divisible by 2 (both 120 and 2 are), $120 + 3$ is divisible by 3 (both 120 and 3 are), $120 + 4$ is divisible by 4 (both 120 and 4 are), and $120 + 5$ is divisible by 5 (both 120 and 5 are).

Think about this a moment: you should convince yourself that a similar trick will always work.[5] If you want n consecutive composite numbers, form $(n + 1)!$, add two (to get $(n + 1)! + 2$), and start there. That number will be the first in a string of n consecutive composites.

Another example: if you wanted 13 consecutive composites (making a gap of at least 14 between primes), then form $14! + 2$ and start there. Now, $14! = 1 \cdot 2 \cdot 3 \cdots 14 = 87178291200$, so we start at 87178291202, and then the numbers from 87178291202 to 87178291214 will all be composite.

So, we have an algorithm which works for any number, so we can construct a gap of any size, without limit, between successive primes. (QED)

Confession: this algorithm doesn't exactly give the "optimal" answer (there may well be a string of 13 consecutive composites using much smaller numbers!); what it does do is give us a guaranteed answer, where no further searching is necessary. For instance, if you wanted 5 consecutive composites, the algorithm would produce ($6! + 2 = 722$) $722, 723, 724, 725, 726$, but from the table we can see that $24, 25, 26, 27, 28$ would also do. But that list is harder to produce without an actual search.

Some folks have asked "why can't we start with $(n+1)!+1$?"; well, the answer is simple: because sometimes it doesn't work, because sometimes that is a prime number (for example $3!+1 = 7$ which is prime). Sometimes it does work; for example, a famous result (often called "Wilson's Theorem") says that for any prime p, p is a divisor of $(p - 1)! + 1$ (so that $(p - 1)! + 1$ is always composite, for any odd prime number p). For example, $6! + 1$ is divisible by 7, and $10! + 1$ is divisible by 11 (check for yourself).

If you try calculating $n!$ for bigger and bigger values of n, you'll see that $n!$ gets very large very fast. You might be curious about how big a number 1000! is. It has 2566 digits, and would take about a page to print. Most hand calculators cannot calculate numbers this large. The approximate value of 1000! (in scientific notation) is 4.02387×10^{2565}. According to Carl Sagan (in his book *Cosmos*) the total number of elementary particles—protons and neutrons and electrons—in the observable universe is about 10^{80}. If the universe were packed solid with neutrons, say, so there were no empty space anywhere, there would still be only about 10^{128} particles in it. 1000! is a number that is about 10^{2485} times as big as the total number of elementary particles in the universe, and about 10^{2440} times as big as the number of neutrons that would fill the universe. It is a large number.

We can calculate it and add 2 to that huge number. The result, and the next 998 numbers after

[4] Usually pronounced "n factorial", though some people like to say "n bang" or "n shriek".

[5] Think about it *another* moment: can you see a more efficient way, which would involve multiplying fewer numbers? I'll mention such an improvement a bit later.

it, will all be composite. There will be a gap of at least 1000 between successive primes.

Of course, we could use a smaller number than 1000! to get a gap of at least 1000 between successive primes. One rather more intelligent (well, more efficient, at least) algorithm would be to multiply together all the primes less than 1000 (rather than all the numbers), and add 2 to the result: that and the next 998 numbers will all be composite. We don't need to check all numbers, merely the primes, to verify that a number is composite. If we used this "simpler" method to get 4 consecutive composite numbers (as we did in our first example), instead of multiplying $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$, we'd just multiply $2 \cdot 3 \cdot 5$ (the primes in that list), to get 30, and then add 2 to get 32, and our claim would be that $32, 33, 34, 35$ are all composite. Somewhat more "efficient" than our previous answer $(122, 123, 124, 125)$.

A remarkable result (originally conjectured by Joseph Bertrand in 1845 and proved a few years later by Pafnuty Chebyshev) that indicates we are never too far from a prime is that for any n, there is always a prime number between n and $2n$.

There are many facts known about prime numbers, and primes are entangled with some of the most challenging areas of mathematics. There are also many conjectured facts which have resisted proof, facts that we're pretty sure must be true (and for which we've run computer calculations to verify as many instances as possible), but for which we (meaning the mathematical community) have not yet found any *proof*. We saw one such conjecture above, that there are infinitely many twin primes. Here is a very famous conjecture going back centuries:

Conjecture: Every even number > 2 is the sum of two primes.

For example, $4 = 2 + 2, 6 = 3 + 3, 8 = 3 + 5, 10 = 3 + 7, 12 = 5 + 7, 14 = 3 + 11 = 7 + 7, 16 = 5 + 11, \ldots$. This is known as the Goldbach Conjecture (there is even a novel about this, *Uncle Petros and Goldbach's Conjecture* by Apostolos Doxiadis, published in 2000—the publisher offered two million dollars if anyone could solve this within two years of the publication of the novel (by 2002): the money was pretty safe, though, as no one has managed to do so since the late eighteenth century, when Goldbach suggested the conjecture, and sure enough, no one claimed the cash!). A lot is actually known about this: for instance, computer checks have shown this is true for a lot of "small" values (at least up to 6×10^{17}, for instance), and it has been proven that every even number > 2 can be written as a sum of at most 6 primes, and that there is some (large) number N for which it's known that every even number $> N$ is the sum of two numbers, one of which is prime, and the other is the product of at most two primes. Close, but still no cherry! Goldbach's Conjecture is probably true, if these results are anything to go by, but we still have no *proof* of that claim. And it's deductive proof that counts in mathematics.

8.1.3 Perfect numbers and Mersenne primes

There is a famous class of prime numbers, called the Mersenne primes after Marin Mersenne (1588-1648) who first pointed out a number of primes of the form $2^n - 1$. In fact, he claimed that for the values of $n = 2, 3, 5, 7, 13, 17, 19, 31, 67, 127$ and 257, $2^n - 1$ would be prime, and not for any other values of $n < 257$. He wasn't entirely correct, although it took several centuries to settle this completely. It was eventually shown that for $n = 67$ and 257 you don't get primes, and for $n = 61, 89, 107$ you do. Many more Mersenne primes have been discovered since.[6]

There are two significant facts about Mersenne primes. The first connects Mersenne primes to primes in general:

[6] These days, newly discovered primes tend to be Mersenne primes, and it's possible to join in the hunt for new primes by running a background computer task—check this at `http://www.mersenne.org/prime.htm` if you're curious.

If $2^n - 1$ is a prime number then n is also a prime number.

It's important to note that the converse isn't true: there are many prime numbers n for which $2^n - 1$ is not prime (we gave the examples of $n = 67$ and 257 above).

The second fact relates Mersenne primes to what are called *perfect* numbers: a number is perfect if it equals the sum of its proper divisors (divisors less than itself). For instance, 6 is perfect, since $6 = 1 + 2 + 3$ (and $1, 2, 3$ are the proper divisors of 6). 28 is the next perfect number: $28 = 1 + 2 + 4 + 7 + 14$. The connection with Mersenne numbers is this:

If $2^n - 1$ is a prime number then $2^{n-1}(2^n - 1)$ is a perfect number.

For instance, if $n = 3$, notice that $2^3 - 1 = 7$ is (a Mersenne) prime, and $2^{n-1}(2^n - 1) = 2^2(2^3 - 1) = 4 \cdot 7 = 28$ is perfect. (Check that $n = 2$ gives the perfect number 6.)

Here is a table of the first 8 perfect numbers (and the corresponding Mersenne primes) generated by this principle.

n	$2^n - 1$	$2^{n-1}(2^n - 1)$
2	3	6
3	7	28
5	31	496
7	127	8128
13	8191	33 550 336
17	13 1071	8 589 869 056
19	524 287	137 438 691 328
31	2 147 483 647	2 305 843 008 139 952 128

(These numbers get very large indeed!) There are several questions about perfect numbers one might ask. For example, the ones generated by Mersenne primes are all even; are there any even ones which don't come from Mersenne primes by this formula? The answer is "no": it's been proven that this formula generates all the even perfect numbers. What about odd perfect numbers? Are there any? So far none have ever been found, and if there are any, they must be very large (bigger than about 10^{50}, since it's been shown that there are no odd perfect numbers less than that—I wouldn't bet any money on the existence of odd perfect numbers!).

8.1.4 Exercises on divisibility and primes

Prove the following facts about positive natural numbers.

1. Suppose $a = bq + r$: prove that if $d|b$ and $d|r$ then $d|a$. Also prove that if $d|a$ and $d|b$ then $d|r$.

2. Find all the divisors of $2^3 \cdot 5^4$.

3. The divisors of 1 more than a product of primes are always "new" primes: what "new" primes do you get from $2 \cdot 11 \cdot 13 + 1$?

4. Find[7] the prime divisors of $2 \cdot 3 \cdot 5 \cdot 7 + 1$.

5. Find the prime factors of $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 + 1$ ($= 7! + 1$, not to be confused with $8!$). Is $7! + 1$ prime? Check that $7!, 7! + 2, 7! + 3, 7! + 4, 7! + 5, 7! + 6, 7! + 7$ are all composite.

[7] Remember that in looking for divisors of a number n, you need look no further than $\sqrt{n}$, since any larger factors must be turned up by their "smaller mates".

6. Construct a number that has the property that the 7 numbers that follow it are all composite.

7. If $d|a$ and $d|b$, prove that $d|a+b$ and $d^2|ab$.

8. If a prime p divides $a+b$, must it divide a (or b)? (This is the converse of the first part of the previous exercise: is it true?)

9. Some primes can be written in the form n^2+1 for some natural number n. For example, $2^2+1=5$, and $4^2+1=17$. Find three more primes of the form n^2+1. (The conjecture that there are infinitely many such primes has never been proved or disproved.)

10. Some primes are one less than a square: they can be written in the form n^2-1 for some natural number n. For example, $2^2-1=3$. Can you find other primes of the form n^2-1? Make a conjecture and prove it.

8.2 Mathematical Induction

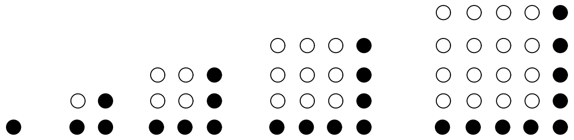
8.2.1 Playing with numbers

Mathematics is not just calculation or computation. Mathematicians enjoy mathematics as one might enjoy performing difficult music or playing a sophisticated game—for the beauty and the pleasure of the pursuit itself. Most mathematicians dislike calculation and computation just as much as any high school student. One reason for doing mathematics is to find ways *not* to do calculations—this was actually one of my main motivations, in fact.

In the development of a science, one: (1) gathers data and looks for patterns in the data; (2) creates classifications and definitions; (3) proposes general (abstract) conjectures and hypotheses; (4) confirms or proves the conjectures: not all that different from what one does in mathematics.

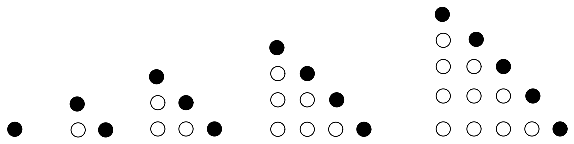
Ancient Greeks seem to have been the first people to study numbers, to explore their properties and discover new patterns, and most importantly, to seek *proofs* of their conjectures. We start this section with some simple examples of the sorts of number patterns they enjoyed.

For example, the Greeks discovered that they could arrange sets of pebbles in different kinds of patterns depending on the size of the set. Some sets made squares, as:



Whether a set of pebbles can be arranged into a square pattern depends on the cardinality of the set of pebbles: the number of pebbles in the set. The cardinalities of sets of pebbles that could make squares they called the "square numbers" (from whence comes our usage "5 squared").

Other sets could be arranged into different patterns. Triangles, for example:



and the numbers associated with these sets they called "triangular numbers".

Finding patterns in the pebbles is abstraction. Extending the process of abstraction leads to the idea of patterns in the numbers themselves. Playing with the pebbles ushers in mathematical questions. Is there a biggest square number? A biggest triangular number? A biggest number? Is there a number bigger than 1 that is both square and triangular? Is there a square number between 16 and 25?

One can arrange square or triangular numbers in a sequence from small to large (as in the diagrams above) and think about the relation between successive numbers. How many pebbles do we have to add to a square number to get the next square number? How many are added to a triangular number to get the next triangular number?[8] What is the sequence of numbers of added pebbles?

Before we go further, you might like to try your hand at this. Imagine it's a pleasant evening, you're sitting in an Athenian café, playing with pebbles on the sand at your feet, a glass of retsina beside you. See what patterns you can find.

- Write the number of pebbles in each square number in the diagrams above, in sequence. Do the same for the triangular numbers.
- Describe (in words) the sequences you wrote down, and try to predict the next few entries.
- Formulate a conjecture or hypothesis about the relation between successive square numbers. What does one add to a square number to get the next square number? Can you create a *formula* which will tell you what the n^{th} square number is? Test your formula on the 5^{th} square number.
- Formulate a conjecture or hypothesis about the relation between successive triangular numbers. What does one add to a triangular number to get the next triangular number? Can you create a *formula* which will tell you what the n^{th} triangular number is? Test your formula on the 5^{th} triangular number.
- Are you sure that your conjectures will hold as the numbers get bigger? Can you *prove* that your conjectures are true for all square and triangular numbers? How? (This is the question to which we shall turn now.)
- Square patterns could be relaxed a bit to form "rectangular numbers"; convince yourself that these are just what we've been calling "composite numbers" (apart from 1, which isn't much of a rectangle!).

8.2.2 Mathematical induction

The natural numbers have many properties, but the one which is most characteristic, the one that virtually *characterizes* the natural numbers, is called **mathematical induction**.[9] This should not be confused with *scientific induction*, also known as *empirical induction*, and I shall say more about that soon, but first, just what is mathematical induction?

Mathematical induction is based on the observation that if you wish to reach a natural number, by counting, you only need one operation ("+1") and a starting place (0). In other words, if you start at 0 and repeatedly add 1, then eventually you will reach whatever natural number you wish,

[8] I have given you a big hint in the shading of pebbles!

[9] Some writers—not too many!—like to call this "perfect induction", contrasting it with the "imperfect" scientific induction. The distinction between these two types of induction is indeed important, but I think the name "perfect induction" is redundant: after all, mathematics is by its nature obviously perfect(!).

be it 5 or 10^{50} (although it may take a little time). This may be used as a principle of *proof.* Suppose $P(x)$ is some statement about numbers x, and suppose you want to know that $P(n)$ is true for all natural numbers n: it would suffice to show your statement P is true for the number 0, and that *whenever* P is true for some number, it is *also* true for the next number (obtained by adding 1). For then, whatever number you choose, you can verify your statement at your chosen value by verifying it for all the numbers up to and including your number, starting with 0.

For example, suppose you had proved the premises of mathematical induction for a particular statement P, and now want to conclude $P(n)$ for some particular n, as mathematical induction concludes you may; suppose you wanted to verify your statement was true for $n = 5$, say. You could start by saying "$P(0)$ is true" ✓. Add 1: Since $P(n) \longrightarrow P(n+1)$, this guarantees $P(0+1)$, so $P(1)$ is true ✓. Again, add 1: a similar use of $P(n) \longrightarrow P(n+1)$ tells you $P(2)$ is true ✓. Continue in this way, and in 3 more steps you'll reach $P(5)$, at which point you know your statement is true indeed for $n = 5$, as you wanted. What made this work was (a) you had a starting point where you knew the statement was true, and (b) you had a way to verify the statement for a particular number, knowing it true for the previous number. This is mathematical induction. (It's really just counting!)

> The **principle of mathematical induction** is:
> If a statement about natural numbers can be proved to be true for the number 0, and
> if it can be proved that, whenever the statement is true for an arbitrarily-chosen natural number n it must be true for $n + 1$,
> then it must be true for all $n \in \mathbb{N}$.

We can state this principle symbolically as follows. Suppose $P(x)$ is a statement (predicate) (about numbers x). Then mathematical induction is this (where the universe of discourse is $\mathbb{N}$):

$$P(0), \forall n[P(n) \longrightarrow P(n+1)] \ \vdash \ \forall n P(n)$$

A note concerning terminology: In proving something by mathematical induction, we have two things to prove: the case where $n = 0$, and the conditional statement that the case for n implies the case for $n+1$. We often refer to the $n = 0$ case as the *base case*, and the conditional statement is often called the *induction step.*

The term "mathematical induction" might seem to be confusing, being so similar to the term "scientific induction", and indeed, in every-day English, "induction" usually means "scientific induction". The difference between these two types of induction is striking: scientific induction only gives a plausibility argument. The fact that the sun has risen every morning within human memory suggests that it is *very* likely that it will rise tomorrow, but that isn't guaranteed fact, with the same certainty that mathematics strives for. After all, maybe tonight the Romulans will launch a murderous attack on our solar system, destroying the sun and all the planets, so that tomorrow the sun won't rise. To be sure, there's no point in worrying about this, but how could one say this is impossible? Merely the observation that something "always" seems to happen cannot be taken as *proof* that it will in fact always happen. Such likelihood is the nature of scientific "truth", but it fails the test of rigour required by mathematical proof.

Mathematical induction, on the other hand, is a rigorous method of proof, leaving nothing to doubt or chance. It depends on a defining characteristic property of the natural numbers (that one can count to any number one wishes, and that there is no number that cannot be reached by such counting). Although it *seems* to argue from the particular (a proof about the particular number 0 and about an arbitrary other particular number n and its successor) to the general (a statement that applies to all natural numbers), in fact that is only in appearance. The point is that the arbitrary number n isn't really a "particular" number, but a variable, which can represent *any* number, so that one deductive step actually represents an infinity of particular steps, enough to justify the conclusion for any particular number, and so enough to justify the conclusion for all numbers. Unlike scientific induction, the reasoning is *deductive*. That is, if the premises (the statement about 0 and the conditional statement about n and $n+1$) are *true*, the conclusion *cannot* be false. The truth of the conclusion is *guaranteed* by the truth of the premises.

Before we look at examples, there is one comment we might make. Mathematical induction requires a starting point, which we've taken to be 0. But in fact, if you think about it a bit, we could use any other starting point (*e.g.* 1 or 2, or whatever number you like), and then induction would give you the truth of your statement P, *starting at the number you chose to start with.* We'll often do that, starting at some other number than 0, using a different base case, such as 1; we'll appropriately adjust our conclusion to claim the truth of our proposition only starting at the value we used.

An example of mathematical induction

Let's look at an example. You probably noticed that the triangular numbers are the sums of the numbers in the sequence $1, 2, 3, 4, \ldots$. That is, the first triangular number is 1, which is the "sum" of the first 1 term(s). The second triangular number (3) is the sum of the first 2 terms $(1 + 2)$. The third (6) is the sum of the first 3 terms $(1 + 2 + 3)$, and so on. The n^{th} triangular number is the sum of the first n natural numbers.[10] Now, the question is, what do these numbers add up to? What formula can we derive for the sum, and so what formula can we derive for the n^{th} triangular number?

There are two ways you could answer this. You could guess the result (you may have already done so), or you could try to find a clever way to figure it out. We shall do both, but in the first case, we'll *prove* our guess is correct by induction.[11]

Let's use some notation here. We shall write $T(n)$ for the n^{th} triangular number, so, by looking at the pebble pictures, we have $T(1) = 1$, $T(2) = 3$, $T(3) = 6$, $T(4) = 10$, $T(5) = 15$, *etc.* One thing that is obvious from the pebble pictures is that $T(n + 1) = T(n) + (n + 1)$; in words, to get "the next" triangular number $(T(n + 1))$, you add "the next" number $(n + 1)$ to "the current" triangular number $(T(n))$.

So we shall take this as our *definition* of the triangular numbers:

$$T(1) = 1 \text{ and } T(n + 1) = T(n) + (n + 1)$$

Notice that $1 \cdot 2 = 2$, $2 \cdot 3 = 6$, $3 \cdot 4 = 12$, $4 \cdot 5 = 20$, $5 \cdot 6 = 30$, and half these products gives $1, 3, 6, 10, 15$, exactly the triangular numbers. This suggests (try a few other cases if you're not convinced!) that the n^{th} triangular number is given by $T(n) = \frac{1}{2}n(n + 1)$. (What an odd formula!—we'll see a way to generate this naturally soon.)

But now that we have a conjectured formula, can we prove it? Yes, with mathematical induction. We take as our statement $P(n)$ the claim "$T(n) = \frac{1}{2}n(n + 1)$". We start with the case $n = 1$: the statement, $P(1)$, we must prove is $T(1) = \frac{1}{2} \cdot 1 \cdot 2 = 1$, which is (by definition) true.

[10] Here is a "philosophical" remark (which is actually mathematically correct too). If you "add" no numbers, your sum should be the number which does nothing when you add it to other numbers: namely, the sum of 0 numbers is 0. (Check if you understand the point: the *product* of 0 numbers should be 1, the number which does nothing when you *multiply* by it.) So, we *define* the empty sum to be 0 (and the empty product to be 1); this would suggest that the 0^{th} triangular number is 0. Which it is!

[11] Unless otherwise stated, "induction" will always refer to "mathematical induction" from now on.

Comment: if you prefer, you could start at 0—that means you'd have to verify $T(0) = \frac{1}{2} \cdot 0 \cdot 1 = 0$, which is true.

Next we do the "induction step", meaning we assume the statement is true for n and use that to prove it for $n+1$. So, *assume* that $T(n) = \frac{1}{2}n(n+1)$. Now *prove* $T(n+1) = \frac{1}{2}(n+1)(n+2)$; this is what the statement $P(n+1)$ becomes when you translate it (meaning, you replace n with $n+1$ wherever you see it—in this process, the expression $n+1$ itself becomes $(n+1)+1$ or just $n+2$). So here goes:

$$\begin{aligned}
T(n+1) &= T(n) + (n+1) \text{ (the basic property of the triangular numbers)} \\
&= \tfrac{1}{2}n(n+1) + (n+1) \text{ (the induction assumption)} \\
&= (\tfrac{1}{2}n+1)(n+1) \text{ (by distributivity)} \\
&= \tfrac{1}{2}(n+2)(n+1) \text{ (by distributivity)} \\
&= \tfrac{1}{2}(n+1)(n+2) \text{ (by commutativity)}
\end{aligned}$$

The only tricky step was where we took that $\frac{1}{2}$ outside the bracket—the hint that this might be a good idea came from the form of the result we wanted: $\frac{1}{2}(n+1)(n+2)$. That suggested the fraction $\frac{1}{2}$ ought to be *outside* the bracket, not *inside*. So we took it outside, and made whatever adjustment needed inside. The nice thing about induction is you know what you want to prove, so you can use that expectation to give you hints about what algebraic steps might be helpful: in this case, the use of distributivity to first get the factor $(n+1)$, and second, the factor $\frac{1}{2}$.

So what's our conclusion?: That the statement (formula) $P(n)$ is true for all $n \geq 1$ (for all n if we chose to start at $n = 0$); in other words, $T(n) = \frac{1}{2}n(n+1)$.

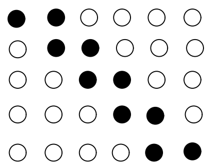
By the way: we could have set this problem up another way: it is clear (is it?) from the way the triangular numbers are defined that

$$T(n) = 1 + 2 + 3 + \cdots + (n-2) + (n-1) + n$$

and so we have just shown that $1 + 2 + 3 + \cdots + (n-2) + (n-1) + n = \frac{1}{2}n(n+1)$.

Remark: Mathematical induction provides a very useful way to prove a result, but it has one serious drawback: it gives no hints as to how to get the result in the first place. If we hadn't "guessed" the formula for $T(n) = \frac{1}{2}n(n+1)$, induction wouldn't have been of much use. So usually induction is used when some *other* method suggests the result wanted, but does so without a rigorous proof.

In the case of the present example, there is another method we could have used to get the formula, using a simple diagrammatic trick, in fact. Take the triangular figure from which the triangular numbers get their name, and double the triangle.



As you can see, this now forms a rectangle, whose sides are n by $n+1$. Since this is *twice* the appropriate triangle, we can see that the original triangle has $\frac{1}{2}n(n+1)$ dots.

