Introduction to the Theory of Computation
Computability, Complexity,
And the Lambda Calculus
Some Notes for CIS262

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

© Jean Gallier

Please, do not reproduce without permission of the author

April 14, 2020
# Contents

## RAM Programs, Turing Machines
1. Partial Functions and RAM Programs ........................................ 10
2. Definition of a Turing Machine ........................................... 15
3. Computations of Turing Machines ....................................... 17
4. Equivalence of RAM programs And Turing Machines .............. 20
5. Listable Languages and Computable Languages ........................ 21
6. A Simple Function Not Known to be Computable .................... 22
7. The Primitive Recursive Functions ....................................... 25
8. Primitive Recursive Predicates .......................................... 33
9. The Partial Computable Functions ....................................... 35

## Universal RAM Programs and the Halting Problem
1. Pairing Functions ............................................................. 41
2. Equivalence of Alphabets .................................................... 48
3. Coding of RAM Programs; The Halting Problem ..................... 50
4. Universal RAM Programs ................................................. 54
5. Indexing of RAM Programs ................................................ 59
6. Kleene’s T-Predicate ....................................................... 60
7. A Non-Computable Function; Busy Beavers ........................... 62

## Elementary Recursive Function Theory
1. Acceptable Indexings ....................................................... 67
2. Undecidable Problems ....................................................... 70
3. Reducibility and Rice’s Theorem ......................................... 73
4. Listable (Recursively Enumerable) Sets ................................ 76
5. Reducibility and Complete Sets ......................................... 82

## The Lambda-Calculus
1. Syntax of the Lambda-Calculus ......................................... 89
2. β-Reduction and β-Conversion; the Church–Rosser Theorem .......... 94
3. Some Useful Combinators .................................................. 98
4.4 Representing the Natural Numbers ........................................... 100
4.5 Fixed-Point Combinators and Recursively Defined Functions .......... 106
4.6 \(\lambda\)-Definability of the Computable Functions ....................... 108
4.7 Definability of Functions in Typed Lambda-Calculi ...................... 114
4.8 Head Normal-Forms and the Partial Computable Functions ............ 121

5 Recursion Theory; More Advanced Topics ................................. 127
5.1 The Recursion Theorem ....................................................... 127
5.2 Extended Rice Theorem ....................................................... 133
5.3 Creative and Productive Sets; Incompleteness ........................... 136

6 Listable and Diophantine Sets; Hilbert’s Tenth ............................. 143
6.1 Diophantine Equations; Hilbert’s Tenth Problem ......................... 143
6.2 Diophantine Sets and Listable Sets ....................................... 146
6.3 Some Applications of the DPRM Theorem .................................. 150
6.4 Gödel’s Incompleteness Theorem .......................................... 153

7 The Post Correspondence Problem; Applications ......................... 159
7.1 The Post Correspondence Problem ......................................... 159
7.2 Some Undecidability Results for CFG’s ................................... 165
7.3 More Undecidable Properties of Languages ............................... 168
7.4 Undecidability of Validity in First-Order Logic ......................... 169

8 Computational Complexity; \(P\) and \(NP\) .................................... 173
8.1 The Class \(P\) ......................................................................... 173
8.2 Directed Graphs, Paths ......................................................... 175
8.3 Eulerian Cycles ..................................................................... 176
8.4 Hamiltonian Cycles ............................................................... 177
8.5 Propositional Logic and Satisfiability ...................................... 178
8.6 The Class \(NP\), \(NP\)-Completeness ........................................... 183
8.7 The Bounded Tiling Problem is \(NP\)-Complete .............................. 189
8.8 The Cook-Levin Theorem ...................................................... 193
8.9 Satisfiability of Arbitrary Propositions and CNF ....................... 196

9 Some \(NP\)-Complete Problems .................................................. 201
9.1 Statements of the Problems .................................................... 201
9.2 Proofs of \(NP\)-Completeness .................................................. 212
9.3 Succinct Certificates, co\(NP\), and \(EXP\) ...................................... 229

10 Primality Testing is in \(NP\) ...................................................... 235
10.1 Prime Numbers and Composite Numbers ................................. 235
10.2 Methods for Primality Testing ................................................. 236
10.3 Modular Arithmetic, the Groups \(Z/nZ, (Z/nZ)^*\) ......................... 239
CONTENTS

10.4 The Lucas Theorem ......................................................... 247
10.5 Lucas Trees ................................................................. 250
10.6 Algorithms for Computing Powers Modulo $m$ ......................... 253
10.7 PRIMES is in $NP$ ......................................................... 255

11 Polynomial-Space Complexity; $PS$ and $NPS$ ............................ 259
  11.1 The Classes $PS$ (or PSPACE) and $NPS$ (NPSPACE) ................. 259
  11.2 Savitch’s Theorem: $PS = NPS$ ........................................ 261
  11.3 A Complete Problem for $PS$: QBF ................................... 262
  11.4 Provability in Intuitionistic Propositional Logic ....................... 270

Bibliography ............................................................................ 277
Symbol Index .......................................................................... 281
Index ..................................................................................... 283
Chapter 1

RAM Programs, Turing Machines, and the Partial Computable Functions

See the scanned version of this chapter found in the web page for CIS511:


In this chapter we address the fundamental question

*What is a computable function?*

Nowadays computers are so pervasive that such a question may seem trivial. Isn’t the answer that a function is computable if we can write a program computing it!

This is basically the answer so what more can be said that will shed more light on the question?

The first issue is that we should be more careful about the kind of functions that we are considering. Are we restricting ourselves to total functions or are we allowing partial functions that may not be defined for some of their inputs? It turns out that if we consider functions computed by programs, then partial functions must be considered. In fact, we will see that “deciding” whether a program terminates for all inputs is impossible. But what does deciding mean?

To be mathematically precise requires a fair amount of work. One of the key technical points is the ability to design a program $U$ that takes other programs $P$ as input, and then executes $P$ on any input $x$. In particular, $U$ should be able to take $U$ itself as input!

Of course a compiler does exactly the above task. But fully describing a compiler for a “real” programming language such as JAVA, PYTHON, C++, etc. is a complicated and lengthy task. So a simpler (still quite complicated) way to proceed is to develop a toy programming language and a toy computation model (some kind of machine) capable of executing programs written in our toy language. Then we show how programs in this toy language can be coded so that they can be given as input to other programs. Having done
CHAPTER 1. RAM PROGRAMS, TURING MACHINES

this we need to demonstrate that our language has \emph{universal computing power}. This means that we need to show that a “real” program, say written in JAVA, could be translated into a possibly much longer program written in our toy language. This step is typically an act of faith, in the sense that the details that such a translation can be performed are usually not provided.

A way to be precise regarding universal computing power is to define mathematically a family of functions that should be regarded as “obviously computable,” and then to show that the functions computed by the programs written either in our toy programming language or in any modern programming language are members of this mathematically defined family of computable functions. This step is usually technically very involved, because one needs to show that executing the instructions of a program can be mimicked by functions in our family of computable functions. Conversely, we should prove that every computable function in this family is indeed computable by a program written in our toy programming language or in any modern programming language. Then we will be have the assurance that we have captured the notion of universal computing power.

Remarkably, Herbrand, Gödel, and Kleene defined such a family of functions in 1934-1935. This is a family of numerical functions $f : \mathbb{N}^m \rightarrow \mathbb{N}$ containing a subset of very simple functions called base functions, and this family is the smallest family containing the base functions closed under three operations:

1. Composition
2. Primitive recursion

Historically, the first two models of computation are the $\lambda$-calculus of Church (1935) and the \emph{Turing machine} (1936) of Turing. Kleene proved that the $\lambda$-definable functions are exactly the (total) computable functions in the sense of Herbrand–Gödel–Kleene in 1936, and Turing proved that the functions computed by Turing machines are exactly the computable functions in the sense of Herbrand–Gödel–Kleene in 1937. Therefore, the $\lambda$-calculus and Turing machines have the same “computing power,” and both compute exactly the class of computable functions in the sense of Herbrand–Gödel–Kleene. In those days these results were considered quite surprising because the formalism of the $\lambda$-calculus has basically nothing to do with the formalism of Turing machines.

Once again we should be more precise about the kinds of functions that we are dealing with. Until Turing (1936), only numerical functions $f : \mathbb{N}^m \rightarrow \mathbb{N}$ were considered. In order to compute numerical functions in the $\lambda$-calculus, Church had to encode the natural numbers are certain $\lambda$-terms, which can be viewed as iterators.

Turing assumes that what he calls his \emph{a-machines} (for automatic machines) make use of the symbols 0 and 1 for the purpose of input and output, and if the machine stops, then the output is a \emph{string} of 0s and 1s. Thus a Turing machine can be viewed as computing a
function \( f : (\{0,1\}^*)^m \rightarrow \{0,1\}^* \) on strings. By allowing a more general alphabet \( \Sigma \), we see that a Turing machine computes a function \( f : (\Sigma^*)^m \rightarrow \Sigma^* \) on strings over \( \Sigma \).

At first glance it appears that Turing machines compute a larger class of functions, but this is not so because there exist mutually invertible computable coding functions \( C : \Sigma^* \rightarrow \mathbb{N} \) and decoding functions \( D : \mathbb{N} \rightarrow \Sigma^* \). Using these coding and decoding functions, it suffices to consider numerical functions.

However, Turing machines can also very naturally be viewed as devices for defining computable languages in terms of acceptance and rejection; some kinds of generalized DFA’s of NFA’s. In this role, it would be very awkward to limit ourselves to sets of natural numbers, although this is possible in theory.

We should also point out that the notion of computable language can be handled in terms of a computation model for functions by considering the characteristic functions of languages. Indeed, a language \( A \) is computable (we say decidable) iff its characteristic function \( \chi_A \) is computable.

The above considerations motivate the definition of the computable functions in the sense of Herbrand–Gödel–Kleene to functions \( f : (\Sigma^*)^m \rightarrow \Sigma^* \) operating on strings. However, it is technically simpler to work out all the undecidability results for numerical functions of for subsets of \( \mathbb{N} \). Since there is no loss of generality in doing so in view of the computable bijections \( C : \Sigma^* \rightarrow \mathbb{N} \) and \( D : \mathbb{N} \rightarrow \Sigma^* \), we will do so.

Nevertheless, in order to deal with languages, it is important to develop a fair amount of computability theory about functions computing on strings, so we will present another computation model, the RAM program model, which computes functions defined on strings. This model was introduced around 1963 (although it was introduced earlier by Post in a different format). It has the advantage of being closer to actual computer architecture, because the RAM model consists of programs operating on a fixed set of registers. This model is equivalent to the Turing machine model, and the translations, although tedious, are not that bad.

The RAM program model also has the technical advantage that coding up a RAM program as a natural number is not that complicated.

The \( \lambda \)-calculus is a very elegant model but it is more abstract than the RAM program model and the Turing machine model so we postpone discussing it until Chapter 4.

Another very interesting computation model particularly well suited to deal with decidable sets of natural numbers is Diophantine definability. This model, arising from the work involved in proving that Hilbert’s tenth problem is undecidable, will be discussed in Chapter 6.

In the following sections we will define the RAM program model, the Turing machine model, and then argue without proofs (relegated to an appendix) that there are algorithms to convert RAM programs into Turing machines, and conversely. Then we define the class of computable functions in the sense of Herbrand–Gödel–Kleene, both for numerical functions
(defined on \( \mathbb{N} \)) and functions defined on strings. This will require explaining what is primitive recursion, which is a restricted form of recursion which guarantees that if it is applied to total functions, then the resulting function is total. Intuitively, primitive recursion corresponds to writing programs that only use for loops (loops where the number of iterations is known ahead of time and fixed).

1.1 Partial Functions and RAM Programs

In this section we define an abstract machine model for computing functions

\[
f: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_n \to \Sigma^*,
\]

where \( \Sigma = \{a_1, \ldots, a_k\} \) is some input alphabet.

Numerical functions \( f: \mathbb{N}^n \to \mathbb{N} \) can be viewed as functions defined over the one-letter alphabet \( \{a_1\} \), using the bijection \( m \mapsto a_1^m \).

Since programs are not guaranteed to terminate for all inputs, we are forced to deal with partial functions so we recall their definition.

**Definition 1.1.** A binary relation \( R \subseteq A \times B \) between two sets \( A \) and \( B \) is *functional* iff, for all \( x \in A \) and \( y, z \in B \),

\[
(x, y) \in R \quad \text{and} \quad (x, z) \in R \quad \text{implies that} \quad y = z.
\]

A *partial function* is a triple \( f = \langle A, G, B \rangle \), where \( A \) and \( B \) are arbitrary sets (possibly empty) and \( G \) is a functional relation (possibly empty) between \( A \) and \( B \), called the *graph* of \( f \).

Hence, a partial function is a functional relation such that every argument has at most one image under \( f \).

The graph of a function \( f \) is denoted as \( \text{graph}(f) \). When no confusion can arise, a function \( f \) and its graph are usually identified.

A partial function \( f = \langle A, G, B \rangle \) is often denoted as \( f: A \to B \).

The *domain* \( \text{dom}(f) \) of a partial function \( f = \langle A, G, B \rangle \) is the set

\[
\text{dom}(f) = \{ x \in A \mid \exists y \in B, (x, y) \in G \}.
\]

For every element \( x \in \text{dom}(f) \), the unique element \( y \in B \) such that \( (x, y) \in \text{graph}(f) \) is denoted as \( f(x) \). We say that \( f(x) \) is *defined*, also denoted as \( f(x) \downarrow \).

If \( x \in A \) and \( x \not\in \text{dom}(f) \), we say that \( f(x) \) is *undefined*, also denoted as \( f(x) \uparrow \).
Intuitively, if a function is partial, it does not return any output for any input not in its domain. This corresponds to an *infinite computation*. It is important to note that two partial functions \( f : A \rightarrow B \) and \( f' : A' \rightarrow B' \) are equal iff \( A = A' \), \( B = B' \), and for all \( a \in A \), either both \( f(a) \) and \( f'(a) \) are defined and \( f(a) = f'(a) \), or both \( f(a) \) and \( f'(a) \) are undefined.

A partial function \( f : A \rightarrow B \) is a *total function* iff \( \text{dom}(f) = A \). It is customary to call a total function simply a function.

We now define a model of computation known as the *RAM programs* or *Post machines*.

RAM programs are written in a sort of assembly language involving simple instructions manipulating strings stored into registers.

Every RAM program uses a fixed and finite number of *registers* denoted as \( R_1, \ldots, R_p \), with no limitation on the size of strings held in the registers.

RAM programs can be defined either in flowchart form or in linear form. Since the linear form is more convenient for coding purposes, we present RAM programs in linear form.

A RAM program \( P \) (in linear form) consists of a finite sequence of *instructions* using a finite number of registers \( R_1, \ldots, R_p \).

Instructions may optionally be labeled with line numbers denoted as \( N_1, \ldots, N_q \).

It is neither mandatory to label all instructions, nor to use distinct line numbers! Thus the same line number can be used in more than one line. As we will see later on, this makes it easier to concatenate two different programs without performing a renumbering of line numbers.

Every instruction has *four fields*, not necessarily all used. The main field is the *op-code*. Here is an example of a RAM program to concatenate two strings \( x_1 \) and \( x_2 \).

\[
\begin{align*}
R3 & \leftarrow R1 \\
R4 & \leftarrow R2 \\
N0 & \quad \text{jmp}_a \quad N1b \\
R4 & \quad \text{jmp}_b \quad N2b \\
 & \quad \text{jmp} \quad N3b \\
N1 & \quad \text{add}_a \quad R3 \\
 & \quad \text{tail} \quad R4 \\
 & \quad \text{jmp} \quad N0a \\
N2 & \quad \text{add}_b \quad R3 \\
 & \quad \text{tail} \quad R4 \\
 & \quad \text{jmp} \quad N0a \\
N3 & R1 \leftarrow R3 \\
 & \quad \text{continue}
\end{align*}
\]
Definition 1.2. RAM programs are constructed from seven types of instructions shown below:

(1) \[ N \quad add_j \quad Y \]
(2) \[ N \quad tail \quad Y \]
(3) \[ N \quad clr \quad Y \]
(4) \[ N \quad Y \quad \leftarrow \quad X \]
(5a) \[ N \quad jmp \quad N1a \]
(5b) \[ N \quad jmp \quad N1b \]
(6a) \[ N \quad Y \quad jmp_j \quad N1a \]
(6b) \[ N \quad Y \quad jmp_j \quad N1b \]
(7) \[ N \quad continue \]

1. An instruction of type (1) concatenates the letter \(a_j\) to the right of the string held by register \(Y\) \((1 \leq j \leq k)\). The effect is the assignment

\[ Y := Ya_j. \]

2. An instruction of type (2) deletes the leftmost letter of the string held by the register \(Y\). This corresponds to the function \(tail\), defined such that

\[
\begin{align*}
tail(\epsilon) &= \epsilon, \\
tail(a_ju) &= u.
\end{align*}
\]

The effect is the assignment

\[ Y := tail(Y). \]

3. An instruction of type (3) clears register \(Y\), i.e., sets its value to the empty string \(\epsilon\). The effect is the assignment

\[ Y := \epsilon. \]

4. An instruction of type (4) assigns the value of register \(X\) to register \(Y\). The effect is the assignment

\[ Y := X. \]

5. An instruction of type (5a) or (5b) is an unconditional jump.

The effect of (5a) is to jump to the closest line number \(N1\) occurring above the instruction being executed, and the effect of (5b) is to jump to the closest line number \(N1\) occurring below the instruction being executed.

6. An instruction of type (6a) or (6b) is a conditional jump. Let \(head\) be the function defined as follows:

\[
\begin{align*}
head(\epsilon) &= \epsilon, \\
head(a_ju) &= a_j.
\end{align*}
\]
The effect of $(6_ja)$ is to jump to the closest line number $N1$ occurring above the instruction being executed iff $head(Y) = a_j$, else to execute the next instruction (the one immediately following the instruction being executed).

The effect of $(6_jb)$ is to jump to the closest line number $N1$ occurring below the instruction being executed iff $head(Y) = a_j$, else to execute the next instruction.

When computing over $\mathbb{N}$, instructions of type $(6_ja)$ or $(6_jb)$ jump to the closest $N1$ above or below iff $Y$ is nonnull.

7. An instruction of type $(7)$ is a no-op, i.e., the registers are unaffected. If there is a next instruction, then it is executed, else, the program stops.

Obviously, a program is syntactically correct only if certain conditions hold.

**Definition 1.3.** A *RAM program* $P$ is a finite sequence of instructions as in Definition 1.2, and satisfying the following conditions:

1. For every jump instruction (conditional or not), the line number to be jumped to must exist in $P$.

2. The last instruction of a RAM program is a `continue`.

The reason for allowing multiple occurrences of line numbers is to make it easier to concatenate programs without having to perform a renaming of line numbers.

The technical choice of jumping to the closest address $N1$ above or below comes from the fact that it is easy to search up or down using primitive recursion, as we will see later on.

For the purpose of computing a function $f: \Sigma^* \times \cdots \times \Sigma^* \rightarrow \Sigma^*$ using a RAM program $P$, we assume that $P$ has at least $n$ registers called *input registers*, and that these registers $R1, \ldots, Rn$ are initialized with the input values of the function $f$. We also assume that the *output is returned in register R1*.