This pictorial proof can be presented algebraically as well. We start with

$$T(n) = 1 + 2 + 3 + \cdots + (n-2) + (n-1) + n$$

and then reverse this expression, giving

$$T(n) = n + (n-1) + (n-2) + \cdots + 3 + 2 + 1$$

Add these "vertically":

$$\begin{aligned} T(n) + T(n) &= (1+n) + (2+n-1) + (3+n-2) + \cdots + (n-2+3) + (n-1+2) + (n+1) \\ 2T(n) &= (n+1) + \quad (n+1) \quad + \quad (n+1) \quad + \cdots + \quad (n+1) \quad + \quad (n+1) \quad + (n+1) \\ 2T(n) &= \ n \cdot (n{+}1) \end{aligned}$$

and then dividing by 2 gives the desired formula $T(n) = \frac{1}{2}n(n+1)$.

Another example of induction

Look at the square numbers again: each is obtained from the previous by adding a new *odd* number, so that the sequence may be thought of as 1, $4 = 1+3$, $9 = 1+3+5$, $16 = 1+3+5+7$, $25 = 1+3+5+7+9$, *etc.* (This should be evident from the black dots in the diagram we had before.) So if we denote the n^{th} square number by $S(n)$, then we have the formula $S(n) = 1+3+5+7+\cdots+(2n-1)$. Check this: if $n=1$ then $2n-1=1$, so we add odd numbers up to 1, *i.e.* $S(1)=1$. If $n=2$ then $2n-1=3$, and so we add odd numbers up to 3, *i.e.* $S(2) = 1+3$. Verify that we've got $S(3)$ and $S(4)$ right. The definition of $S(n)$ has as an immediate consequence that $S(n+1) = S(n) + (2n+1)$: the "next" square number $S(n+1)$ is equal to the current one $S(n)$ *plus* the "next" odd number $2n+1$. (Do you see why $2n+1$ is the next odd number after $2n-1$?)

The picture of square numbers then suggests that $S(n) = n^2$, and this is what we shall prove by induction. We shall start with $n=1$, and so we must "prove" $S(1) = 1^2 = 1$; but this is immediate.[12]

For the induction step, we *assume* the n case: that $S(n) = n^2$, and we aim to *prove* the $n+1$ case from that assumption: that $S(n+1) = (n+1)^2$. Here's how this might go:

$$\begin{aligned} S(n+1) &= S(n) + (2n+1) \text{ (the basic property of the square numbers)} \\ &= n^2 + (2n+1) \text{ (the induction assumption)} \\ &= (n+1)^2 \text{ (by factoring the expression)} \end{aligned}$$

So we have proved the base case, and the induction step, and hence we have proved the formula $S(n) = n^2$ for all $n \geq 1$.[13]

Yet another example

Here is an example, without all the chit-chat, so you can see the structure of the method more clearly. (It's really just the triangular numbers example again, but doubled; can you see that?)

Show that $2+4+6+\cdots+2n = n(n+1)$ by mathematical induction.

First verify case $n=1$: $2 = 1(1+1)$ is clearly true.

Next: Assume case n:

$$2+4+6+\cdots+2n = n(n+1)$$

[12] You could start with $n=0$, using the same convention as we mentioned with the triangular numbers: if $n=0$, $S(0)=0$, so the base case is ok.

[13] Or for all n, if you are happy with the base case being $n=0$.

and **prove** case $n+1$:

$$2+4+6+\cdots+2(n+1) \overset{?}{=} (n+1)(n+2)$$

This follows from the following calculation:

$$\begin{aligned} &2+4+6+\cdots+2(n+1) \\ &\quad = (2+4+6+\cdots+2n)+2(n+1) && \textit{(notice the previous case is part of the current one)} \\ &\quad = n(n+1)+2(n+1) && \textit{(use the assumed formula)} \\ &\quad = (n+1)(n+2) && \textit{(take out the common factor)} \end{aligned}$$

(QED)

Sigma notation

Although it is not essential for our use, you might like to know that there is a nice notation for sums of this sort. If a_k is an expression which contains a variable k, representing a natural number, such as $a_k = 2k-1$, then we denote a sum of such expressions as follows:

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + a_3 + \cdots + a_n$$

For example, if $a_k = 2k-1$, then notice that if $k=1$, $a_k = a_1 = 2\cdot 1 - 1 = 1$, and similarly (if $k=2$) $a_2 = 3$, (if $k=3$) $a_3 = 5$, and so on. So the sum of the first n odd numbers would be written as

$$\sum_{k=1}^{n}(2k-1) = 1+3+5+\cdots+(2n-1)$$

Similarly, the sum of the first n positive integers would be written as

$$\sum_{k=1}^{n} k = 1+2+3+\cdots+n$$

With this notation, our results about the triangular and square numbers could be written as

$$\sum_{k=1}^{n} k = \tfrac{1}{2}n(n+1) \quad \text{and} \quad \sum_{k=1}^{n}(2k-1) = n^2$$

8.2.3 Exercises on induction

Use Mathematical Induction to prove the following facts about positive natural numbers.[14]

1. Prove that for all n: $1+4+7+\cdots+(3n-2) = \frac{1}{2}n(3n-1)$.
2. Prove that for all n: $3+7+11+\cdots+(4n-1) = n(2n+1)$.
3. Prove that for all n: $1\cdot 2 + 2\cdot 3 + 3\cdot 4 + \cdots + n(n+1) = \frac{1}{3}n(n+1)(n+2)$.
4. Prove that for all n: $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \dfrac{1}{3\cdot 4} + \cdots + \dfrac{1}{n(n+1)} = \dfrac{n}{n+1}$.

[14] For simplicity's sake, you may assume all these are for $n \geq 1$, unless otherwise stated; in many of these, one could start with $n = 0$; I'll let you explore that possibility.

5. Prove that for all n: $1+2+2^2+2^3+\cdots+2^n = 2^{n+1}-1$

6. Prove that for all n, if a set X has n elements, then $\mathcal{P}(X)$ has 2^n elements.

7. Prove that for all n: 2 divides 3^n-1.

8. Prove that for all n: 5 divides 8^n-3^n.

9. Prove that for all n: 4 divides 6^n-2^n.

10. Prove that for all n: 7 divides 15^n-8^n. (Do you see a pattern here? Can you guess a conjecture about such divisibility statements?)

11. Prove that for all $n \geq 2$: 6 divides n^3-n.
(*Hint: You might want to use the fact that* $n(n+1)$ *is always even.*)

12. Prove that for all n: $1+x+x^2+x^3+\ldots+x^n = \dfrac{x^{n+1}-1}{x-1}$ (provided $x \neq 1$).

13. Prove that for all $n>3$: $n^2>2n+1$

14. Prove that for all $n>4$: $2^n>n^2$ (*Hint: You may want to use #13 to help with #14.*)

15. Look at the following facts:

$$\begin{aligned} 1 &= 1^3 \\ 3+5 &= 2^3 \\ 7+9+11 &= 3^3 \\ 13+15+17+19 &= 4^3 \\ 21+23+25+27+29 &= 5^3 \\ \textit{etc.} & \end{aligned}$$

Adding the left hand sides gives the sum of the first $T(n)$ odd numbers, which is $S(T(n)) = T(n)^2 = \left(\frac{1}{2}n(n+1)\right)^2 = \frac{1}{4}n^2(n+1)^2$, and adding the right hand sides gives $1^3+2^3+3^3+\cdots+n^3$, so we are led to the equation

$$1^3+2^3+3^3+\cdots+n^3 = \tfrac{1}{4}n^2(n+1)^2$$

prove this equation by induction.

16. Prove (for $n \geq 1$) $1^3+3^3+5^3+\cdots+(2n-1)^3 = n^2(2n^2-1)$.

17. Look at the sequence of squares. We can try to calculate the sums of the first n squares. The sequence of sums is $1^2=1$, $1^2+2^2=5$, $1^2+2^2+3^2=14$, $1^2+2^2+3^2+4^2=30$, and so on. Use mathematical induction to prove the conjecture that

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

18. Use mathematical induction to prove that the largest binary (base 2) number that can be represented by n binary digits (bits) is 2^n-1.

19. Use mathematical induction to prove that the largest n-digit decimal number is 10^n-1.

20. Prove that the sum of the interior angles of an n-sided convex polygon[15] is $180(n-2)$ degrees (or $(n-2)\pi$ radians, if you know what they are), for $n \geq 3$.

21. Prove (for $n \geq 1$) $\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} < 2\sqrt{n}$.

Optional Extras

22. If a group of n people all want to shake hands with one another (without repetition), prove that (for all $n \geq 2$) the number of handshakes necessary is $\frac{1}{2}n(n-1)$.
(*Hint: if $n = 2$, then there is just one handshake between the two people; if a third joins them (so $n = 3$), then the new person will shake hands with each of the other two, making the new total number of necessary handshakes = 3. Check this fits the formula given, and see if it tells you how to go from a group of size n to a group of size $n+1$.*)

23. Suppose an ATM machine has \$20 and \$50 bills (only, but it has **lots** of them—an unlimited supply). Show that for any $n \geq 4$, the machine can give an exact payment of \$$10n$ using just \$20s and \$50s.
(*Hint: For the induction step, you might want to consider two cases: if the previous payout was all in \$20s or not.*)

24. (Fibonacci Numbers) Define the following number sequence:[16]

$$f_0 = 1 \ , \ f_1 = 1 \ , \ f_{n+2} = f_n + f_{n+1} \quad \text{for } n \geq 0$$

(so the sequence begins $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$). Show that the following are true:
$f_1 + f_3 + f_5 + \ldots + f_{2n-1} = f_{2n}$
$f_2 + f_4 + f_6 + \ldots + f_{2n} = f_{2n+1} - 1$
$f_0^2 + f_1^2 + f_2^2 + \ldots + f_n^2 = f_n f_{n+1}$
$f_n < 2^n$ *[This is a bit tricky, and is easier if you try instead to prove that $f_k < 2^k \ \forall k \leq n$.]*

25. $1^2 + 3^2 + 5^2 + \cdots + (2n-1)^2 = \frac{1}{3}n(2n-1)(2n+1)$.

26. $2 + 9 + 16 + \cdots + (7n-5) = \frac{1}{2}n(7n-3)$.

27. $\dfrac{1}{1 \cdot 3} + \dfrac{1}{3 \cdot 5} + \dfrac{1}{5 \cdot 7} + \cdots + \dfrac{1}{(2n-1)(2n+1)} = \dfrac{n}{2n+1}$.

28. For each of the following, what can you conclude from the information given about a proposition $P(n)$?
For example, if you are told that $P(7) \wedge \forall n(P(n) \longrightarrow P(n+1))$, then you could conclude $\forall n \geq 7 \ P(n)$.
(a) $P(4) \wedge \forall n(P(n) \longrightarrow P(n+1))$
(b) $\neg P(10) \wedge \forall n(P(n) \longrightarrow P(n+1))$
(c) $P(1) \wedge \neg\forall n(P(n) \longrightarrow P(n+1))$
(d) $P(1) \wedge P(2) \wedge \cdots \wedge P(1000)$
(e) $P(1) \wedge \forall n(P(n) \longrightarrow P(n+2))$
(f) $P(40) \wedge \forall n(P(n) \longrightarrow P(n-1))$
(g) $P(1) \wedge P(2) \wedge \forall n(P(n) \longrightarrow P(n+2))$
(h) $P(1) \wedge P(2) \wedge \forall n(P(n) \wedge P(n+1) \longrightarrow P(n+2))$
(i) $P(1) \wedge \forall n(P(n) \longrightarrow P(5n))$
(j) $P(1) \wedge \forall n((P(n) \longrightarrow P(5n)) \wedge (P(n) \longrightarrow P(n-1)))$

[15] In a convex polygon, every interior angle is less than 180 degrees (*i.e.* π radians). This exercise can be modified to also allow for non-convex polygons, without altering the formula. Try that if you feel up to it.

[16] This sequence answers a famous question of Leonardo Fibonacci (c.1170-1250): Beginning with a single pair of newborn rabbits, if every month each productive pair bears a new pair, which becomes productive when they are 2 months old, how many rabbits will there be after n months? f_n is the number of pairs of rabbits in month n, assuming rabbits are immortal. There is a wealth of information about and pattern derived from these simple numbers—an interesting example may be found on the course webpage ("Fibonacci & Phyllotaxis"). Google for other connections with life, the universe, and everything.

8.3 The Fundamental Theorem of Arithmetic

We saw previously (the Prime Factorization Theorem) that every natural number > 1 is either a prime, or a product of primes. We shall regard a prime number as a product of primes, by allowing "products" of just one number, so with this usage, we can say every number > 1 is a product of primes. The question we turn to now is whether this can be done in more than one way. Is it possible to have two different products of primes, where either the number of primes or some of the actual primes themselves are not the same? Or does

$$p_1 p_2 p_3 \cdots p_j = q_1 q_2 q_3 \cdots q_k$$

where all the ps and qs are prime, mean that $j = k$ and exactly the same primes appear in the two lists $p_1, p_2, p_3, \ldots, p_j$, and $q_1, q_2, q_3, \ldots, q_k$? It turns out the answer is that there is one and only one way to factor any number into prime factors:

Theorem: Any natural number $n > 1$ can be represented as a product of primes in one and only one way.

Remarks:

1. A minor detail: 1 can also be represented as a product of primes, if we allow the empty product. Note that we do **not** want to alter the definition of "prime" to include 1 as a prime, as that would destroy the truth of the theorem (as it would allow arbitrarily many extra factors 1).

2. Our proof will use the following two important properties of natural numbers: first, that *any non-empty set* of natural numbers always has a smallest (or least) element,[17]

3. and second, that if a particular number m has *only one* prime factorization, then that factorization must contain *all* the prime factors of m. For if p were a prime factor of m, then $m = pk$ for some k, and factoring k into primes $k := q_1 q_2 \ldots q_j$, we get a prime factorization of $m := pq_1 q_2 \ldots q_j$. Since m only has one such, this must have been the original factorization, which therefore contained p.

Proof of the Theorem:[18] The proof we give here will be by contradiction. We start by assuming that there are some numbers which do admit of two (or more) different prime factorizations, and so we can choose the smallest such number: let's call that number n. (Note that n itself must be composite, and not prime.) So, we can assume, for any number $m < n$, that m has only one prime factorization, which contains all the prime factors of m.

Since we are supposing that n has two (or more) different prime factorizations, let's represent them as follows:

$$n = p_1 p_2 p_3 \ldots p_j = q_1 q_2 q_3 \ldots q_k$$

where $p_1, p_2, p_3, \ldots, p_j, q_1, q_2, q_3, \ldots, q_k$ are all primes. Note that no prime can occur in both products, for if it did we could cancel it out, getting a smaller number than n with more than one representation as a product of primes, contradicting our assumption that n was the smallest such number. Note also this means none of the numbers $n, p_1, p_2, p_3, \ldots, p_j, q_1, q_2, q_3, \ldots, q_k$ can be even.

[17] This "simple" fact is in fact equivalent to mathematical induction. See section 8.5.

[18] This proof seems to have been first presented by H. Hasse in 1928; the presentation here comes from Davenport, H. *The Higher Arithmetic: An Introduction to the Theory of Numbers*, Cambridge University Press, shown me by my friend and colleague Bill Boshuck.

We can suppose that the primes are listed in increasing order (with whatever repetitions are necessary), and that p_1 is the smallest prime so listed. Note that each product has at least two factors (since n is composite), so that p_1 and q_1 cannot be greater than $\sqrt{n}$, and hence $p_1 < \sqrt{n}$ and so $p_1 q_1 < n$. Consider $n - p_1 q_1$: it is smaller than n and larger than 1 (since, being the difference of odd numbers, it is even), so admits only one prime factorization. But clearly p_1 and q_1 are factors (remember that if $d|a$ and $d|b$, then $d|(a-b)$ as well), so that factorization must look like

$$n - p_1 q_1 = p_1 q_1 r_1 \ldots r_\ell$$

for primes $r_1, \ldots, r_\ell$. This implies $p_1 q_1$ is also a factor of n (because if $d|(a-b)$ and $d|b$, then $d|a$ as well); *i.e.* $p_1 q_1$ is a factor of $p_1 p_2 p_3 \ldots p_j$. Cancelling p_1 shows that q_1 is a factor of $p_2 p_3 \ldots p_j$; since $p_2 p_3 \ldots p_j < n$ it has a unique factorization, consisting of all its factors, and so q_1 is one of those factors. But this contradicts the fact that q_1 is supposed to be a prime not occurring among $p_2, p_3, \ldots, p_j$.

So this contradiction shows there cannot be any number $n > 1$ with more than one prime factorization. (QED)

Remark:[19] There is a noticeable fact about this proof: it is more complicated than the proof that any number admits (at least one) prime factorization. (For that, we only had to use the definition of composite number as being not prime, and prime numbers as admitting only one factorization at all, namely 1 × the number itself.) This proof uses subtleties about divisors and subtraction, for example. There is a good reason for that: without such properties, the result is simply false. Here is an illustration of this fact.

We shall restrict ourselves to a subsystem of the natural numbers, namely, the numbers

$$1, 5, 9, 13, 17, 21, 25, 29, \ldots$$

which are all of the form 1 + a multiple of 4 (*i.e.* numbers of the form $4k+1$: we say these numbers are "$\equiv 1 \bmod 4$"). You can multiply such numbers together, and get products which are still in our subsystem (still $\equiv 1 \bmod 4$). This system has "pseudo-primes": numbers which admit only the trivial factorization **in this system**. Check that among the numbers listed above, $1, 5, 9, 13, 17, 21, 29$ are all pseudo-prime, but 25 is composite (since it $= 5 \times 5$).[20] In this system the usual proof that every number can be factored as a product of pseudo-primes is valid, but in this system, not all such factorizations are unique. For example, 693 is composite, and may be factored as 9×77 or as 21×33, and $9, 21, 33, 77$ are all pseudo-prime. (Of course, in the full number system, these two factorizations may be broken down further into real primes, and then we do get a unique factorization—but that further breakdown takes us outside the collection of numbers $\equiv 1 \bmod 4$).

What is the point? What has gone wrong? Well, it's simple: in the subsystem of numbers $\equiv 1 \bmod 4$, although you can multiply easily enough, what you **cannot** do is **add** or **subtract**! For example, $1 + 5 = 6$ and 6 is no longer in our subsystem. In other words, this subsystem is not "closed under" addition and subtraction. This blocks the proof above from working (and indeed, blocks any other proof from working, since the unique prime factorization theorem is not true in this system). So any proof of unique prime factorization is bound to use more than just the definitions of prime, composite, and multiplication.

[19] This example is due to Hilbert, about whom we shall hear more in Chapter 9.

[20] If you are tempted to say "Hey! 21 isn't (pseudo-)prime, it is composite! It is 3×7", then remember that 3 and 7 are not numbers in this system, and in this system, 21 only has itself and 1 as divisors.

FWIW: I don't know of any proof of the Fundamental Theorem of Arithmetic which is any simpler than the one given here—but you can judge for yourself: on the course webpage you can find a description of a proof based on Euclid's original ideas, (with some helpful comments thrown in as well), as it appeared in the text written by the previous teacher of this course.

8.3.1 Special numbers

Prime numbers are defined in terms of the numbers that divide *into* them: a number is prime if only 1 and itself are divisors. One may say this is a definition that "looks down" (at smaller numbers). There is another way to characterize primes, however, one that "looks up" (at larger numbers, more specifically, at the numbers into which the prime divides).

> **Definition**: A natural number $n > 1$ is **special** if (and only if) it satisfies the following property: whenever it divides a product ($n|ab$), it must divide at least one factor ($n|a$ or $n|b$). In other words:
>
> $$n \text{ is special } \leftrightarrow \forall a \forall b \, [\, n|ab \longrightarrow (n|a \vee n|b) \,]$$

Our point is that primes are special, and *vice versa*, so that this definition provides another way to characterize primes.[21]

Lemma: For any natural numbers $a, b > 1$, suppose the prime factorizations of a and b are given by

$$a = p_1 p_2 \cdots p_m \quad \text{and} \quad b = q_1 q_2 \cdots q_n$$

then the prime factorization of ab is the product of these prime factorizations:

$$ab = p_1 p_2 \cdots p_m \, q_1 q_2 \cdots q_n$$

Proof: Clearly $ab = p_1 p_2 \cdots p_m \, q_1 q_2 \cdots q_n$ is a prime factorization of ab, and since prime factorizations are unique, it is *the* prime factorization of ab. Notice that this also tells you that if you are given the prime factorization of ab, the prime factorizations of a and b must be given by the same primes, just redistributed, some to a and some to b.

Theorem: If a prime p divides ab, then it must either divide a or it must divide b. In other words, primes are special.

Proof: Suppose p divides ab: that means p appears in the prime factorization of ab, and hence in particular p must appear in one of the prime factorizations, either for a or for b, which means it divides either a or b. (QED)

Theorem: If n is special, it must be a prime.

Proof: This is easiest to see if we take the contrapositive and prove that if n is *not* prime, then it cannot be special. So suppose n is not prime, and so it is composite: $n = ab$. Notice that both a, b are > 1 and $< n$, and so n cannot divide either a or b. But this contradicts the property of being special, so n is not special. (QED)

An example might help here: consider the number 6. It is composite: $6 = 2 \cdot 3$. But $6 \not| \, 2$ and $6 \not| \, 3$. So 6 is not special.

[21] But only in a setting, such as the natural numbers, where the Fundamental Theorem of Arithmetic is valid. In fact, in the system of numbers $\equiv 1 \bmod 4$, we saw that 21 is a (pseudo-)prime, but it is not special, since it divides $693 = 9 \times 77$, but it does not divide either 9 or 77.

Remarks

1. If a number p is special, then it has an apparently stronger property than explicitly given by the definition of being special: if p divides a product $abc\cdots$, then it must divide at least one of the factors.[22] The proof of this is straightforward (can you prove it without looking at the next sentence?). For, if p divides $abc\cdots$, then p divides $a(bc\cdots)$, so by the property of being special, p divides a or $bc\cdots$. If it divides a, we're done, and if not then it divides $b(c\cdots)$, so (by being special), p divides b or p divides $c\cdots$. If it divides b, we're done, and if not ... well, I hope you get the idea: we just keep using the binary property over and over till we exhaust the factors. At some point, we'll have found one (at least) that p divides.

2. If we knew that primes are special, then the proof of the Fundamental Theorem would be even simpler: we could argue as follows. Suppose a number had two prime factorizations: $n = pqr\ldots = p'q'r'\ldots$. Since $p|n$, $p|p'q'r'\ldots$, and so p must divide (and hence equal) one of the primes $p', q', r', \ldots$. But then we could divide both factorizations by p, and get a smaller number with two prime factorizations. In this way, we get the same contradiction as before, and so prime factorizations must be unique.

   On the other hand, we saw above that the Fundamental Theorem implies that primes are special. So, what this shows is that the fact that primes are special is *equivalent* to the Fundamental Theorem of Arithmetic.

3. If you turn back to our proof that $\sqrt{2}$ is irrational, you will see that this only depended on the fact that 2 is special. This means that $\sqrt{p}$ is irrational for any prime, and indeed, it then follows that the only way a composite number can have a rational square root is if in the prime factorization of the number, every prime appears an even number of times. Hence if a number is not a perfect square, its square root must be irrational:

   > **Proposition**: For any natural number n, $\sqrt{n}$ must either be a natural number or an irrational number.

Exercise: using the proof that $\sqrt{2}$ is irrational as a model, try to prove that $\sqrt{3}$ is irrational. Can you modify the proof to prove $\sqrt{6}$ is irrational? What about $\sqrt{9}$? Be sure your proof for $\sqrt{6}$ doesn't equally work for $\sqrt{9}$. (Why?)

(If you feel up to it, you could also try to modify your proof to show that any root (cube root, fourth root, *etc.*) of a natural number is either a natural number or is irrational, but never a non-integral rational number.)

8.4 Answers to the exercises

Exercises 8.1.4

1. (i) $b = dx$ and $r = dy$ so $a = bq + r = dxq + dy = d(xq + y)$, so $d|a$. (ii) $a = dx$ and $b = dy$ so $r = a - bq = dx - dyq = d(x - yq)$ so $d|r$.

[22] Why do I say this is stronger? Because the definition only said this for binary products, but in fact it is true for any finite product, no matter how many factors there are.

2. Divisors: $1, 2, 4, 8, 5, 25, 125, 625, 10, 50, 250, 1250, 20, 100, 500, 2500, 40, 200, 1000, 5000$ (The FTA tells you the only primes that can make up factors are 2 and 5, and you can only use up to three 2s and up to four 5s.)

3. $2 \cdot 11 \cdot 13\ + 1 = 287 = 7 \cdot 41$, so the new primes are 7 and 41.

4. $2 \cdot 3 \cdot 5 \cdot 7 + 1 = 211$ which is prime itself.

5. $7!\ + 1 = 5041 = 71^2$. I will leave it to you to check the rest.

6. Start with $8!\ + 2 = 40322$: it and the next 6 numbers (making 7 in all) are all composite. (So the answer is 40321, since it's the 7 numbers *after* it that are composite. Actually, it's composite too, as are all the numbers from 40289 to 40343, which are both prime.)

7. If $a = dx$, $b = dy$, then $a + b = d(x + y)$ and $ab = d^2xy$, as required.

8. No. For example, $3|4 + 5$, but 3 does not divide either 4 or 5.

9. and 10. I'll leave these to you.

Exercises 8.2.3

1. $1 = \frac{1}{2} \cdot 1 \cdot (3 - 1)$; $1 + 4 + \cdots + (3n - 2) + (3n + 1) = \frac{1}{2}n(3n - 1) + 3n + 1 = \frac{3}{2}n^2 + \frac{5}{2}n + 1 = \frac{1}{2}(n + 1)(3n + 2)$.

2. $3 = 1(2+1)$; $3+7+\cdots+(4n-1)+(4n+3) = n(2n+1)+(4n+3) = 2n^2+5n+3 = (n+1)(2n+3)$.

3. $1\cdot 2 = \frac{1}{3}\cdot 1\cdot 2\cdot 3$; $1\cdot 2+2\cdot 3+3\cdot 4+\cdots+(n+1)(n+2) = \frac{1}{3}n(n+1)(n+2)+(n+1)(n+2) = (\frac{1}{3}n + 1)(n + 1)(n + 2) = \frac{1}{3}(n + 1)(n + 2)(n + 3)$.

4. $\frac{1}{1\cdot 2} = \frac{1}{2}$; $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \dfrac{1}{3\cdot 4} + \cdots + \dfrac{1}{n(n+1)} + \dfrac{1}{(n+1)(n+2)} = \dfrac{n}{n+1} + \dfrac{1}{(n+1)(n+2)} = \dfrac{n(n+2)+1}{(n+1)(n+2)} = \dfrac{(n+1)^2}{(n+1)(n+2)} = \dfrac{n+1}{n+2}$.

5. (We can start at $n = 0$ here if we like:) $1 = 2^1 - 1$ (you can also start at $n = 1$). $1 + 2 + 2^2 + 2^3 + \cdots + 2^n + 2^{n+1} = 2^{n+1} - 1 + 2^{n+1} = 2 \cdot 2^{n+1} - 1 = 2^{n+2} - 1$.

6. If $\#X = 1$, $X = \{x\}$, and $\mathcal{P}(X) = \{\emptyset, \{x\}\}$ has $2^1 = 2$ elements. If $\#X = n+1$, $X = Y \cup \{x\}$ where $\#Y = n$; then $\mathcal{P}(X) = \mathcal{P}(Y) \cup \{A \cup \{x\} \mid A \subseteq Y\}$, which has twice as many elements as $\mathcal{P}(Y)$: so it has $2 \cdot 2^n = 2^{n+1}$ elements.

7. $2|0$; $3^{n+1} - 1 = (2 + 1) \cdot 3^n - 1 = 2 \cdot 3^n + (3^n - 1)$ and 2 divides each of these.

8. $5|8^1 - 3^1 = 5$; $8^{n+1} - 3^{n+1} = 8 \cdot 8^n - 3^{n+1} = (5 + 3)8^n - 3 \cdot 3^n = 5 \cdot 8^n + 3 \cdot (8^n - 3^n)$ and 5 divides each of these terms.

9. and 10. are almost "identical" (just replace the appropriate numerical values): what all three problems have in common is that the difference in case $n = 1$ gives the divisor. So, if $a-b = c$, then $c|a^n - b^n$ for all n.

11. $6|(2^3-2)$; $(n+1)^3-(n+1) = n^3+3n^2+3n+1-n-1 = (n^3-n)+3(n^2+n) = (n^3-n)+3n(n+1)$ and 6 divides each of these (6 divides $3n(n + 1)$ because 2 divides $(n(n + 1))$).

12. $1 = \frac{x-1}{x-1}$; $1 + x + x^2 + x^3 + \ldots + x^n + x^{n+1} = \frac{x^{n+1}-1}{x-1} + x^{n+1} = \frac{x^{n+1}-1}{x-1} + \frac{x^{n+1}(x-1)}{x-1} = \frac{x^{n+1}-1+x^{n+2}-x^{n+1}}{x-1} = \frac{x^{n+2}-1}{x-1}$

13. $4^2 > 2 \cdot 4 + 1$; $(n+1)^2 = n^2 + 2n + 1 > 2 + 2n + 1 = 2n + 3 = 2(n+1) + 1$ ($n^2 > 2$ since $n > 2$)

14. $2^5 > 5^2$; $2^{n+1} = 2^n + 2^n > n^2 + n^2 > n^2 + 2n + 1 = (n+1)^2$

15. $1^3 = \frac{1}{4}1^2 2^2 = 1$;
$1^3 + 2^3 + 3^3 + \ldots n^3 + (n+1)^3 = \frac{1}{4}n^2(n+1)^2 + (n+1)^3 = (n+1)^2(\frac{1}{4}n^2 + n + 1) = \frac{1}{4}(n+1)^2(n^2 + 4n + 4) = \frac{1}{4}(n+1)^2(n+2)^2$

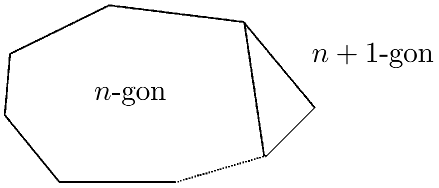
16. $1^3 = 1^2(2 \cdot 1^2 - 1) = 1 \cdot 1$;
$1^3 + 3^3 + 5^3 + \cdots + (2(n+1) - 1)^3 = n^2(2n^2 - 1) + (2n+1)^3 = 2n^4 + 8n^3 + 11n^2 + 6n + 1 = (n+1)^2(2(n+1)^2 - 1)$; (you just have to multiply these out and see that they are equal as claimed).

17. $1^2 = \frac{1}{6}1(1+1)(2+1) = 1$;
$1^2 + 2^2 + \ldots + n^2 + (n+1)^2 = \frac{1}{6}n(n+1)(2n+1) + (n+1)^2 = (n+1)(\frac{1}{6}n(2n+1) + n + 1) = \frac{1}{6}(n+1)(2n^2 + 7n + 6) = \frac{1}{6}(n+1)(n+2)(2n+3)$

18. The largest binary number with n digits is $1 + 2 + 2^2 + 2^3 + \ldots + 2^{n-1}$ ($= 1111\ldots1_2$, with n 1s). So the induction goes thus: $1 = 2^1 - 1$; $1 + 2 + 2^2 + 2^3 + \ldots + 2^{n-1} + 2^n = 2^n - 1 + 2^n = 2 \cdot 2^n - 1 = 2^{n+1} - 1$.
(Note that Mersenne primes are of this form.)

19. The proof for base 10 is similar: the largest number with n digits (all 9s) is $9 + 9 \cdot 10 + 9 \cdot 10^2 + 9 \cdot 10^3 + \ldots + 9 \cdot 10^{n-1}$. So the induction goes thus: $9 = 10^1 - 1$; $9 + 9 \cdot 10 + 9 \cdot 10^2 + 9 \cdot 10^3 + \ldots + 9 \cdot 10^{n-1} + 9 \cdot 10^n = 10^n - 1 + 9 \cdot 10^n = (1+9)10^n - 1 = 10^{n+1} - 1$.

20. $n = 3$: The angles of a triangle add to 180 degrees. Assuming the result for n, we must prove that an $n+1$-sided convex polygon has the sum of interior angles $= 180(n-1)$ degrees. But consider the following picture, where we take an $n+1$-gon and imagine replacing two sides by a single side joining their endpoints, thus creating an n-gon inside the $n+1$-gon. The n-gon's angles add up to $180(n-2)$ degrees, and the additional triangle adds a further 180 degrees, giving a total for the $n+1$-gon equal to $180(n-1)$ degrees.



21. $\frac{1}{\sqrt{1}} < 2\sqrt{1}$; $\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n+1}} < 2\sqrt{n} + \frac{1}{\sqrt{n+1}} < 2\sqrt{n+1}$. The tricky part is showing the last step: $2\sqrt{n} + \frac{1}{\sqrt{n+1}} < 2\sqrt{n+1}$. Here is the calculation, easiest to understand backwards (*i.e.* we start with what we want, play with it a bit, ending with something obviously true, then imagine reversing our steps; here the last statement *is* obviously true, so

just read from the bottom up, and we get the required result).

$$\begin{array}{rcl}
2\sqrt{n}+\dfrac{1}{\sqrt{n+1}} & < & 2\sqrt{n+1} \\
2\sqrt{n}\sqrt{n+1}+1 & < & 2n+2 \\
2\sqrt{n}\sqrt{n+1} & < & 2n+1 \\
\sqrt{n}\sqrt{n+1} & < & n+\frac{1}{2} \\
n(n+1) & < & (n+\frac{1}{2})^2 \\
n^2+n & < & n^2+n+\frac{1}{4}
\end{array}$$

The final exercise:

Prove that $\sqrt{3}$ is an irrational number.

Proof: The key property used is that 3, being prime, is special. As with the similar proof for $\sqrt{2}$, we prove the result by contradiction. So, assume that there is a rational number equal to $\sqrt{3}$: *i.e.* there are integers a and b such that

$$3 = \left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}$$

we may also assume that $\frac{a}{b}$ is a fraction in reduced form (*i.e.*, a and b are relatively prime), for if it is not, then we replace a, b with appropriate integers so that the fraction *is* in reduced form, (so in particular, a, b are not *both* multiples of 3). Multiplying both sides of the equation $3 = \frac{a^2}{b^2}$ by b^2, we get $3b^2 = a^2$. But $3b^2$ is a multiple of 3, so a^2 must also be a multiple of 3. However $a^2 = a \cdot a$ and so a must be a multiple of 3, since 3 is special. Since a is a multiple of 3, there is some number k such that $a = 3k$. Substituting into $3b^2 = a^2$, we get $3b^2 = (3k)^2 = 9k^2$. Dividing both sides by 3 gives $b^2 = 3k^2$. Reasoning as above, b^2 must be a multiple of 3 and hence b is also a multiple of 3. So we have shown that a and b must have a common factor of 3, contradicting our assumption that a and b are relatively prime. So there is no rational number whose square is three. (QED)

Compare this proof to the proof in Chapter 7 that $\sqrt{2}$ is irrational—it is virtually identical, merely replacing 2 with 3. In a similar fashion, we can show that $\sqrt{p}$ is irrational for any prime p. I'll leave it to you to try this for other numbers which are not perfect squares; for example, try to prove $\sqrt{6}$ is irrational—you will want to consider that $6 = 2 \times 3$. Notice that you have to be a little subtle: since 6 is not special, you cannot simply say that if $6|a^2$ then $6|a$. You must use another argument: instead of using 6, you can use a prime factor of 6, provided that prime factor appears an odd number of times in the prime factorization of the number 6 (*e.g.* you could use 2 or 3). (If a number's prime factorization only has even powers of primes, then the number is a perfect square, like 9 is, so its square root *is* rational.)

Here is an illustration, contrasting $\sqrt{12}$ which is irrational, with $\sqrt{36}$ which is rational.

Proof that $\sqrt{12}$ is irrational. Suppose $\sqrt{12} = \frac{a}{b}$, a, b in lowest terms. Then $12b^2 = a^2$, so $12 \mid a^2$, and hence $3 \mid a^2$, so $3 \mid a$. Since then $a = 3x$ for some x, $a^2 = 9x^2 = 12b^2$, so $3x^2 = 4b^2$, so $3 \mid b^2$ and hence $3 \mid b$, giving the contradiction we wanted.

Why does this fail for $\sqrt{36}$? Follow the same pattern: $\sqrt{36} = \frac{a}{b}$, a, b in lowest terms. Then $36b^2 = a^2$, so $36 \mid a^2$, and hence $3 \mid a^2$, so $3 \mid a$. Then $a = 3x$ for some x, $a^2 = 9x^2 = 36b^2$, or just that $x^2 = 4b^2$; we do not get $3 \mid b$. So we fail to get our contradiction.

The point? This failure would happen whenever the number involved is a perfect square, so that its prime factorization consists only of even powers of primes. The proof for $\sqrt{12}$ depended on having an odd power of a prime (namely 3) in the prime factorization of 12.

8.5 Historical Remark: The method of infinite descent

Before the nineteenth century, mathematical induction was more familiarly known as "the method of infinite descent", and was worded a little differently: for any property $P(n)$ of natural numbers n, $P(n)$ is true for all n if, for any n, $P(n)$ is false only if there is some smaller number $m < n$ for which $P(m)$ is also false.[23] In symbols:

$$\forall n(\neg P(n) \rightarrow \exists m < n\ \neg P(m)) \rightarrow \forall n P(n)$$

Think a bit about this: try to convince yourself it is equivalent to the principle of induction.

Infinite descent is not obviously equivalent to induction as we've stated it—a seemingly stronger version of induction is necessary, but one which is actually equivalent to the usual version. "Strong induction", as it's often called, is the following: for any property $P(n)$, if for all n, $P(n)$ is true whenever $P(m)$ is true for all $m < n$, then $P(n)$ must be true for all n:

$$\forall n(\forall m < n P(m) \rightarrow P(n)) \rightarrow \forall n P(n)$$

It's not too hard to show this is equivalent to the usual principle of induction, and it's easy to show strong induction is equivalent to infinite descent. An example where strong induction is helpful was seen in exercise 8.2.3 #24. See the hint there: that hint also shows how one recasts a proof using strong induction as a proof using ordinary induction, by replacing the aim of proving $P(n)$ with the aim of proving $\forall m \leq n\ P(m)$.

From the equivalence of infinite descent (or strong induction) and ordinary induction, we can also see that induction is equivalent to the statement that every non-empty set S of natural numbers has a least element (this statement is usually called "the well-orderedness" of the natural numbers). First, it's easy to show that infinite descent implies well-orderedness: take $P(n)$ to mean $n \notin S$, so $\neg P(n)$ means $n \in S$. Then if S has no least element, $\forall n(\neg P(n) \rightarrow \exists m < n\ \neg P(m))$, and hence $\forall n P(n)$, *i.e.* S is empty. On the other hand, well-orderedness implies ordinary induction: suppose that well-orderedness is true yet mathematical induction fails, *i.e.* that $P(0) \wedge \forall n(P(n) \rightarrow P(n+1))$, and yet $\forall n P(n)$ is false. Let S be the set of n for which $\neg P(n)$, and notice S is not empty by assumption. So let m be the least element of S: m cannot be 0, since $P(0)$, so let $n = m - 1$: $n \notin S$ must be true (since m is the least element of S), and hence $P(n)$. But then $P(n+1)$, *i.e.* $P(m)$, which contradicts $m \in S$. So $\forall n P(n)$ must be true, and so induction is valid.

With this perspective, we can redo the proof we gave of the Fundamental Theorem of Arithmetic to avoid the method of contradiction, and to use strong induction, or equivalently infinite descent, instead. If you feel up to it, give this a try. From this you can see that the FTA really uses all the essential structure of the natural numbers.

[23] You may be able to see why this principle is "obviously true": the fact that any number where P fails to be true produces a smaller number where P also fails to be true means that if P is ever false, you'd start an "infinite descent", counting down from one failure to ever smaller ones—but that would mean counting down from one number without ever stopping. That is impossible, as eventually you'd want to use negative numbers, and $\mathbb{N}$ has none of those. That contradiction means P can never be false, so is always true.

Chapter 9

The Axiomatic Method

Introduction

The main object of mathematics is the study of structure. Structure manifests itself in many ways and in many places: in number patterns, in geometric forms, in computation, in music, painting, architecture, sculpture, and other arts, in literature (both prose and poetry), in language, in logic, ... , to name but a few, and all of these are the subject of mathematical investigation. You may have thought before this course that maths dealt with just numbers, algebra, and geometry—you now know it also deals with logic, AKA "the laws of thought", and I hope we'll have time to see some mathematical approaches to linguistics and maybe even music. There are many many other applications—it's probably safe to say nothing is maths-proof!

One thing that strikes a mathematician particularly forcefully is when s/he[1] notices "the same pattern" occurring in several, seemingly unrelated, situations. The usual reaction is to try to understand "what's really going on here": in other words, to get at the underlying structural pattern that is being displayed in the various situations. One example we have already seen in this course may serve as an example: the deduction rules for logic seem to be similar to basic relations among simple sets, and this connection seems to go very deep, as illustrated by the similarity between the de Morgan equivalences $\neg(A \vee B) \equiv \neg A \wedge \neg B$ and $(A \cup B)^{\mathsf{C}} = A^{\mathsf{C}} \cap B^{\mathsf{C}}$, *etc.*

One way to analyse the underlying pattern is to isolate key properties of the pattern, and see what structure follows from those properties. This leads a mathematician to define a structure in terms of its fundamental notions: basic terms are defined, and basic relationships between these terms are postulated (these are usually called *axioms*). Then one would verify that the original situations are *models* of these axioms. Usually one then wants to see what other models there are, and if any interesting general story can be told about the totality of such models.

It is probably safe to say that this "axiomatic approach" to mathematics, begun in the 19th century, but really only strongly exploited from the early 20th century on, has become the primary tool of research (theoretical) mathematics. However, in several "popular" accounts of modern mathematics, the nature of the axiomatic method is distorted, I think: the suggestion is made that the language and axioms of a mathematical theory are "without meaning", that formal mathematics is something like a game, where one only follows arbitrary "rules", operating with meaningless symbols. I want to make it clear that I disagree with this sentiment in the strongest terms, and most mathematicians I've spoken to about the subject feel pretty much the same way. Moreover, it simply does not reflect how we actually do mathematics. I suspect that most writers of such statements are not practicing mathematicians, but are merely observers—I have known very few mathematicians

[1] I'll use the masculine form for both from now on—it is a curious fact at present that most mathematicians are men, although there seems to be no good "innate" reason for this (there are several possible "societal" ones). Pity.

who really merely manipulate symbols without some idea of their intended meaning.

Although an axiom system is often treated abstractly, and although one is not to use "facts" not explicitly stated in (or derived from) the axioms, there is nothing "arbitrary" about important axiom systems. They are intended to illuminate real patterns in real situations, and their importance comes from that fact. If one isn't to use "outside facts", it's not because one is "playing a game", but because one wants to be sure to capture what is common to all the situations one has in mind, and not to allow one situation to dominate in an unwanted way. Frequently, a mathematician **will** add further axioms that are true only in some particular models he's interested in, if that suits his purpose. As long as one keeps clear what assumptions are being made, this is harmless, and may even be helpful.

As a start, we shall look at some history. The first axiomatic presentation of a mathematical structure is Euclid's geometry. His axioms don't measure up to today's standards of rigour, but they did set a style and precedent that were an inspiration for later work. Euclid's axioms were intended to reflect reality in a literal sense: his geometry was intended to be the geometry of the universe in which we live. But one of his axioms was controversial from the start, and investigating it led in the nineteenth century to an unexpected "crisis", when non-Euclidean geometries were discovered. This threw the mathematical world into confusion, and was one of the factors that led to a more careful examination of the fundamental principles on which the subject rested, and to a more central role for the axiomatic approach to mathematics.

Then we shall look in some detail at a few structures related to sets and logic, beginning with what are called Boolean algebras. A slightly more general structure, Heyting algebras, will be useful if we want to consider a logic without the double-negation rule ($\neg\neg\ E$): $\neg\neg A \vdash A$, and this also happens to capture the structure of some more "special" sets, structure not captured by Boolean algebras. An even more general logical structure will allow us also to capture the structure of English (and French, and Italian, *etc.*) grammar (next chapter).

9.1 Euclidean Geometry

An axiomatic system begins with definitions of terms plus a set of axioms. One constructs valid deductive arguments using the definitions and axioms as premises. If the axioms are true, then every conclusion derived from them must be true as well.

The earliest axiomatic mathematics is the 13 books of Euclid's *Elements*. The *Elements* begins with 23 definitions, five "postulates", and five "common notions". The "postulates" and "common notions" (now usually simply called "axioms") were considered to be "self-evident truths", so obviously true as to need no proof. If the postulates and common notions are true, then every conclusion derived from them by valid deductive argument must be true.

Euclid's five postulates of plane geometry were:

P1. A straight line may be drawn between any two points.

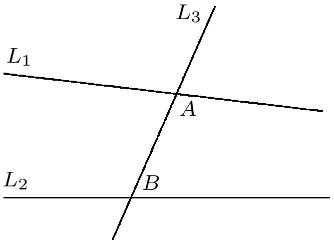
P2. Any straight line may be extended indefinitely.

P3. A circle may be drawn with any point as centre and any radius.

P4. All right angles are equal.

P5. If two straight lines lying in a plane are crossed by another straight line, and if the sum of the internal angles on one side is less than two right angles, then the straight lines will meet if extended sufficiently on the side on which the sum of the angles is less than two right angles.

9.1.1 A Problem with Euclid's Postulates



Postulates P1 through P4 are easy to understand. They may even be "self-evident". However, postulate P5 is much more complicated than the others. In the diagram, the lines labelled L_1 and L_2 are the original "two straight lines lying in a plane". Line L_3 is "another straight line" that crosses them. Angles A and B are the "internal angles on one side" of line L_3. The postulate says that if the sum of angles $A + B$ is less than twice the size of a right angle (180°), then if we extend lines L_1 and L_2 indefinitely, they'll meet somewhere off to the right. The definition of "parallel lines" says "*parallel straight lines* are straight lines which, being in the same plane and produced indefinitely in both directions, do not meet one another in either direction". So L_1, L_2 as shown are not parallel. The effect of P5 is to say that if L_1, L_2 are parallel, then the sum of the angles A, B must be *exactly* two right angles.

Even when the meaning of the postulate is clear, its truth is not self-evident. In the diagram, L_1 and L_2 are clearly inclined toward each other. But what if the sum of angles $A + B$ were only very slightly less than two right angles, so that the lines would have to be drawn millions of miles long before they met? Would they meet?

For two thousand years, geometers tried to derive the parallel postulate from the other postulates. They failed, but they did prove that the parallel postulate was logically equivalent to other statements. Some statements that are equivalent to P5 are:

P5-1. If a straight line intersects one of two parallel lines, it will intersect the other.

P5-2. Straight lines parallel to the same straight line are parallel to each other.

P5-3. Two straight lines that intersect one another cannot be parallel to the same line.

P5-4. Given a line L and a point P in a plane, where P is not on L, there is one and only one line through P which is parallel to L.

If we could prove any of these, then (by equivalence) we would have proved postulate P5. They failed to find a proof using direct methods. They tried proof by contradiction: if one derives a contradiction from the assumption that postulate P5 is *false*, then one will have proved that the postulate is *true*.

Occasionally somebody would claim to have proved a contradiction, but either the "proof" was invalid or the "contradiction" was not really a contradiction. On several occasions, one or another mathematician came close to realizing that in fact $\neg$P5 could not be proved inconsistent with the other postulates, and so there might actually be other geometries, but they (and the world!) were not ready to accept so radical a notion (yet!).

9.1.2 Different geometries

For the fact is, if there is no contradiction, we could invent new plane geometries by replacing Euclid's fifth postulate with its denial. Lobachevskian geometry replaces postulate P5-4 with "Given a line L and a point P in a plane, where P is not on L, there is more than one line through P which is parallel to L". Riemannian geometry uses the statement "Given a line L and a point P in a plane, where P is not on L, there is no line through P which is parallel to L".

Each of these statements was proved to be logically independent of the other postulates. A statement is logically independent of other statements if it can be false when the other statements are all true. That entails that none of them can be proved by deductive argument using the other postulates as premises. It was also shown that if Euclidean geometry is consistent then Riemannian geometry and Lobachevskian geometry are consistent too. By all purely mathematical and logical standards, Riemannian and Lobachevskian geometries are "just as good" as Euclid's.

A geometry (one of the many different geometries) came to be seen as one particular set of undefined terms and basic postulates. The postulates specify the relations between the terms.