**Example 1.1.** The following RAM program concatenates two strings $x_1$ and $x_2$ held in registers $R1$ and $R2$. Since $\Sigma = \{a, b\}$, for more clarity, we wrote `jmp_a` instead of `jmp_1`, `jmp_b` instead of `jmp_2`, `add_a` instead of `add_1`, and `add_b` instead of `add_2`. 
Definition 1.4. A RAM program \( P \) computes the partial function \( \varphi: (\Sigma^*)^n \to \Sigma^* \) if the following conditions hold: For every input \((x_1, \ldots, x_n) \in (\Sigma^*)^n\), having initialized the input registers \(R_1, \ldots, R_n\) with \(x_1, \ldots, x_n\), the program eventually halts iff \(\varphi(x_1, \ldots, x_n)\) is defined, and if and when \(P\) halts, the value of \(R_1\) is equal to \(\varphi(x_1, \ldots, x_n)\). A partial function \(\varphi\) is \textit{RAM-computable} iff it is computed by some RAM program.

For example, the following program computes the\textit{ erase function} \(E\) defined such that
\[
E(u) = \epsilon
\]
for all \(u \in \Sigma^*:\)
\[
\text{clr} \quad \text{R1}
\text{continue}
\]

The following program computes the\textit{ jth successor function} \(S_j\) defined such that
\[
S_j(u) = ua_j
\]
for all \(u \in \Sigma^*:\)
\[
\text{add}_j \quad \text{R1}
\text{continue}
\]

The following program (with \(n\) input variables) computes the\textit{ projection function} \(P^n_i\)
defined such that
\[
P^n_i(u_1, \ldots, u_n) = u_i,
\]
where \(n \geq 1\), and \(1 \leq i \leq n: \)
\[
R1 \leftarrow Ri
\text{continue}
\]

Note that \(P^1_1\) is the identity function.
Having a programming language, we would like to know how powerful it is, that is, we would like to know what kind of functions are RAM-computable. At first glance, it seems that RAM programs don’t do much, but this is not so. Indeed, we will see shortly that the class of RAM-computable functions is quite extensive.

One way of getting new programs from previous ones is via composition. Another one is by primitive recursion. We will investigate these constructions after introducing another model of computation, Turing machines.

Remarkably, the classes of (partial) functions computed by RAM programs and by Turing machines are identical. This is the class of partial computable functions in the sense of Herbrand–Gödel–Kleene, also called partial recursive functions, a term which is now considered old-fashioned. We will present the definition of the so-called $\mu$-recursive functions (due to Kleene).

The following proposition will be needed to simplify the encoding of RAM programs as numbers.

**Proposition 1.1.** Every RAM program can be converted to an equivalent program only using the following type of instructions:

1. $N \quad \text{add}_j \quad Y$
2. $N \quad \text{tail} \quad Y$
3. $N \quad \text{jmp}_j \quad N \quad Y \quad N1a$
4. $N \quad \text{jmp}_j \quad N \quad Y \quad N1b$
5. $N \quad \text{continue}$

The proof is fairly simple. For example, instructions of the form $R_i \leftarrow R_j$ can be eliminated by transferring the contents of $R_j$ into an auxiliary register $R_k$, and then by transferring the contents of $R_k$ into $R_i$ and $R_j$.

### 1.2 Definition of a Turing Machine

We define a Turing machine model for computing functions

$$f: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_n \rightarrow \Sigma^*,$$

where $\Sigma = \{a_1, \ldots, a_k\}$ is some input alphabet. In this section, since we are primarily interested in computing functions we only consider deterministic Turing machines.

There are many variants of the Turing machine model. The main decision that needs to be made has to do with the kind of tape used by the machine. We opt for a single finite
tape that is both an input and a storage mechanism. This tape can be viewed as a string over tape alphabet \( \Gamma \) such that \( \Sigma \subseteq \Gamma \). There is a read/write head pointing to some symbol on the tape, symbols on the tape can be overwritten, and the read/write head can move one symbol to the left or one symbol to the right, also causing a state transition. When the write/read head attempts to move past the rightmost or the leftmost symbol on the tape, the tape is allowed to grow. To accommodate such a move, the tape alphabet contains some special symbol \( B \notin \Sigma \), the blank, and this symbol is added to the tape as the new leftmost or rightmost symbol on the tape.

A common variant uses a tape which is infinite at both ends, but only has finitely many symbols not equal to \( B \), so effectively it is equivalent to a finite tape allowed to grow at either ends. Another variant uses a semi-infinite tape infinite to the right, but with a left end. We find this model cumbersome because it requires shifting right the entire tape when a left move is attempted from the left end of the tape.

Another decision that needs to be made is the format of the instructions. Does an instruction cause both a state transition and a symbol overwrite, or do we have separate instructions for a state transition and a symbol overwrite. In the first case, an instruction can be specified as a quintuple, and in the second case by a quadruple. We opt for quintuples. Here is our definition.

**Definition 1.5.** A (deterministic) Turing machine (or TM) \( M \) is a sextuple \( M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0) \), where

- \( K \) is a finite set of states;
- \( \Sigma \) is a finite input alphabet;
- \( \Gamma \) is a finite tape alphabet, s.t. \( \Sigma \subseteq \Gamma \), \( \Gamma \cap \Gamma = \emptyset \), and with blank \( B \notin \Sigma \);
- \( q_0 \in K \) is the start state (or initial state);
- \( \delta \) is the transition function, a (finite) set of quintuples

\[
\delta \subseteq K \times \Gamma \times \Gamma \times \{L, R\} \times K,
\]

such that for all \((p, a) \in K \times \Gamma\), there is at most one triple \((b, m, q) \in \Gamma \times \{L, R\} \times K\) such that \((p, a, b, m, q) \in \delta\).

A quintuple \((p, a, b, m, q) \in \delta\) is called an instruction. It is also denoted as

\[
p, a \rightarrow b, m, q.
\]

The effect of an instruction is to switch from state \( p \) to state \( q \), overwrite the symbol currently scanned \( a \) with \( b \), and move the read/write head either left or right, according to \( m \).
Example 1.2. Here is an example of a Turing machine specified by

\[ K = \{q_0, q_1, q_2, q_3\}; \Sigma = \{a, b\}; \Gamma = \{a, b, B\}; \]

The instructions in \( \delta \) are:

- \( q_0, B \rightarrow B, R, q_3 \)
- \( q_0, a \rightarrow b, R, q_1 \)
- \( q_0, b \rightarrow a, R, q_1 \)
- \( q_1, a \rightarrow b, R, q_1 \)
- \( q_1, b \rightarrow a, R, q_1 \)
- \( q_1, B \rightarrow B, L, q_2 \)
- \( q_2, a \rightarrow a, L, q_2 \)
- \( q_2, b \rightarrow b, L, q_2 \)
- \( q_2, B \rightarrow B, R, q_3 \)

1.3 Computations of Turing Machines

To explain how a Turing machine works, we describe its action on *instantaneous descriptions*. We take advantage of the fact that \( K \cap \Gamma = \emptyset \) to define instantaneous descriptions.

**Definition 1.6.** Given a Turing machine

\[ M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0), \]

an *instantaneous description* (for short an ID) is a (nonempty) string in \( \Gamma^*K\Gamma^+ \), that is, a string of the form

\[ upav, \]

where \( u, v \in \Gamma^*, p \in K, \) and \( a \in \Gamma. \)

The intuition is that an ID \( upav \) describes a snapshot of a TM in the current state \( p \), whose tape contains the string \( uav \), and with the read/write head pointing to the symbol \( a \). Thus, in \( upav \), the state \( p \) is just to the left of the symbol presently scanned by the read/write head.

We explain how a TM works by showing how it acts on ID’s.

**Definition 1.7.** Given a Turing machine

\[ M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0), \]

the *yield relation (or compute relation)* \( \vdash \) is a binary relation defined on the set of ID’s as follows. For any two ID’s \( ID_1 \) and \( ID_2 \), we have \( ID_1 \vdash ID_2 \) iff either
(1) \((p, a, b, R, q) \in \delta\), and either
   
   (a) \(ID_1 = upacv\), \(c \in \Gamma\), and \(ID_2 = ubqcv\), or
   
   (b) \(ID_1 = upa\) and \(ID_2 = ubqB\);

or

(2) \((p, a, b, L, q) \in \delta\), and either
   
   (a) \(ID_1 = ucpav\), \(c \in \Gamma\), and \(ID_2 = uqcbv\), or
   
   (b) \(ID_1 = pav\) and \(ID_2 = qBbv\).

Note how the tape is extended by one blank after the rightmost symbol in case (1)(b), and by one blank before the leftmost symbol in case (2)(b).

As usual, we let \(\vdash^+\) denote the transitive closure of \(\vdash\), and we let \(\vdash^*\) denote the reflexive and transitive closure of \(\vdash\). We can now explain how a Turing machine computes a partial function

\[
f: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_n \to \Sigma^*.
\]

Since we allow functions taking \(n \geq 1\) input strings, we assume that \(\Gamma\) contains the special delimiter \(\_,\) not in \(\Sigma\), used to separate the various input strings.

It is convenient to assume that a Turing machine “cleans up” its tape when it halts before returning its output. What this means is that when the Turing machine halts, the output should be clearly identifiable, so all symbols not in \(\Sigma \cup \{B\}\) that may have been used during the computation must be erased. Thus when the TM stops the tape must consist of a string \(w \in \Sigma^*\) possibly surrounded by blanks (the symbol \(B\)). Actually, if the output is \(\epsilon\), the tape must contain a nonempty string of blanks. To achieve this technically, we define proper ID’s.

**Definition 1.8.** Given a Turing machine

\[
M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0),
\]

where \(\Gamma\) contains some delimiter \(\_,\), not in \(\Sigma\) in addition to the blank \(B\), a **starting ID** is of the form

\[
q_0 w_1, w_2, \ldots, w_n
\]

where \(w_1, \ldots, w_n \in \Sigma^*\) and \(n \geq 2\), or \(q_0 w\) with \(w \in \Sigma^+\), or \(q_0 B\).

A **blocking (or halting) ID** is an ID \(upav\) such that there are no instructions \((p, a, b, m, q) \in \delta\) for any \((b, m, q) \in \Gamma \times \{L, R\} \times K\).

A **proper ID** is a halting ID of the form

\[
B^h pwB^l,
\]

where \(w \in \Sigma^*\), and \(h, l \geq 0\) (with \(h \geq 1\) when \(w = \epsilon\)).
Computation sequences are defined as follows.

**Definition 1.9.** Given a Turing machine

\[ M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0), \]

a *computation sequence (or computation)* is a finite or infinite sequence of ID’s

\[ ID_0, ID_1, \ldots, ID_i, ID_{i+1}, \ldots, \]

such that \( ID_i \vdash ID_{i+1} \) for all \( i \geq 0 \).

A computation sequence *halts* iff it is a finite sequence of ID’s, so that

\[ ID_0 \vdash^* ID_n, \]

and \( ID_n \) is a halting ID.

A computation sequence *diverges* if it is an infinite sequence of ID’s.

We now explain how a Turing machine computes a partial function.

**Definition 1.10.** A Turing machine

\[ M = (K, \Sigma, \Gamma, \{L, R\}, \delta, q_0) \]

*computes the partial function*

\[ f: (\Sigma^*)^n \rightarrow \Sigma^* \]

iff the following conditions hold:

1. For every \( w_1, \ldots, w_n \in \Sigma^* \), given the starting ID

\[ ID_0 = q_0w_1w_2\ldots w_n \]

or \( q_0w \) with \( w \in \Sigma^+ \), or \( q_0B \), the computation sequence of \( M \) from \( ID_0 \) halts in a proper ID iff \( f(w_1, \ldots, w_n) \) is defined.

2. If \( f(w_1, \ldots, w_n) \) is defined, then \( M \) halts in a proper ID of the form

\[ ID_n = B^hpf(w_1, \ldots, w_n)B^l, \]

which means that it computes the right value.

A function \( f \) (over \( \Sigma^* \)) is *Turing computable* iff it is computed by some Turing machine \( M \).
Note that by (1), the TM $M$ may halt in an improper ID, in which case $f(w_1, \ldots, w_n)$ must be undefined. This corresponds to the fact that we only accept to retrieve the output of a computation if the TM has cleaned up its tape, i.e., produced a proper ID. In particular, intermediate calculations have to be erased before halting.

**Example 1.3.** Consider the Turing machine specified by $K = \{q_0, q_1, q_2, q_3\}$; $\Sigma = \{a, b\}$; $\Gamma = \{a, b, B\}$.

The instructions in $\delta$ are:

- $q_0, B \rightarrow B, R, q_3,$
- $q_0, a \rightarrow b, R, q_1,$
- $q_0, b \rightarrow a, R, q_1,$
- $q_1, a \rightarrow b, R, q_1,$
- $q_1, b \rightarrow a, R, q_1,$
- $q_1, B \rightarrow B, L, q_2,$
- $q_2, a \rightarrow a, L, q_2,$
- $q_2, b \rightarrow b, L, q_2,$
- $q_2, B \rightarrow B, R, q_3.$

The reader can easily verify that this machine exchanges the $a$'s and $b$'s in a string. For example, on input $w = aaababb$, the output is $bbbabaa$.

### 1.4 Equivalence of RAM Programs And Turing Machines

Turing machines can simulate RAM programs, and as a result, we have the following Theorem.

**Theorem 1.2.** Every RAM-computable function is Turing-computable. Furthermore, given a RAM program $P$, we can effectively construct a Turing machine $M$ computing the same function.

The idea of the proof is to represent the contents of the registers $R_1, \ldots, R_p$ on the Turing machine tape by the string

$$\#r_1\#r_2\# \cdots \#r_p\#,$$

where $\#$ is a special marker and $r_i$ represents the string held by $R_i$. We also use Proposition 1.1 to reduce the number of instructions to be dealt with.
The Turing machine $M$ is built of blocks, each block simulating the effect of some instruction of the program $P$. The details are a bit tedious, and can be found in an appendix or in Machtay and Young [25].

RAM programs can also simulate Turing machines.

**Theorem 1.3.** Every Turing-computable function is RAM-computable. Furthermore, given a Turing machine $M$, one can effectively construct a RAM program $P$ computing the same function.

The idea of the proof is to design a RAM program containing an encoding of the current ID of the Turing machine $M$ in register $R_1$, and to use other registers $R_2, R_3$ to simulate the effect of executing an instruction of $M$ by updating the ID of $M$ in $R_1$.

The details are tedious and can be found in an appendix.

Another proof can be obtained by proving that the class of Turing computable functions coincides with the class of partial computable functions (formerly called partial recursive functions), to be defined shortly. Indeed, it turns out that both RAM programs and Turing machines compute precisely the class of partial recursive functions. For this, we will need to define the primitive recursive functions.

Informally, a primitive recursive function is a total recursive function that can be computed using only for loops, that is, loops in which the number of iterations is fixed (unlike a while loop). A formal definition of the primitive functions is given in Section 1.7. For the time being we make the following provisional definition.

**Definition 1.11.** Let $\Sigma = \{a_1, \ldots, a_k\}$. The class of partial computable functions also called partial recursive functions is the class of partial functions (over $\Sigma^*$) that can be computed by RAM programs (or equivalently by Turing machines).

The class of computable functions also called recursive functions is the subset of the class of partial computable functions consisting of functions defined for every input (i.e., total functions).

Turing machines can also be used as acceptors to define languages so we introduce the basic relevant definitions. A more detailed study of these languages will be provided in Chapter 3.

### 1.5 Listable Languages and Computable Languages

We define the computably enumerable languages, also called listable languages, and the computable languages. The old-fashion terminology for listable languages is recursively enumerable languages, and for computable languages is recursive languages.

When operating as an acceptor, a Turing machine takes a single string as input and either goes on forever or halts with the answer “accept” or “reject.” One way to deal with
acceptance or rejection is to assume that the TM has a set of final states. Another way more consistent with our view that machines compute functions is to assume that the TM’s under consideration have a tape alphabet containing the special symbols 0 and 1. Then acceptance is signaled by the output 1, and rejection is signaled by the output 0.

Note that with our convention that in order to produce an output a TM must halt in a proper ID, the TM must erase the tape before outputing 0 or 1.

**Definition 1.12.** Let $\Sigma = \{a_1, \ldots, a_k\}$. A language $L \subseteq \Sigma^*$ is (Turing) listable or (Turing) computably enumerable (for short, a c.e. set) (or recursively enumerable (for short, a r.e. set)) iff there is some TM $M$ such that for every $w \in L$, $M$ halts in a proper ID with the output 1, and for every $w \notin L$, either $M$ halts in a proper ID with the output 0 or it runs forever.

A language $L \subseteq \Sigma^*$ is (Turing) computable (or recursive) iff there is some TM $M$ such that for every $w \in L$, $M$ halts in a proper ID with the output 1, and for every $w \notin L$, $M$ halts in a proper ID with the output 0.

Thus, given a computably enumerable language $L$, for some $w \notin L$, it is possible that a TM accepting $L$ runs forever on input $w$. On the other hand, for a computable (recursive) language $L$, a TM accepting $L$ always halts in a proper ID.

When dealing with languages, it is often useful to consider nondeterministic Turing machines. Such machines are defined just like deterministic Turing machines, except that their transition function $\delta$ is just a (finite) set of quintuples

$$\delta \subseteq K \times \Gamma \times \Gamma \times \{L, R\} \times K,$$

with no particular extra condition.

It can be shown that every nondeterministic Turing machine can be simulated by a deterministic Turing machine, and thus, nondeterministic Turing machines also accept the class of c.e. sets. This is a very tedious simulation, and very few books actually provide all the details!

It can be shown that a computably enumerable language is the range of some computable (recursive) function; see Section 3.4. It can also be shown that a language $L$ is computable (recursive) iff both $L$ and its complement are computably enumerable; see Section 3.4. There are computably enumerable languages that are not computable (recursive); see Section 3.4.

### 1.6 A Simple Function Not Known to be Computable

The “$3n + 1$ problem” proposed by Collatz around 1937 is the following:
Given any positive integer \( n \geq 1 \), construct the sequence \( c_i(n) \) as follows starting with \( i = 1 \):

\[
c_1(n) = n \\
c_{i+1}(n) = \begin{cases} 
\frac{c_i(n)}{2} & \text{if } c_i(n) \text{ is even} \\
3c_i(n) + 1 & \text{if } c_i(n) \text{ is odd}.
\end{cases}
\]

Observe that for \( n = 1 \), we get the infinite periodic sequence

\[
1 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1 \Rightarrow \cdots,
\]

so we may assume that we stop the first time that the sequence \( c_i(n) \) reaches the value 1 (if it actually does). Such an index \( i \) is called the *stopping time* of the sequence. And this is the problem:

**Conjecture (Collatz):**

For any starting integer value \( n \geq 1 \), the sequence \( (c_i(n)) \) always reaches 1.

Starting with \( n = 3 \), we get the sequence

\[
3 \Rightarrow 10 \Rightarrow 5 \Rightarrow 16 \Rightarrow 8 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1.
\]

Starting with \( n = 5 \), we get the sequence

\[
5 \Rightarrow 16 \Rightarrow 8 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1.
\]

Starting with \( n = 6 \), we get the sequence

\[
6 \Rightarrow 3 \Rightarrow 10 \Rightarrow 5 \Rightarrow 16 \Rightarrow 8 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1.
\]

Starting with \( n = 7 \), we get the sequence

\[
7 \Rightarrow 22 \Rightarrow 11 \Rightarrow 34 \Rightarrow 17 \Rightarrow 52 \Rightarrow 26 \Rightarrow 13 \Rightarrow 40 \\
\Rightarrow 20 \Rightarrow 10 \Rightarrow 25 \Rightarrow 16 \Rightarrow 8 \Rightarrow 4 \Rightarrow 2 \Rightarrow 1.
\]

One might be surprised to find that for \( n = 27 \), it takes 111 steps to reach 1, and for \( n = 97 \), it takes 118 steps. I computed the stopping times for \( n \) up to \( 10^7 \) and found that the largest stopping time, 686 (685 steps) is obtained for \( n = 8400511 \). The terms of this sequence reach values over \( 1.5 \times 10^{11} \). The graph of the sequence \( c(8400511) \) is shown in Figure 1.1.

We can define the partial computable function \( C \) (with positive integer inputs) defined by

\[
C(n) = \text{the smallest } i \text{ such that } c_i(n) = 1 \text{ if it exists.}
\]
Then the Collatz conjecture is equivalent to asserting that the function \( C \) is (total) computable. The graph of the function \( C \) for \( 1 \leq n \leq 10^7 \) is shown in Figure 1.2.

So far, the conjecture remains open. It has been checked by computer for all integers less than or equal to \( 87 \times 2^{60} \).

We now return to the computability of functions. Our goal is to define the partial computable functions in the sense of Herbrand–Gödel–Kleene. This class of functions is defined from some base functions in terms of three closure operations:

1. Composition

2. Primitive recursion

1.7. THE PRIMITIVE RECURSIVE FUNCTIONS

Historically the primitive recursive functions were defined for numerical functions (computing on the natural numbers). Since one of our goals is to show that the RAM-computable functions are partial recursive, we define the primitive recursive functions as functions \( f : (\Sigma^*)^m \rightarrow \Sigma^* \), where \( \Sigma = \{a_1, \ldots, a_k\} \) is a finite alphabet. As usual, by assuming that \( \Sigma = \{a_1\} \), we can deal with numerical functions \( f : \mathbb{N}^m \rightarrow \mathbb{N} \).

The class of primitive recursive functions is defined in terms of base functions and two closure operations.
Definition 1.13. Let $\Sigma = \{a_1, \ldots, a_k\}$. The base functions over $\Sigma$ are the following functions:

1. The erase function $E$, defined such that $E(w) = \epsilon$, for all $w \in \Sigma^*$;
2. For every $j$, $1 \leq j \leq k$, the $j$-successor function $S_j$, defined such that $S_j(w) = wa_j$, for all $w \in \Sigma^*$;
3. The projection functions $P^n_i$, defined such that

$$P^n_i(w_1, \ldots, w_n) = w_i,$$

for every $n \geq 1$, every $i$, $1 \leq i \leq n$, and for all $w_1, \ldots, w_n \in \Sigma^*$.

Note that $P^1_1$ is the identity function on $\Sigma^*$. Projection functions can be used to permute, duplicate, or drop the arguments of another function.

In the special case where we are only considering numerical functions ($\Sigma = \{a_1\}$), the function $E: \mathbb{N} \to \mathbb{N}$ is the zero function given by $E(n) = 0$ for all $n \in \mathbb{Z}$, and it is often denoted by $Z$. There is a single successor function $S_{a_1}: \mathbb{N} \to \mathbb{N}$ usually denoted $S$ (or $\text{Succ}$) given by $S(n) = n + 1$ for all $n \in \mathbb{N}$.

Even though in this section we are primarily interested in total functions, later on, the same closure operations will be applied to partial functions so we state the definition of the closure operations in the more general case of partial functions. The first closure operation is (extended) composition.

Definition 1.14. Let $\Sigma = \{a_1, \ldots, a_k\}$. For any partial or total function

$$g: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_m \to \Sigma^*,$$

and any $m \geq 1$ partial or total functions

$$h_i: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_n \to \Sigma^*, \quad n \geq 1,$$

the composition of $g$ and the $h_i$ is the partial function

$$f: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_n \to \Sigma^*,$$

denoted as $g \circ (h_1, \ldots, h_m)$, such that

$$f(w_1, \ldots, w_n) = g(h_1(w_1, \ldots, w_n), \ldots, h_m(w_1, \ldots, w_n)),$$

for all $w_1, \ldots, w_n \in \Sigma^*$. If $g$ and all the $h_i$ are total functions, then $g \circ (h_1, \ldots, h_m)$ is obviously a total function. But if $g$ or any of the $h_i$ is a partial function, then the value $(g \circ (h_1, \ldots, h_m))(x_1, \ldots, x_n)$ is defined if and only if all the values $h_i(x_1, \ldots, x_n)$ are defined for $i = 1, \ldots, m$, and $g(h_1(w_1, \ldots, w_n), \ldots, h_m(w_1, \ldots, w_n))$ is defined.
Thus even if $g$ “ignores” some of its inputs, in computing $g(h_1(w_1,\ldots,w_n),\ldots,h_m(w_1,\ldots,w_n))$, all arguments $h_1(w_1,\ldots,w_n)$ must be evaluated.

As an example of a composition, $f = g \circ (P^2_2, P^2_1)$ is such that

$$f(w_1, w_2) = g(P^2_2(w_1, w_2), P^2_1(w_1, w_2)) = g(w_2, w_1).$$

The second closure operation is primitive recursion. First we define primitive recursion for numerical functions because it is simpler.

**Definition 1.15.** Given any two partial or total functions $g : \mathbb{N}^{m-1} \to \mathbb{N}$ and $h : \mathbb{N}^{m+1} \to \mathbb{N}$ ($m \geq 2$), the partial or total function $f : \mathbb{N}^m \to \mathbb{N}$ is defined by **primitive recursion from $g$ and $h$** if $f$ is given by

$$f(0, x_2, \ldots, x_m) = g(x_2, \ldots, x_m),$$

$$f(n + 1, x_2, \ldots, x_m) = h(x, f(n, x_2, \ldots, x_m), x_2, \ldots, x_m),$$

for all $n, x_2, \ldots, x_m \in \mathbb{N}$. When $m = 1$, we have

$$f(0) = b,$$

$$f(n + 1) = h(x, f(x)), \quad \text{for all } n \in \mathbb{N},$$

for some fixed natural number $b \in \mathbb{N}$.

If $g$ and $h$ are total functions, it is easy to show that $f$ is also a total function. If $g$ or $h$ is partial, obviously $f(0, x_2, \ldots, x_m)$ is defined iff $g(x_2, \ldots, x_m)$ is defined, and $f(n + 1, x_2, \ldots, x_m)$ is defined iff $f(n, x_2, \ldots, x_m)$ is defined and $h(n, f(n, x_2, \ldots, x_m), x_2, \ldots, x_m)$ is defined.

Definition 1.15 is quite a straightjacket in the sense that $n + 1$ must be the first argument of $f$, and the definition only applies if $h$ has $m + 1$ arguments, but in practice a “natural” definition often ignores the argument $n$ and some of the arguments $x_2, \ldots, x_m$. This is where the projection functions come into play to drop, duplicate, or permute arguments.

For example, a “natural” definition of the predecessor function $\text{pred}$ is

$$\text{pred}(0) = 0$$

$$\text{pred}(m + 1) = m,$$

but this is not a legal primitive recursive definition. To make it a legal primitive recursive definition we need the function $h = P^2_1$, and a legal primitive recursive definition for $\text{pred}$ is

$$\text{pred}(0) = 0$$

$$\text{pred}(m + 1) = P^2_1(m, \text{pred}(m)).$$
Addition, multiplication, exponentiation, and super-exponentiation, can be defined by
primitive recursion as follows (being a bit loose, for supexp we should use some projections ...

\[
\begin{align*}
\text{add}(0, n) &= P^1_1(n) = n, \\
\text{add}(m + 1, n) &= S \circ P^3_2(m, \text{add}(m, n), n) \\
&= S(\text{add}(m, n)) \\
\text{mult}(0, n) &= E(n) = 0, \\
\text{mult}(m + 1, n) &= \text{add} \circ (P^3_2, P^3_3)(m, \text{mult}(m, n), n) \\
&= \text{add}(\text{mult}(m, n), n), \\
\text{rexp}(0, n) &= S \circ E(n) = 1, \\
\text{rexp}(m + 1, n) &= \text{mult}(\text{rexp}(m, n), n), \\
\text{exp}(m, n) &= \text{rexp} \circ (P^2_2, P^2_1)(m, n), \\
\text{supexp}(0, n) &= 1, \\
\text{supexp}(m + 1, n) &= \text{exp}(n, \text{supexp}(m, n)).
\end{align*}
\]

We usually write \(m + n\) for \(\text{add}(m, n)\), \(m \star n\) or even \(mn\) for \(\text{mult}(m, n)\), and \(m^n\) for \(\text{exp}(m, n)\).

There is a minus operation on \(\mathbb{N}\) named \textit{monus}. This operation denoted by \(\div\) is defined by

\[
m \div n = \begin{cases} 
& m - n \quad \text{if } m \geq n \\
& 0 \quad \text{if } m < n.
\end{cases}
\]

Then \textit{monus} is defined by

\[
m \div 0 = m \\
m \div (n + 1) = \text{pred}(m \div n),
\]

except that the above is not a legal primitive recursion. For one thing, recursion should be performed on \(m\), not \(n\). We can define \textit{rmonus} as

\[
\text{rmonus}(n, m) = m \div n,
\]

and then \(m \div n = (\text{rmonus} \circ (P^2_2, P^2_1))(m, n)\), and

\[
\begin{align*}
\text{rmonus}(0 \div m) &= P^1_1(m) \\
\text{rmonus}(n + 1, m) &= \text{pred} \circ P^2_2(n, \text{rmonus}(n, m)).
\end{align*}
\]

The following functions are also primitive recursive:

\[
\text{sg}(n) = \begin{cases} 
& 1 \quad \text{if } n > 0 \\
& 0 \quad \text{if } n = 0,
\end{cases}
\]
1.7. THE PRIMITIVE RECURSIVE FUNCTIONS

\[ \overline{sg}(n) = \begin{cases} 
0 & \text{if } n > 0 \\
1 & \text{if } n = 0,
\end{cases} \]

as well as

\[ abs(m, n) = |m - n| = m \div n + n \div m, \]

and

\[ eq(m, n) = \begin{cases} 
1 & \text{if } m = n \\
0 & \text{if } m \neq n.
\end{cases} \]

Indeed

\[ sg(0) = 0 \]
\[ sg(n + 1) = S \circ E \circ P_1^2(n, sg(n)) \]
\[ \overline{sg}(n) = S(E(n)) \div sg(n) = 1 \div sg(n), \]

and

\[ eq(m, n) = sg(|m - n|). \]

Finally, the function

\[ cond(m, n, p, q) = \begin{cases} 
p & \text{if } m = n \\
q & \text{if } m \neq n,
\end{cases} \]

is primitive recursive since

\[ cond(m, n, p, q) = eq(m, n) \star p + \overline{sg}(eq(m, n)) \star q. \]

We can also design more general version of \( cond \). For example, define \( compare_{\leq} \) as

\[ compare_{\leq}(m, n) = \begin{cases} 
1 & \text{if } m \leq n \\
0 & \text{if } m > n,
\end{cases} \]

which is given by

\[ compare_{\leq}(m, n) = 1 \div sg(m \div n). \]

Then we can define

\[ cond_{\leq}(m, n, p, q) = \begin{cases} 
p & \text{if } m \leq n \\
q & \text{if } m > n,
\end{cases} \]

with

\[ cond_{\leq}(m, n, n, p) = compare_{\leq}(m, n) \star p + \overline{sg}(compare_{\leq}(m, n)) \star q. \]

The above allows to define functions by cases.

We now generalize primitive recursion to functions defined on strings (in \( \Sigma^* \)). The new twist is that instead of the argument \( n + 1 \) of \( f \), we need to consider the \( k \) arguments \( u a_i \) of \( f \) for \( i = 1, \ldots, k \) (with \( u \in \Sigma^* \)), so instead of a single function \( h \), we need \( k \) functions \( h_i \) to define primitive recursively what \( f(u a_i, w_2, \ldots, w_m) \) is.
Definition 1.16. Let $\Sigma = \{a_1, \ldots, a_k\}$. For any partial or total function
\[
g : \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_{m-1} \rightarrow \Sigma^*,
\]
where $m \geq 2$, and any $k$ partial or total functions
\[
h_i : \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_{m+1} \rightarrow \Sigma^*,
\]
the partial function
\[
f : \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_{m} \rightarrow \Sigma^*,
\]
is defined by \textit{primitive recursion from $g$ and $h_1, \ldots, h_k$}, if
\[
f(\epsilon, w_2, \ldots, w_m) = g(w_2, \ldots, w_m),
\]
\[
f(ua_1, w_2, \ldots, w_m) = h_1(u, f(u, w_2, \ldots, w_m), w_2, \ldots, w_m),
\]
\[
\cdots = \cdots
\]
\[
f(ua_k, w_2, \ldots, w_m) = h_k(u, f(u, w_2, \ldots, w_m), w_2, \ldots, w_m),
\]
for all $u, w_2, \ldots, w_m \in \Sigma^*$.

When $m = 1$, for some fixed $w \in \Sigma^*$, we have
\[
f(\epsilon) = w,
\]
\[
f(ua_1) = h_1(u, f(u)),
\]
\[
\cdots = \cdots
\]
\[
f(ua_k) = h_k(u, f(u)),
\]
for all $u \in \Sigma^*$.

Again, if $g$ and the $h_i$ are total, it is easy to see that $f$ is total.

As an example over $\{a, b\}^*$, the following function $g : \Sigma^* \times \Sigma^* \rightarrow \Sigma^*$, is defined by primitive recursion:
\[
g(\epsilon, v) = P_1^1(v),
\]
\[
g(ua_i, v) = S_i \circ P_2^i(u, g(u, v), v),
\]
where $1 \leq i \leq k$. It is easily verified that $g(u, v) = vu$. Then,
\[
\text{con} = g \circ (P_2^2, P_1^2)
\]
computes the concatenation function, \textit{i.e.}, $\text{con}(u, v) = uv$. Here are some primitive recursive functions that often appear as building blocks for other primitive recursive functions.
The delete last function $dell$ given by

$$
dell(\epsilon) = \epsilon
$$

$$
dell(ua_i) = u, \quad 1 \leq i \leq k, \; u \in \Sigma^*
$$
is defined primitive recursively by

$$
dell(\epsilon) = \epsilon
$$

$$
dell(ua_i) = P_1^2(u, dell(u)), \quad 1 \leq i \leq k, \; u \in \Sigma^*.
$$

For every string $w \in \Sigma^*$, the constant function $c_w$ given by

$$
c_w(u) = w \quad \text{for all} \; u \in \Sigma^*
$$
is defined primitive recursively by induction on the length of $w$ by

$$
c_\epsilon = E
$$

$$
c_{va_i} = S_i \circ c_v, \quad 1 \leq i \leq k.
$$

The sign function $sg$ given by

$$
sg(x) = \begin{cases} 
\epsilon & \text{if } x = \epsilon \\
 a_1 & \text{if } x \neq \epsilon 
\end{cases}
$$
is defined primitive recursively by

$$
sg(\epsilon) = \epsilon
$$

$$
sg(ua_i) = (c_{a_1} \circ P_1^2)(u, sg(u)).
$$

The anti-sign function $\bar{sg}$ given by

$$
\bar{sg}(x) = \begin{cases} 
a_1 & \text{if } x = \epsilon \\
 \epsilon & \text{if } x \neq \epsilon 
\end{cases}
$$
is primitive recursive. The proof is left an an exercise.

The function $end_j$ $(1 \leq j \leq k)$ given by

$$
end_j(x) = \begin{cases} 
a_1 & \text{if } x \text{ ends with } a_j \\
 \epsilon & \text{otherwise}
\end{cases}
$$
is primitive recursive. The proof is left an an exercise.

The reverse function $rev: \Sigma^* \rightarrow \Sigma^*$ given by $rev(u) = u^R$ is primitive recursive, because

$$
rev(\epsilon) = \epsilon
$$

$$
rev(ua_i) = (con \circ (c_{a_i} \circ P_1^2, P_2^2))(u, rev(u)), \quad 1 \leq i \leq k.
$$
The *tail function* `tail` given by

\[
\begin{align*}
\text{tail}(\epsilon) &= \epsilon \\
\text{tail}(a_iu) &= u
\end{align*}
\]

is primitive recursive, because

\[
tail = \text{rev} \circ \text{dell} \circ \text{rev}.
\]

The *last function* `last` given by

\[
\begin{align*}
\text{last}(\epsilon) &= \epsilon \\
\text{last}(ua_i) &= a_i
\end{align*}
\]

is primitive recursive, because

\[
\begin{align*}
\text{last}(\epsilon) &= \epsilon \\
\text{last}(ua_i) &= c_{a_i} \circ P_1^2(u, \text{last}(u)).
\end{align*}
\]

The *head function* `head` given by

\[
\begin{align*}
\text{head}(\epsilon) &= \epsilon \\
\text{head}(a_iu) &= a_i
\end{align*}
\]

is primitive recursive, because

\[
\text{head} = \text{last} \circ \text{rev}.
\]

We are now ready to define the class of primitive recursive functions.

**Definition 1.17.** Let \( \Sigma = \{a_1, \ldots, a_k\} \). The class of primitive recursive functions is the smallest class of (total) functions (over \( \Sigma^* \)) which contains the base functions and is closed under composition and primitive recursion.

In the special where \( k = 1 \), we obtain the class of numerical primitive recursive functions.

The class of primitive recursive functions may not seem very big, but it contains all the total functions that we would ever want to compute. Although it is rather tedious to prove, the following theorem can be shown.

**Theorem 1.4.** For any alphabet \( \Sigma = \{a_1, \ldots, a_k\} \), every primitive recursive function is RAM computable, and thus Turing computable.

The proof is given in an appendix.

In order to define new functions it is also useful to use predicates.
1.8 Primitive Recursive Predicates

Primitive recursive predicates will be used in Section 2.3.

Definition 1.18. An $n$-ary predicate $P$ over $\mathbb{N}$ is any subset of $\mathbb{N}^n$. We write that a tuple $(x_1, \ldots, x_n)$ satisfies $P$ as $(x_1, \ldots, x_n) \in P$ or as $P(x_1, \ldots, x_n)$. The characteristic function of a predicate $P$ is the function $C_P: \mathbb{N}^n \to \{0, 1\}$ defined by

$$C_P(x_1, \ldots, x_n) = \begin{cases} 1 & \text{iff } P(x_1, \ldots, x_n) \text{ holds} \\ 0 & \text{iff } \neg P(x_1, \ldots, x_n). \end{cases}$$

A predicate $P$ (over $\mathbb{N}$) is primitive recursive iff its characteristic function $C_P$ is primitive recursive.

More generally, an $n$-ary predicate $P$ (over $\Sigma^*$) is any subset of $(\Sigma^*)^n$. We write that a tuple $(x_1, \ldots, x_n)$ satisfies $P$ as $(x_1, \ldots, x_n) \in P$ or as $P(x_1, \ldots, x_n)$. The characteristic function of a predicate $P$ is the function $C_P: (\Sigma^*)^n \to \{a_1\}^*$ defined by

$$C_P(x_1, \ldots, x_n) = \begin{cases} a_1 & \text{iff } P(x_1, \ldots, x_n) \text{ holds} \\ \epsilon & \text{iff } \neg P(x_1, \ldots, x_n). \end{cases}$$

A predicate $P$ (over $\Sigma^*$) is primitive recursive iff its characteristic function $C_P$ is primitive recursive.

Since we will only need to use primitive recursive predicates over $\mathbb{N}$ in the following chapters, for simplicity of exposition we will restrict ourselves to such predicates. The general case is treated in Machtey and Young [25].