Why "undefined terms"? Euclid's definitions of "point", "line", "surface", and so on are not very clear anyway. What does it mean to define "point" as "that which has no part", or "line" as "breadthless length", as Euclid did? These definitions require us to go on to define "part" and "breadth" and "length". The words used in those definitions require definition also. We're always going to have undefined terms. Why not stop at "point" and "line", and leave them undefined?

So now one views Euclid's definitions in a more "abstract" way (they have also been "tightened up", since he actually used intuitive principles which were not part of the explicit axioms he stated). The postulates were sanitized to use undefined terms. For example, Euclid's first postulate says, "A straight line may be drawn between any two points", which is equivalent to "Given any two points, there is at least one straight line that contains them". To keep this abstract, we shouldn't talk about "straight line", because people may interpret that to mean something like the path of a ray of light or the streak of ink left by a pen following a ruler. Abstract geometry is not about the paths of rays of light or streaks of ink. Using an undefined word like *luggle* would be better. Similarly with "point", where we might say *puggle* instead. The relation "contains" is tied to the intuitive interpretation of "point" and "line" so we use another undefined word (e.g., *cuggle*). Now we can restate the postulate as "Given any two *puggles* there is at least one *luggle* such that both *puggles* have a *cuggle* relation to that *luggle*". Better yet, we can replace the verbal formulation with a symbolic one, like:

$$\forall x \forall y((P(x) \wedge P(y)) \longrightarrow \exists z(L(z) \wedge (C(x,z) \wedge C(y,z))))$$

The parallel postulate P5-4 can be replaced with "Given any *luggle* L_1 and a *puggle* P_1 that does not have the *cuggle* relation to L_1, there is one and only one *luggle* L_2 that has the *cuggle* relation to P_1 such that, no matter how much we *exuggle*[2] L_1 and L_2, there will never be a *puggle* that has the relation *cuggle* to both L_1 and L_2".

What does it mean to say that something is a *puggle*? What does it mean to say that some x has the property L or that there is a C relation between two things? At this level of abstraction, it doesn't matter: these things have the meaning given by the axioms which are stated about them, no more, no less. We can say that the theorems (the statements we can prove from the axioms) of a geometry are true of any things, properties, operations and relations that satisfy the postulates. If we can interpret the basic notions so that the postulates state truths about some set of things, then the theorems will also be true of that set of things under that interpretation. The mathematician can concentrate on deriving theorems without caring about what sort of things he is dealing with. He can also focus on the properties of the system, in abstraction from questions about the things the system describes.

David Hilbert created such an abstract axiomatic geometry. In the process, he showed that some of Euclid's proofs were invalid. Euclid had assumed things that were not "contained" in the premises (axioms and postulates) he was supposed to be using.

One can interpret Hilbert's abstract geometry in terms of Euclidean lines and points. Under that interpretation it is a cleaned-up and corrected version of Euclid's geometry. One can also interpret it differently (*e.g.*, if we interpreted *puggle* (*i.e.* $P(x)$) to mean line and *luggle* (*i.e.* $L(x)$)

[2] We do not say "extend", since that is likely to be given intuitive meaning beyond what the axioms specify.

to mean point) to give different geometries. Similar abstraction and interpretation could be done with the non-Euclidean geometries. You are probably familiar with the fact that on the globe, the shortest path between two points is a great circle (a circle whose center coincides with the center of the sphere), and so it may not be a surprise that in spherical geometry, the appropriate notion that replaces "straight line" in plane geometry is "great circle". (The usual airplane routes, for example, from Montreal to Paris, are along great circles.) So in spherical geometry, *luggle* (*i.e.* $L(x)$) would be interpreted as "great circle".

In these alternative geometries, some familiar "facts" are rather different from what one learnt about plane (Euclidean) geometry. For instance, in plane geometry the angles of a triangle add up to 180°,[3] but in spherical geometry, they add up to more, and not always to the same sum: consider for example a triangle on the surface of the globe made up of two lines of longitude (meeting at the north pole) and part of the equator: the angles at the equator are each 90°, and the angle at the north pole could be anything up to 360°.

In this way, we can create all sorts of abstract "geometries" unrelated to anything that had been done before. Following the success of this way of viewing geometry, the use of axiomatic presentations of mathematical structures became more common in the late nineteenth, early twentieth century, so that it is now the normal way mathematicians view the structures they work with.

9.1.3 Axiomatic systems

Modern pure (or abstract) mathematics usually concerns itself with the investigation of axiomatic systems, systems which are intended to focus on important properties of situations where patterns are worth close study.

An axiomatic system consists of several features. One must have a *language*, meaning a set of predicate and function symbols, including constants, (which may be *typed*, in the sense we discussed previously), and a set of *axioms*, or premises one always assumes in derivations. These axioms establish the fundamental relationships between the basic constants and predicates. Sometimes we refer to the collection of theorems (logical consequences) of an axiomatic system as a *theory* (in the given language); it is one of the peculiarities of language that mathematicians use the word *theory* to refer to a body of definitely proven truths, almost the opposite of everyday English. (The scientific use of the word is closer to the mathematical usage, a fact that causes a lot of misunderstanding in the general public, seen most forcefully when political or religious doctrine is involved, as with the "debates" over evolution.)

Our presentation of propositional logic was virtually as an axiomatic system—ignore all the motivational prose, focus on just the formation and derivation rules, and you have the essence of an axiomatic system. In such a context, we would be best dropping the word "true", and stick to saying "provable".

A couple of "toy" axiomatic systems may be found on the course website—play around with them to get the feel of how one may derive abstract statements, just by using the usual rules of logic and the axioms as premises. We shall now turn to an axiomatic system which has deep links with much of what we've done so far this semester.

[3] In fact, this "fact" is equivalent to the parallel postulate P5. Pythagoras' theorem is another "fact" that's equivalent to P5.

9.2 Boolean and Heyting Algebras

9.2.1 Boolean algebras

As an example of how the axiomatic method often (usually) works in practice, we shall present a structure which has the main features shared by propositional logic and simple sets. We start by recalling that those structures did have a lot in common—a feature we remarked upon, and actually used, in showing some equalities in sets, by using the similar structure in logic as reflected in the set definitions. Recall, for example, that we proved $(A \cup B) \cup C = A \cup (B \cup C)$ by invoking the fact that $(p \vee q) \vee r \leftrightarrow p \vee (q \vee r)$ (Exercises 6.4.7). Exactly what the salient features of these two structures are is a matter of taste and experience—one possibility is explored in this section (and some "hints" at other possible axiomatizations will appear in the comments and exercises).

Our language will admit one sort of entity (which we shall not name, but feel free to think of them as "sets" if you wish, as long as you don't use anything you know about sets other than what is asserted here!). We shall have two constants $0, 1$, a unary operation $-$ (which produces $-x$) and two binary operations $+, \cdot$ (which produce $x + y, x \cdot y$). And finally, we shall have the following axioms, which are supposed true for all x, y, z:

[B1a]	$x + y = y + x$	[B1b]	$x \cdot y = y \cdot x$	(commutativity)
[B2a]	$x + (y + z) = (x + y) + z$	[B2b]	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$	(associativity)
[B3a]	$x + (y \cdot z) = (x + y) \cdot (x + z)$	[B3b]	$x \cdot (y + z) = (x \cdot y) + (x \cdot z)$	(distributivity)
[B4a]	$x + (x \cdot y) = x$	[B4b]	$x \cdot (x + y) = x$	(absorption)
[B5a]	$x + (-x) = 1$	[B5b]	$x \cdot (-x) = 0$	(complements)

If these axioms seem weird to you, keep in mind the two intended interpretations (which are the ones that caused one[4] to come up with the notion of Boolean algebra in the first place). In Set Theory, equality is ordinary equality of sets; in Propositional Logic, equality is interpreted as provable equivalence, so *eg.* $x = y$ would become $\vdash X \leftrightarrow Y$, where X and Y are the interpretations of x, y respectively.

Boolean Algebra	Set Theory	Propositional Logic
1	U	$\top$
0	$\emptyset$	$\bot$
$-x$	A^{c}	$\neg P$
$x + y$	$A \cup B$	$P \vee Q$
$x \cdot y$	$A \cap B$	$P \wedge Q$

The next step in developing an axiomatic system, after "deciding" on the axioms (usually this is a "work-in-progress" for a while, seeing how well the axioms capture the intended structure—even trying to best understand what is really "intended"—and modifying them as needed) is to see if they capture other essential aspects of the desired structure. In other words, to explore the consequences of the axioms. We'll do some of that now. But first, we must check that all the axioms actually are true in the intended interpretations.

You should be able to do the following exercises.

[4] Naturally enough, Boolean algebras were first conceived of by George Boole, but he used a slightly different notion of "+" than we do, and so had somewhat different axioms. His formulation is essentially that given in BAFact 5. Here are some relevant webpages: `http://www.maa.org/devlin/devlin_01_04.html`, `http://www.gutenberg.org/files/15114/15114-pdf.pdf`, as linked on the course webpage.

Exercise BA0: Prove all the basic axioms are true in each of the intended models.

Exercise BA1a: Prove that for all x, $x + 0 = x$ and $x \cdot 1 = x$.

Exercise BA1b (Characterizing negation): If $x + y = 1$ and $x \cdot y = 0$, then $y = -x$.

You will find this helpful in doing exercise BA2.4 and BA2.5 next.

Exercise BA2: Prove the following equations are consequences of the axioms for Boolean algebras

1. $x + x = x \quad x \cdot x = x$ 2. $x + 1 = 1 \quad x \cdot 0 = 0$
3. $-0 = 1 \quad -1 = 0$ 4. $-(x + y) = -x \cdot -y \quad -(x \cdot y) = -x + -y$
5. $- - x = x$

BAFact 1 (Duality) Notice that any equation involving the operations and constants of Boolean algebra may be "dualized": interchange + and $\cdot$ and interchange 0 and 1, to get another equation. If either one is true, then so is the other one. So, you only need to verify half the equations in most of the Exercises, the other half following by duality.

This is because each axiom comes in two versions, one dual to the other, so any proof using some of the axioms will yield a proof of the dual result using the dual axioms.

Once the basic structure of an axiom system is "clear", one usually wants to make sure that the system has **all** the structure one associates with the intended models, including structure that may not be explicit in the axioms. In our cases, logic and sets have more structure than we've seen so far: in logic, the notion of entailment ($p \vdash q$) is fundamental, and in sets, the notion of subset ($A \subseteq B$) is also fundamental, neither of which appear in the axioms or structure above. Do we have to extend our axiomatisation to include this? In some situations, such extensions prove necessary, but not here: this structure can be *defined* in any Boolean algebra, only using the structure we already have.

Exercise BA3: Define an order on a Boolean algebra as follows:

$$x \leq y \text{ if and only if } x = x \cdot y$$

Interpret this order in the two main models: what does it mean in each?

Prove the following:

1. $x \leq y$ iff[5] $y = x + y$
2. $\leq$ has the usual properties of a "partial order": for all x, y, z: $x \leq x$, $(x \leq y \wedge y \leq x) \longrightarrow x = x$, $(x \leq y \wedge y \leq z) \longrightarrow x \leq z$. (We say the order is "reflexive", "antisymmetric", and "transitive".)
3. $0 \leq x$ and $x \leq 1$ for all x. (So 0 is the least element and 1 the greatest element of the Boolean algebra.)
4. $x + y$ is the smallest element z satisfying $x \leq z$ and $y \leq z$ (we call such an element a *least upper bound*).
5. $x \cdot y$ is the largest element z satisfying $z \leq x$ and $z \leq y$ (we call such an element a *greatest lower bound*).

[5] Recall that "iff" means "if and only if".

Note that under duality, $\leq$ becomes $\geq$: $x \geq y$ iff $x = x + y$ iff $y = x \cdot y$ iff $y \leq x$.

After setting out the basic properties of the desired structure one's axioms are intended to capture, one then wants to see if there are other models (and perhaps even to see the relevance of those models to the general theory being developed). Here are two simple models of our theory.

Exercise BA4: Show that the set $\{0, 1\}$ is a Boolean algebra, in the only[6] way possible, given the equations in Exercise BA2. Explicitly write out what $0+0$, $0+1$, $1+1$, $0 \cdot 0$, $0 \cdot 1$, $1 \cdot 1$, -0, and -1 are. Verify all the axioms. (This Boolean algebra is often called 2.)

Exercise BA5: Show that for any positive square-free integer n (*i.e.* n has no divisors of the form k^2), the set of positive divisors of n forms a Boolean algebra, with order relation $a \leq b$ iff $a|b$. Identify what 0, 1, $-a$, $a+b$, and $a \cdot b$ are in this Boolean algebra, and verify that the order defined in Exercise BA3 coincides with the order given by $a|b$ (in other words, show that $a|b$ iff $b = a + b$ for the definition of $a + b$ you give here).

We could go further with this; here are a few possible extensions. Often one can find a model (or small class of models) that in some way characterize all models: $\{0, 1\}$ and sets $\mathcal{P}(X)$ have that property. Also, often the axiomatised structure can be "re-axiomatised" in an alternative way, which emphasises slightly different aspects of the intended interpretation, illuminating the structure in a different way. For example, Boolean algebras can be presented using the order relation. And we end our treatment of Boolean algebras by pointing out that Boole himself had a slightly different axiomatisation, based on exclusive "or" rather than the inclusive "or" (disjunction) we have used in this course. In the next section we shall see another way an axiomatic system may be modified, in order to capture a similar but different sort of model: Heyting algebras are given by a modification of the axioms for Boolean algebras, and they model logic without the double negation rule ($\neg\neg E$) in the same way Boolean algebras model ordinary (classical) logic.

The following "facts" are intended as "enrichment" material: you will not be examined on their contents, but they should help better understand some of the structure of Boolean algebras. Needless to say, this is only the tip of an iceberg ...

BAFact 2: The two-element Boolean algebra $\{0, 1\}$ (Exercise BA4) has a special property: any equation using the constants and operations of Boolean algebras which is true in $\{0, 1\}$ is also true in all Boolean algebras. (This means you can use truth tables to verify equations true in all Boolean algebras.)

BAFact 3: Every Boolean algebra may be "represented" as an algebra of subsets; in fact, every finite Boolean algebra may be "represented" as one of the form $\mathcal{P}(X)$ for some set X. Find an X so that the two-element Boolean algebra $\{0, 1\}$ is (*i.e.* may be interpreted as) $\mathcal{P}(X)$.

BAFact 4: Boolean algebras may also be defined in terms of the order relation $\leq$ (Exercise BA3): we suppose we have an order $\leq$ which is reflexive ($a \leq a$ for all a), antisymmetric ($a \leq b$ and $b \leq a$ implies $a = b$, for all a, b), and transitive ($a \leq b$ and $b \leq c$ implies $a \leq c$, for all a, b, c), with the properties that every pair of elements a, b has a least upper bound $a \sqcup b$ and a greatest lower bound $a \sqcap b$; that there is a least element 0 (so $0 \leq a$ for all a) and a greatest element 1 (so $a \leq 1$ for all a); the structure so far is often called a *bounded lattice*.

[6] Actually, there is another (somewhat perverse!) way this set may be given Boolean algebra structure, but the result is isomorphic to the obvious way given in the solutions. Can you see what that other way is?

We saw in Exercise BA3 that every Boolean algebra is naturally a bounded lattice. To capture the full structure of a Boolean algebra, we require further that each of the two operations $\sqcap, \sqcup$ are distributive with each other (so we have equations $a \sqcup (b \sqcap c) = (a \sqcup b) \sqcap (a \sqcup c)$ and $a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c)$); and that every element a has a unique complement $a^{\perp}$ (so $a \sqcap a^{\perp} = 0$ and $a \sqcup a^{\perp} = 1$). These additional properties are summarized by saying we have a *bounded distributive complemented lattice*; our claim is this is equivalent to being a Boolean algebra.

You might like to show that the definition of $\leq$ from Exercise BA3 shows how to go between this definition of a Boolean algebra and our original one; with this translation, $\sqcup, \sqcap$ correspond to $+, \cdot$ and $a^{\perp}$ corresponds to $-a$. The point of this is that we can replace the original axioms of Boolean algebras with more order-specific axioms, which for some models are easier to verify.

BAFact 5: Boolean algebras have another description, in terms of what are called "rings". A (commutative) ring (with 1) is a set with operations $\oplus$ and $\cdot$ which are commutative and associative, which have units 0 (for $\oplus$, so $a \oplus 0 = a$ for all a) and 1 (for $\cdot$, so $a \cdot 1 = a$ for all a), for which the distributive law $a \cdot (b \oplus c) = (a \cdot b) \oplus (a \cdot c)$ holds (for all a, b, c), and for which every element a has an additive inverse $-a$ (so $a \oplus (-a) = 0$).

The key idea here is that a Boolean algebra has an operation $\oplus$, given by $a \oplus b = (a \cdot -b) + (b \cdot -a)$. (What does this correspond to in the Boolean algebra of sets?) It is with respect to this modified form of "plus" that the elements of a Boolean algebra form a ring (using the usual $\cdot$). Rings formed this way have an extra property; they are "idempotent": $a \cdot a = a$ for every a. In fact, rings with this property are exactly Boolean algebras: we recapture the Boolean operation $+$ with the equation: $a + b = a \oplus b \oplus (a \cdot b)$. Notice, by the way, that regarding $\cdot$ as "and", and $+$ as (inclusive) "or", then $\oplus$ is exclusive "or".

9.2.2 Heyting algebras

In the early 20th century, some philosophically-minded mathematicians and logicians began to query the notion that for any formula P, $P \vee \neg P$ is true. The suggestion was that for one to know $P \vee \neg P$, you ought to know which was true. Similarly, if one claims that $\exists n P(n)$ is true, they ought to be able to tell you which n it is that justifies the claim. (This view of logic has become particularly relevant now, as it is related to the computational aspects of logic and hence to computer science.)

The resulting "intuitionist" logic is easy for us to formulate, because it has the same language as the logic we studied (which is now called "classical"), but omits just one of the derivation rules we used, namely the $(\neg\neg\ E)$ rule (the only rule to explicitly use $\neg\neg P \equiv P$). Many of the derived rules we had will also disappear, since they used the $(\neg\neg\ E)$ rule (for example, some of the de Morgan equivalences use it for one of the directions). In the resulting logic (without additional axioms or rules), it will be true that the only way to prove $P \vee Q$ is to prove either P or Q first, and the only way to prove $\exists n P(n)$ in intuitionistic arithmetic is to identify an actual numeral n and prove $P(n)$.

Fact: A statement P is provable in classical propositional logic if and only if $\neg\neg P$ is provable in intuitionist propositional logic. (And hence formulas of the form $\neg\neg P$ are provable in the one system iff they are provable in the other. In general, if P is provable intuitionistically, it's provable classically, but possibly not the other way round.)

(This is not true for predicate logic, but there is a translation of classical predicate logic into intuitionist predicate logic which allows a somewhat similar result to be stated.)

So, the question is: what algebraic system handles propositional intuitionist logic the way Boolean algebras handle classical propositional logic? The answer is Heyting algebras: using the notions introduced in BAFact 4 above, we can say a Heyting algebra is a bounded lattice (just as Boolean algebras are), which also has "relative pseudo-complements": for any x, y, there is an element $(x \Rightarrow y)$ with the property that it is the largest element z satisfying $x \sqcap z \leq y$. So $x \sqcap z \leq y$ iff $z \leq (x \Rightarrow y)$. The notation suggests that $x \Rightarrow y$ corresponds to the logical formula $P \longrightarrow Q$, and that is indeed exactly what it does, but of course in intuitionist logic.

HAFact 1: In a Heyting algebra, each element x has a "pseudo-complement" $\neg x = (x \Rightarrow 0)$. (We say an element z is a pseudo-complement of x if it is the greatest element with the property $x \sqcap z = 0$.) An element of a Heyting algebra is called "regular" if it satisfies $x = \neg\neg x$. Not too surprisingly, a Heyting algebra in which every element is regular is a Boolean algebra, and every Boolean algebra is a Heyting algebra (in which every element is regular). (Exercise: show that the bounded lattice structure of Boolean algebras gives the same structure required by a Heyting algebra: in other words, $\sqcup, \sqcap$ are given by $+, \cdot$ in Boolean algebras. What are $(x \Rightarrow y)$ and $\neg x$ in Boolean algebras?)

HAFact 2: Heyting algebras share some (but not all) properties of Boolean algebras; for example, one distributive law holds, $a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c)$, but not the other. Each de Morgan equivalence holds in one direction, but not the other; for example, $\neg P \vee \neg Q \ \vdash \ \neg(P \wedge Q)$ is true intuitionistically (and so when interpreted in Heyting algebras), since its proof doesn't use the $(\neg\neg\ E)$ rule (Exercise!), but the reverse direction does use that rule, so is only true classically (*i.e.* in Boolean algebras). So, Heyting algebras satisfy $\neg x \sqcup \neg y \leq \neg(x \sqcap y)$, but not the reverse.

HAFact 3: There are set models of Heyting algebras, which are not also Boolean algebras, but we have to "select" carefully which sets to use, usually using a "topological" criterion to select the sets. (Sorry—I won't define that term here!) Here is an example: take all "open" subsets of the real numbers, where an open set is a set with the property that for each element x in the set, there is an open interval containing x which is also in the set. This is not a Boolean algebra (since the complement of an open set isn't usually open—think of the complement of an open interval, for instance). But open sets do have open pseudo-complements (the interior of the usual set complement), and we do have a Heyting algebra.

9.2.3 Solutions to the exercises

Exercise BA0.

I leave this to you! You may use Venn diagrams for the set theory equations, and truth tables (or tableaux) for the propositional logic equivalences.

Exercise BA1a.

$x = x \cdot (x + (-x)) = x \cdot 1$ (and dual)

Exercise BA1b.

Remark: If $x + y = 1$ and $x \cdot y = 0$, then $y = -x$ (in words: "If you add two elements and get 1 and multiplying them gives you 0, then the elements must be 'negatives'.") Since this is true for x and $-x$, this property characterizes negation in terms of two equations it must satisfy. This is an important fact about negation. In BA3 we shall see similar characterizations of $+, \cdot, 0, 1$ in terms of inequalities they must satisfy, so all the structure of Boolean algebras may be described in terms of properties the various operators satisfy.

Proof: $y = y + 0 = y + (x \cdot (-x)) = (y + x) \cdot (y + (-x)) = 1 \cdot (y + (-x)) = y + (-x) = (-x + y) \cdot (-x + x) = (x \cdot y) + (-x) = -x$

Exercise BA2.

1. $x = x \cdot (x + 0) = x \cdot x$ (and dual)
2. $x + 1 = x + (x + (-x)) = x + (-x) = 1$ (and dual)
3. $-0 = 0 + (-0) = 1$
4. We use BA1b:

   $(x + y) + ((-x) \cdot (-y)) = x + ((y + (-x)) \cdot (y + (-y))) = x + (-x) + y = 1$ and
   $(x + y) \cdot ((-x) \cdot (-y)) = ((x \cdot (-x)) + (y \cdot (-x))) \cdot (-y) = y \cdot (-x \cdot (-y)) = y \cdot (-y) \cdot (-x) = 0.$

   So $-(x + y) = (-x) \cdot (-y)$. (The dual is dual!)
5. (Use BA1b and commutativity.)

Exercise BA3.

1. Suppose $x = x \cdot y$: then $x + y = x \cdot y + y = y$.

   And suppose $y = x + y$: then $x \cdot y = x \cdot (x + y) = x$
2. In order: $x = x \cdot x$; if $x = x \cdot y$ and $y = y \cdot x$, then $x = x \cdot y = y \cdot x = y$; if $x = x \cdot y$ and $y = y \cdot z$ then $x = x \cdot (y \cdot z) = (x \cdot y) \cdot z = x \cdot z$.
3. $0 = 0 \cdot x$ and $x = x \cdot 1$
4. This essentially means [$z = x + z$ and $z = y + z$] iff $z = x + y + z$. Suppose $z = x + z$ and $z = y + z$: then $z = x + z = x + (y + z)$. On the other hand, suppose $z = x + y + z$: then $x + z = x + x + y + z = x + y + z = z$ and similarly $y + z = z$.
5. is dual to 4.

Exercise BA4.

Notice that $0 + 0 = 0$, $0 + 1 = 1 + 1 = 1$, $0 \cdot 0 = 0 \cdot 1 = 0$, $1 \cdot 1 = 1$, $-0 = 1$, $-1 = 0$; verifying this satisfies the 10 axioms is a simple matter I'll leave to you.

Exercise BA5.

Again, just the basics: 0 is the number 1. 1 is the number n. $-a$ is the number n/a (it is an integer since $a|n$). $a + b$ is the least common multiple of a, b; $a \cdot b$ is the highest common factor of a, b. And so $a = a \cdot b$ means $a|b$, so the order is the same with the Boolean algebra definition as with the definition of this exercise.

9.3 Groups

Next we shall look at some more "mathematical" (or "algebraic") systems, specifically at permutations and groups.

Algebra is usually taught as if it were just symbolic arithmetic. That is, one learns to use letters as variables to stand for numbers. Modern algebra is more abstract. The things the variables stand for don't have to be numbers. We can explore arithmetic operations and relations without thinking about numeric interpretations. The freedom that results from seeing mathematics as the construction of postulational systems allows us to construct algebras of a quite general nature.

An algebraic structure or algebraic system (or simply "an algebra") is a system consisting of a set of elements or objects, together with operations on and relations between those objects. Ordinary arithmetic on integers is one algebra. Changing the set of objects to rational numbers gives a different (but similar) algebra. Defining different arithmetic operators or relations gives other algebras. Mathematicians explore the properties of various abstract algebras, including the properties that many different algebras share.

Although an algebraic system can be freely invented with no particular interpretation, it is more common, and more useful, to construct an algebraic system based on something significant. We can then illustrate how one explores the properties of the system that results.

One very important structure is the structure of *operations*; we shall start with a concrete example, *permutations*, which are operations on sets which merely alter the *order* of the elements.

9.3.1 Permutations

To make sense of this, we shall consider *ordered* sets, or ordered lists, rather than (ordinary) sets. If an ordered set has n elements, we often call it an "n-tuple" (for example, a "2-tuple" is just an ordered pair).

Since order matters, two ordered lists are distinct if the elements are listed in different orders. Each distinct way of ordering the n elements of an n-tuple is called a *permutation* of those elements. "Permutation" is also the word for the operation of permuting the elements of an ordered n-tuple into a different ordering.

How many permutations of two elements a and b can we distinguish? The elements can be put into ordered pairs as $\langle a, b\rangle$ or as $\langle b, a\rangle$. If there are three elements a, b and c, they can be ordered in six distinct ways, so there are six permutations of three elements: $\langle a, b, c\rangle$, $\langle a, c, b\rangle$, $\langle b, a, c\rangle$, $\langle b, c, a\rangle$, $\langle c, a, b\rangle$, and $\langle c, b, a\rangle$.

We develop a notation. Let's say that (21) specifies a permutation of an ordered pair that puts the first element into the second place and the second element into the first place. Let's call that permutation A: $A = (21)$. When we apply the permutation A to the ordered pair $\langle a, b\rangle$, we get $\langle b, a\rangle$; we write this with function notation: $A\langle a, b\rangle = \langle b, a\rangle$. Another permutation on ordered pairs would be (12), denoted $\mathbf{1}$. This puts the first item in the pair into the first position, and puts the second item into the second position—in other words, it does nothing(!). We call this the identity permutation; it leaves the order unchanged: $\mathbf{1}\langle a, b\rangle = \langle a, b\rangle$.

This is an example of *reification* ("thingification"). We started with ordered pairs and treated them as things. Then we treated operations on ordered pairs as things, and gave the operations names, calling them $\mathbf{1}$ and A.

Working on triples (3-tuples) is more interesting. There are 6 permutations of triples, corresponding to the 6 different ways one can reorder a triple. Let's use the same notation as for permutations on pairs, where we indicate where each element of a triple is moved to (so, for instance, the permutation $A = (132)$ means move the first element to the first position, the second element to the third position, and the third element to the second position). We can list all the permutations of 3-tuples as: $1 = (123)$, $A = (132)$, $B = (213)$, $C = (231)$, $D = (312)$ and $E = (321)$. If we apply the E permutation to $\langle a, b, c\rangle$ we get $\langle c, b, a\rangle$. If we apply the E permutation to $\langle c, b, a\rangle$ we get $\langle a, b, c\rangle$. So applying the E permutation twice is like applying the identity permutation. Also, $C\langle a, b, c\rangle = \langle c, a, b\rangle$; $C\langle c, a, b\rangle = \langle b, c, a\rangle$, so, by applying C twice to $\langle a, b, c\rangle$ we get $\langle b, c, a\rangle$, which is the same thing we'd get by applying D to $\langle a, b, c\rangle$. We can describe this as $CC\langle a, b, c\rangle = D\langle a, b, c\rangle$.

What happens if you apply E to the result of applying A to $\langle a, b, c\rangle$? Applying A to $\langle a, b, c\rangle$ gives you $\langle a, c, b\rangle$. Applying E to $\langle a, c, b\rangle$ gives $\langle b, c, a\rangle$. This is just what you'd get from $D\langle a, b, c\rangle$. So $EA\langle a, b, c\rangle = D\langle a, b, c\rangle$.

The point here is that applying a permutation to the result of applying a permutation to a triple is an operation (as addition and multiplication are operations): it assigns a new permutation to two given permutations. We call this operation "composition" of permutations. We use the $\circ$ symbol to represent composition of permutations,[7] so $F \circ G$ is defined by $(F \circ G)\langle a, b, c\rangle = FG\langle a, b, c\rangle$ for any triple $\langle a, b, c\rangle$. We define equality of permutations by $F = G$ if $F\langle a, b, c\rangle = G\langle a, b, c\rangle$ for any triple $\langle a, b, c\rangle$. Then, for example, in the previous paragraph we showed that $E \circ A = D$.

We could summarize the statements about the results of all possible compositions of permutations on triples in a table, rather like a multiplication table. In this way, we would define the behavior of the composition operation, and in effect create an algebraic system consisting of six elements 1, A, B, C, D, and E, and the operation $\circ$. Its creation was motivated by the notion of permutations of ordered triples, but now that we've got it, we can study it on its own, unrelated to its origins.

We said that our operator symbol $\circ$ was "like $+$ or $\times$ in ordinary arithmetic". How much (or little) does it resemble those common arithmetic operators? For example, is our $\circ$ operator associative? Now we're treating the operation $\circ$ as a thing! If you are feeling energetic, you could write down all possible three-fold products and verify that $\circ$ is associative. But there is a good conceptual reason for believing that it is, which would save you the trouble: permutations are functions on ordered triples, and $\circ$ is ordinary composition of functions, which you already know is associative. This consideration should convince you that another common property is *not* true of $\circ$: it is not likely to be commutative, since ordinary function composition isn't commutative.

Another property $\circ$ shares with multiplication is the existence of a unit: 1 has the property that $1 \circ F = F = F \circ 1$ for any permutation F. And every permutation has an inverse: just reverse the rearrangement corresponding to the permutation.

We could do the same thing for any length of ordered n-tuples; indeed, we could do this for infinite ordered sets as well: a permutation is a rearrangement of the elements in an ordered set (a one-to-one correspondence of the set with itself, in effect), and as such may be treated like a function; permutations may be composed, just as functions may be composed. There is an identity permutation 1, and every permutation has an inverse. These are properties true of the collection of permutations on any ordered set, and are true of "reversible operations" in general, and so are probably worth abstracting to an algebraic system. Such structures are called *groups*.

9.3.2 Groups

Definition: A group is an algebraic system $\langle G, \circ\rangle$ consisting of a nonempty set G of "elements" and one binary operation $\circ$ on G satisfying the following four axioms:

[7] Actually, frequently one drops the use of the symbol $\circ$, simply using juxtaposition, just as we do with multiplication in high school algebra (where ab represents the product $a \cdot b$). You can see why by looking at the previous paragraphs.

G1: To every pair of elements a and b of G, given in the stated order, there corresponds a definite unique element of G, denoted by $a \circ b$.[8]

G2: For all $a, b, c \in G$, $(a \circ b) \circ c = a \circ (b \circ c)$.[9]

G3: There exists an element ι in G such that, for any a in G, $a \circ \iota = \iota \circ a = a$. The element ι is called an *identity element* of the group. We can prove that a group will never contain more than one identity element.

G4: For each element a of G there is an element a' of G such that $a \circ a' = a' \circ a = \iota$. The element a' is called the *inverse* of a. We can prove that an element a of a group possesses only one inverse element.

If the following axiom is also satisfied,[10] the group is called a *commutative* or *Abelian* group.

G5: For all $a, b \in G$, $a \circ b = b \circ a$. This is the commutative law for the operation $\circ$.

A group for which axiom G5 does *not* hold is called a *non-Abelian* group. If the set G of a group contains only a finite number of distinct elements, the group is called a finite group; otherwise it is called an infinite group.

Permutations of n-tuples (for $n \in \mathbb{N}$) are examples of finite groups. Another way of saying this is to say that they are interpretations of the axioms that define a finite group. The identity element in each case is the element we called 1.

We can check that composition of permutations on pairs represents an Abelian group, but composition of permutations on triples does not. So permutations on ordered pairs is an Abelian group, but permutations on ordered triples is non-Abelian (as is the set of permutations (with composition) on any finite set of cardinality greater than 2).

9.3.3 Exercise on groups

1. Prove that a group can never have more than one identity element.

2. Prove that an element of a group cannot have more than one inverse.

3. Consider a set of two objects $\{E, O\}$ (think of E as standing for "even" and O as standing for "odd"). Make a table for an operation $\circ$ on $\{E, O\}$, where the "product" is odd or even depending on whether the sum of the two operands is odd or even. For example, the sum of two even numbers is an even number, so $E \circ E = E$, and the sum of an even plus an odd number is odd, so $E \circ O = O$. Does this table specify a group (*i.e.*, does $\langle \{E, O\}, \circ \rangle$ satisfy the definition of "group")? Is this like (isomorphic to) any group we have already considered?

4. ("Clock arithmetic") Imagine a "clock" having the numbers 0, 1, 2, 3, 4, 5, 6 equally spaced on its face. To find the sum of any two numbers in that set of numbers, we start at the first number and move clockwise around the dial a number of spaces equal to the second number. Thus, the "sum" of $2 + 3$ is found by starting at 2 and moving clockwise three places (to 3, then 4, then 5). The "sum" is 5. $4 + 5$ is found by starting at the 4 position and moving through 5, 6, 0, 1, 2, giving a "sum" of 2. Write out the table of clock-arithmetic "sums" formed from the set $\{0, 1, 2, 3, 4, 5, 6\}$. Does this table define a group? Why or why not?

[8] This says that G is "closed" under the operation $\circ$: for any $a, b \in G$, $a \circ b \in G$.

[9] This is the *associative law* for the operation $\circ$.

[10] That is, if the operation $\circ$ is commutative.

5. Each of the following algebraic systems satisfies the definition of an infinite group. In each example, name the identity element and specify the inverse x' for each element x.
   (a) $\langle \mathbb{Z}, + \rangle$ (*i.e.*, the set of integers and the operation of ordinary arithmetic addition).
   (b) $\langle \mathbb{Q}^{-}, \cdot \rangle$ (*i.e.*, the set of rational numbers *excluding* 0 and the operation of ordinary arithmetic multiplication). Why was it necessary to exclude 0?

6. Explain why the following algebraic systems are *not* infinite groups:
   (a) $\langle \mathbb{Z}, \cdot \rangle$ (*i.e.*, the set of integers and the operation of ordinary arithmetic multiplication).
   (b) $\langle \mathbb{N}, + \rangle$ (*i.e.*, the set of natural numbers with ordinary addition).

In this exercise we saw that parts of our ordinary arithmetic systems can be studied as interpretations of the abstract algebraic systems called groups. We can also dream up new "arithmetics" that are groups.

We can go on to study the abstraction itself. We can derive theorems from the postulates, and know that any system that satisfies the definition of a group will be such that all the theorems will be true for that system. Many books and articles in scientific journals are devoted to the study of groups. In fact, it has been a major field of active research in mathematics for most of the past century.

One result of "group theory" is the "representation theorem": every group is isomorphic to a (sub)group of permutations on some set. So our motivating example of permutations was more than merely motivation: it turns out that there are no other essentially different groups.

Beyond this, mathematicians study more complicated abstract algebraic systems like rings, subrings, ideals, integral domains, ordered integral domains, and fields. These involve more than one operator and additional postulates and properties. One or the other system will describe almost any arithmetic system, including those that we invent with our new kinds of numbers. In studying such abstract systems, we can study our ordinary arithmetic at a level behind the mere numbers, and discover isomorphisms between arithmetic systems and other deductive structures.

Chapter 10

Endgame

We shall end the course with one of the remaining topics. Constraints of time will probably mean only one.

10.1 Gödel's Incompleteness Theorems

Introduction

Late in the nineteenth century, early in the twentieth century, a number of logical and mathematical paradoxes revealed that the logical or foundational underpinnings of the mathematical universe had some structural cracks in the fabric, or so it seemed. (Some of the paradoxes are discussed in Chapter 1.) It was hoped that perhaps the axiomatic method might help settle these worries, by developing an axiom system for mathematics itself, and then proving the axioms to be adequate for the job, in that all their logical consequences would account for all mathematical truth. This was one of the most pressing mathematical questions at the time, in the view of a substantial part of the mathematical community. At an international conference in Paris in 1900, David Hilbert (one of the leading mathematicians of the day) posed 23 problems that he thought would be the most significant mathematical challenges for the 20th century; the first on the list was the "continuum hypothesis" (mentioned in Chapter 7), and the second was the consistency of the axioms for arithmetic. Exactly what "axioms" he had in mind is not entirely clear, but before 1930 it's probably not too unreasonable to imagine that most mathematicians who thought about the problem would have agreed that the axioms of Peano, as presented in the 1910 Whitehead–Russell system of *Principia Mathematica*, would have been acceptable. The hope was that they might provide a firm foundation for mathematics *via* a consistent set of axioms, so that "truth" might be understood as simply "provability" in a formal system. But this hope was to prove in vain! In 1931, a young Austrian mathematical logician, Kurt Gödel, proved that this was simply not possible.

What is to be proved[1]

Gödel proved that first order arithmetic cannot be effectively, consistently, and completely axiomatized; that is, no effectively specified axiom system for first-order arithmetic can be both sound (consistent) and complete. Gödel used the system *Principia Mathematica, PM* for short, but any other suitable system S would do.

[1] Based on notes for COM3412, a course on Logic and Computation at the University of Exeter. There's a link to those notes on my webpage.

To be definite, we'll take the following system S:[2] its language will include a constant 0, a unary symbol s (so we can write terms like $ssss0$), the usual logical symbols and logical rules, and the following axioms.

$$\begin{array}{l} \forall x(\neg\; sx = 0) \\ \forall x \forall x'(sx = sx' \to x = x') \\ \varphi(0) \wedge \forall x(\varphi(x) \to \varphi(sx)) \to \forall x \varphi(x) \end{array}$$

for any well-formed formula $\varphi(x)$ written in the language of S.

The essential idea is as follows.[3] A system S for first-order arithmetic has a standard interpretation: to every formula φ of S there is associated a proposition $\lceil \varphi \rceil$ of arithmetic. 0 is interpreted as the number 0, and s is "successor" or "1+", *i.e.* sx is interpreted as the number following the interpretation of x. So, for example, the first axiom above would be interpreted as the proposition "$0 \neq n+1$ for any natural number n". It's easy (in principle) to show that this system is sufficient to express all the usual operations and statements about arithmetic. For instance we can define + and × inductively:

$$\begin{array}{rclcrcl} n+0 &=& n & \qquad & n+sm &=& s(n+m) \\ n \times 0 &=& 0 & & n \times sm &=& n \times m + n \end{array}$$

and we can then prove theorems in S which correspond to the standard theorems of arithmetic, like $k \times (m+n) = (k \times m) + (k \times n)$, and many more. (For example, the system is sufficient to prove formal versions of all the theorems we proved in class, such as the Fundamental Theorem of Arithmetic.)

We say S is *sound* (or *consistent*) if whenever S proves a formula φ, the standard interpretation $\lceil \varphi \rceil$ is actually true: $S \vdash \varphi$ implies $\lceil \varphi \rceil$. We say S is *complete* if S proves all formulas which are true under the standard interpretation: $\lceil \varphi \rceil$ implies $S \vdash \varphi$. (So S is sound and complete means it proves exactly true statements, no more, no less.) Most mathematicians would regard it as a fact that S is sound, although that could be taken as a statement of faith. Clearly, if it isn't then we've got lots of potential problems with our use of numbers and all that depends on them! However, completeness is a totally different matter—although mathematicians before Gödel hoped a system like S might be complete, they had no proof, nor any real reason to believe it (other than wishful thinking!). Following Gödel's theorem, we now know that it is not complete, for completeness would imply arithmetic was inconsistent.

The tricky thing Gödel introduced was the construction of an alternative interpretation: to any appropriate formula φ, he constructed a statement $[\![\varphi]\!]$, a statement actually about the system S itself, in such a way that $[\![\varphi]\!]$ is true if and only if $\lceil \varphi \rceil$ is true. Moreover, he did this in a way that allowed him to construct a formula g in S (and so formally g is just a formula of first-order arithmetic, just about numbers) whose interpretation $[\![\mathsf{g}]\!]$ is "g is not provable in S". (g is often called a Gödel formula for the system S.)

Here's the point: g has both the standard interpretation $\lceil \mathsf{g} \rceil$ as well as the Gödelian interpretation $[\![\mathsf{g}]\!]$; $\lceil \mathsf{g} \rceil$ may be true or false. If it's true, then $[\![\mathsf{g}]\!]$ is also true, which means that g is not provable in S: in short, we've got a formula which is true (in its standard interpretation) but is not provable, so S is not complete. On the other hand, if $\lceil \mathsf{g} \rceil$ is false, then $[\![\mathsf{g}]\!]$ is also false, which means that g is provable: so we've got a formula that's false (in the standard interpretation) but

[2] These axioms are usually called the Peano axioms for the natural numbers, after the Italian mathematician who first proposed them.

[3] I must make a disclaimer: there are many informalities in my text; some points are over-simplified in an attempt to make the gist easier to follow. One such is that in his original paper, Gödel actually needed a slightly more complicated notion than *consistency*; that problem was quickly sorted out in the next few years, by J.B. Rosser, so that ordinary consistency sufficed.

provable; in other words, S isn't sound. (It is reasonable in the case of our system S to say this proves $\ulcorner \mathsf{g} \urcorner$ is in fact true.)

The hard part of all this is getting things set up so the Gödelian interpretation is possible. That's the core of Gödel's proof, and we'll sketch the main details next.

10.1.1 Gödel's interpretation

Here's what Gödel has to say (with some informality in the translation!).

> The formulas of a formal system (such as PM) may be viewed as finite sequences of basic symbols ... and one can state precisely *which* such sequences are well-formed formulas. Similarly proofs are just finite sequences of formulas (with appropriately specified properties). Since it is irrelevant just what symbols are used, we may use natural numbers. So a formula will be a finite sequence of natural numbers, and a proof scheme will be a finite sequence of such finite sequences of natural numbers. In this way, meta-mathematical concepts or theorems become concepts or theorems about natural numbers, which makes them (at least partially) expressible in the language of PM. In particular, we can show that the concepts "formula", "proof scheme", "provable formula" all become expressible in the system PM, so we can, for example, construct a formula $F(v)$ with one free variable v whose semantic interpretation is "v is a provable formula". We then construct an undecidable proposition of the system PM, that is, a formula A for which neither A nor $\neg A$ is provable, as follows.
>
> Let's call formulas of PM with exactly one free variable "class-symbols". List the class-symbols in some way, denoting the n^{th} one by $\mathcal{R}_n$. (This is all definable in PM.) Let $\mathcal{A}$ be a class-symbol: by $\mathcal{A}(n)$ we denote the result of substituting n for the free variable in $\mathcal{A}$. The relation $\mathcal{B} \equiv \mathcal{A}(n)$ is also expressible in PM. So, we can define a class K of natural numbers as follows:
>
> $$K = \{n \in \mathbb{N} \mid \neg\mathsf{provable}(\mathcal{R}_n(n))\}$$
>
> where $\mathsf{provable}(x)$ means x is a provable formula in PM. Since all these notions can be defined in PM, so is K, and so there's a class-symbol $\mathcal{K}$ so that $\mathcal{K}(n)$ states that $n \in K$. But $\mathcal{K}$ must be identical to one of the $\mathcal{R}_q$, for some number q, so
>
> $$\mathcal{K} \equiv \mathcal{R}_q$$
>
> $\mathcal{R}_q(q)$ is undecidable in PM, for if $\mathcal{R}_q(q)$ were provable, then it would be also true. But in that case, by the definitions above, $q \in K$, and so $\neg\mathsf{provable}(\mathcal{R}_q(q))$, contradicting our assumption. If on the other hand, $\neg\mathcal{R}_q(q)$ were provable, then $q \notin K$, and so $\mathsf{provable}(\mathcal{R}_q(q))$. This would mean $\mathcal{R}_q(q)$ and its negation would both be provable, which is impossible.

(Note that $\mathcal{R}_q(q)$ is the Gödel formula g mentioned previously.) The rest of Gödel's paper is devoted to proving these claims.

10.1.2 Gödel numbering

We start by coding the symbols, formulas, and sequences of formulas of arithmetic as follows.

The symbol:	0	s	$=$	$\neg$	$\vee$	$\forall$	$($	$)$	x	$'$
is coded as:	1	3	5	7	9	11	13	15	17	19

Note that these symbols are sufficient for all the logic and arithmetic of S. For example, we can use de Morgan equivalences to represent $\wedge$, $\longrightarrow$, $\exists$, and we can use variables $x', x'', x''', x'''', \ldots$ instead of the more user-friendly $x, y, z, m, n, \ldots$.

Strings of symbols are then encoded by using these numerical codes as exponents of successive prime numbers, so, for example, $\forall x(\neg\ sx = 0)$ would be encoded as $2^{11}3^{17}5^{13}7^{7}11^{3}13^{17}17^{5}19^{1}23^{15}$:

$$\begin{array}{ccccccccc} \forall & x & (& \neg & s & x & = & 0 &) \\ 2^{11} & 3^{17} & 5^{13} & 7^{7} & 11^{3} & 13^{17} & 17^{5} & 19^{1} & 23^{15} \end{array}$$

(this is a **very** large number, approximately 1.55×10^{70}).

In the same way, sequences of such strings may be encoded. If $G(\varphi)$ is the number coding the formula φ, then a sequence $\varphi_1, \varphi_2, \varphi_3, \ldots$ may be encoded by

$$2^{G(\varphi_1)} \times 3^{G(\varphi_2)} \times 5^{G(\varphi_3)} \times \ldots .$$

Note that we can distinguish between the codes of single formulas and the codes of strings of formulas, since the latter are all perfect squares (since all codes $G(\varphi)$ are even numbers), whereas the former are never perfect squares (since all the basic codes give odd exponents). Note that given any number, using the Fundamental Theorem of Arithmetic we can factor it into powers of primes, and so we can see if it corresponds to a formula of arithmetic, or to a sequence of such formulas. In other words, coded formulas or sequences of formulas can be decoded as well.

The key point now is that logical properties of formulas and sequences of formulas (such as a formula being well-formed or a sequence of formulas being a proof of some formula) are translated into arithmetical properties of its Gödel number (*i.e.* of the number which codes it). A simple example: a formula starts with $\neg$ if and only if its Gödel number is divisible by 2^7 and is not divisible by 2^8. What took most of the work in Gödel's proof was establishing such correspondences for logical properties such as "formula φ is provable from the axioms of S". But by doing that, he managed to make it possible for system S to talk about itself, in addition to talking about arithmetic, and so he set the scene for the self-referential paradox that proves undecidability.

10.1.3 Some details

First, one constructs formulas which say such things as "x is a well-formed formula", "x is a substitution instance of one of the axioms of S", "x is a pair of well-formed formulas, the second of which follows from the first", "y is a sequence of formulas which comprises a valid proof in S of a formula φ". The main objective is thus accomplished: one constructs a formula $P(x, y)$ (of system S) which says that x is the Gödel number of a formula $\varphi(z)$ containing one free variable, and that y is the Gödel number of a proof of the formula $\varphi(x)$ obtained by substituting x for the free variable z in $\varphi(z)$.