It is easily shown that if $P$ and $Q$ are primitive recursive predicates (over $(\mathbb{N}^n)$, then $P \lor Q$, $P \land Q$ and $\neg P$ are also primitive recursive.

As an exercise, the reader may want to prove that the predicate, prime($n$) iff $n$ is a prime number, is a primitive recursive predicate.

For any fixed $k \geq 1$, the function

$$\text{ord}(k, n) = \text{exponent of the } k\text{th prime in the prime factorization of } n,$$

is a primitive recursive function.

We can also define functions by cases.

Proposition 1.5. If $P_1, \ldots, P_n$ are pairwise disjoint primitive recursive predicates (which means that $P_i \cap P_j = \emptyset$ for all $i \neq j$) and $f_1, \ldots, f_{n+1}$ are primitive recursive functions, the function $g$ defined below is also primitive recursive:

$$g(\bar{x}) = \begin{cases} f_1(\bar{x}) & \text{iff } P_1(\bar{x}) \\ \vdots \\ f_n(\bar{x}) & \text{iff } P_n(\bar{x}) \\ f_{n+1}(\bar{x}) & \text{otherwise}. \end{cases}$$

Here we write $\bar{x}$ for $(x_1, \ldots, x_n)$. 
It is also useful to have bounded quantification and bounded minimization. Recall that we are restricting our attention to numerical predicates and functions, so all variables range over \( \mathbb{N} \). Proofs of the results stated below can be found in Machtey and Young [25].

**Definition 1.19.** If \( P \) is an \((n+1)\)-ary predicate, then the **bounded existential predicate** \((\exists y \leq x)P(y, z)\) holds iff some \( y \leq x \) makes \( P(y, z) \) true.

The **bounded universal predicate** \((\forall y \leq x)P(y, z)\) holds iff every \( y \leq x \) makes \( P(y, z) \) true.

**Proposition 1.6.** If \( P \) is an \((n+1)\)-ary primitive recursive predicate, then \((\exists y \leq x)P(y, z)\) and \((\forall y \leq x)P(y, z)\) are also primitive recursive predicates.

As an application, we can show that the equality predicate, \( u = v \), is primitive recursive. The following slight generalization of Proposition 1.6 will be needed in Section 2.3.

**Proposition 1.7.** If \( P \) is an \((n+1)\)-ary primitive recursive predicate and \( f : \mathbb{N}^n \rightarrow \mathbb{N} \) is a primitive recursive function, then \((\exists y \leq f(z))P(y, z)\) is also a primitive recursive predicate.

**Definition 1.20.** If \( P \) is an \((n+1)\)-ary predicate, then the **bounded minimization of** \( P \), \( \text{min}(y \leq x)P(y, z) \), is the function defined such that \( \text{min}(y \leq x)P(y, z) \) is the least natural number \( y \leq x \) such that \( P(y, z) \) if such a \( y \) exists, \( x + 1 \) otherwise.

The **bounded maximization of** \( P \), \( \text{max}(y \leq x)P(y, z) \), is the function defined such that \( \text{max}(y \leq x)P(y, z) \) is the largest natural number \( y \leq x \) such that \( P(y, z) \) if such a \( y \) exists, \( x + 1 \) otherwise.

**Proposition 1.8.** If \( P \) is an \((n+1)\)-ary primitive recursive predicate, then \( \text{min}(y \leq x)P(y, z) \) and \( \text{max}(y \leq x)P(y, z) \) are primitive recursive functions.

So far the primitive recursive functions do not yield all the Turing-computable functions. The following proposition also shows that restricting ourselves to total functions is too limiting.

Let \( \mathcal{F} \) be any set of total functions that contains the base functions and is closed under composition and primitive recursion (and thus, \( \mathcal{F} \) contains all the primitive recursive functions).

**Definition 1.21.** We say that a function \( f : \Sigma^* \times \Sigma^* \rightarrow \Sigma^* \) is **universal** for the one-argument functions in \( \mathcal{F} \) iff for every function \( g : \Sigma^* \rightarrow \Sigma^* \) in \( \mathcal{F} \), there is some \( n \in \mathbb{N} \) such that

\[
f(a_1^n, u) = g(u)
\]

for all \( u \in \Sigma^* \).

**Proposition 1.9.** For any countable set \( \mathcal{F} \) of total functions containing the base functions and closed under composition and primitive recursion, if \( f \) is a universal function for the functions \( g : \Sigma^* \rightarrow \Sigma^* \) in \( \mathcal{F} \), then \( f \notin \mathcal{F} \).
1.9. THE PARTIAL COMPUTABLE FUNCTIONS

Proof. Assume that the universal function $f$ is in $\mathcal{F}$. Let $g$ be the function such that

$$g(u) = f(a_1^{|u|}, u)a_1$$

for all $u \in \Sigma^*$. We claim that $g \in \mathcal{F}$. It is enough to prove that the function $h$ such that

$$h(u) = a_1^{|u|}$$

is primitive recursive, which is easily shown.

Then, because $f$ is universal, there is some $m$ such that

$$g(u) = f(a_1^m, u)$$

for all $u \in \Sigma^*$. Letting $u = a_1^m$, we get

$$g(a_1^m) = f(a_1^m, a_1^m) = f(a_1^m, a_1^m)a_1,$$

a contradiction. \qed

Thus, either a universal function for $\mathcal{F}$ is partial, or it is not in $\mathcal{F}$.

In order to get a larger class of functions, we need the closure operation known as minimization.

1.9 The Partial Computable Functions

Minimization can be viewed as an abstract version of a while loop. First let us consider the simpler case of numerical functions.

Consider a function $g: \mathbb{N}^{m+1} \rightarrow \mathbb{N}$, with $m \geq 0$. We would like to know if for any fixed $n_1, \ldots, n_m \in \mathbb{N}$, the equation

$$g(n, n_1, \ldots, n_m) = 0 \quad \text{with respect to } n \in \mathbb{N}$$

has a solution $n \in \mathbb{N}$, and if so, we return the smallest such solution. Thus we are defining a (partial) function $f: \mathbb{N}^m \rightarrow \mathbb{N}$ such that

$$f(n_1, \ldots, n_m) = \min\{n \in \mathbb{N} \mid g(n, n_1, \ldots, n_m) = 0\},$$

with the understanding that $f(n_1, \ldots, n_m)$ is undefined otherwise. If $g$ is computed by a RAM program, computing $f(n_1, \ldots, n_m)$ corresponds to the while loop

$n := 0;$
while $g(n, n_1, \ldots, n_m) \neq 0$ do
$n := n + 1;$
endwhile
let $f(n_1, \ldots, n_m) = n.$
**Definition 1.22.** For any function $g: \mathbb{N}^{m+1} \to \mathbb{N}$, where $m \geq 0$, the function $f: \mathbb{N}^m \to \mathbb{N}$ is defined by minimization from $g$, if the following conditions hold for all $n_1, \ldots, n_m \in \mathbb{N}$:

1. $f(n_1, \ldots, n_m)$ is defined iff there is some $n \in \mathbb{N}$ such that $g(p, n_1, \ldots, n_m)$ is defined for all $p$, $0 \leq p \leq n$, and
   \[ g(n, n_1, \ldots, n_m) = 0; \]

2. When $f(n_1, \ldots, n_m)$ is defined,
   \[ f(n_1, \ldots, n_m) = n, \]
   where $n$ is such that $g(n, n_1, \ldots, n_m) = 0$ and $g(p, n_1, \ldots, n_m) \neq 0$ for every $p$, $0 \leq p \leq n - 1$. In other words, $n$ is the smallest natural number such that $g(n, n_1, \ldots, n_m) = 0$.

Following Kleene, we write
\[ f(n_1, \ldots, n_m) = \mu n [g(n, n_1, \ldots, n_m) = 0]. \]

**Remark:** When $f(n_1, \ldots, n_m)$ is defined, $f(n_1, \ldots, n_m) = n$, where $n$ is the smallest natural number such that condition (1) holds. It is very important to require that all the values $g(p, n_1, \ldots, n_m)$ be defined for all $p$, $0 \leq p \leq n$, when defining $f(n_1, \ldots, n_m)$. Failure to do so allows non-computable functions.

Minimization can be generalized to functions defined on strings as follows. Given a function $g: (\Sigma^*)^{m+1} \to \Sigma^*$, for any fixed $w_1, \ldots, w_m \in \Sigma^*$, we wish to solve the equation
\[ g(u, w_1, \ldots, w_m) = \epsilon \]
with respect to $u \in \Sigma^*$,

and return the “smallest” solution $u$, if any. The only issue is, what does smallest solution mean. We resolve this issue by restricting $u$ to be a string of $a_j$’s, for some fixed letter $a_j \in \Sigma$. Thus there are $k$ variants of minimization corresponding to searching for a shortest string in $\{a_j\}^*$, for a fixed $j$, $1 \leq j \leq k$.

Let $\Sigma = \{a_1, \ldots, a_k\}$. For any function
\[ g: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_{m+1} \to \Sigma^*, \]
where $m \geq 0$, for every $j$, $1 \leq j \leq k$, the function
\[ f: \underbrace{\Sigma^* \times \cdots \times \Sigma^*}_{m} \to \Sigma^* \]
looks for the shortest string $u$ over $\{a_j\}^*$ (for a fixed $j$) such that
\[ g(u, w_1, \ldots, w_m) = \epsilon : \]
This corresponds to the following while loop:

\begin{verbatim}
u := \epsilon;
while g(u, w_1, \ldots, w_m) \neq \epsilon do
u := u a_j;
endwhile
let f(w_1, \ldots, w_m) = u
\end{verbatim}

The operation of minimization (sometimes called minimalization) is defined as follows.

**Definition 1.23.** Let \( \Sigma = \{a_1, \ldots, a_k\} \). For any function

\[ g: \Sigma^* \times \cdots \times \Sigma^* \rightarrow \Sigma^*, \]

where \( m \geq 0 \), for every \( j, 1 \leq j \leq k \), the function

\[ f: \Sigma^* \times \cdots \times \Sigma^* \rightarrow \Sigma^*, \]

is defined by *minimization over \( \{a_j\}^* \) from \( g \), if the following conditions hold for all \( w_1, \ldots, w_m \in \Sigma^* \):

1. \( f(w_1, \ldots, w_m) \) is defined iff there is some \( n \geq 0 \) such that \( g(a_j^n, w_1, \ldots, w_m) \) is defined for all \( p, 0 \leq p \leq n \), and \( g(a_j^n, w_1, \ldots, w_m) = \epsilon \).

2. When \( f(w_1, \ldots, w_m) \) is defined,

\[ f(w_1, \ldots, w_m) = a_j^n, \]

where \( n \) is such that

\[ g(a_j^n, w_1, \ldots, w_m) = \epsilon \]

and

\[ g(a_j^p, w_1, \ldots, w_m) \neq \epsilon \]

for every \( p, 0 \leq p \leq n - 1 \).

We write

\[ f(w_1, \ldots, w_m) = \min_j u[g(u, w_1, \ldots, w_m) = \epsilon]. \]

**Note:** When \( f(w_1, \ldots, w_m) \) is defined,

\[ f(w_1, \ldots, w_m) = a_j^n, \]

where \( n \) is the smallest natural number such that condition (1) holds. It is very important to require that all the values \( g(a_j^p, w_1, \ldots, w_m) \) be defined for all \( p, 0 \leq p \leq n \), when defining \( f(w_1, \ldots, w_m) \). Failure to do so allows non-computable functions.
Remark: Inspired by Kleene’s notation in the case of numerical functions, we may use the
\(\mu\)-notation:

\[
f(w_1, \ldots, w_m) = \mu_j u [g(u, w_1, \ldots, w_m) = \epsilon].
\]

The class of partial computable functions is defined as follows.

**Definition 1.24.** Let \(\Sigma = \{a_1, \ldots, a_k\}\). The class of partial computable functions (in the sense of Herbrand–Gödel–Kleene), also called partial recursive functions is the smallest class of partial functions (over \(\Sigma^*\)) which contains the base functions and is closed under composition, primitive recursion, and minimization.

The class of computable functions also called recursive functions is the subset of the class of partial computable functions consisting of functions defined for every input (i.e., total functions).

One of the major results of computability theory is the following theorem.

**Theorem 1.10.** For an alphabet \(\Sigma = \{a_1, \ldots, a_k\}\), every partial computable function (partial recursive function) is RAM-computable, and thus Turing-computable. Conversely, every Turing-computable function is a partial computable function (partial recursive function). Similarly, the class of computable functions (recursive functions) is equal to the class of Turing-computable functions that halt in a proper ID for every input.

First we prove that every partial computable function is RAM-computable, which is not that difficult because composition, primitive recursion, and minimization are easily implemented using RAM programs. By Theorem 1.2, every RAM program can be converted to a Turing machine, so every partial computable function is Turing-computable.

For the converse, one can show that given a Turing machine, there is a primitive recursive function describing how to go from one ID to the next. Then minimization is used to guess whether a computation halts. The proof shows that every partial computable function needs minimization at most once. The characterization of the computable functions in terms of TM’s follows easily. Details are given in an appendix. See also Machtey and Young [25] and Kleene I.M. [21] (Chapter XIII).

We will prove in Section 2.3 that every RAM-computable function (over \(\mathbb{N}\)) is partial computable. This will be done by encoding RAM programs as natural numbers.

There are computable functions (recursive functions) that are not primitive recursive. Such an example is given by Ackermann’s function.

**Ackermann’s function** is the function \(A: \mathbb{N} \times \mathbb{N} \to \mathbb{N}\) which is defined by the following recursive clauses:

\[
\begin{align*}
A(0, y) & = y + 1, \\
A(x + 1, 0) & = A(x, 1), \\
A(x + 1, y + 1) & = A(x, A(x + 1, y)).
\end{align*}
\]
It turns out that $A$ is a computable function which is not primitive recursive. This is not easy to prove. It can be shown that:

\[
A(0, x) = x + 1, \\
A(1, x) = x + 2, \\
A(2, x) = 2x + 3, \\
A(3, x) = 2^{x+3} - 3,
\]

and

\[
A(4, x) = 2^{2^{\ldots^{2^{16}}}} - 3,
\]

with $A(4, 0) = 16 - 3 = 13$.

For example

\[
A(4, 1) = 2^{16} - 3, \quad A(4, 2) = 2^{216} - 3.
\]

Actually, it is not so obvious that $A$ is a total function, but it is.

**Proposition 1.11.** Ackermann’s function $A$ is a total function.

**Proof.** This is shown by induction, using the lexicographic ordering $\preceq$ on $\mathbb{N} \times \mathbb{N}$, which is defined as follows:

\[
(m, n) \preceq (m', n') \quad \text{iff either} \\
\quad m = m' \text{ and } n = n', \text{ or} \\
\quad m < m', \text{ or} \\
\quad m = m' \text{ and } n < n'.
\]

We write $(m, n) \prec (m', n')$ when $(m, n) \preceq (m', n')$ and $(m, n) \neq (m', n')$.

We prove that $A(m, n)$ is defined for all $(m, n) \in \mathbb{N} \times \mathbb{N}$ by complete induction over the lexicographic ordering on $\mathbb{N} \times \mathbb{N}$.

In the base case, $(m, n) = (0, 0)$, and since $A(0, n) = n + 1$, we have $A(0, 0) = 1$, and $A(0, 0)$ is defined.

For $(m, n) \neq (0, 0)$, the induction hypothesis is that $A(m', n')$ is defined for all $(m', n') \prec (m, n)$. We need to conclude that $A(m, n)$ is defined.

If $m = 0$, since $A(0, n) = n + 1$, $A(0, n)$ is defined.

If $m \neq 0$ and $n = 0$, since

\[
(m - 1, 1) \prec (m, 0),
\]

by the induction hypothesis, $A(m - 1, 1)$ is defined, but $A(m, 0) = A(m - 1, 1)$, and thus $A(m, 0)$ is defined.
If $m \neq 0$ and $n \neq 0$, since
\[(m, n - 1) \prec (m, n),\]
by the induction hypothesis, $A(m, n - 1)$ is defined. Since
\[(m - 1, A(m, n - 1)) \prec (m, n),\]
by the induction hypothesis, $A(m - 1, A(m, n - 1))$ is defined. But $A(m, n) = A(m - 1, A(m, n - 1))$, and thus $A(m, n)$ is defined.

Thus, $A(m, n)$ is defined for all $(m, n) \in \mathbb{N} \times \mathbb{N}$. \qed

It is possible to show that $A$ is a computable (recursive) function, although the quickest way to prove it requires some fancy machinery (the recursion theorem; see Section 5.1). Proving that $A$ is not primitive recursive is even harder.

A further study of the partial recursive functions requires the notions of pairing functions and of universal functions (or universal Turing machines).
Chapter 2

Universal RAM Programs and
Undecidability of the Halting Problem

The goal of this chapter is to prove three of the main results of computability theory:

(1) The undecidability of the halting problem for RAM programs (and Turing machines).

(2) The existence of universal RAM programs.

(3) The existence of the Kleene T-predicate.

All three require the ability to code a RAM program as a natural number. Gödel pioneered the technique of encoding objects such as proofs as natural numbers in his famous paper on the (first) incompleteness theorem (1931). One of the technical issues is to code (pack) a tuple of natural numbers as a single natural number, so that the numbers being packed can be retrieved. Gödel designed a fancy function whose definition does not involve recursion (Gödel’s β function; see Kleene [21] or Shoenfield [33]). For our purposes, a simpler function $J$ due to Cantor packing two natural numbers $m$ and $n$ as a single natural number $J(m, n)$ suffices.

Another technical issue is the fact it is possible to reduce most of computability theory to numerical functions $f : \mathbb{N}^m \to \mathbb{N}$, and even to functions $f : \mathbb{N} \to \mathbb{N}$. Indeed, there are primitive recursive coding and decoding functions $D_k : \Sigma^* \to \mathbb{N}$ and $C_k : \mathbb{N} \to \Sigma^*$ such that $C_k \circ D_k = \text{id}_{\Sigma^*}$, where $\Sigma = \{ a_1, \ldots, a_k \}$. It is simpler to code programs (or Turing machines) taking natural numbers as input.

Unfortunately, these coding techniques are very tedious so we advise the reader not to get bogged down with technical details upon first reading.

2.1 Pairing Functions

Pairing functions are used to encode pairs of integers into single integers, or more generally, finite sequences of integers into single integers. We begin by exhibiting a bijective pairing
function \( J: \mathbb{N}^2 \to \mathbb{N} \). The function \( J \) has the graph partially showed below:

\[
\begin{array}{cccccc}
y & 4 & 10 \\
3 & 6 & 11 & \\
2 & 3 & 7 & 12 & \\
1 & 1 & 4 & 8 & 13 & \\
0 & 0 & 2 & 5 & 9 & 14 & \\
0 & 1 & 2 & 3 & 4 & x
\end{array}
\]

The function \( J \) corresponds to a certain way of enumerating pairs of integers \((x, y)\). Note that the value of \( x + y \) is constant along each descending diagonal, and consequently, we have

\[
J(x, y) = 1 + 2 + \cdots + (x + y) + x,
\]

\[
= ((x + y)(x + y + 1) + 2x)/2,
\]

\[
= ((x + y)^2 + 3x + y)/2,
\]

that is,

\[
J(x, y) = ((x + y)^2 + 3x + y)/2.
\]

For example, \( J(0, 3) = 6, \ J(1, 2) = 7, \ J(2, 2) = 12, \ J(3, 1) = 13, \ J(4, 0) = 14 \).