[There is a technical point I should make here: for this to be possible, there must be effective algorithms which make it possible to decide the corresponding properties of system S. For example, it must be possible to effectively decide (regarding the logic) if a given sequence of (real) formulas constitutes a valid proof in S of a given formula. To deal with this, Gödel had to more or less invent a field of mathematics dealing with the question of exactly just what it means to be an effective algorithm.]

Next, one looks at the formula $\forall y \neg P(x, y)$. (It has just one free variable x, y being bound.) This formula says in essence that the formula $\varphi(x)$ cannot be proved in system S, where φ is the formula whose Gödel number is x. Suppose the Gödel number of $\forall y \neg P(x, y)$ is g, and now consider $\forall y \neg P(g, y)$. Think about this a bit (the idea is discussed in the various references I've given you, if you need a bit of help!). In essence it says that there is no y which is the code of a proof of

$\varphi(g)$, where φ is the formula whose code is g: *i.e.* φ is $\forall y \neg P(x, y)$ itself, and so $\varphi(g)$ is $\forall y \neg P(g, y)$. In other words, the formula is saying that it itself is not provable. It is the Gödel formula g I mentioned at the start.

So, we can conclude (by the argument given at the start of these notes) that S isn't both sound and complete.

Coda 1: There is a further twist to this result: it not only shows that there is a sentence g which is not provable, but also it follows from this that the consistency of system S is not provable within system S itself (unless S is inconsistent, in which case anything and everything is provable in the system!). Here's why.

We have just shown that *if S is consistent, then* g *is not provable, nor is* $\neg\mathsf{g}$. The coding we developed is expressive enough for us to actually encode this statement into a formula of S. We can code "S is consistent" by saying $\neg\mathsf{provable}(0 = 1)$, and so we might code the entire statement something like this:

$$\neg\mathsf{provable}(0 = 1) \longrightarrow \neg\mathsf{provable}(\mathsf{g}) \wedge \neg\mathsf{provable}(\neg\mathsf{g})$$

Our point in the previous discussion is that this statement is in fact provable in S. But now suppose that $\neg\mathsf{provable}(0 = 1)$ (*i.e.* "S is consistent") is also provable in S. Then (by *Modus Ponens*, our old friend $(\longrightarrow E)$) we would have also shown that $\neg\mathsf{provable}(\mathsf{g}) \wedge \neg\mathsf{provable}(\neg\mathsf{g})$, and hence $\neg\mathsf{provable}(\mathsf{g})$ are provable. But the interpretation of g is just $\neg\mathsf{provable}(\mathsf{g})$, and (assuming the consistency of S) we know that this then is *not* provable. This contradiction gives us our conclusion: if S is consistent, then S *cannot* prove that fact: $\neg\mathsf{provable}(0 = 1)$ is not provable in S. This result is often called "Gödel's second incompleteness theorem".

A simple extension of this result shows that S cannot prove the consistency of any other theory strong enough to prove S's consistency, so there really is no way out of this situation. If you want to prove the consistency of a system like S, you really need something that truly goes beyond S, in an essential way.

Coda 2 (Tarski's theorem): A variant of this method (indeed, a simplification thereof) also shows a result usually attributed to Tarski, namely that system S cannot have an internal notion of *truth*, in the following sense: there cannot be a one-variable predicate true with the property that for any sentence φ,

$$\mathsf{true}([\![\varphi]\!]) \leftrightarrow \varphi$$

holds in S. In other words, "truth" for arithmetic cannot be defined within arithmetic itself. This result can be proved by a method ("diagonalization") similar to the preceeding proof of incompleteness, and like that result, holds more generally than merely within system S. In a sense, Tarski's result is simpler, requiring only the existence of a Gödel numbering and a logic including negation. And just as Gödel's result may be regarded as the internalization of "this sentence cannot be proved", Tarski's result is the internalization of the original liar paradox "this statement is false".

To show how to prove Tarski's theorem, we shall modify the presentation from subsection 10.1.1, imagining there is an enumeration of one-variable predicates ("class-symbols") $\mathcal{R}_n$, and a class L (analogous to K)

$$L = \{n \in \mathbb{N} \mid \neg\mathsf{true}(\left[\!\!\left[\mathcal{R}_n(n)\right]\!\!\right])\}$$

of certain "untrue" predicates. As with K, L is a class in S (since true is assumed to be a predicate of S), and so there must be a class-symbol (*i.e.* a predicate) $\mathcal{L}$ so that $\mathcal{L}(n)$ if and only if $n \in L$, as well as a number q so that $\mathcal{L} \equiv \mathcal{R}_q$. But then we have a contradiction, for either $\mathcal{R}_q(q)$ is true or $\neg\mathcal{R}_q(q)$ is true. If $\mathcal{R}_q(q)$ is true, then $q \in L$ by definition of L, and so $\neg\mathsf{true}(\left[\!\!\left[\mathcal{R}_q(q)\right]\!\!\right])$, *i.e.* $\neg\mathcal{R}_q(q)$ is true (a contradiction). The reverse direction is similar. So the existence of the true predicate is impossible.

10.1.4 So . . .

What does this mean? There are plenty of things it doesn't mean (I've given a link on my webpage dealing with this!), such as "mathematics is indeterminate, you cannot determine what's true and what's not using mathematical methods". Many folks have tried to use Gödel's theorem to justify claims such as "mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth", and that "relativity theory and Gödel's theorems show that even in physics and mathematics there's no objective truth and rationality". Gödel himself saw nothing of the sort in his theorems—he was no post-modernist, seeing uncertainty and relativity everywhere. On the contrary, he was a mathematical Platonist, believing that mathematical truth was grounded in reality. And his proof showed that what we can know as true exceeds what we can represent by formal methods. His methods were mathematical in the best sense (in the only sense?); they are not mere philosophical hand-waving. They allow for no doubt about the truth of his conclusions: you would have to deny the validity of logic itself (at which point, anything you say is cast into doubt, including your denial itself!).

For further information, there are a number of links on the course webpage on various related topics. You should definitely consult some of them.

10.1.5 Exercises

1. Using the definitions of $+, \times$ given for system S above, prove (by using mathematical induction on n) that $k + (m + n) = (k + m) + n$ for any natural numbers k, m, n.

2. Similarly, show that $k \times (m + n) = (k \times m) + (k \times n)$ for any natural numbers k, m, n.

3. (Optional) Show associativity of $\times$ and commutativity of $+, \times$.

4. Prove (by induction) that in system S, every element is either 0 or a successor (*i.e.* has the form sx for some x).

10.1.6 Answers

The point of this exercise is to be careful about what equations you are allowed to use; you may *only* use the equations that *define* $+$ and $\times$ and the induction assumption in the induction step.

For instance, the equations defining $+$ are:

$$m + 0 = m, \quad m + (n + 1) = (m + n) + 1$$

These are true *for all* m, n (so you can replace m, n by any other expressions as needed—for example, if I replace m with $(k + m)$, then the first defining equation says $(k + m) + 0 = (k + m)$, which I shall use below).

So here's the proof for question 1: note that we treat k, m as parameters, and only do the induction on n.

Base case ($n = 0$):

$$\begin{aligned} k + (m + 0) &= k + m \quad [\text{by definition } m + 0 = m] \\ &= (k + m) + 0 \quad [\text{by definition } (k + m) + 0 = (k + m)] \end{aligned}$$

Induction step: Assume $k+(m+n)=(k+m)+n$ and prove $k+(m+(n+1))=(k+m)+(n+1)$.

$$\begin{array}{rcll}
k+(m+(n+1)) & = & k+((m+n)+1) & \text{[by definition } m+(n+1)=(m+n)+1\text{]} \\
 & = & (k+(m+n))+1 & \text{[by definition } k+((m+n)+1)=(k+(m+n))+1\text{]} \\
 & = & ((k+m)+n)+1 & \text{[by induction assumption]} \\
 & = & (k+m)+(n+1) & \text{[by definition } (k+m)+(n+1)=((k+m)+n)+1\text{]}
\end{array}$$

(Note again that I have replaced other expressions for the simple k, m in the defining equations for $+$, as needed.) (QED)

I will leave you to do the second one yourself. As for the optional ones (if you choose): associativity for × is straightforward, though it uses distributivity, but to prove commutativity for $+, \times$, one must be a bit more clever. For instance, you can use induction, but you will also have to use induction again in both the base case and the induction step, amounting to a sort of "double induction".

And the last question? Here's a hint: it is really easy! There is really nothing to prove; set up the induction steps, and you're done.

10.2 Categorial Grammar

Introduction

We conclude the course with an application of some of the ideas we have seen to linguistics. Earlier (in the previous chapter), we considered two structures presented axiomatically, each dealing with the algebraic structure of logic: Boolean algebras, which correspond to classical propositional logic, and Heyting algebras, which correspond to intuitionist propositional logic, which may be thought of as propositional logic *without* the "double negation" rule ($\neg\neg\, E$). We also met with the notion of partial order which underlies these structures (Exercise BA3). In this section, we shall use a certain kind of partial order, called a pregroup, to analyse sentence structure in natural language; this will illustrate how the structure of logic appears naturally in linguistics. As a further illustration of the use of pregroups in grammar, we shall sketch an application to music.

Then we shall remark that there is an algebraic structure which shares some of the features of orders, groups, and deductive logic, and shall axiomatize this common structure to obtain the notion of a *category*. We shall consider some special kinds of categories, which capture the structure of some curious logics, and end by showing that the pregroup analysis of sentence structure can also be done using an appropriate categorical logic.

Disclaimer: In these notes I am aiming more at giving the "flavour" of the subject, sometimes at the expense of precision.

You might like to review Chapter 9, particularly the section on groups.

10.2.1 Pre-orders, partial orders, and pregroups

A pre-order is a set of entities equipped with a relation $\leq$, which is reflexive and transitive:

$$x \leq x$$
$$x \leq y \text{ and } y \leq z \text{ implies } x \leq z$$

Note we do **not** assume any property concerning symmetry or antisymmetry: we may have $x \leq y$ with, or without, also having $y \leq x$, and it is possible for $x \leq y$, $y \leq x$, without then having $x = y$. A pre-order is called a partial order if it also satisfies the antisymmetric axiom:

$$x \leq y \text{ and } y \leq x \text{ implies } x = y$$

Every Boolean algebra, and also every Heyting algebra, is a pre-order, with the canonical order as defined in the previous chapter. Ordinary sets form a pre-order, with $X \leq Y$ meaning $X \subseteq Y$. The natural numbers, and likewise the real numbers, is a pre-order with the standard order. These are also partial orders, since each satisfies antisymmetry. However, you have already seen an example of a pre-order that is not a partial order (that doesn't satisfy antisymmetry): consider ordinary sets, but now with the order given by cardinality. So define $X \leq Y$ to mean $\#X \leq \#Y$, *i.e.* that the cardinality of X is less than or equal to the cardinality of Y. (Recall that by our earlier definition of cardinality, this just means it is possible to construct a 1-1 correspondence between X and a subset of Y, but not necessarily (though possibly) *vice versa.*) Note that if $X \leq Y$ and $Y \leq X$ in this sense, then $\#X = \#Y$; however, this does not mean $X = Y$, so antisymmetry is not satisfied.

A pregroup[4] is a partial order equipped with a binary operation, called "product" (which could be denoted by some appropriate symbol such as $\cdot$, but which we shall denote by juxtaposition,

[4] The name comes from the fact that this notion shares some of the properties of a partial order, and some of the properties of a group.

thus: xy), and two unary operations, called “left and right adjoints” (which we shall denote with superscripts ℓ and r, thus: x^ℓ, x^r), and a constant (denoted 1). The constant 1 is a unit for the binary product operation, which is associative. This means we have the following equations (for all x, y, z in the pregroup):

$$\begin{aligned} 1x &= x \\ x1 &= x \\ x(yz) &= (xy)z \end{aligned}$$

(The observant reader will recognise these as the first three group axioms G1, G2, G3. We shall not assume G4 nor G5: we replace G4 with the following assumption.) In addition, the order must be compatible with the product operation, and the unary “left and right adjoints” must satisfy the following inequalities for all a, x, y (using the given order):

$$\text{If } \ a \leq b \text{ , then } \ xay \leq xby$$
$$\begin{aligned} x^\ell x &\leq 1 \\ 1 &\leq xx^\ell \\ xx^r &\leq 1 \\ 1 &\leq x^r x \end{aligned}$$

Exercise PG: Show that from these axioms the following are true in any pregroup.

1. Left and right adjoints are unique: for any x, if an object y has the properties $yx \leq 1$ and $1 \leq xy$, then $y = x^\ell$, and similarly if an object z satisfies $xz \leq 1$ and $1 \leq zx$, then $z = x^r$.
2. The adjoints are contravariant: for any x, y, if $x \leq y$ then $y^\ell \leq x^\ell$ and $y^r \leq x^r$.
3. For any x, y, $(xy)^\ell = y^\ell x^\ell$, and similarly $(xy)^r = y^r x^r$. Furthermore $x^{\ell r} = x^{r\ell} = x$.
4. $1^\ell = 1^r = 1$.

(These facts are established following essentially the same pattern as similar facts for groups and Boolean algebras; the adjoints play a role not unlike group inverses and Boolean negation.)

There are mathematical examples of pregroups (which I might, time permitting, describe in class), but for our main aim, what is really important is that they give a simple algebraic system for “calculating sentences”, as described below. For this, the equations and inequalities above are the main thing to remember, especially the “contraction” or “reduction” inequalities $x^\ell x \leq 1$ and $xx^r \leq 1$. In fact we shall not use the other (“expansion”) inequalities $1 \leq xx^\ell$, $1 \leq x^r x$; this is not a coincidence, but a consequence of the fact that the pregroups we consider are in a sense “free”, and for such pregroups Lambek showed the expansion inequalities can be avoided.

Note: It is also important to keep left and right clear in your mind (or to refer to the definition frequently, if you are—as I am!—left/right challenged!), since pregroups are **not** commutative: xy is not the same as yx. As a mnemonic, you might think that x^ℓ “cancels” x on the **l**eft, and x^r “cancels” x on the **r**ight:

$$x^\ell x \ \leq \ 1 \qquad\qquad xx^r \ \leq \ 1$$

Non-commutativity is important for our application to linguistics, and amounts to the observation that in (*e.g.*) English the phrase *John works* is not grammatically the same as **works John.*[5] Order matters.

[5] It is traditional in linguistics to indicate **non**-sentences by an asterisk.

10.2.2 Pregroups and natural language processing

In an article that appeared in 1958[6], Joachim Lambek used a logical syntax not unlike part of propositional logic as a tool to analyse sentence generation in natural languages, such as English, French, *etc.* Specifically, he was interested in obtaining an algorithm (or rule) for distinguishing sentences from non-sentences in such languages.[7] In the 1990s, he developed a new approach to this question using pregroups. We shall outline how this works, but as his original article is fairly readable,[8] I suggest you look at that article, which is linked on my webpage. Over the years a lot of research has been done on this style of linguistics; I have also linked two other articles you may find useful, *viz* an introduction to a book on the subject (I am a coauthor of that introduction, and a coeditor of the volume), as well as a more recent survey article[9] by Lambek on his pregroup approach to the question of sentence generation, an approach he regards as far more successful. It is this more recent approach we shall look at now. (By the way, he once told me that he thought the new approach *via* pregroups was also far easier, and would be a better introduction for a class such as this one. We'll look at the original approach later in the appendix to this section—I'd appreciate any feedback you might have on the two approaches.)

Briefly, here's the main idea. To every word in a language (I'll use English for now) we associate a "*syntactic type*", which designates what part of speech (to use a more traditional phrase—though the concepts aren't entirely identical) the word has. (A word might actually have several syntactic types, depending on context, and there might be relationships between the types—more of that later.) To start with, we would want two atomic types, s (for sentences) and n (for nouns, including class-nouns, such as *milk* or *rice*, and for names, such as *John*). We might want to refine this, for example to distinguish between various types of sentences or types of nouns, but we'll ignore this option for the moment.

Then we can give other words compound types, built up from the atomic ones using the pregroup operations. So we'd have types like nn^ℓ, $n^r s$, $n^r s s^\ell n$, and so on. The idea then is to give types to words in potential sentences, and his algorithm to determine whether your phrase is in fact a sentence or not is to see if the product of the types of the words in your phrase can be reduced to the type s.

For example, intransitive verbs such as *works* generally receive the type $n^r s$ (because they need a noun in front to make a sentence), so a phrase like *John works* would be typed $n(n^r s)$, since *John* is type n and *works* is type $n^r s$, and their product is therefore $n(n^r s)$. To see this is a sentence, we try to "reduce" this type to the simple type s, using the equations and inequalities of a pregroup:

$$n(n^r s) \;=\; (nn^r)s \;\leq\; 1s \;=\; s$$

So we can conclude *John works* is a sentence, since the product of the types of the words in the phrase is indeed $\leq s$.

In a similar way, we can type most adjectives, like *poor*, as nn^ℓ, because they must precede a noun to create another noun (or rather a noun phrase). For example, *Poor John works* would be

[6] *The mathematics of sentence structure*, by J. Lambek, American Mathematical Monthly (65) 1958.

[7] This is to be understood in a purely syntactic way: the meaning (or semantics) is *not* what is being considered here, merely the way the words are combined. Chomsky's famous example, *Colorless green ideas sleep furiously* is a grammatically correct sentence, although it is nonsense semantically. **Furiously sleep ideas green colorless* is not grammatically correct, as well as being nonsense. Lambek's analysis shows the first to be a sentence, and the second not to be one.

[8] At least those parts that deal with the linguistics more than the underlying maths should not be beyond anyone who's managed to survive this course so far!

[9] *Pregroups and natural language processing*, J. Lambek, Mathematical Intelligencer, 2006.

typed $(nn^\ell)n(n^r s)$, and then we can calculate:

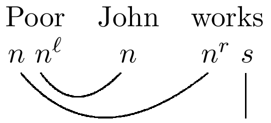
$$(nn^\ell)n(n^r s) \;=\; n(n^\ell n)(n^r s) \;\leq\; n1(n^r s) \;=\; n(n^r s) \;=\; (nn^r)s \;\leq\; 1s \;=\; s$$

and so *Poor John works* is also a sentence.

Notice how in these calculations we use the associativity and unit equations, and especially the adjoint inequalities $nn^r \leq 1$ and $n^\ell n \leq 1$. It is tempting to abbreviate this process by ignoring associativity (*i.e.* dropping brackets) and the unit (dropping reference to 1), and focussing on the adjoint inequalities (which are sometimes referred to as "cuts", since they "cut" certain types from the expression). In this way, we could represent the calculation for *Poor John works* thus:

$$nn^\ell n\, n^r s \;\leq\; nn^r s \;\leq\; s$$

each inequality corresponding to a cut of some formulas. But we can go one step further, making this both simpler (dropping the inequalities) and more explicit (showing what cuts are being made) with the following notation. Under each word in the phrase, we give its type. Then we draw a "typing graph" which links the types which may be cut: an $x^\ell x$ will be linked by joining the x^ℓ to the x, and similarly an xx^r will be linked by joining the x to the x^r. As long as no cut link actually intersects another, and as long as the only unlinked type is s (which we shall indicate by a vertical line so it is easily seen), then the phrase is a sentence. Here is *Poor John works* with its typing graph:



The two cut links correspond to the adjoint inequalities $n^\ell n \leq 1$ and $nn^r \leq 1$, and so indicate where types may be eliminated, and after that is done, all that remains is the simple type s, indicating that we have a sentence. Note that one cut is "inside" another: this means that in the sequence of inequalities represented by this graph, the inside one must be done first before the outside one is "legal" (*i.e.* before the n and the n^r are next to each other, ready to cut). This is why one may not have cut links intersect (since then one would not have the right types next to each other for a cut).

You might like to do the (very simple!) exercise of typing and graphing *John works* to check that you have the idea. (The typing graph will have just one cut link.) Notice that in the *Poor John works* example, the structure of the analysis naturally groups *Poor John* together, as a new noun phrase to be combined with *works*; in other words, the sentence is naturally parsed as *(Poor John) works* and not as *Poor (John works).* The types reflect the natural grouping by the way they reduce to s.

The point of this sort of analysis is that the same sort of calculation works equally well for other sentences, and that if you try to do similar calculations for non-sentences, the typing will block your attempts to reduce to s. For example, **John poor works* is not a sentence, and its type $n\, nn^\ell n^r s$ *cannot* be reduced to s. This is because there is no cut that can eliminate the n^ℓ, since n^ℓ needs an n on its right to form a cut. Equally bad, the n^r cannot be cut, since n^r needs an n on its left, and the n^ℓ gets in the way (and as there's no way to eliminate it, that is a problem for n^r as well).

*John	poor	works	
n	$n\ n^\ell$	$n^r\ s$	$(\not\leq s)$

In the following exercises you will find some examples to try—notice how the typing of different parts of speech is naturally suggested by known sentences, but also how those types block non-sentences. Ultimately, the power of this analysis may really only be seen by seeing how it works for many different examples in many different languages, and how it accounts for what is and what is not grammatically correct in those languages. In this section we merely scrape the surface with these simple examples, with only a couple of more complicated ones to indicate the flavour of the pregroup approach to linguistics. More complete accounts may be found in these books: *From Word to Sentence: A Computational Algebraic Approach to Grammar* by Lambek, Polimetrica (2008), and *Computational Algebraic Approaches to Natural Language* by Casadio & Lambek, Polimetrica (2008).

10.2.3 Exercises

1. By considering the phrase (sentence) *John works here*, what must the type of *here* be? Notice how this type indicates that *here* follows a sentence to create a new sentence.

   I'll give you (a hint for) the answer (but give it a try first!): *John works* types as s, as we've seen, so *here* must have a type x which, when it follows s, must produce the type s again. So the s of *John works* must be cut, and the only type that cuts it on the right is s^r, and since we want s left over, *here* must have the type ... (try it yourself before reading further!!) ... $s^r s$. (Did you get that?) Verify this works with the typing graph. Notice that the graph suggests that *here* modifies *works*, since *works here* has the same type as *works*, allowing the sentence to parse as *John (works here).* In the rest of the examples/exercises, see how the typing reflects the "deep structure" of the sentence, its *syntax* as opposed to its *semantics*.

2. By considering the phrase *John often works*, what must the type of *often* be? Once you have your type for *often*, show it is correct by giving the typing graph for *John often works*, showing the types reduce to s.

   Again, a hint to get you started (if you need it!): *often* follows *John*, and so its type must cut the n of *John*, and so must begin with n^r. (Why not n^ℓ ?) Its type must end with something to cut the $n^r s$ of *works*: that would require $(n^r s)^\ell$ (since we want a type which cuts the $n^r s$ which follows it). You must figure out what $(n^r s)^\ell$ can be simplified to. (Refer to Exercise PG3.) Finally, between these two elements we want the s, which will be left over when the various cuts are made.

3. What is the type of *and*, as in *John runs and Jane watches* ? Show the typing graph of *John runs and Jane watches.* Note that *John* and *Jane* both have type n, and *runs* and *watches* have the type $n^r s$, like *works* (as do all intransitive verbs).

4. What is the type of *for*, as in *John works for Jane* ? Show the typing graph for *John works for Jane.*

5. (Transitive verbs) By considering *John likes Jane*, what is the type of the (transitive) verb *likes* ? Show the typing graph for *John likes Jane.*

6. What is the type of an adjective like *fresh*, as in *John likes fresh milk* ? Note that *milk* has type n. Show the typing graph of *John likes fresh milk.* Note that the graph shows how *fresh* modifies *milk*, so that essentially the sentence is parsed as *John likes (fresh milk).*

7. (Pronouns) What is the type of *he* ? Use the sentence *He works* as an example; do not be satisfied with the type n however (we'll see why soon!). Show the typing graph of *He works.*

8. Show the typing graph of *He likes Jane.* Use the same type (not n) of *He* you used in the previous exercise.

9. What is the type of *him* ? Again, do not be satisfied with the type n. Use *Jane likes him* as an example, and show its typing graph.

   Notice these types for *he* and *him* are not the same—nor should they be, since if they were the same, then we would get satisfactory typing graphs for non-sentences like **Him likes Jane* and **Jane likes he.* It is precisely the different typing of *he* and *him* that forces *he* into the "subject" position (left of the verb) and *him* into the "object" position (right of the verb).