Let \( K: \mathbb{N} \to \mathbb{N} \) and \( L: \mathbb{N} \to \mathbb{N} \) be the projection functions onto the axes, that is, the unique functions such that

\[
K(J(a, b)) = a \quad \text{and} \quad L(J(a, b)) = b,
\]

for all \( a, b \in \mathbb{N} \). For example, \( K(11) = 1 \), and \( L(11) = 3 \); \( K(12) = 2 \), and \( L(12) = 2 \); \( K(13) = 3 \) and \( L(13) = 1 \).

**Definition 2.1.** The *pairing function* \( J: \mathbb{N}^2 \to \mathbb{N} \) is defined by

\[
J(x, y) = ((x + y)^2 + 3x + y)/2 \quad \text{for all } x, y \in \mathbb{N}.
\]

The functions \( K: \mathbb{N} \to \mathbb{N} \) and \( L: \mathbb{N} \to \mathbb{N} \) are the *projection functions* onto the axes, that is, the unique functions such that

\[
K(J(a, b)) = a \quad \text{and} \quad L(J(a, b)) = b,
\]

for all \( a, b \in \mathbb{N} \).
2.1. PAIRING FUNCTIONS

The functions $J, K, L$ are called Cantor’s pairing functions. They were used by Cantor to prove that the set $\mathbb{Q}$ of rational numbers is countable.

Clearly, $J$ is primitive recursive, since it is given by a polynomial. It is not hard to prove that $J$ is injective and surjective, and that it is strictly monotonic in each argument, which means that for all $x, x', y, y' \in \mathbb{N}$, if $x < x'$ then $J(x, y) < J(x', y)$, and if $y < y'$ then $J(x, y) < J(x, y')$.

The projection functions can be computed explicitly, although this is a bit tricky. We only need to observe that by monotonicity of $J$,

$$x \leq J(x, y) \quad \text{and} \quad y \leq J(x, y),$$

and thus,

$$K(z) = \min(x \leq z)(\exists y \leq z)[J(x, y) = z],$$

and

$$L(z) = \min(y \leq z)(\exists x \leq z)[J(x, y) = z].$$

Therefore, by the results of Section 1.8, $K$ and $L$ are primitive recursive. It can be verified that $J(K(z), L(z)) = z$, for all $z \in \mathbb{N}$.

More explicit formulae can be given for $K$ and $L$. If we define

$$Q_1(z) = \lfloor(\lfloor \sqrt{8z + 1} \rfloor + 1)/2 \rfloor - 1,$$

$$Q_2(z) = 2z - (Q_1(z))^2,$$

then it can be shown that

$$K(z) = \frac{1}{2}(Q_2(z) - Q_1(z)),$$

$$L(z) = Q_1(z) - \frac{1}{2}(Q_2(z) - Q_1(z)).$$

In the above formula, the function $m \mapsto \lfloor \sqrt{m} \rfloor$ yields the largest integer $s$ such that $s^2 \leq m$. It can be computed by a RAM program.

The pairing function $J(x, y)$ is also denoted as $\langle x, y \rangle$, and $K$ and $L$ are also denoted as $\Pi_1$ and $\Pi_2$. The notation $\langle x, y \rangle$ is “intentionally ambiguous,” in the sense that it can be interpreted as the actual ordered pair consisting of the two numbers $x$ and $y$, or as the number $\langle x, y \rangle = J(x, y)$ that encodes the pair consisting of the two numbers $x$ and $y$. The context should make it clear which interpretation is intended. In this chapter and the next, it is the number (code) interpretation.

We can define bijections between $\mathbb{N}^n$ and $\mathbb{N}$ by induction for all $n \geq 1$.

**Definition 2.2.** The function $\langle - , \ldots , - \rangle_n : \mathbb{N}^n \to \mathbb{N}$ called an extended pairing function is defined as follows. We let

$$\langle z \rangle_1 = z,$$

$$\langle x_1, x_2 \rangle_2 = \langle x_1, x_2 \rangle,$$
and
\[
\langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1} = \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n,
\]
for all \( z, x_2, \ldots, x_{n+1} \in \mathbb{N} \).

Again we stress that \( \langle x_1, \ldots, x_n \rangle \) is a \textit{natural number}. For example,
\[
\begin{align*}
\langle x_1, x_2, x_3 \rangle_3 &= \langle x_1, \langle x_2, x_3 \rangle \rangle_2 \\
&= \langle x_1, \langle x_2, x_3 \rangle \rangle \\
\langle x_1, x_2, x_3, x_4 \rangle_4 &= \langle x_1, x_2, \langle x_3, x_4 \rangle \rangle_3 \\
&= \langle x_1, \langle x_2, \langle x_3, x_4 \rangle \rangle \rangle \\
\langle x_1, x_2, x_3, x_4, x_5 \rangle_5 &= \langle x_1, x_2, x_3, \langle x_4, x_5 \rangle \rangle_4 \\
&= \langle x_1, \langle x_2, \langle x_3, \langle x_4, x_5 \rangle \rangle \rangle \rangle_3.
\end{align*}
\]

It can be shown by induction on \( n \) that
\[
\langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1} = \langle x_1, \langle x_2, \ldots, x_{n+1} \rangle \rangle_n.
\] (*)

Observe that if \( z = \langle x_1, \ldots, x_n \rangle_n \), then \( x_1 = \Pi_1(z) \), \( x_2 = \Pi_1(\Pi_2(z)) \), \( x_3 = \Pi_1(\Pi_2(\Pi_2(z))) \), \( x_4 = \Pi_1(\Pi_2(\Pi_2(\Pi_2(z)))) \), \( x_5 = \Pi_2(\Pi_2(\Pi_2(\Pi_2(z)))) \).

We can also define a uniform projection function \( \Pi: \mathbb{N}^3 \to \mathbb{N} \) with the following property: if \( z = \langle x_1, \ldots, x_n \rangle \), with \( n \geq 2 \), then
\[
\Pi(i, n, z) = x_i \quad \text{for all } i, \text{ where } 1 \leq i \leq n.
\]

The idea is to view \( z \) as an \( n \)-tuple, and \( \Pi(i, n, z) \) as the \( i \)-th component of that \( n \)-tuple, but if \( z \), \( n \) and \( i \) do not fit this interpretation, the function must be still be defined and we give it a “crazy” value by default using some simple primitive recursive clauses.

**Definition 2.3.** The uniform projection function \( \Pi: \mathbb{N}^3 \to \mathbb{N} \) is defined by cases as follows:

\[
\begin{align*}
\Pi(i, 0, z) &= 0, \quad \text{for all } i \geq 0, \\
\Pi(i, 1, z) &= z, \quad \text{for all } i \geq 0, \\
\Pi(i, 2, z) &= \Pi_1(z), \quad \text{if } 0 \leq i \leq 1, \\
\Pi(i, 2, z) &= \Pi_2(z), \quad \text{for all } i \geq 2,
\end{align*}
\]

and for all \( n \geq 2 \),
\[
\Pi(i, n+1, z) = \begin{cases} 
\Pi(i, n, z) & \text{if } 0 \leq i < n, \\
\Pi_1(\Pi(n, n, z)) & \text{if } i = n, \\
\Pi_2(\Pi(n, n, z)) & \text{if } i > n.
\end{cases}
\]
2.1. PAIRING FUNCTIONS

By the results of Section 1.8, this is a legitimate primitive recursive definition. If \( z \) is the code \( \langle x_1, \ldots, x_{n+1} \rangle_{n+1} \) for the \((n+1)\)-tuple \((x_1, \ldots, x_{n+1})\) with \( n \geq 2 \), then for \( 0 \leq i < n \), the clause of Definition 2.3 that applies is

\[
\Pi(i, n + 1, z) = \Pi(i, n, z),
\]

and since

\[
\langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1} = \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n,
\]

we have

\[
\Pi(i, n + 1, \langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1}) = \Pi(i, n + 1, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n)
= \Pi(i, n, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n),
\]

and since \( \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n \) codes an \( n \)-tuple, for \( i = 1, \ldots, n-1 \), the value returned is indeed \( x_i \). If \( i = n \), then the clause that applies is

\[
\Pi(n, n + 1, z) = \Pi_1(\Pi(n, n, z)),
\]

so we have

\[
\Pi(n, n + 1, \langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1}) = \Pi(n, n + 1, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n)
= \Pi_1(\Pi(n, n, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n))
= \Pi_1(\langle x_n, x_{n+1} \rangle)
= x_n.
\]

Finally, if \( i = n + 1 \), then the clause that applies is

\[
\Pi(n + 1, n + 1, z) = \Pi_2(\Pi(n, n, z)),
\]

so we have

\[
\Pi(n + 1, n + 1, \langle x_1, \ldots, x_n, x_{n+1} \rangle_{n+1}) = \Pi(n + 1, n + 1, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n)
= \Pi_2(\Pi(n, n, \langle x_1, \ldots, x_{n-1}, \langle x_n, x_{n+1} \rangle \rangle_n))
= \Pi_2(\langle x_n, x_{n+1} \rangle)
= x_{n+1}.
\]

When \( i = 0 \) or \( i > n + 1 \), we get “bogus” values.

**Remark:** One might argue that it would have been preferable to order the arguments of \( \Pi \) as \((n, i, z)\) rather than \((i, n, z)\). We use the order \((i, n, z)\) in conformity with Machtey and Young [25].

Some basic properties of \( \Pi \) are given as exercises. In particular, the following properties are easily shown:
(a) \([0, \ldots, 0]_n = 0, \langle x, 0 \rangle = \langle x, 0, \ldots, 0 \rangle_n;\)
(b) \(\Pi(0, n, z) = \Pi(1, n, z)\) and \(\Pi(i, n, z) = \Pi(n, n, z)\), for all \(i \geq n\) and all \(n, z \in \mathbb{N}\);
(c) \(\Pi(1, n, z), \ldots, \Pi(n(n, z))_n = z\), for all \(n \geq 1\) and all \(z \in \mathbb{N}\);
(d) \(\Pi(i, n, z) \leq z\), for all \(i, n, z \in \mathbb{N}\);
(e) There is a primitive recursive function \(\text{Large}\), such that,
\[
\Pi(i, n + 1, \text{Large}(n + 1, z)) = z,
\]
for \(i, n, z \in \mathbb{N}\).

As a first application, we observe that we need only consider partial computable functions (partial recursive functions)\(^1\) of a single argument. Indeed, let \(\phi : \mathbb{N}^n \to \mathbb{N}\) be a partial computable function of \(n \geq 2\) arguments. Let \(\bar{\phi} : \mathbb{N} \to \mathbb{N}\) be the function given by
\[
\bar{\phi}(z) = \phi(\Pi(1, n, z), \ldots, \Pi(n(n, z))),
\]
for all \(z \in \mathbb{N}\). Then \(\bar{\phi}\) is a partial computable function of a single argument, and \(\phi\) can be recovered from \(\bar{\phi}\), since
\[
\phi(x_1, \ldots, x_n) = \bar{\phi}(\langle x_1, \ldots, x_n \rangle).
\]
Thus, using \(\langle -, \cdots, - \rangle\) and \(\Pi\) as coding and decoding functions, we can restrict our attention to functions of a single argument.

Pairing functions can also be used to prove that certain functions are primitive recursive, even though their definition is not a legal primitive recursive definition. For example, consider the Fibonacci function defined as follows:
\[
f(0) = 1,
\]
\[
f(1) = 1,
\]
\[
f(n + 2) = f(n + 1) + f(n),
\]
for all \(n \in \mathbb{N}\). This is not a legal primitive recursive definition, since \(f(n + 2)\) depends both on \(f(n + 1)\) and \(f(n)\). In a primitive recursive definition, \(g(y + 1, \bar{x})\) is only allowed to depend upon \(g(y, \bar{x})\), where \(\bar{x}\) is an abbreviation for \((x_2, \ldots, x_m)\).

**Definition 2.4.** Given any function \(f : \mathbb{N}^n \to \mathbb{N}\), the function \(\bar{f} : \mathbb{N}^{n+1} \to \mathbb{N}\) defined such that
\[
\bar{f}(y, \bar{x}) = \langle f(0, \bar{x}), \ldots, f(y, \bar{x}) \rangle_{y+1}
\]
is called the course-of-value function for \(f\).

The following lemma holds.

\(^1\)The term *partial recursive* is now considered old-fashion. Many researchers have switched to the term *partial computable.*
2.1. PAIRING FUNCTIONS

**Proposition 2.1.** Given any function \( f : \mathbb{N}^n \to \mathbb{N} \), if \( f \) is primitive recursive, then so is \( \overline{f} \).

**Proof.** First it is necessary to define a function \( \text{con} \) such that if \( x = \langle x_1, \ldots, x_m \rangle \) and \( y = \langle y_1, \ldots, y_n \rangle \), where \( m, n \geq 1 \), then
\[
\text{con}(m, x, y) = \langle x_1, \ldots, x_m, y_1, \ldots, y_n \rangle.
\]
This fact is left as an exercise. Now, if \( f \) is primitive recursive, let
\[
\overline{f}(0, x) = f(0, x),
\]
\[
\overline{f}(y + 1, x) = \text{con}(y + 1, \overline{f}(y, x), f(y + 1, x)),
\]
showing that \( \overline{f} \) is primitive recursive. Conversely, if \( \overline{f} \) is primitive recursive, then
\[
f(y, x) = \Pi(y + 1, y + 1, \overline{f}(y, x)),
\]
and so, \( f \) is primitive recursive. \( \square \)

*Remark:* Why is it that \( \overline{f}(y + 1, x) = \langle \overline{f}(y, x), f(y + 1, x) \rangle \) does not work? Check the definition of \( \langle x_1, \ldots, x_n \rangle \).

We define *course-of-value recursion* as follows.

**Definition 2.5.** Given any two functions \( g : \mathbb{N}^n \to \mathbb{N} \) and \( h : \mathbb{N}^{n+2} \to \mathbb{N} \), the function \( f : \mathbb{N}^{n+1} \to \mathbb{N} \) is defined by *course-of-value recursion* from \( g \) and \( h \) if
\[
f(0, x) = g(x),
\]
\[
f(y + 1, x) = h(y, \overline{f}(y, x), x).
\]

The following lemma holds.

**Proposition 2.2.** If \( f : \mathbb{N}^{n+1} \to \mathbb{N} \) is defined by course-of-value recursion from \( g \) and \( h \) and \( g, h \) are primitive recursive, then \( f \) is primitive recursive.

**Proof.** We prove that \( \overline{f} \) is primitive recursive. Then by Proposition 2.1, \( f \) is also primitive recursive. To prove that \( \overline{f} \) is primitive recursive, observe that
\[
\overline{f}(0, x) = g(x),
\]
\[
\overline{f}(y + 1, x) = \text{con}(y + 1, \overline{f}(y, x), h(y, \overline{f}(y, x), x)).
\]

When we use Proposition 2.2 to prove that a function is primitive recursive, we rarely bother to construct a formal course-of-value recursion. Instead, we simply indicate how the value of \( f(y + 1, x) \) can be obtained in a primitive recursive manner from \( f(0, x) \) through \( f(y, x) \). Thus, an informal use of Proposition 2.2 shows that the Fibonacci function is primitive recursive. A rigorous proof of this fact is left as an exercise.

Next we show that there exist coding and decoding functions between \( \Sigma^* \) and \( \{a_1\}^* \), and that partial computable functions over \( \Sigma^* \) can be recoded as partial computable functions over \( \{a_1\}^* \). Since \( \{a_1\}^* \) is isomorphic to \( \mathbb{N} \), this shows that we can restrict our attention to functions defined over \( \mathbb{N} \).
2.2 Equivalence of Alphabets

Given an alphabet $\Sigma = \{a_1, \ldots, a_k\}$, strings over $\Sigma$ can be ordered by viewing strings as numbers in a number system where the digits are $a_1, \ldots, a_k$. In this number system, which is almost the number system with base $k$, the string $a_1$ corresponds to zero, and $a_k$ to $k - 1$. Hence, we have a kind of shifted number system in base $k$. The total order on $\Sigma^*$ induced by this number system is defined so that $u$ precedes $v$ if $|u| < |v|$, and if $|u| = |v|$, then $u$ comes before $v$ in the lexicographic ordering. For example, if $\Sigma = \{a, b, c\}$, a listing of $\Sigma^*$ in the ordering corresponding to the number system begins with

$$a, b, c, aa, ab, ac, ba, bb, bc, ca, cb, cc,$$

$$aaa, aab,aac,aba,abb,abc,\ldots$$

This ordering induces a function from $\Sigma^*$ to $\mathbb{N}$ which is a bijection. Indeed, if $u = a_{i_1} \cdots a_{i_n}$, this function $f : \Sigma^* \to \mathbb{N}$ is given by

$$f(u) = i_1 k^{n-1} + i_2 k^{n-2} + \cdots + i_{n-1} k + i_n.$$

Since we also want a decoding function, we define the coding function $C_k : \Sigma^* \to \Sigma^*$ as follows:

$$C_k(\epsilon) = \epsilon,$$
$$C_k(xa_i) = C_k(x)a_i \quad \text{if} \; i \neq k,$$
$$C_k(xa_k) = C_k(x).$$

The function $C_k$ is primitive recursive, because

$$C_k(\epsilon) = \epsilon,$$
$$C_k(xa_i) = C_k(x)a_i.$$

The inverse of $C_k$ is a function $D_k : \{a_1\}^* \to \Sigma^*$. However, primitive recursive functions are total, and we need to extend $D_k$ to $\Sigma^*$. This is easily done by letting

$$D_k(x) = D_k(a_1^{|x|})$$

for all $x \in \Sigma^*$. It remains to define $D_k$ by primitive recursion over $\Sigma^* = \{a_1, \ldots, a_k\}^*$. For this, we introduce three auxiliary functions $p, q, r$, defined as follows. Let

$$p(\epsilon) = \epsilon,$$
$$p(xa_i) = xa_i \quad \text{if} \; i \neq k,$$
$$p(xa_k) = p(x).$$

Note that $p(x)$ is the result of deleting consecutive $a_k$'s in the tail of $x$. Let

$$q(\epsilon) = \epsilon,$$
$$q(xa_i) = q(x)a_1.$$
2.2. EQUIVALENCE OF ALPHABETS

Note that \( q(x) = a^{[x]}_1 \). Finally, let

\[
\begin{align*}
    r(\epsilon) &= a_1, \\
    r(xa_i) &= xa_{i+1}, & \text{if } i \neq k, \\
    r(xa_k) &= xa_k.
\end{align*}
\]

The function \( r \) is almost the successor function for the ordering. Then the trick is that \( D_k(xa_i) \) is the successor of \( D_k(x) \) in the ordering so usually \( D_k(xa_i) = r(D_k(x)) \), except if

\[
D_k(xa_i) = ya_ja^n_k
\]

with \( j \neq k \), since the successor of \( ya_ja^n_k \) is \( ya_{j+1}a^n_k \). Thus, we have

\[
\begin{align*}
    D_k(\epsilon) &= \epsilon, \\
    D_k(xa_i) &= r(p(D_k(x)))q(D_k(x) - p(D_k(x))), & a_i \in \Sigma.
\end{align*}
\]

Then both \( C_k \) and \( D_k \) are primitive recursive, and \( D_k \circ C_k = \text{id} \). Here

\[
u - v = \begin{cases} 
\epsilon & \text{if } |u| \leq |v| \\
w & \text{if } u = xw \text{ and } |x| = |v|.
\end{cases}
\]

In other words, \( u - v \) is \( u \) with its first \(|v| \) letters deleted. We can show that this function can be defined by primitive recursion by first defining \( \text{rdiff}(u, v) \) as \( v \) with its first \(|u| \) letters deleted, and then

\[
u - v = \text{rdiff}(v, u).
\]

To define \( \text{rdiff} \), we use \( \text{tail} \) given by

\[
\begin{align*}
    \text{tail}(\epsilon) &= \epsilon \\
    \text{tail}(a_iu) &= u, & a_i \in \Sigma, u \in \Sigma^*.
\end{align*}
\]

We proved in Section 1.7 that \( \text{tail} \) is primitive recursive. Then

\[
\begin{align*}
    \text{rdiff}(\epsilon, v) &= v \\
    \text{rdiff}(ua_i, v) &= \text{rdiff}(u, \text{tail}(v)), & a_i \in \Sigma.
\end{align*}
\]

We leave as an exercise to put all these definitions into the proper format of primitive recursion using projections.

Let \( \varphi: (\Sigma^*)^n \to \Sigma^* \) be a partial function over \( \Sigma^* \), and let \( \varphi^+: (\{a_1\}^*)^n \to \{a_1\}^* \) be the function given by

\[
\varphi^+(x_1, \ldots, x_n) = C_k(\varphi(D_k(x_1), \ldots, D_k(x_n)))).
\]

Also, for any partial function \( \psi: (\{a_1\}^*)^n \to \{a_1\}^* \), let \( \psi^+: (\Sigma^*)^n \to \Sigma^* \) be the function given by

\[
\psi^+(x_1, \ldots, x_n) = D_k(\psi(C_k(x_1), \ldots, C_k(x_n)))).
\]
We claim that if $\psi$ is a partial computable function over $(\{a_1\}^*)^n$, then $\psi^\sharp$ is partial computable over $(\Sigma^*)^n$, and that if $\varphi$ is a partial computable function over $(\Sigma^*)^n$, then $\varphi^+$ is partial computable over $(\{a_1\}^*)^n$.

The function $\psi$ can be extended to $(\Sigma^*)^n$ by letting

$$
\psi(x_1, \ldots, x_n) = \psi(a_1^{\lfloor x_1 \rfloor}, \ldots, a_1^{\lfloor x_n \rfloor})
$$

for all $x_1, \ldots, x_n \in \Sigma^*$, and so, if $\psi$ is partial computable, then so is the extended function, by composition. It follows that if $\psi$ is partial (or primitive) recursive, then so is $\psi^\sharp$.

This seems equally obvious for $\varphi$ and $\varphi^+$, but there is a difficulty. The problem is that $\varphi^+$ is defined as a composition of functions over $\Sigma^*$. We have to show how $\varphi^+$ can be defined directly over $\{a_1\}^*$ without using any additional alphabet symbols. This is done in Machtey and Young [25], see Section 2.2, Lemma 2.2.3.

### 2.3 Coding of RAM Programs; The Halting Problem

In this section we present a specific encoding of RAM programs which allows us to treat programs as integers. This encoding will allow us to prove one of the most important results of computability theory first proven by Turing for Turing machines (1936-1937), the undecidability of the halting problem for RAM programs (and Turing machines).

Encoding programs as integers also allows us to have programs that take other programs as input, and we obtain a universal program. Universal programs have the property that given two inputs, the first one being the code of a program and the second one an input data, the universal program simulates the actions of the encoded program on the input data. A coding scheme is also called an indexing or a Gödel numbering, in honor to Gödel, who invented this technique.

From results of the previous chapter, without loss of generality, we can restrict out attention to RAM programs computing partial functions of one argument over $\mathbb{N}$. Furthermore, we only need the following kinds of instructions, each instruction being coded as shown below. Since we are considering functions over the natural numbers, which corresponds to a one-letter alphabet, there is only one kind of instruction of the form $\text{add}$ and $\text{jmp}$ ($\text{add}$ increments by 1 the contents of the specified register $R_j$).

Recall that a conditional jump causes a jump to the closest address $Nk$ above or below iff $R_j$ is nonzero, and if $R_j$ is null, the next instruction is executed. We assume that all lines in a RAM program are numbered. This is always feasible, by labeling unnamed instructions with a new and unused line number.
Definition 2.6. Instructions of a RAM program (operating on \( \mathbb{N} \)) are coded as follows:

- \( Ni \ add \ Rj \) \( code = \langle 1, i, j, 0 \rangle \)
- \( Ni \ tail \ Rj \) \( code = \langle 2, i, j, 0 \rangle \)
- \( Ni \ continue \) \( code = \langle 3, i, 1, 0 \rangle \)
- \( Ni \ Rj \ jmp \ Nka \) \( code = \langle 4, i, j, k \rangle \)
- \( Ni \ Rj \ jmp \ Nkb \) \( code = \langle 5, i, j, k \rangle \)

The code of an instruction \( I \) is denoted as \( #I \).

To simplify the notation, we introduce the following decoding primitive recursive functions \( Typ, LNum, Reg, \) and \( Jmp \), defined as follows:

\[
\begin{align*}
Typ(x) &= \Pi(1, 4, x), \\
LNum(x) &= \Pi(2, 4, x), \\
Reg(x) &= \Pi(3, 4, x), \\
Jmp(x) &= \Pi(4, 4, x).
\end{align*}
\]

The functions yield the type, line number, register name, and line number jumped to, if any, for an instruction coded by \( x \). Note that we have no need to interpret the values of these functions if \( x \) does not code an instruction.

We can define the primitive recursive predicate \( INST \), such that \( INST(x) \) holds iff \( x \) codes an instruction. First, we need the connective \( \Rightarrow \) (implies), defined such that

\[ P \Rightarrow Q \text{ iff } \neg P \lor Q. \]

Definition 2.7. The predicate \( INST(x) \) is defined primitive recursively as follows:

\[
\begin{align*}
[1 \leq Typ(x) \leq 5] & \land [1 \leq Reg(x)] \land \\
[Typ(x) \leq 3 \Rightarrow Jmp(x) = 0] & \land \\
[Typ(x) = 3 \Rightarrow Reg(x) = 1].
\end{align*}
\]

The predicate \( INST(x) \) says that if \( x \) is the code of an instruction, say \( x = \langle c, i, j, k \rangle \), then \( 1 \leq c \leq 5 \), \( j \geq 1 \), if \( c \leq 3 \), then \( k = 0 \), and if \( c = 0 \) then we also have \( j = 1 \).

Definition 2.8. Program are coded as follows. If \( P \) is a RAM program composed of the \( n \) instructions \( I_1, \ldots, I_n \), the code of \( P \), denoted as \( #P \), is

\[ #P = \langle n, #I_1, \ldots, #I_n \rangle. \]

Recall from Property \((*)\) in Section 2.1 that

\[ \langle n, #I_1, \ldots, #I_n \rangle = \langle n, \langle #I_1, \ldots, #I_n \rangle \rangle. \]

Also recall that

\[ \langle x, y \rangle = \frac{(x + y)^2 + 3x + y}{2}. \]
Example 2.1. Consider the following program \texttt{Padd2} computing the function \texttt{add2}: \( \mathbb{N} \to \mathbb{N} \) given by \( \text{add2}(n) = n + 2 \).

\texttt{Padd2}:

\begin{align*}
I_1 : & \quad 1 \quad \text{add} \quad R1 \\
I_2 : & \quad 2 \quad \text{add} \quad R1 \\
I_3 : & \quad 3 \quad \text{continue}
\end{align*}

We have

\begin{align*}
\#I_1 &= \langle 1, 1, 1, 0 \rangle_4 = \langle 1, \langle 1, 1, 0 \rangle \rangle = 37 \\
\#I_2 &= \langle 1, 2, 1, 0 \rangle_4 = \langle 1, \langle 2, 1, 0 \rangle \rangle = 92 \\
\#I_3 &= \langle 3, 3, 1, 0 \rangle_4 = \langle 3, \langle 3, 1, 0 \rangle \rangle = 234
\end{align*}

and

\[
\#Padd2 = \langle 3, \#I_1, \#I_2, \#I_3 \rangle_4 = \langle 3, \langle 37, 92, 234 \rangle \rangle = 1018748519973070618.
\]

The codes get big fast!

We define the primitive recursive functions \( \text{Ln} \), \( \text{Pg} \), and \( \text{Line} \), such that:

\[
\begin{align*}
\text{Ln}(x) &= \Pi(1, 2, x), \\
\text{Pg}(x) &= \Pi(2, 2, x), \\
\text{Line}(i, x) &= \Pi(i, \text{Ln}(x), \text{Pg}(x)).
\end{align*}
\]

The function \( \text{Ln} \) yields the length of the program (the number of instructions), \( \text{Pg} \) yields the sequence of instructions in the program (really, a code for the sequence), and \( \text{Line}(i, x) \) yields the code of the \( i \)th instruction in the program. Again, if \( x \) does not code a program, there is no need to interpret these functions. However, note that by a previous exercise, it happens that

\[
\text{Line}(0, x) = \text{Line}(1, x), \quad \text{and} \quad \text{Line}(\text{Ln}(x), x) = \text{Line}(i, x), \quad \text{for all} \ i \geq \text{Ln}(x).
\]

The primitive recursive predicate \( \text{PROG} \) is defined such that \( \text{PROG}(x) \) holds iff \( x \) codes a program. Thus, \( \text{PROG}(x) \) holds if each line codes an instruction, each jump has an instruction to jump to, and the last instruction is a \text{continue}.

**Definition 2.9.** The primitive recursive predicate \( \text{PROG}(x) \) is given by

\[
\forall i \leq \text{Ln}(x) [ i \geq 1 \Rightarrow \\
(\text{INST}(	ext{Line}(i, x)) \land \text{Typ}(	ext{Line}(\text{Ln}(x), x)) = 3 \\
\land \text{Typ}(	ext{Line}(i, x)) = 4 \Rightarrow \\
\exists j \leq i - 1 [ j \geq 1 \land \text{LNum}(	ext{Line}(j, x)) = \text{Jmp}(	ext{Line}(i, x)) ] ] \\
\land \\
(\text{Typ}(	ext{Line}(i, x)) = 5 \Rightarrow \\
\exists j \leq \text{Ln}(x) [ j > i \land \text{LNum}(	ext{Line}(j, x)) = \text{Jmp}(	ext{Line}(i, x)) ] ]].
\]

Note that we have used Proposition 1.7 which states that if \( f \) is a primitive recursive function and if \( P \) is a primitive recursive predicate, then \( \exists x \leq f(y) P(x) \) is primitive recursive.

The last instruction \( \text{Line}(\text{Ln}(x), x) \) in the program must be a **continue**, which means that \( \text{Typ}(	ext{Line}(\text{Ln}(x), x)) = 3 \). When the \( i \)th instruction coded by \( \text{Line}(i, x) \) of the program coded by \( x \) has its first field \( \text{Typ}(	ext{Line}(i, x)) = 4 \), this instruction is a jump above, and there must be an instruction in line \( j \) above instruction in line \( i \), which means that \( 1 \leq j \leq i - 1 \), and the line number \( \text{LNum}(	ext{Line}(j, x)) \) of the \( j \)th instruction must be equal to the jump address \( \text{Jmp}(	ext{Line}(i, x)) \) of the \( i \)th instruction. When \( \text{Typ}(	ext{Line}(i, x)) = 5 \), this instruction is a jump below, and the analysis is similar.

We are now ready to prove a fundamental result in the theory of algorithms. This result points out some of the limitations of the notion of algorithm.

**Theorem 2.3.** *(Undecidability of the halting problem)* There is no RAM program \( \text{Decider} \) which halts for all inputs and has the following property when started with input \( x \) in register \( R_1 \) and with input \( i \) in register \( R_2 \) (the other registers being set to zero):

1. **Decider** halts with output 1 iff \( i \) codes a program that eventually halts when started on input \( x \) (all other registers set to zero).

2. **Decider** halts with output 0 in \( R_1 \) iff \( i \) codes a program that runs forever when started on input \( x \) in \( R_1 \) (all other registers set to zero).

3. If \( i \) does not code a program, then **Decider** halts with output 2 in \( R_1 \).

**Proof.** Assume that **Decider** is such a RAM program, and let \( Q \) be the following program with a single input:

\[
\begin{align*}
\text{Program } Q \text{ (code } q) \quad & \left\{ \begin{array}{c}
R_2 \leftarrow R_1 \\
P \\
\text{continue} \\
N_1 \text{ jmp } N_1a \\
R_1 \text{ continue}
\end{array} \right. \\
\end{align*}
\]

Let \( i \) be the code of some program \( P \). The key point is that *the termination behavior of \( Q \) on input \( i \) is exactly the opposite of the termination behavior of **Decider** on input \( i \) and code \( i \).*
(1) If Decider says that program $P$ coded by $i$ halts on input $i$, then $R1$ just after the continue in line $N1$ contains 1, and $Q$ loops forever.

(2) If Decider says that program $P$ coded by $i$ loops forever on input $i$, then $R1$ just after continue in line $N1$ contains 0, and $Q$ halts.

The program $Q$ can be translated into a program using only instructions of type 1, 2, 3, 4, 5, described previously, and let $q$ be the code of the program $Q$.

Let us see what happens if we run the program $Q$ on input $q$ in $R1$ (all other registers set to zero).

Just after execution of the assignment $R2 ← R1$, the program Decider is started with $q$ in both $R1$ and $R2$. Since Decider is supposed to halt for all inputs, it eventually halts with output 0 or 1 in $R1$. If Decider halts with output 1 in $R1$, then $Q$ goes into an infinite loop, while if Decider halts with output 0 in $R1$, then $Q$ halts. But then, because of the definition of Decider, we see that Decider says that $Q$ halts when started on input $q$ iff $Q$ loops forever on input $q$, and that $Q$ loops forever on input $q$ iff $Q$ halts on input $q$, a contradiction. Therefore, Decider cannot exist. \square

The argument used in the proof of 2.3 is quite similar in spirit to “Russell’s Paradox.” If we identify the notion of algorithm with that of a RAM program which halts for all inputs, the above theorem says that there is no algorithm for deciding whether a RAM program eventually halts for a given input. We say that the halting problem for RAM programs is undecidable (or unsolvable).

The above theorem also implies that the halting problem for Turing machines is undecidable. Indeed, if we had an algorithm for solving the halting problem for Turing machines, we could solve the halting problem for RAM programs as follows: first, apply the algorithm for translating a RAM program into an equivalent Turing machine, and then apply the algorithm solving the halting problem for Turing machines.

The argument is typical in computability theory and is called a “reducibility argument.”

Our next goal is to define a primitive recursive function that describes the computation of RAM programs.

### 2.4 Universal RAM Programs

To describe the computation of a RAM program, we need to code not only RAM programs but also the contents of the registers. Assume that we have a RAM program $P$ using $n$ registers $R1, \ldots, Rn$, whose contents are denoted as $r_1, \ldots, r_n$. We can code $r_1, \ldots, r_n$ into a single integer $⟨r_1, \ldots, r_n⟩$. Conversely, every integer $x$ can be viewed as coding the contents of $R1, \ldots, Rn$, by taking the sequence $Π(1, n, x), \ldots, Π(n, n, x)$.
Actually, it is not necessary to know \( n \), the number of registers, if we make the following observation:

\[
\text{Reg(Line}(i, x)) \leq \text{Line}(i, x) \leq \text{Pg}(x) < x
\]

for all \( i, x \in \mathbb{N} \). If \( x \) codes a program, then \( R_1, \ldots, R_x \) certainly include all the registers in the program. Also note that from a previous exercise,

\[
\langle r_1, \ldots, r_n, 0, \ldots, 0 \rangle = \langle r_1, \ldots, r_n, 0 \rangle.
\]

We now define the primitive recursive functions Nextline, Nextcont, and Comp, describing the computation of RAM programs. There are a lot of tedious technical details that the reader should skip upon first reading. However, to be rigorous, we must spell out all these details.

**Definition 2.10.** Let \( x \) code a program and let \( i \) be such that \( 1 \leq i \leq \text{Ln}(x) \). The following functions are defined:

1. \( \text{Nextline}(i, x, y) \) is the number of the next instruction to be executed after executing the \( i \)th instruction (the current instruction) in the program coded by \( x \), where the contents of the registers is coded by \( y \).

2. \( \text{Nextcont}(i, x, y) \) is the code of the contents of the registers after executing the \( i \)th instruction in the program coded by \( x \), where the contents of the registers is coded by \( y \).

3. \( \text{Comp}(x, y, m) = \langle i, z \rangle \), where \( i \) and \( z \) are defined such that after running the program coded by \( x \) for \( m \) steps, where the initial contents of the program registers are coded by \( y \), the next instruction to be executed is the \( i \)th one, and \( z \) is the code of the current contents of the registers.

**Proposition 2.4.** The functions Nextline, Nextcont, and Comp are primitive recursive.

**Proof.** (1) \( \text{Nextline}(i, x, y) = i + 1 \), unless the \( i \)th instruction is a jump and the contents of the register being tested is nonzero:

\[
\text{Nextline}(i, x, y) = \begin{cases} 
\max \{ j \leq \text{Ln}(x) | j < i \land \text{LNum(Line}(j, x)) = \text{Jmp(Line}(i, x)) \} 
& \text{if } \text{Typ(Line}(i, x)) = 4 \land \Pi(\text{Reg(Line}(i, x)), x, y) \neq 0 \\
\min \{ j \leq \text{Ln}(x) | j > i \land \text{LNum(Line}(j, x)) = \text{Jmp(Line}(i, x)) \} 
& \text{if } \text{Typ(Line}(i, x)) = 5 \land \Pi(\text{Reg(Line}(i, x)), x, y) \neq 0 \\
i + 1 & \text{otherwise.}
\end{cases}
\]

For example, if the \( i \)th instruction of the program coded by \( x \) is a jump above, namely \( \text{Typ(Line}(i, x)) = 4 \), then the register being tested is \( \text{Reg(Line}(i, x)) \), and its contents must be nonzero for a jump to occur, so the contents of this register, which is obtained from the code \( y \) of all registers as \( \Pi(\text{Reg(Line}(i, x)), x, y) \) (remember that we may assume that there are \( x \) registers, by padding with zeros) must be nonzero.
Note that according to this definition, if the \( i \)th line is the final `continue`, then `Nextline` signals that the program has halted by yielding

\[
\text{Nextline}(i, x, y) > \text{Ln}(x).
\]

(2) We need two auxiliary functions `Add` and `Sub` defined as follows.