10. Show the typing graph of *He likes him.* Show that **Him likes he* is not a sentence according to the typing.

11. Show the typing graphs of the following sentences:

    (a) *John loves Jane and she loves him.*

    (b) *John likes her mother.*

    (c) *She likes his mother.*

    (d) *His mother likes her.*

    (e) *John works and he often plays.*

12. Show the typing graph of *Colorless green ideas sleep furiously.* (Yes! You have enough typing information to do this.)

13. (An entertainment.) Show that there are two different typing graphs for *Time flies*, corresponding to two rather different meanings.

10.2.4 More subtle analysis

In his 2006 survey article, Lambek illustrates pregroup analysis with a richer set of atomic types, taking into account such structure as tenses, questions, participles, and much else. In this richer setting, we must add some inequalities among the types, corresponding to how one type might include another as a special case. (Other, more general, inequalities are also permissible in pregroup analysis, reflecting some of the flexibility of natural language.)

Here's an example (you should refer to the article for the details—I am still aiming at just the "flavour" here), using these basic types: s_2 for sentences in the simple past tense, $\overline{q}$ for questions, q for yes-or-no questions (where tense is irrelevant), q_1 for yes-or-no questions in the present tense, $\hat{o}$ and o for direct objects (for technical reasons he uses two distinct types here), p_2 for past participle of intransitive verbs, π_1, π_3 for first and third person subject pronouns ("I" and "he", "she", or "it"), and π for pronouns in general, with reductions $q_1 \leq q$, $q \leq \overline{q}$, $\hat{o} \leq o$, and $\pi_k \leq \pi$.

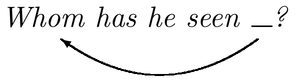
With this, one can type *I* as π_1, *saw* as $\pi^r s_2 o^l$, and *her* as o; (*saw* is a transitive verb, so needs both a subject such as a pronoun on the left and an object on the right to make a sentence: hence gets the typing $\pi^r s_2 o^\ell$). Then *I saw her* becomes $\pi_1(\pi^r s_2 o^l)o$ which reduces to s_2. (**Exercise:** construct the typing graph.)

$$\pi(\pi^r s o^\ell)o \;\leq\; (\pi\pi^r)s(o^\ell o) \;\leq\; 1s1 \;\leq\; s$$

Questions are traditionally a bit trickier to analyse, especially "wh-questions". Consider the sentence *Whom has he seen?.* This is typed as follows:

$$\begin{array}{cccc} \text{Whom} & \text{has} & \text{he} & \text{seen?} \\ (\overline{q}\hat{o}^{\ell\ell}q^\ell) & (q_1 p_2^\ell \pi_3^\ell) & \pi_3 & (p_2 o^\ell) \end{array}$$

and the equations imposed on the theory of pregroups allow one to compute this expression, reducing it to $\overline{q}$; this is indeed a question. (**Exercise:** verify this does reduce to $\overline{q}$ by constructing the typing graph.) One of the features Lambek liked about this approach is how well it handles what Chomsky calls a "trace": in a question such as the one above, since *seen* usually takes an object (*see* is a transitive verb), Chomsky imagined that there was a "ghost" of that object following the word *seen*, which pointed back to the word *whom* at the start of the question. In a sense it is as if the question really were



In Lambek's pregroup analysis, such traces correspond to instances of double adjoints $x^{\ell\ell}$ or x^{rr}, like the $\hat{o}^{\ell\ell}$ we saw above in the type of *whom*. In fact, in this case the link between $\hat{o}^{\ell\ell}$ and $\hat{o}^{\ell}$ looks rather like the trace arrow.

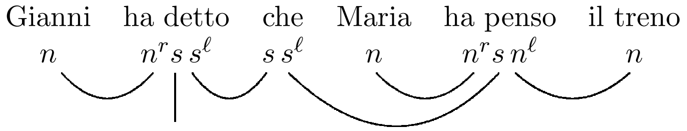
Other languages

Pregroup analysis has been made of many languages, including Burushaski (an isolated language spoken in Pakistan), Arabic, Chinese, German, French, Italian, and many more. Here are two examples: first, a simple example in Italian,[10] followed by an example of Lambek's in French.[11]

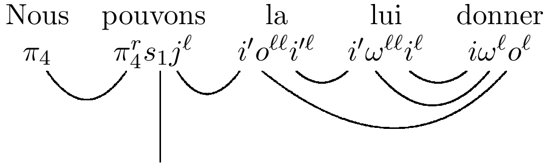
The Italian example:

Gianni ha detto che Maria ha perso il treno.
Gianni said that Mary had missed the train.

We'll simplify this a bit by assuming some resultant types of some compound phrases: "ha detto" has type $n^r ss^{ell}$, "ha penso" has type $n^r sn^\ell$, and "il treno" has type n. Then the sentence is typed thus:



Next, the French example: to illustrate a more "interesting" (and so complicated) grammar, Lambek introduces such types as π_4 (first person plural subject pronoun in nominative case, like *nous*), s_1 (for declarative sentences in the present tense), o (direct object), ω (indirect object), i, i', j (three different types of infinitives, with the relations $i < i' < j$), among others. Then he analyses *Nous pouvons la lui donner* as follows. *Nous* has the type π_4; *pouvons* the type $\pi_4^r s_1 j^\ell$; *la* the type $i' o^{\ell\ell} i'^\ell$; *lui* the type $i' \omega^{\ell\ell} i^\ell$; and *donner* the type $i\omega^\ell o^\ell$. Then we obtain the following typing graph for *Nous pouvons la lui donner*, showing it is a sentence (in the present tense).



[10] Taken from *The Lambek Program*, by C. Casadio, P.J. Scott, R.A.G. Seely. The example is due to our first author, Claudia Casadio.

[11] J. Lambek, *Exploring Feature Agreement in French with Parallel Pregroup Computations*, J Log Lang Inf **19** (2010), 75–88.

The only reduction that might cause concern is the cut on $j^\ell i'$: but recall that we supposed $i' < j$ and hence (by Exercise PG2) $j^\ell < i'^\ell$, and so $j^\ell i' \le i'^\ell i' \le 1$, as required.

Analysis of musical chords

Just to illustrate that pregroups may be used for "languages" in quite a general sense, we sketch (briefly) how they might be used to analyse chords in music.[12]

Terrat's basic idea is that one can build chords from pitches, much as one builds sentences from words, and that non-dissonant chords may be specified by a suitable rule (as an example, he gives the rule "a chord is a sequence of at least 3 pitches such that the distance between two successive pitches of the chord is at least 3 semitones"). He does not distinguish chords with the same interval structure but based on different roots (so all major chords, for example, are considered the same in this simple example). So all that needs to be specified are the intervals between pitches, and to achieve non-dissonance, there must be some way to guarantee that successive pitches are at least three semitones apart.

Here's how he does this. Define the basic types as C for (non-dissonant) chord, $P(i)$ for the top pitch, $P(i, j)$ for intermediate pitches, where i, j are positive integers, i giving the distance (in semitones) above the root, and j the position of the pitch in the chord. Suppose we have these inequalities:

$$P(i, j) \le P(k, j) \text{ if and only if } i \le k$$

$$P(i, j) \le P(i)$$

for all i, j, k. Then, given a chord, in terms of the distances of its pitches above the root note, we assign types as follows: $P(i, 1)$ for the first pitch, where i is the number of semitones above the root ($i \ge 3$); $P(i-3, k)^r P(i, k+1)$ for each intermediate pitch in position k ($i \ge 3(k+1)$); $P(i-3)^r C$ for the last pitch ($i \ge 6$). Then if the various values of i, k satisfy the condition in brackets, the type of the chord will reduce to C, indicating the chord is non-dissonant (according to the definition), but otherwise that will not be the case.

An example might help: Consider an "m7♭9" chord[13] (you could consider Cm7♭9, whose root is C, and whose basic notes are then C, E♭, G, B♭, and D♭: Cm7 with a "flat 9th"). We'll represent such a chord by this integer sequence: 3,7,10,13. With each integer we associate a term, as follows:

$$\begin{array}{rl}
3 & P(3,1) \\
7 & P(4,1)^r P(7,2) \\
10 & P(7,2)^r P(10,3) \\
13 & P(10)^r C
\end{array}$$

so that the term for the chord is $P(3,1)P(4,1)^r P(7,2)P(7,2)^r P(10,3)P(10)^r C$. So: is this a (non-dissonant) chord? Well, we have $P(3,1) \le P(4,1)$, so $P(4,1)^r \le P(3,1)^r$ and so $P(3,1)P(4,1)^r \le 1$, and the rest is easy. (**Exercise:** draw the typing graph for this calculation to show it is a non-dissonant chord. Show that if we lower the third note, to get the chord C, E♭, F, B♭, then the resulting chord is dissonant: the resulting type does not reduce to C.)

Terrat extends this simple example to get rather more mileage, and observes that different types of music might allow different types and type relationships; feel free to explore these ideas yourself.

[12] This discussion is based on the paper *Pregroup grammars for chords*, by Richard Terrat (2004); a copy may be found on my webpage, along with a sequel paper.

[13] Don't worry if you know nothing of music theory; it's not really necessary to follow the example. There is a link on my webpage to a page which allows you to listen to various chords.

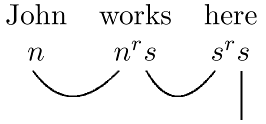
10.2.5 Solutions to the exercises

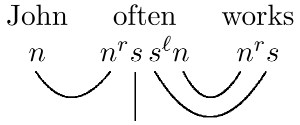
Exercise PG in Section 10.2.1

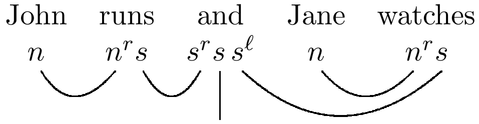
1. If $yx \leq 1$, $1 \leq xy$, then $y = y1 \leq y(xx^\ell) = (yx)x^\ell \leq 1x^\ell = x^\ell$, and $x^\ell = x^\ell 1 \leq x^\ell(xy) = (x^\ell x)y \leq 1y = y$; *i.e.* $y \leq x^\ell$ and $x^\ell \leq y$, and so $y = x^\ell$. The proof for r is similar.
2. If $x \leq y$ then $y^\ell = y^\ell 1 \leq y^\ell(xx^\ell) \leq y^\ell(yx^\ell) = (y^\ell y)x^\ell \leq 1x^\ell = x^\ell$. The proof for r is similar.
3. To show $(xy)^\ell = y^\ell x^\ell$, we can use (1): we just need to verify that $(y^\ell x^\ell)(xy) \leq 1$ and $1 \leq (xy)(y^\ell x^\ell)$. But these are easy; for example $(y^\ell x^\ell)(xy) = y^\ell(x^\ell x)y \leq y^\ell 1y = y^\ell y \leq 1$, and similarly for the expansion inequality. The case with r is similar. To show $x^{r\ell} = x$, we also use (1): here, since $x^{r\ell} = (x^r)^\ell$, we only need to show that $x^r x \leq 1$ and $1 \leq xx^r$, which is true. The equation $x^{\ell r} = x$ is similar.
4. Again we use (1): this time the result is really easy, since it follows from $1 \cdot 1 = 1$.

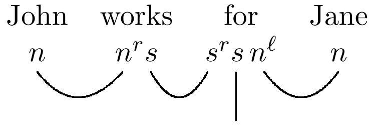
Exercises in Section 10.2.3

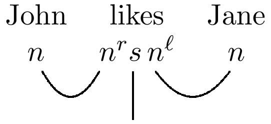
I've just given the typing graphs—the rest of the answers can be deduced from them.

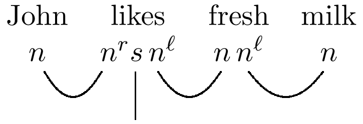
1.


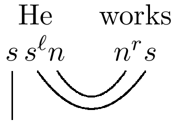
2.


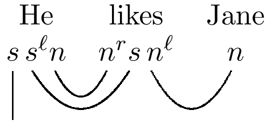
3.

John runs and Jane watches
n $n^r s$ $s^r s\, s^\ell$ n $n^r s$

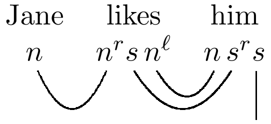
4.


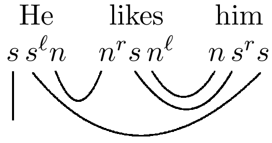
5.


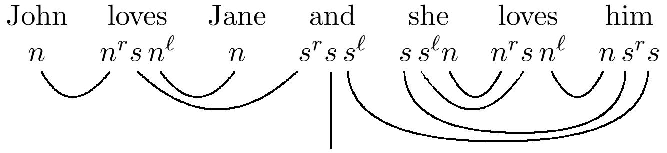
6.


7.


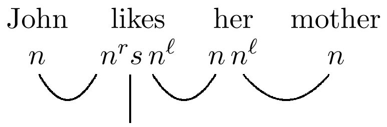
8.


9.


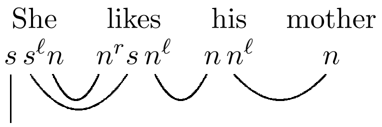
10.


11. (a)

John loves Jane and she loves him
n $n^r s\, n^\ell$ n $s^r s\, s^\ell$ $s\, s^\ell n$ $n^r s\, n^\ell$ $n\, s^r s$

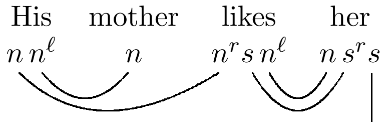
(b)



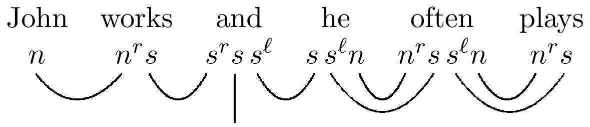
(c)



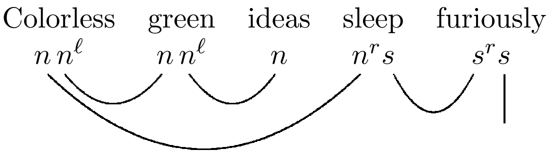
(d)



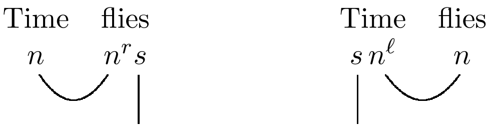
(e)


John works and he often plays
n $n^r s$ $s^r s s^\ell$ $s s^\ell n$ $n^r s s^\ell n$ $n^r s$

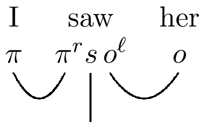
12.

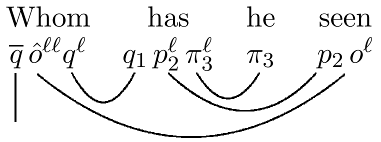

Colorless green ideas sleep furiously
$n n^\ell$ $n n^\ell$ n $n^r s$ $s^r s$

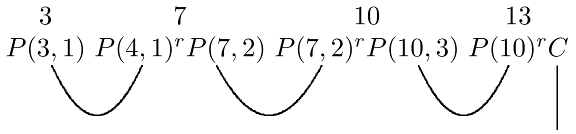
13.



The “in-text” exercises in Section 10.2.4




Whom has he seen
$\overline{q}\,\hat{o}^{\ell\ell} q^\ell$ $q_1 p_2^\ell \pi_3^\ell$ π_3 $p_2 o^\ell$



10.2.6 Appendix: Categories and Categorial Grammar

For comparison, we'll turn now to Lambek's original approach to linguistics, *via* the *syntactic calculus.* We start with a tiny detour, however, to put this approach into a context which is close to the deductive systems we've considered earlier in the semester. We intend to draw attention to the fact that there is some structure common to groups and to pre-orders, although it may be less than obvious because of the notation we use. We shall begin with a different structure entirely, and then see if we can spot it in groups as well as in pre-orders.

Consider the collection of sets and maps (functions) between them. You may think of a function as a "little black box" which considers things from one set and assigns them to another set, such as the function "mother", which looks at a person and tells who their mother is: LOURDES ↦ MADONNA; or the function x^2 which looks at a number and tells you what its square is: $5 \mapsto 25$.

The point about functions is that you can compose them: if one function f gives the square of a number: $f(x) = x^2$, and another adds 1: $g(x) = x + 1$, then you can do one operation after the other: $g(f(x)) = x^2 + 1$. Note this is not the same as doing them in the other order: $f(g(x)) = (x + 1)^2 = x^2 + 2x + 1$, so composition is not a commutative operation. It is however associative, and it has a unit element, namely the identity function. (Think about that for a moment, and convince yourself that it's true.)

You saw the same structure when you looked at permutations (section 9.3.1). At that time, we abstracted the structure to get the structure of a group in the following way: we supposed that there is an operation $\circ$ combining two elements of a group. But this omits the fact that different operations or functions may operate on different sets. For instance, we can have a function which to every person assigns their current age (as a natural number: round down!), and another function which to each natural number, adds 1. Let's call these $A(x) :=$ Age of x and $S(n) := n + 1$. It's is traditional to denote these this way (where P is the set of people, and $\mathbb{N}$ is the set of natural numbers):

$$A : P \longrightarrow \mathbb{N} \qquad\qquad S : \mathbb{N} \longrightarrow \mathbb{N}$$

The composite function $S(A(x))$ tells you how old someone will be on their next birthday. We regard this as an operation, "composition", on functions: $(S \circ A)(x) = S(A(x))$

$$A : P \longrightarrow \mathbb{N} \text{ and } S : \mathbb{N} \longrightarrow \mathbb{N} \quad \mapsto \quad S \circ A : P \longrightarrow \mathbb{N}$$

But we cannot compose functions where the values are of the wrong type: for example, if M is the function "mother of", so that $M($LOURDES$) =$ MADONNA, so that $M : P \longrightarrow P$ assigns to each person the appropriate person (his/her mother), then we cannot compose M with S: it makes no sense to talk of $S(M($LOURDES$)) =$ MADONNA $+ 1$. So in composing functions, we need to pay attention to their "domains" (the collection of things they apply to) and their "codomains" (or "targets", being the collection of things they might produce as "values").

The arrow notation makes it clear what compositions work. When we write $A : P \longrightarrow \mathbb{N}$, $S : \mathbb{N} \longrightarrow \mathbb{N}$, it's clear that we can perform A and then S on the output of A, since S needs an input of type $\mathbb{N}$, and A produces output of that type, ready for S to use. Notationally, the arrows "fit together": $P \xrightarrow{A} \mathbb{N} \xrightarrow{S} \mathbb{N}$. But when we try to fit the arrow S to M, we don't get a match: $P \xrightarrow{M} P$ but $\mathbb{N} \xrightarrow{S} \mathbb{N}$. (Note that this also shows we couldn't have reversed the order of A and S: $\mathbb{N} \xrightarrow{S} \mathbb{N}$ followed by $P \xrightarrow{A} \mathbb{N}$ doesn't fit: just what sense does it make to ask for the age of the number $n + 1$ anyway?)

We'll give the definition of the algebraic structure that captures this notion, and then show how it relates to groups and pre-orders.

Definition: A category $\mathbf{C}$ consists of two collections of things (called *objects* and *arrows*), together with some operations: an operation ι which assigns an arrow ι_A to each object A, two operations d, c which assign objects d_f, c_f to each arrow f, and a ("partial") operation $\circ$ which assigns an arrow $g \circ f$ to each pair f, g of arrows satisfying the condition $c_f = d_g$. All this must satisfy the following axioms. To simplify reading this, we use the convention that, for any arrow f, we write $f : A \longrightarrow B$ to mean that $d_f = A$ and $c_f = B$.

$$\begin{array}{lll}
\iota_A : A \longrightarrow A & & \\
g \circ f : A \longrightarrow C & \text{if} & f : A \longrightarrow B\ ,\ g : B \longrightarrow C \\
\iota_B \circ f = f = f \circ \iota_A & \text{if} & f : A \longrightarrow B \\
h \circ (g \circ f) = (h \circ g) \circ f & \text{if} & f : A \longrightarrow B\ ,\ g : B \longrightarrow C\ ,\ \text{and } h : C \longrightarrow D
\end{array}$$

(Think of ι_A as an "identity function", d_f, c_f as the "domain and codomain" of f, and $\circ$ as composition of functions. Indeed, functions between sets do form a category in just this way.)

Examples of categories are not hard to find—as we indicated above, ordinary sets and functions form one, and in a sense, that is the example which will initially guide our intuition. But be warned!: most categories don't look anything like sets and functions! Here are two examples: pre-orders, and groups.

Every pre-order is a category in a natural way: the objects are just the entities which make up the pre-order, and there is an arrow $x \longrightarrow y$ if and only if $x \leq y$. So there is at most one arrow $x \longrightarrow y$ between any two objects x, y; in fact, any category with this property, that there is at most one arrow $x \longrightarrow y$ between any two objects x, y, is a pre-order.

Exercise: Prove the claims in the preceding paragraph.

Every group is also a category, but in a different way. Given a group $\mathbf{G}$, we define a category with exactly one object (which we shall call G, but note the different typeface); the arrows $a : G \longrightarrow G$ are just the group elements $a \in \mathbf{G}$.

Exercise: Prove that this does define a category.

Remark: We say an arrow $f : A \longrightarrow B$ in a category $\mathbf{C}$ is an *isomorphism* if there is an arrow g in $\mathbf{C}$ with the property $f \circ g = \iota_B$ and $g \circ f = \iota_A$. A group is the same thing as a category with exactly one object all of whose arrows are isomorphisms. You might like to try to prove this if you feel frisky.

Definition: A functor $F : \mathbf{C} \longrightarrow \mathbf{D}$ from a category $\mathbf{C}$ to a category $\mathbf{D}$ consists of functions F which assign to any object C of $\mathbf{C}$ an object $F(C)$ of $\mathbf{D}$, and to any arrow $f : C \longrightarrow C'$ of $\mathbf{C}$ an arrow $F(f) : F(C) \longrightarrow F(C')$ of $\mathbf{D}$, subject to the following equations:

$$\begin{array}{l}
F(\iota_C) = \iota_{F(C)} \\
F(f \circ g) = F(f) \circ F(g)
\end{array}$$

Exercise: Define what one might mean by a "structure preserving map" between groups, and show that if **C, D** are groups, then a functor between them is exactly such a map.
Do the same for pre-orders: a functor between pre-orders is precisely an order-preserving map between them.

In general, we regard functors as structure preserving maps between categories; the examples we shall have in mind will involve categories which represent logics, and the functors will be transformations or operations on the logics which respect the notion of inference.

Categories and logic

Since Boolean and Heyting algebras are pre-orders, they are also categories. So there are lots of categories which have some connection with (propositional) logic. (And there are categories which have connections with predicate logic—you can look into this on my research website if you are so inclined.) For now, we'll consider some simple categorical structures which correspond to very weak logical notions, but which nonetheless have some relevance to other things.

Deductive systems and monoidal categories

We say a category **C** is a *monoidal* category if there is a functor $\otimes$ (pronounced "tensor"):
which assigns an object $A \otimes B$ to any two objects A, B
and an arrow $f \otimes g : A \otimes C \longrightarrow B \otimes D$ to any arrows $f : A \longrightarrow B, g : C \longrightarrow D$
as well as a "unit" object $\top$, satisfying the following axioms[14]:

$$\begin{aligned} &A \otimes \top = A = \top \otimes A \\ &A \otimes (B \otimes C) = (A \otimes B) \otimes C \end{aligned}$$

and similarly for arrows.

Actually, this is what is usually called a "strict" monoidal category, but the distinction is one we shall blur in this course. Usually one only asks for coherent isomorphisms, rather than equalities, in the axioms above, but to spell out just what that means would take more time than I want to spend on the matter. I have not supposed that $\otimes$ is commutative—to do so would force us to take the notion of coherent isomorphism more seriously, as it is just not reasonable to ask for strict commutativity. In any event, the examples we will want to consider will not have commutative tensors, for precisely the same reason we didn't want commutativity for the product in a preorder.

For us, the point about such monoidal structure is that it models some basic properties of conjunction; in fact, one could describe an (admittedly weak) logic which only had that connective and the properties that monoidal categories possess. What are those properties? We may summarize them with the following "deduction rules" (I'll explain the notation below, though it is similar to what we used in Chapter 2):

$$\frac{}{A \to A}\ (\text{Axiom}) \qquad\qquad \frac{\Gamma \to A \quad \Delta, A, \Delta' \to B}{\Delta, \Gamma, \Delta' \to B}\ (\text{Cut})$$

$$\frac{\Gamma, A, B, \Gamma' \to C}{\Gamma, A \otimes B, \Gamma' \to C}\ (\otimes L) \qquad\qquad \frac{\Gamma \to A \quad \Delta \to B}{\Gamma, \Delta \to A \otimes B}\ (\otimes R)$$

$$\frac{\Gamma, \Delta \to A}{\Gamma, \top, \Delta \to A}\ (\top L) \qquad\qquad \frac{}{\to \top}\ (\top R)$$

To emphasise the connection with the categorical structure, we have replaced the entailment sign $\vdash$ used earlier in the course with an arrow (not to be confused with implication!). So you should think of $A, B \longrightarrow C$ (for example) as representing an entailment of C with hypotheses A, B: $A, B \vdash C$. We shall call this logical system "monoidal logic".

In these rules, we use capital Greek characters Γ, Δ to indicate finite sequences (lists) of WFFs, and we imagine WFFs are defined with one constant $\top$ and one binary connective $\otimes$, in a pretty

[14] Again, there is some hidden structure we're assuming—I'll make this clearer in class, but for now, just assume these axioms are to be interpreted "naturally".

obvious manner. The intention of the horizontal lines is to indicate how we can get new valid inferences from existing ones: each deduction rule represents a way to construct a valid argument (the one below the line) from one or more valid arguments (the ones above the line). Each valid argument is represented as an inference, with premises before the arrow, and the conclusion after it. The first two rules just correspond to identity arrows and composition of arrows; interpreted as inferences they amount to the trivial inference of A from itself, and the process of combining inferences (if one can infer A from some premises, then one can use that in another inference, replacing a premise A with those premises from which one inferred A). The rule $(\otimes L)$ merely expresses the idea that we represent tensors on the left of an inference by commas, and the $(\otimes R)$ rule corresponds to the functoriality of tensor. Remember that we regard $\otimes$ as a weak notion of AND, so it shares some of the properties of $\wedge$. Indeed, interpreting these two rules as inferences, we have two familiar properties of AND from propositional logic: one may replace two premises A, B with a single premise $A \otimes B$, and if we have inferences for each of A and B (as conclusions), then by combining all the premises used, we may arrive at an inference of $A \otimes B$. The last two rules do the same for the constant $\top$: it is a unit for tensor, and is represented on the left by (literally) "nothing". You may think of $\top$ as being a weak version of the logical constant TRUE; it has some (but not all) of the properties $\top$ had in classical propositional logic. (An example of a property $\top$ does *not* have in monoidal logic is that although $A \vdash \top$ is a tautology in classical logic, it is not so in monoidal logic.)

There is a subtlety here we must mention: in the rule $(\otimes R)$ you should note that we repeated the premises in the conclusion; this means that if we use a premise in proving A, and also use it again in proving B, then in our proof of $A \otimes B$, that premise gets listed **twice**. That is no error! We **want** that duplicate listing, as it is related to the "conservation of resources" aspect of this logic, which we'll discuss in class. But here's an illustration of this idea. Tensor is to be thought of as a form of conjunction which pays attention to the use of "resources", so (unlike propositional logic), just because one has "$A \longrightarrow B$" and "$A \longrightarrow C$", one may **not** have "$A \longrightarrow B$ AND C" (where "AND" means $\otimes$), although we do have "A AND $A \longrightarrow B$ AND C". For example: If I have \$5, I can buy a hamburger and If I have \$5 I can buy a milkshake does **not** imply If I have \$5 I can buy a hamburger and a milkshake (you might need \$10, which is what you are guaranteed by I have \$5 and I have \$5). The format of the $(\otimes R)$ rule, with different Γ, Δ, is necessary to capture this idea.

It is worth noticing that these rules are sufficient to prove associativity of tensor, and the unit properties of $\top$. Here are some such derivations (including two different ones of the unit equation; you should try to prove these yourself, once you understand what's going on!):

$$\dfrac{\dfrac{\dfrac{\dfrac{}{A \to A} \quad \dfrac{\dfrac{}{B \to B} \quad \dfrac{}{C \to C}}{B,\, C \to B \otimes C}}{A, B, C \to A \otimes (B \otimes C)}}{A \otimes B,\, C \to A \otimes (B \otimes C)}}{(A \otimes B) \otimes C \to A \otimes (B \otimes C)} \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{A \to A}}{A, \top \to A}}{A \otimes \top \to A} \quad \dfrac{\dfrac{}{\to \top} \quad \dfrac{}{A \to A}}{A \to \top \otimes A}}{A \otimes \top \to \top \otimes A}\,(\text{Cut}) \qquad \dfrac{\dfrac{}{\to \top} \quad \dfrac{\dfrac{\dfrac{}{A \to A}}{A, \top \to A}}{A \otimes \top \to A}}{A \otimes \top \to \top \otimes A}$$

But one must notice that this system does **not** have several "structural rules" that we took for granted before. These are the following (usually called "contraction", "weakening", "exchange"):

Rules **not** valid here: $\dfrac{\Gamma, A, A, \Delta \ \to B}{\Gamma, A, \Delta \ \to B}\,(\text{c}) \qquad \dfrac{\Gamma, \Delta \ \to B}{\Gamma, A, \Delta \ \to B}\,(\text{w}) \qquad \dfrac{\Gamma, A, B, \Delta \ \to C}{\Gamma, B, A, \Delta \ \to C}\,(\text{e})$

Notice that the lack of the contraction rule corresponds to the "conservation of resources" idea mentioned above (as does in a way the lack of weakening: one may not have "unused resources"), and the failure of exchange corresponds to the lack of commutativity for tensor.

Monoidal closed structure

What about implication? We recall that in classical (and intuitionistic) propositional logic, implication "internalized" derivation (or valid argument), in the following sense: there is a bijection between valid arguments with a premise A and conclusion B, and derivations of the WFF[15] $A \Rightarrow B$. In fact, this is true in the presence of other premises as well, so we have the following bijection (indicated by the double horizontal line—you should think of this as meaning you can go either way, from top to bottom, or *vice versa*):

$$\begin{array}{c} \Gamma, A \vdash B \\ \hline\hline \Gamma \vdash A \Rightarrow B \end{array}$$

In the monoidal context, this looks similar, with a minor notational change: to distinguish the implication in a monoidal category, we shall use a different shape for our "if ... then ..." and "... if ..." arrows, namely $\multimap$ and $\multimapinv$. (We need the two directions, since without commutativity we cannot deduce the properties of one from the other. In other words, $A \multimap B$ ("if A then B") is not equivalent to $B \multimapinv A$ ("B if A"). For them to be equivalent would require conjunction (tensor) be commutative. I may say more about this in class.)

So we want to add to our monoidal logical system the following rules:

$$\frac{\Gamma \to A \quad \Delta, B, \Delta' \to C}{\Delta, \Gamma, A \multimap B, \Delta' \to C}\ (\multimap L) \qquad\qquad \frac{A, \Gamma \to B}{\Gamma \to A \multimap B}\ (\multimap R)$$

$$\frac{\Gamma \to A \quad \Delta, B, \Delta' \to C}{\Delta, B \multimapinv A, \Gamma, \Delta' \to C}\ (\multimapinv L) \qquad\qquad \frac{\Gamma, A \to B}{\Gamma \to B \multimapinv A}\ (\multimapinv R)$$

The (L) rules are just generalizations of the entailments $A, A \multimap B \longrightarrow B$ and $B \multimapinv A, A \longrightarrow B$, which are the monoidal versions of *Modus Ponens*, our old friend $(\longrightarrow E)$. (To verify this claim, imagine Γ is A, and that Δ, Δ' are both empty in the $(\multimap L)$ and $(\multimapinv L)$ rules.)

With these rules (and a suitable notion of equivalence of proofs), we can then establish the following bijections, which in effect says that $\multimap$ and $\multimapinv$ properly internalize the notions of valid argument (*i.e.* derivation) in our monoidal logic—we shall continue to call this logic by that name, now with the addition of these "monoidal implications".

$$\begin{array}{c} A, \Gamma \to B \\ \hline\hline \Gamma \to A \multimap B \end{array} \qquad \begin{array}{c} \Gamma, A \to B \\ \hline\hline \Gamma \to B \multimapinv A \end{array}$$

Summary: We have developed a logic with a conjunction and an implication, but with different properties from classical propositional logic (different also from intuitionistic propositional logic). For example, one cannot derive inferences like $A \longrightarrow A \otimes A$ or $A, B \longrightarrow A$, which would be easy in classical (or intuitionistic) logic. (Exercise: verify this.) But this logic does have some similarities with classical and intuitionistic logic; for example, we can derive inferences like these: $A, A \multimap B \longrightarrow B$ and $B \multimapinv A, A \longrightarrow B$. (Exercise: verify this too.)

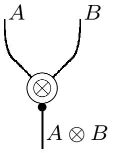
There is a (not totally misleading) way to think of the tensor, which is to impart to it a temporal component: think of $A \otimes B$ as "A then B", so A comes first. For example I had breakfast and I went to school implies that breakfast came first, and would not describe the situation where your breakfast was poutine at the Munch Box: that would be I went to school and I had breakfast. The

[15] We shall use the symbol $\Rightarrow$ for implication in classical propositional logic, so as not to confuse matters with the arrows in categories, which correspond to $\vdash$.

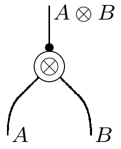
two "lollipop arrows" are needed to allow the two possible ways one might have an implication plus its premise (needed to conclude its conclusion), depending on which comes first. If the premise comes first, we need a $\multimap$ lollipop, as in $A, A \multimap B \longrightarrow B$, but if the implication comes first, we need the $\multimapinv$ lollipop, as in $B \multimapinv A, A \longrightarrow B$.

At the start of the course, I mentioned that logicians have tried to formulate logics that do not have the paradoxical property of classical (and intuitionist) logic that material implication seems to entirely miss the notion of causality, so that If that's a purple unicorn in the corner, then I'm a monkey's uncle is a true statement (simply because there is no unicorn, purple or otherwise, in the corner). Most of the "relevance logics"[16] they have developed are in fact based on monoidal logic. You can see that the "resource-sensitive" nature of monoidal logic means that if $A \multimap B$ is true in the logic, then there must be at least some "connection" between A and B.

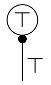
Remark: Several years ago, my coauthor and I developed a graphical representation of derivations in a similar monoidal logic. These pictures look something like the following. (If there's time, I may discuss this in class.)



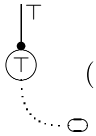
$(\otimes\ I)$



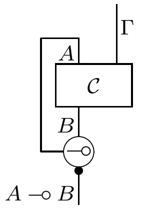
$(\otimes\ E)$



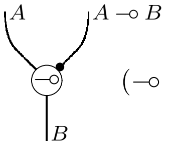
$(\top\ I)$



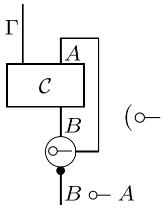
$(\top\ E)$



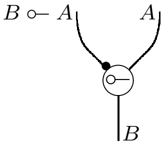
$(\multimap\ I)$



$(\multimap\ E)$



$(\multimapinv\ I)$



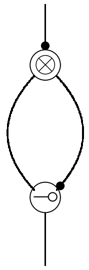
$(\multimapinv\ E)$

You should read these "top-down", with premises at the top and the conclusion at the bottom. With these we can represent general derivations in monoidal logic as more complicated graphs. For example, here are several graphs which represent some derivations, including the ones given earlier of associativity of $\otimes$ and of the unit law for $\top$ (the two derivations given for the unit law are in fact represented by the same graph, which gives a hint of the nature of the "equivalence of proofs"

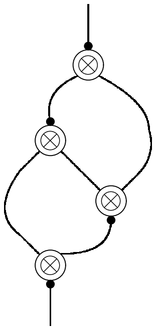
[16] A lovely (and useful) such logic is "linear logic", developed in the mid 1980s by Jean-Yves Girard; you can read more about it in the papers on my research website and the references given there.

mentioned before).

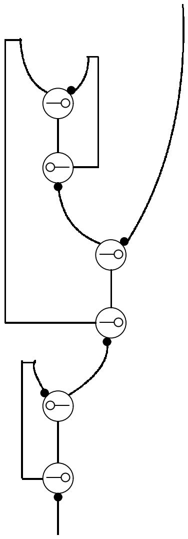
(1)



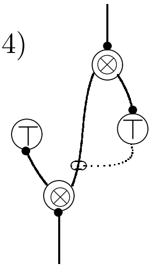
(2)



(3)



(4)



Optional Exercise: Identify what derivations these might be. (Answers: (1) $A \otimes (A \multimap B) \longrightarrow B$; (2) $(A \otimes B) \otimes C \longrightarrow A \otimes (B \otimes C)$ (this is the derivation given earlier of this form); (3) $(B \multimapinv (A \multimap B)) \multimap B \longrightarrow (B \multimapinv (A \multimap B)) \multimap B$ (it is a non-trivial problem to decide under what conditions this is just the identity arrow, and solving this problem, and others like it, was one of the main objectives of our early research into such monoidal logics—we succeeded!); (4) $A \otimes \top \longrightarrow \top \otimes A$ (this corresponds to both the derivations given earlier of this form).)

An application to linguistics

We are now in a position to consider the approach to categorial grammar that appeared in Lambek's 1958 paper. He used a logical syntax like the one we've developed here to analyse sentence generation. The main idea is very similar to the pregroup approach: to every word in a language (I'll still use English) we associate a syntactic type, designating its part of speech. Again, in our examples we'll start with the two atomic types, s (for sentences) and n (for nouns), as before.

Then we give other words compound types, built up from the atomic ones using the connectives of monoidal logic. So we'd have types like $n \multimapinv n$, $n \multimap s$, and so on. (Actually, Lambek used a much more compact notation, which you'll see in his paper, and which I'll mention later.) Then as before, the idea is to give types to words in potential sentences, and to see if the type s may be derived from the tensor of all the types from the words in your phrase.

We'll take the same examples as we had before, so you can easily compare the two approaches. So, intransitive verbs such as *works* generally receive the type $n \multimap s$ (because they need a noun in front to make up a sentence), so a phrase like *John works* would be typed $n \otimes (n \multimap s)$, and we've already seen that there is in fact a derivation $n \otimes (n \multimap s) \longrightarrow s$, so we'd conclude *John works* is a sentence. Using a "display style" presentation (now using a horizontal line to represent the "in-line" arrow meaning a valid inference, and dropping the use of the tensor), this would look like this:

$$\begin{array}{cc} \text{John} & \text{works} \\ \end{array}$$
$$\frac{n \qquad n \multimap s}{s}$$

Such “type derivations” correspond to the typing graphs we used in the pregroup approach. This works for other languages as well; here is the example from Italian we saw before:

Gianni ha detto che Maria ha perso il treno.
Gianni said that Mary had missed the train.

We analyse this this way (again using horizontal lines to indicate inferences, rather than our “in-line” notation with arrows):

Gianni ha detto che Maria ha perso il treno

$$\dfrac{n \qquad \dfrac{(n \multimap s) \multimapinv s \qquad \dfrac{s \multimapinv s \qquad \dfrac{n \qquad \dfrac{(n \multimap s) \multimapinv n \qquad n}{n \multimap s}}{s}}{s}}{n \multimap s}}{s}$$

As with the *John works* example, I have put the type of each component in the Italian sentence directly under the component. To simplify the example, some words (“ha detto” and “ha perso”) have been treated as a single component.

Some of our other simple examples may be seen in Figure 10.1; in each case, I have used the “display style” presentation, as above: underneath each word of the sentence appears the type of the word, and the horizontal lines indicate inferences from the tensor of the types above the line (as premises) to the type (the conclusion) that results from them. In each case the final type (at the bottom) is s, indicating that the words do in fact form a grammatically correct sentence.

Exercise 1: Show that (for any objects x, y, z in a monoidal closed category)

$$(x \multimap y) \multimapinv z \ \leftrightarrow \ x \multimap (y \multimapinv z)$$

(The use of this result is illustrated in Figure 10.1.)

We shall often write these equivalent forms without brackets: $x \multimap y \multimapinv z$. The point is that this equivalence allows us to analyse a sentence in any sensible way, with the same outcome. Let’s see some more examples.

Our first uses $n \multimap (s \multimapinv (n \multimap s)) \leftrightarrow (n \multimap s) \multimapinv (n \multimap s)$ to analyse *John often works*, two ways:

John often works

$$\dfrac{\dfrac{n \quad n \multimap (s \multimapinv (n \multimap s))}{s \multimapinv (n \multimap s)} \qquad n \multimap s}{s}$$

John often works

$$\dfrac{n \qquad \dfrac{(n \multimap s) \multimapinv (n \multimap s) \quad n \multimap s}{n \multimap s}}{s}$$

John works

$$\frac{n \qquad n \multimap s}{s}$$

(Poor John) works

$$\frac{\dfrac{n \multimapinv n \quad n}{n} \qquad n \multimap s}{s}$$

(John works) here

$$\frac{\dfrac{n \quad n \multimap s}{s} \qquad s \multimap s}{s}$$

(John works) (for Jane)

$$\frac{\dfrac{n \quad n \multimap s}{s} \qquad \dfrac{(s \multimap s) \multimapinv n \quad n}{s \multimap s}}{s}$$

John (likes Jane)

$$\frac{n \qquad \dfrac{(n \multimap s) \multimapinv n \quad n}{n \multimap s}}{s}$$

(John likes) Jane

$$\frac{\dfrac{n \quad n \multimap (s \multimapinv n)}{s \multimapinv n} \qquad n}{s} \qquad (\dagger)$$

John (likes (fresh milk))

$$\frac{n \qquad \dfrac{n \multimap s \multimapinv n \quad \dfrac{n \multimapinv n \quad n}{n}}{n \multimap s}}{s}$$

Time flies

$$\frac{n \quad n \multimap s}{s}$$

$$\frac{s \multimapinv n \quad n}{s}$$

[†] Note in this and the previous example we have given "likes" different types: $(n \multimap s) \multimapinv n$ and $n \multimap (s \multimapinv n)$. We might expect $(x \multimap y) \multimapinv z$ is equivalent to $x \multimap (y \multimapinv z)$. In fact, we can prove this (Exercise 1), and so we can simply write $n \multimap s \multimapinv n$ as the type of "likes", which we do in the next example (in fact, it is the type of any *transitive* verb, a verb which takes a subject and an object).

Figure 10.1: Examples: the syntactic calculus

Exercise 2: Try this one for yourself: *John runs and Susan watches.* You know that *John* and *Susan* are type n, that *runs* and *watches* are type $n \multimap s$ (since they are intransitive verbs), and it's pretty easy to see that *and* is of type $s \multimap s \multimapinv s$ (it takes a sentence on either side and produces a sentence).

Now, what about pronouns? We start with some familiar simple examples.

$$\dfrac{\overset{\text{He}}{s \multimapinv (n \multimap s)} \quad \overset{\text{works}}{n \multimap s}}{s}$$

$$\dfrac{\overset{\text{He}}{s \multimapinv (n \multimap s)} \qquad \dfrac{\overset{\text{likes}}{n \multimap s \multimapinv n} \quad \overset{\text{Jane}}{n}}{n \multimap s}}{s}$$

$$\dfrac{\dfrac{\overset{\text{Jane}}{n} \quad \overset{\text{likes}}{n \multimap s \multimapinv n}}{s \multimapinv n} \qquad \overset{\text{him}}{(s \multimapinv n) \multimap s}}{s}$$

Remarks: *likes*, being a transitive verb, has type $n \multimap s \multimapinv n$ (bracketed either way). You might be tempted to give both *he* and *him* type n, but then a phrase like *Him works* would be a sentence, just as *He works* is. This is not correct—*he* and *him* have to have different types, reflecting that they are used differently. This means *he* must be typed $s \multimapinv (n \multimap s)$, and *him* must be typed $(s \multimapinv n) \multimap s$. It's important to note that there is no equivalence between $s \multimapinv (n \multimap s)$ and $(s \multimapinv n) \multimap s$—this time the arrows point the wrong way. But notice that the bracketing is demanding that *he* is followed by something to make a sentence, and that *him* is preceded by something to make a sentence. This fits the idea that *he* is a subject and *him* is an object of a sentence.

Now consider the following sentence: *He likes him.* We expect to be able to reduce $[s \multimapinv (n \multimap s)] \otimes [n \multimap s \multimapinv n] \otimes [(s \multimapinv n) \multimap s]$ to s, but there is no evident arrow that does this. However, if we had either of the following, then we could get to s:

$$\begin{aligned} (x \multimap y) \otimes (y \multimap z) &\longrightarrow x \multimap z \\ (z \multimapinv y) \otimes (y \multimapinv x) &\longrightarrow z \multimapinv x \end{aligned}$$

Exercise 3: These are consequences of the deduction rules defining monoidal closed categories, and so we do have entailments of this sort in our monoidal logic. Show how to derive them.

Then we can finish the analysis of *He likes him*, for example as follows.

$$\dfrac{\dfrac{\overset{\text{He}}{s \multimapinv (n \multimap s)} \quad \overset{\text{likes}}{(n \multimap s) \multimapinv n}}{s \multimapinv n} \qquad \overset{\text{him}}{(s \multimapinv n) \multimap s}}{s}$$

Exercise 4: do the analysis the other way, as *He (likes him)*.

These are simple examples; in the full article, Lambek illustrates the idea with more complicated examples, and indicates where his approach has problems (many of which were resolved in later work, by him and by others, and many of which he feels his current approach handles more successfully—such is the nature of research!). I suggest you read at least sections 1-3, 5, 6, and if you feel like some more mathematical material, section 7; the rest of the paper will probably be fairly heavy-going at a first attempt. One warning when you look at his article, however: his notation is different from ours. He drops the tensor, just writing it as "concatenation", so $x \otimes y$ would just be written xy, and the "lollipop arrows" become slashes, so $x \multimap y$ becomes $x\backslash y$ and $x \multimapinv y$ becomes x/y. This more compact notation has some advantages when it comes to annotating words with types (but it does lose the flavour of "conjunction" and "implication", which we wanted for our logic, replacing that intuition with one based on multiplication and division).

Solutions to the exercises in section 10.2.6

Exercise 1:

Remember the defining bijections which establish the meaning of $\multimap$ and $\multimapinv$, which we use as follows:

$$\begin{array}{c} u \longrightarrow (x \multimap y) \multimapinv z \\ \hline\hline u, z \longrightarrow x \multimap y \\ \hline\hline x, u, z \longrightarrow y \\ \hline\hline x, u \longrightarrow y \multimapinv z \\ \hline\hline u \longrightarrow x \multimap (y \multimapinv z) \end{array}$$

and letting either the top or the bottom entailment be the identity gives us the two equivalent entailments we want.

Exercise 2:

$$\begin{array}{ccc} \text{John runs} & \text{and} & \text{Susan watches} \end{array}$$

$$\cfrac{\cfrac{n \quad n \multimap s}{s} \qquad \cfrac{s \multimap s \multimapinv s \qquad \cfrac{n \quad n \multimap s}{s}}{s \multimap s}}{s}$$

Exercise 3:

Use

$$\begin{array}{c} (x \multimap y) \otimes (y \multimap z) \longrightarrow x \multimap z \\ \hline\hline x \otimes (x \multimap y) \otimes (y \multimap z) \longrightarrow z \end{array}$$

and then get the bottom arrow as follows:

$$x \otimes (x \multimap y) \otimes (y \multimap z) \longrightarrow y \otimes (y \multimap z) \longrightarrow z$$

(The other is similar.)

Exercise 4:

$$\begin{array}{ccc} \text{He} & \text{likes} & \text{him} \end{array}$$

$$\cfrac{s \multimapinv (n \multimap s) \qquad \cfrac{n \multimap (s \multimapinv n) \qquad (s \multimapinv n) \multimap s}{n \multimap s}}{s}$$

A coda:

Monoidal closed categories and pregroups are both special cases of a more general structure (linear bicategory)[17], which turns up in many other contexts, illustrating the mathematical principal that one should find common meaningful structures in seemingly unrelated contexts, and see what's "really going on" in the common underlying structure.

[17] See *Introduction to linear bicategories*, J.R.B. Cockett, J. Koslowski, R.A.G. Seely, Mathematical Structures in Computer Science (10), 2002. This paper was dedicated to Jim Lambek, in honour of his contributions to the subject.