`Add(j, x, y)` is the number coding the contents of the registers used by the program coded by \( x \) after register \( R_j \) coded by \( \Pi(j, x, y) \) has been increased by 1, and

`Sub(j, x, y)` codes the contents of the registers after register \( R_j \) has been decremented by 1 (\( y \) codes the previous contents of the registers). It is easy to see that

\[
\text{Sub}(j, x, y) = \min z \leq y \left[ \Pi(j, x, z) = \Pi(j, x, y) - 1 \wedge \forall k \leq x [0 < k \neq j \Rightarrow \Pi(k, x, z) = \Pi(k, x, y)] \right].
\]

The definition of `Add` is slightly more tricky. We leave as an exercise to the reader to prove that:

\[
\text{Add}(j, x, y) = \min z \leq \text{Large}(x, y + 1) \left[ \Pi(j, x, z) = \Pi(j, x, y) + 1 \wedge \forall k \leq x [0 < k \neq j \Rightarrow \Pi(k, x, z) = \Pi(k, x, y)] \right],
\]

where the function `Large` is the function defined in an earlier exercise. Then

\[
\text{Nextcont}(i, x, y) =
\begin{align*}
\text{Add(Reg(Line}(i, x), x, y) \quad &\text{if } \text{Typ(Line}(i, x)) = 1 \\
\text{Sub(Reg(Line}(i, x), x, y) \quad &\text{if } \text{Typ(Line}(i, x)) = 2 \\
y \quad &\text{if } \text{Typ(Line}(i, x)) \geq 3.
\end{align*}
\]

(3) Recall that \( \Pi_1(z) = \Pi(1, 2, z) \) and \( \Pi_2(z) = \Pi(2, 2, z) \). The function `Comp` is defined by primitive recursion as follows:

\[
\text{Comp}(x, y, 0) = \langle 1, y \rangle \\
\text{Comp}(x, y, m + 1) = \langle \text{Nextline}(\Pi_1(\text{Comp}(x, y, m)), x, \Pi_2(\text{Comp}(x, y, m))), \text{Nextcont}(\Pi_1(\text{Comp}(x, y, m)), x, \Pi_2(\text{Comp}(x, y, m))) \rangle.
\]

If \( \text{Comp}(x, y, m) = \langle i, z \rangle \), then \( \Pi_1(\text{Comp}(x, y, m)) = i \) is the number of the next instruction to be executed and \( \Pi_2(\text{Comp}(x, y, m)) = z \) codes the current contents of the registers, so

\[
\text{Comp}(x, y, m + 1) = \langle \text{Nextline}(i, x, z), \text{Nextcont}(i, x, z) \rangle,
\]
as desired. \( \square \)
We can now reprove that every RAM computable function is partial computable. Indeed, assume that \( x \) codes a program \( P \).

We would like to define the partial function \( \text{End} \) so that for all \( x, y \), where \( x \) codes a program and \( y \) codes the contents of its registers, \( \text{End}(x, y) \) is the number of steps for which the computation runs before halting, if it halts. If the program does not halt, then \( \text{End}(x, y) \) is undefined.

If \( y \) is the value of the register \( R1 \) before the program \( P \) coded by \( x \) is started, recall that the contents of the registers is coded by \( \langle y, 0 \rangle \). Noticing that 0 and 1 do not code programs, we note that if \( x \) codes a program, then \( x \geq 2 \), and \( \Pi_1(z) = \Pi(1, x, z) \) is the contents of \( R1 \) as coded by \( z \).

Since \( \text{Comp}(x, y, m) = \langle i, z \rangle \), we have
\[
\Pi_1(\text{Comp}(x, y, m)) = i,
\]
where \( i \) is the number (index) of the instruction reached after running the program \( P \) coded by \( x \) with initial values of the registers coded by \( y \) for \( m \) steps. Thus, \( P \) halts if \( i \) is the last instruction in \( P \), namely \( \text{Ln}(x) \), iff
\[
\Pi_1(\text{Comp}(x, y, m)) = \text{Ln}(x).
\]
This suggests the following definition.

**Definition 2.11.** The predicate \( \text{End}(x, y) \) is defined by
\[
\text{End}(x, y) = \min m [\Pi_1(\text{Comp}(x, y, m)) = \text{Ln}(x)].
\]

Note that \( \text{End} \) is a partial computable function; it can be computed by a RAM program involving only one while loop searching for the number of steps \( m \). The function involved in the minimization is primitive recursive. However, in general, \( \text{End} \) is not a total function.

If \( \varphi \) is the partial computable function computed by the program \( P \) coded by \( x \), then we claim that
\[
\varphi(y) = \Pi_1(\Pi_2(\text{Comp}(x, \langle y, 0 \rangle, \text{End}(x, \langle y, 0 \rangle)))).
\]
This is because if \( m = \text{End}(x, \langle y, 0 \rangle) \) is the number of steps after which the program \( P \) coded by \( x \) halts on input \( y \), then
\[
\text{Comp}(x, \langle y, 0 \rangle, m) = \langle \text{Ln}(x), z \rangle,
\]
where \( z \) is the code of the register contents when the program stops. Consequently
\[
z = \Pi_2(\text{Comp}(x, \langle y, 0 \rangle, m))
\]
\[
z = \Pi_2(\text{Comp}(x, \langle y, 0 \rangle, \text{End}(x, \langle y, 0 \rangle)))).
\]
The value of the register \( R1 \) is \( \Pi_1(z) \), that is
\[
\varphi(y) = \Pi_1(\Pi_2(\text{Comp}(x, \langle y, 0 \rangle, \text{End}(x, \langle y, 0 \rangle)))).
\]
The above fact is worth recording as the following proposition which is a variant of a result known as the Kleene normal form.
Proposition 2.5. (Kleene normal form for RAM programs) If $\varphi$ is the partial computable function computed by the program $P$ coded by $x$, then we have

$$\varphi(y) = \Pi_1(\Pi_2(\text{Comp}(x, \langle y, 0 \rangle), \text{End}(x, \langle y, 0 \rangle))) \text{ for all } y \in \mathbb{N}.$$ 

Observe that $\varphi$ is written in the form $\varphi = g \circ \min f$, for some primitive recursive functions $f$ and $g$. It will convenient to denote the function $\varphi$ computed by the RAM program coded by $x$ as $\varphi_x$.

We can also exhibit a partial computable function which enumerates all the unary partial computable functions. It is a universal function.

Abusing the notation slightly, we will write $\varphi(x, y)$ for $\varphi(\langle x, y \rangle)$, viewing $\varphi$ as a function of two arguments (however, $\varphi$ is really a function of a single argument). We define the function $\varphi_{\text{univ}}$ as follows:

$$\varphi_{\text{univ}}(x, y) = \begin{cases} \Pi_1(\Pi_2(\text{Comp}(x, \langle y, 0 \rangle), \text{End}(x, \langle y, 0 \rangle))) & \text{if PROG}(x), \\ \text{undefined} & \text{otherwise}. \end{cases}$$

The function $\varphi_{\text{univ}}$ is a partial computable function with the following property: for every $x$ coding a RAM program $P$, for every input $y$,

$$\varphi_{\text{univ}}(x, y) = \varphi_x(y),$$

the value of the partial computable function $\varphi_x$ computed by the RAM program $P$ coded by $x$. If $x$ does not code a program, then $\varphi_{\text{univ}}(x, y)$ is undefined for all $y$.

By Proposition 1.9, the partial function $\varphi_{\text{univ}}$ is not computable (recursive). Indeed, being an enumerating function for the partial computable functions, it is an enumerating function for the total computable functions, and thus, it cannot be computable. Being a partial function saves us from a contradiction.

The existence of the universal function $\varphi_{\text{univ}}$ is sufficiently important to be recorded in the following proposition.

Proposition 2.6. (Universal RAM program) For the indexing of RAM programs defined earlier, there is a universal partial computable function $\varphi_{\text{univ}}$ such that, for all $x, y \in \mathbb{N}$, if $\varphi_x$ is the partial computable function computed by $P_x$, then

$$\varphi_x(y) = \varphi_{\text{univ}}(\langle x, y \rangle).$$

The program UNIV computing $\varphi_{\text{univ}}$ can be viewed as an interpreter for RAM programs. By giving the universal program UNIV the “program” $x$ and the “data” $y$, we get the result of executing program $P_x$ on input $y$. We can view the RAM model as a stored program computer.

---

2The term recursive function is now considered old-fashioned. Many researchers have switched to the term computable function.
By Theorem 2.3 and Proposition 2.6, the halting problem for the single program UNIV is undecidable. Otherwise, the halting problem for RAM programs would be decidable, a contradiction. It should be noted that the program UNIV can actually be written (with a certain amount of pain).

The existence of the function \( \varphi_{\text{univ}} \) leads us to the notion of an indexing of the RAM programs.

### 2.5 Indexing of RAM Programs

We can define a listing of the RAM programs as follows. If \( x \) codes a program (that is, if \( \text{PROG}(x) \) holds) and \( P \) is the program that \( x \) codes, we call this program \( P \) the \( x \)th RAM program and denote it as \( P_x \). If \( x \) does not code a program, we let \( P_x \) be the program that diverges for every input:

\[
\begin{align*}
N1 & \quad \text{add} & R1 \\
N1 & \quad R1 & \quad \text{jmp} & \quad N1a \\
N1 & \quad \text{continue}
\end{align*}
\]

Therefore, in all cases, \( P_x \) stands for the \( x \)th RAM program. Thus, we have a listing of RAM programs, \( P_0, P_1, P_2, P_3, \ldots \), such that every RAM program (of the restricted type considered here) appears in the list exactly once, except for the “infinite loop” program. For example, the program \( \text{Padd2} \) (adding 2 to an integer) appears as

\[
P_{1018748519973070618}.
\]

In particular, note that \( \varphi_{\text{univ}} \) being a partial computable function, it is computed by some RAM program UNIV that has a code \( \text{univ} \) and is the program \( P_{\text{univ}} \) in the list.

Having an indexing of the RAM programs, we also have an indexing of the partial computable functions.

**Definition 2.12.** For every integer \( x \geq 0 \), we let \( P_x \) be the RAM program coded by \( x \) as defined earlier, and \( \varphi_x \) be the partial computable function computed by \( P_x \).

For example, the function \( \text{add2} \) (adding 2 to an integer) appears as

\[
\varphi_{1018748519973070618}.
\]

**Remark:** Kleene used the notation \( \{x\} \) for the partial computable function coded by \( x \). Due to the potential confusion with singleton sets, we follow Rogers, and use the notation \( \varphi_x \); see Rogers [32], page 21.

It is important to observe that different programs \( P_x \) and \( P_y \) may compute the same function, that is, while \( P_x \neq P_y \) for all \( x \neq y \), it is possible that \( \varphi_x = \varphi_y \). For example,
the program $P_y$ coded by $y$ may be the program obtained from the program $P_x$ coded by $x$ obtained by adding and subtracting 1 a million times to a register not in the program $P_x$. In fact, it is undecidable whether $\varphi_x = \varphi_y$.

The object of the next section is to show the existence of Kleene’s $T$-predicate. This will yield another important normal form. In addition, the $T$-predicate is a basic tool in recursion theory.

### 2.6 Kleene’s $T$-Predicate

In Section 2.3, we have encoded programs. The idea of this section is to also encode computations of RAM programs. Assume that $x$ codes a program, that $y$ is some input (not a code), and that $z$ codes a computation of $P_x$ on input $y$.

**Definition 2.13.** The predicate $T(x, y, z)$ is defined as follows:

$T(x, y, z)$ holds iff $x$ codes a RAM program, $y$ is an input, and $z$ codes a halting computation of $P_x$ on input $y$.

The code $z$ of a computation packs the consecutive “states” of the computation, namely the pairs $\langle i_j, y_j \rangle$, where $i_j$ is the physical location of the next instruction to be executed and each $y_j$ codes the contents of the registers just before execution of this instruction. We will show that $T$ is primitive recursive.

First we need to encode computations. We say that $z$ codes a computation of length $n \geq 1$ if

$$z = \langle n + 2, \langle 1, y_0 \rangle, \langle i_1, y_1 \rangle, \ldots, \langle i_n, y_n \rangle \rangle,$$

where each $i_j$ is the physical location of the next instruction to be executed and each $y_j$ codes the contents of the registers just before execution of the instruction at the location $i_j$. Also, $y_0$ codes the initial contents of the registers, that is, $y_0 = \langle y, 0 \rangle$, for some input $y$.

We let $L_z(z) = \Pi_1(I_z)$ (not to be confused with $Ln(x)$).

Note that $i_j$ denotes the physical location of the next instruction to be executed in the sequence of instructions constituting the program coded by $x$, and not the line number (label) of this instruction. Thus, the first instruction to be executed is in location 1, $1 \leq i_j \leq Ln(x)$, and $i_{n-1} = Ln(x)$. Since the last instruction which is executed is the last physical instruction in the program, namely, a continue, there is no next instruction to be executed after that, and $i_n$ is irrelevant. Writing the definition of $T$ is a little simpler if we let $i_n = Ln(x) + 1$. 
Definition 2.14. The $T$-predicate is the primitive recursive predicate defined as follows:

\[ T(x, y, z) \iff \text{PROG}(x) \text{ and } (Lz(z) \geq 3) \text{ and } \]

\[ \forall j \leq Lz(z) - 3 [0 \leq j \Rightarrow \]

\[ \text{Nextline}(\Pi(j + 2, Lz(z), z)) = \Pi(j + 3, Lz(z), z) \text{ and } \]

\[ \text{Nextcont}(\Pi(j + 2, Lz(z), z)) = \Pi(j + 3, Lz(z), z) \text{ and } \]

\[ \Pi(\Pi(Lz(z) - 1, Lz(z), z)) = \text{Ln}(x) \text{ and } \]

\[ y = \Pi(\Pi(2, Lz(z), z)) \text{ and } \Pi(\Pi(2, Lz(z), z)) = 0]. \]

The reader can verify that $T(x, y, z)$ holds iff $x$ codes a RAM program, $y$ is an input, and $z$ codes a halting computation of $P_x$ on input $y$. For example, since

\[ z = \langle n + 2, \langle 1, y_0 \rangle, \langle i_1, y_1 \rangle, \ldots, \langle i_n, y_n \rangle \rangle, \]

we have $\Pi(j + 2, Lz(z), z) = \langle i_j, y_j \rangle$ and $\Pi(j + 3, Lz(z), z) = \langle i_j \rangle$, so $\Pi(\Pi(j + 2, Lz(z), z)) = y_j$, and similarly $\Pi(\Pi(Lz(z), z)) = i_j$, $\Pi(\Pi(j + 3, Lz(z), z)) = y_j$, so the $T$ predicate expresses that $\text{Nextline}(i_{j-1}, y_{j-1}) = i_j$ and $\text{Nextcont}(i_{j-1}, y_{j-1}) = y_j$.

In order to extract the output of $P_x$ from $z$, we define the primitive recursive function $\text{Res}$ as follows:

\[ \text{Res}(z) = \Pi(\Pi(\Pi(Lz(z), Lz(z), z))) \]

The explanation for this formula is that if $\Pi(Lz(z), Lz(z), z) = \langle i_n, y_n \rangle$, then $\Pi(\Pi(Lz(z), Lz(z), z)) = y_n$, the code of the registers, and since the output is returned in Register $R1$, $\text{Res}(z)$ is the contents of register $R1$ when $P_x$ halts, that is, $\Pi(\Pi(Lz(z)))$. Using the $T$-predicate, we get the so-called Kleene normal form.

Theorem 2.7. (Kleene Normal Form) Using the indexing of the partial computable functions defined earlier, we have

\[ \varphi_x(y) = \text{Res}[\min z(T(x, y, z))], \]

where $T(x, y, z)$ and $\text{Res}$ are primitive recursive.

Note that the universal function $\varphi_{univ}$ can be defined as

\[ \varphi_{univ}(x, y) = \text{Res}[\min(z(T(x, y, z))]. \]

There is another important property of the partial computable functions, namely, that composition is effective (computable). We need two auxiliary primitive recursive functions. The function $\text{Conprogs}$ creates the code of the program obtained by concatenating the programs $P_x$ and $P_y$, and for $i \geq 2$, $\text{Cumclr}(i)$ is the code of the program which clears registers $R2, \ldots, R_i$. To get $\text{Cumclr}$, we can use the function $\text{clr}(i)$ such that $\text{clr}(i)$ is the code of the program
CHAPTER 2. UNIVERSAL RAM PROGRAMS AND THE HALTING PROBLEM

\[
\begin{array}{ccc}
N1 & \text{tail} & Ri \\
N1 & Ri & \text{jmp} & N1a \\
N & \text{continue}
\end{array}
\]

We leave it as an exercise to prove that clr, Conprogs, and Cumclr, are primitive recursive.

**Theorem 2.8.** There is a primitive recursive function \( c \) such that

\[
\varphi_{c(x,y)} = \varphi_x \circ \varphi_y.
\]

**Proof.** If both \( x \) and \( y \) code programs, then \( \varphi_x \circ \varphi_y \) can be computed as follows: Run \( P_y \), clear all registers but \( R1 \), then run \( P_x \). Otherwise, let loop be the index of the infinite loop program:

\[
c(x, y) = \begin{cases} 
\text{Conprogs}(y, \text{Conprogs}(\text{Cumclr}(y), x)) & \text{if PROG}(x) \text{ and PROG}(y) \\
\text{loop} & \text{otherwise.}
\end{cases}
\]

\( \square \)

### 2.7 A Non-Computable Function; Busy Beavers

Total functions that are not computable must grow very fast and thus are very complicated. Yet, in 1962, Radó published a paper in which he defined two functions \( \Sigma \) and \( S \) (involving computations of Turing machines) that are total and not computable.

Consider Turing machines with a tape alphabet \( \Gamma = \{1, B\} \) with two symbols (\( B \) being the blank). We also assume that these Turing machines have a special final state \( q_F \), which is a blocking state (there are no transitions from \( q_F \)). We do not count this state when counting the number of states of such Turing machines. The game is to run such Turing machines with a fixed number of states \( n \) starting on a blank tape, with the goal of producing the maximum number of (not necessarily consecutive) ones (1).

**Definition 2.15.** The function \( \Sigma \) (defined on the positive natural numbers) is defined as the maximum number \( \Sigma(n) \) of (not necessarily consecutive) 1’s written on the tape after a Turing machine with \( n \geq 1 \) states started on the blank tape halts. The function \( S \) is defined as the maximum number \( S(n) \) of moves that can be made by a Turing machine of the above type with \( n \) states before it halts, started on the blank tape.\(^3\)

**Definition 2.16.** A Turing machine with \( n \) states that writes the maximum number \( \Sigma(n) \) of 1’s when started on the blank tape is called a busy beaver.

\(^3\)The function \( S \) defined here is obviously not the successor function from Definition 1.13.
Busy beavers are hard to find, even for small \( n \). First, it can be shown that the number of distinct Turing machines of the above kind with \( n \) states is \((4(n + 1))^{2^n}\). Second, since it is undecidable whether a Turing machine halts on a given input, it is hard to tell which machines loop or halt after a very long time.

Here is a summary of what is known for \( 1 \leq n \leq 6 \). Observe that the exact value of \( \Sigma(5), \Sigma(6), S(5) \) and \( S(6) \) is unknown.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \Sigma(n) )</th>
<th>( S(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>107</td>
</tr>
<tr>
<td>5</td>
<td>( \geq 4098 )</td>
<td>( \geq 47,176,870 )</td>
</tr>
<tr>
<td>6</td>
<td>( \geq 95,524,079 )</td>
<td>( \geq 8,690,333,381,690,951 )</td>
</tr>
<tr>
<td>6</td>
<td>( \geq 3.515 \times 10^{18267} )</td>
<td>( \geq 7.412 \times 10^{405534} )</td>
</tr>
</tbody>
</table>

The first entry in the table for \( n = 6 \) corresponds to a machine due to Heiner Marxen (1999). This record was surpassed by Pavel Kropitz in 2010, which corresponds to the second entry for \( n = 6 \). The machines achieving the record in 2017 for \( n = 4, 5, 6 \) are shown below, where the blank is denoted \( \Delta \) instead of \( B \), and where the special halting state is denoted \( H \): 

4-state busy beaver:

\[
\begin{array}{cccc}
\Delta & (1, R, B) & (1, L, A) & (1, R, D) \\
1 & (1, L, B) & (\Delta, L, C) & (1, L, D) & (\Delta, R, A)
\end{array}
\]

The above machine output 13 ones in 107 steps. In fact, the output is \( \Delta \Delta 1 \Delta 1 1 1 1 1 1 1 1 1 \Delta \Delta \).

5-state best contender:

\[
\begin{array}{ccccc}
\Delta & (1, R, B) & (1, R, C) & (1, R, D) & (1, L, A) & (1, H) \\
1 & (1, L, C) & (1, R, B) & (\Delta, L, E) & (1, L, D) & (\Delta, L, A)
\end{array}
\]

The above machine output 4098 ones in 47,176,870 steps.

6-state contender (Heiner Marxen):

\[
\begin{array}{cccccc}
\Delta & (1, R, B) & (1, L, C) & (\Delta, R, F) & (1, R, A) & (1, L, H) & (\Delta, L, A) \\
1 & (1, R, A) & (1, L, B) & (1, L, D) & (\Delta, L, E) & (1, L, F) & (\Delta, L, C)
\end{array}
\]
The above machine outputs 96,524,079 ones in 8,690,333,381,690,951 steps.

6-state best contender (Pavel Kropitz):

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ</td>
<td>(1, R, B)</td>
<td>(1, R, C)</td>
<td>(1, L, D)</td>
<td>(1, R, E)</td>
<td>(1, L, A)</td>
<td>(1, L, H)</td>
</tr>
<tr>
<td>1</td>
<td>(1, L, E)</td>
<td>(1, R, F)</td>
<td>(Δ, R, B)</td>
<td>(Δ, L, C)</td>
<td>(Δ, R, D)</td>
<td>(1, R, C)</td>
</tr>
</tbody>
</table>

The above machine output at least $3.515 \times 10^{18267}$ ones!

The reason why it is so hard to compute $\Sigma$ and $S$ is that they are not computable!

**Theorem 2.9.** The functions $\Sigma$ and $S$ are total functions that are not computable (not recursive).

**Proof sketch.** The proof consists in showing that $\Sigma$ (and similarly for $S$) eventually outgrows any computable function. More specifically, we claim that for every computable function $f$, there is some positive integer $k_f$ such that

$$\Sigma(n + k_f) \geq f(n) \quad \text{for all } n \geq 0.$$  

We simply have to pick $k_f$ to be the number of states of a Turing machine $M_f$ computing $f$. Then we can create a Turing machine $M_{n,f}$ that works as follows. Using $n$ of its states, it writes $n$ ones on the tape, and then it simulates $M_f$ with input $1^n$. Since the output of $M_{n,f}$ started on the blank tape consists of $f(n)$ ones, and since $\Sigma(n + k_f)$ is the maximum number of ones that a Turing machine with $n + k_f$ states will output when it stops, we must have

$$\Sigma(n + k_f) \geq f(n) \quad \text{for all } n \geq 0.$$  

Next observe that $\Sigma(n) < \Sigma(n + 1)$, because we can create a Turing machine with $n + 1$ states which simulates a busy beaver machine with $n$ states, and then writes an extra 1 when the busy beaver stops, by making a transition to the $(n + 1)$th state. It follows immediately that if $m < n$ then $\Sigma(m) < \Sigma(n)$. If $\Sigma$ was computable, then so would be the function $g$ given by $g(n) = \Sigma(2n)$. By the above, we would have

$$\Sigma(n + k_g) \geq g(n) = \Sigma(2n) \quad \text{for all } n \geq 0,$$

and for $n > k_g$, since $2n > n + k_g$, we would have $\Sigma(n + n_g) < \Sigma(2n)$, contradicting the fact that $\Sigma(n + n_g) \geq \Sigma(2n)$.

Since by definition $S(n)$ is the maximum number of moves that can be made by a Turing machine of the above type with $n$ states before it halts, $S(n) \geq \Sigma(n)$. Then the same reasoning as above shows that $S$ is not a computable function.

The zoo of computable and non-computable functions is illustrated in Figure 2.1.
2.7. A NON-COMPUTABLE FUNCTION; BUSY BEAVERS

Figure 2.1: Computability Classification of Functions.
Chapter 3

Elementary Recursive Function Theory

3.1 Acceptable Indexings

In Chapter 2, we have exhibited a specific indexing of the partial computable functions by encoding the RAM programs. Using this indexing, we showed the existence of a universal function $\varphi_{\text{univ}}$ and of a computable function $c$, with the property that for all $x, y \in \mathbb{N}$,

$$\varphi_c(x,y) = \varphi_x \circ \varphi_y.$$ 

It is natural to wonder whether the same results hold if a different coding scheme is used or if a different model of computation is used, for example, Turing machines. In other words, we would like to know if our results depend on a specific coding scheme or not.

Our previous results showing the characterization of the partial computable functions being independent of the specific model used, suggests that it might be possible to pinpoint certain properties of coding schemes which would allow an axiomatic development of recursive function theory. What we are aiming at is to find some simple properties of “nice” coding schemes that allow one to proceed without using explicit coding schemes, as long as the above properties hold.

Remarkably, such properties exist. Furthermore, any two coding schemes having these properties are equivalent in a strong sense (called effectively equivalent), and so, one can pick any such coding scheme without any risk of losing anything else because the wrong coding scheme was chosen. Such coding schemes, also called indexings, or Gödel numberings, or even programming systems, are called acceptable indexings.

**Definition 3.1.** An indexing of the partial computable functions is an infinite sequence $\varphi_0, \varphi_1, \ldots$, of partial computable functions that includes all the partial computable functions of one argument (there might be repetitions, this is why we are not using the term enumeration). An indexing is universal if it contains the partial computable function $\varphi_{\text{univ}}$.
such that

\[ \varphi_{univ}(i, x) = \varphi_i(x) \text{ for all } i, x \in \mathbb{N}. \]  

(*univ)

An indexing is acceptable if it is universal and if there is a total computable function \( c \) for composition, such that

\[ \varphi_{c(i,j)} = \varphi_i \circ \varphi_j \text{ for all } i, j \in \mathbb{N}. \]  

(*compos)

An indexing may fail to be universal because it is not “computable enough,” in the sense that it does not yield a function \( \varphi_{univ} \) satisfying (*univ). It may also fail to be acceptable because it is not “computable enough,” in the sense that it does not yield a function \( \varphi_{univ} \) satisfying (*compos).

From Chapter 2, we know that the specific indexing of the partial computable functions given for RAM programs is acceptable. Another characterization of acceptable indexings left as an exercise is the following: an indexing \( \psi_0, \psi_1, \psi_2, \ldots \) of the partial computable functions is acceptable iff there exists a total computable function \( f \) translating the RAM indexing of Section 2.3 into the indexing \( \psi_0, \psi_1, \psi_2, \ldots \), that is,

\[ \varphi_i = \psi_{f(i)} \text{ for all } i \in \mathbb{N}. \]

A very useful property of acceptable indexings is the so-called “s-m-n Theorem”. Using the slightly loose notation \( \varphi(x_1, \ldots, x_n) \) for \( \varphi((x_1, \ldots, x_n)) \), the s-m-n Theorem says the following. Given a function \( \varphi \) considered as having \( m + n \) arguments, if we fix the values of the first \( m \) arguments and we let the other \( n \) arguments vary, we obtain a function \( \psi \) of \( n \) arguments. Then the index of \( \psi \) depends in a computable fashion upon the index of \( \varphi \) and the first \( m \) arguments \( x_1, \ldots, x_m \). We can “pull” the first \( m \) arguments of \( \varphi \) into the index of \( \psi \).

**Theorem 3.1.** (The “s-m-n Theorem”) For any acceptable indexing \( \varphi_0, \varphi_1, \ldots \), there is a total computable function \( s : \mathbb{N}^{n+2} \rightarrow \mathbb{N} \), such that, for all \( i, m, n \geq 1 \), for all \( x_1, \ldots, x_m \) and all \( y_1, \ldots, y_n \), we have

\[ \varphi_{s(i,m,x_1,\ldots,x_m)}(y_1, \ldots, y_n) = \varphi_i(x_1, \ldots, x_m, y_1, \ldots, y_n). \]

**Proof.** First, note that the above identity is really

\[ \varphi_{s(i,m,(x_1,\ldots,x_m))}(y_1, \ldots, y_n) = \varphi_i((x_1, \ldots, x_m, y_1, \ldots, y_n)). \]

Recall that there is a primitive recursive function \( \text{Con} \) such that

\[ \text{Con}(m, (x_1, \ldots, x_m), (y_1, \ldots, y_n)) = (x_1, \ldots, x_m, y_1, \ldots, y_n) \]

for all \( x_1, \ldots, x_m, y_1, \ldots, y_n \in \mathbb{N} \). Hence, a computable function \( s \) such that

\[ \varphi_{s(i,m,x)}(y) = \varphi_i(\text{Con}(m, x, y)) \]
will do. We define some auxiliary primitive recursive functions as follows:
\[ P(y) = \langle 0, y \rangle \quad \text{and} \quad Q(\langle x, y \rangle) = \langle x + 1, y \rangle. \]

Since we have an indexing of the partial computable functions, there are indices \( p \) and \( q \) such that \( P = \varphi_p \) and \( Q = \varphi_q \). Let \( R \) be defined such that
\[
\begin{align*}
R(0) &= p, \\
R(x + 1) &= c(q, R(x)),
\end{align*}
\]
where \( c \) is the computable function for composition given by the indexing. We prove by induction of \( x \) that
\[ \varphi_{R(x)}(y) = \langle x, y \rangle \quad \text{for all} \quad x, y \in \mathbb{N}. \]
For this we use the existence of the universal function \( \varphi_{\text{univ}} \).

For the base case \( x = 0 \), we have
\[
\begin{align*}
\varphi_{R(0)}(y) &= \varphi_{\text{univ}}(\langle R(0), y \rangle) \\
&= \varphi_{\text{univ}}(\langle p, y \rangle) \\
&= \varphi_p(y) = P(y) = \langle 0, y \rangle.
\end{align*}
\]

For the induction step, we have
\[
\begin{align*}
\varphi_{R(x+1)}(y) &= \varphi_{\text{univ}}(\langle R(x + 1), y \rangle) \\
&= \varphi_{\text{univ}}(\langle c(q, R(x)), y \rangle) \\
&= \varphi_{c(q, R(x))}(y) \\
&= (\varphi_q \circ \varphi_{R(x)})(y) \\
&= \varphi_q(\langle x, y \rangle) = Q(\langle x, y \rangle) = \langle x + 1, y \rangle.
\end{align*}
\]

Also, recall that \( \langle x, y, z \rangle = \langle x, \langle y, z \rangle \rangle \), by definition of pairing. Then we have
\[ \varphi_{R(x)} \circ \varphi_{R(y)}(z) = \varphi_{R(x)}(\langle y, z \rangle) = \langle x, y, z \rangle. \]
Finally, let \( k \) be an index for the function Con, that is, let
\[ \varphi_k(\langle m, x, y \rangle) = \text{Con}(m, x, y). \]
Define \( s \) by
\[ s(i, m, x) = c(i, c(k, c(R(m), R(x)))). \]
Then we have
\[ \varphi_{s(i,m,x)}(y) = \varphi_i \circ \varphi_k \circ \varphi_{R(m)} \circ \varphi_{R(x)}(y) = \varphi_i(\text{Con}(m, x, y)), \]
as desired. Notice that if the composition function \( c \) is primitive recursive, then \( s \) is also primitive recursive. In particular, for the specific indexing of the RAM programs given in Section 2.3, the function \( s \) is primitive recursive. \( \square \)
In practice, when using the s-m-n Theorem we usually denote the function \( s(i, m, x) \) simply as \( s(x) \).

As a first application of the s-m-n Theorem, we show that any two acceptable indexings are effectively inter-translatable, that is, computably inter-translatable.

**Theorem 3.2.** Let \( \varphi_0, \varphi_1, \ldots \) be a universal indexing, and let \( \psi_0, \psi_1, \ldots \), be any indexing with a total computable s-1-1 function, that is, a function \( s \) such that

\[
\psi_{s(i, 1, x)}(y) = \psi_i(x, y)
\]

for all \( i, x, y \in \mathbb{N} \). Then there is a total computable function \( t \) such that \( \varphi_i = \psi_{t(i)} \).

**Proof.** Let \( \varphi_{\text{univ}} \) be a universal partial computable function for the indexing \( \varphi_0, \varphi_1, \ldots \). Since \( \psi_0, \psi_1, \ldots \), is also an indexing \( \varphi_{\text{univ}} \) occurs somewhere in the second list, and thus, there is some \( k \) such that \( \varphi_{\text{univ}} = \psi_k \). Then we have

\[
\psi_{s(k, 1, i)}(x) = \psi_k(i, x) = \varphi_{\text{univ}}(i, x) = \varphi_i(x),
\]

for all \( i, x \in \mathbb{N} \). Therefore, we can take the function \( t \) to be the function defined such that

\[
t(i) = s(k, 1, i)
\]

for all \( i \in \mathbb{N} \). \( \square \)

Using Theorem 3.2, if we have two acceptable indexings \( \varphi_0, \varphi_1, \ldots \), and \( \psi_0, \psi_1, \ldots \), there exist total computable functions \( t \) and \( u \) such that

\[
\varphi_i = \psi_{t(i)} \quad \text{and} \quad \psi_i = \varphi_{u(i)}
\]

for all \( i \in \mathbb{N} \).

Also note that if the composition function \( c \) is primitive recursive, then any s-m-n function is primitive recursive, and the translation functions are primitive recursive. Actually, a stronger result can be shown. It can be shown that for any two acceptable indexings, there exist total computable injective and surjective translation functions. In other words, any two acceptable indexings are recursively isomorphic (Roger’s isomorphism theorem); see Machtey and Young [25]. Next we turn to algorithmically unsolvable, or undecidable, problems.

### 3.2 Undecidable Problems

We saw in Section 2.3 that the halting problem for RAM programs is undecidable. In this section, we take a slightly more general approach to study the undecidability of problems, and give some tools for resolving decidability questions.

First, we prove again the undecidability of the halting problem, but this time, for any indexing of the partial computable functions.
Theorem 3.3. (Halting Problem, Abstract Version) Let $\psi_0, \psi_1, \ldots$, be any indexing of the partial computable functions. Then the function $f$ defined such that

$$f(x, y) = \begin{cases} 1 & \text{if } \psi_x(y) \text{ is defined}, \\ 0 & \text{if } \psi_x(y) \text{ is undefined}, \end{cases}$$

is not computable.

Proof. Assume that $f$ is computable, and let $g$ be the function defined such that

$$g(x) = f(x, x)$$

for all $x \in \mathbb{N}$. Then $g$ is also computable. Let $\theta$ be the function defined such that

$$\theta(x) = \begin{cases} 0 & \text{if } g(x) = 0, \\ \text{undefined} & \text{if } g(x) = 1. \end{cases}$$

We claim that $\theta$ is not even partial computable. Observe that $\theta$ is such that

$$\theta(x) = \begin{cases} 0 & \text{if } \psi_x(x) \text{ is undefined}, \\ \text{undefined} & \text{if } \psi_x(x) \text{ is defined}. \end{cases}$$

If $\theta$ was partial computable, it would occur in the list as some $\psi_i$, and we would have

$$\theta(i) = \psi_i(i) = 0 \quad \text{iff} \quad \psi_i(i) \text{ is undefined},$$

a contradiction. Therefore, $f$ and $g$ can’t be computable. \qed

Observe that the proof of Theorem 3.3 does not use the fact that the indexing is universal or acceptable, and thus, the theorem holds for any indexing of the partial computable functions.

Given any set, $X$, for any subset, $A \subseteq X$, of $X$, recall that the characteristic function, $C_A$ (or $\chi_A$), of $A$ is the function, $C_A : X \to \{0, 1\}$, defined so that, for all $x \in X$,

$$C_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

The function $g$ defined in the proof of Theorem 3.3 is the characteristic function of an important set denoted as $K$.

Definition 3.2. Given any indexing $(\psi_i)$ of the partial computable functions, the set $K$ is defined by

$$K = \{ x \mid \psi_x(x) \text{ is defined} \}.$$
Definition 3.3. A subset $A$ of $\Sigma^*$ (or a subset $A$ of $\mathbb{N}$) is computable, or recursive,\textsuperscript{1} or decidable iff its characteristic function, $C_A$, is a total computable function.

Using Definition 3.3, Theorem 3.3 can be restated as follows.

Proposition 3.4. For any indexing $\varphi_0, \varphi_1, \ldots$ of the partial computable functions (over $\Sigma^*$ or $\mathbb{N}$), the set $K = \{x \mid \varphi_x(x) \text{ is defined}\}$ is not computable (not recursive).

Computable (recursive) sets allow us to define the concept of a decidable (or undecidable) problem. The idea is to generalize the situation described in Section 2.3 and Section 2.6, where a set of objects, the RAM programs, is encoded into a set of natural numbers, using a coding scheme. For example, we would like to discuss the notion of computability of sets of trees or sets of graphs.

Definition 3.4. Let $C$ be a countable set of objects, and let $P$ be a property of objects in $C$. We view $P$ as the set
\[
\{ a \in C \mid P(a) \}.
\]
A coding-scheme is an injective function $\# : C \to \mathbb{N}$ that assigns a unique code to each object in $C$. The property $P$ is decidable (relative to $\#$) iff the set $\{ \#(a) \mid a \in C \text{ and } P(a) \}$ is computable (recursive). The property $P$ is undecidable (relative to $\#$) iff the set $\{ \#(a) \mid a \in C \text{ and } P(a) \}$ is not computable (not recursive).

Observe that the decidability of a property $P$ of objects in $C$ depends upon the coding scheme $\#$. Thus, if we are cheating in using a non-effective (i.e. not computable by a computer program) coding scheme, we may declare that a property is decidable even though it is not decidable in some reasonable coding scheme. Consequently, we require a coding scheme $\#$ to be effective in the following sense. Given any object $a \in C$, we can effectively (i.e. algorithmically) determine its code $\#(a)$. Conversely, given any integer $n \in \mathbb{N}$, we should be able to tell effectively if $n$ is the code of some object in $C$, and if so, to find this object. In practice, it is always possible to describe the objects in $C$ as strings over some (possibly complex) alphabet $\Sigma$ (sets of trees, graphs, etc). In such cases, the coding schemes are computable functions from $\Sigma^*$ to $\mathbb{N} = \{a_1\}^*$.

For example, let $C = \mathbb{N} \times \mathbb{N}$, where the property $P$ is the equality of the partial functions $\varphi_x$ and $\varphi_y$. We can use the pairing function $\langle -, - \rangle$ as a coding function, and the problem is formally encoded as the computability (recursiveness) of the set
\[
\{ \langle x, y \rangle \mid x, y \in \mathbb{N}, \varphi_x = \varphi_y \}.
\]

In most cases, we don’t even bother to describe the coding scheme explicitly, knowing that such a description is routine, although perhaps tedious.

We now show that most properties about programs (except the trivial ones) are undecidable.

\textsuperscript{1}Since 1996, the term recursive has been considered old-fashioned by many researchers, and the term computable has been used instead.
3.3 Reducibility and Rice’s Theorem

First, we show that it is undecidable whether a RAM program halts for every input. In other words, it is undecidable whether a procedure is an algorithm. We actually prove a more general fact.

Proposition 3.5. For any acceptable indexing \( \varphi_0, \varphi_1, \ldots \) of the partial computable functions, the set

\[
\text{TOTAL} = \{ x \mid \varphi_x \text{ is a total function} \}
\]

is not computable (not recursive).

Proof. The proof uses a technique known as reducibility. We try to reduce a set \( A \) known to be noncomputable (nonrecursive) to TOTAL via a computable function \( f : A \to \text{TOTAL} \), so that

\[
x \in A \iff f(x) \in \text{TOTAL}.
\]

If TOTAL were computable (recursive), its characteristic function \( g \) would be computable, and thus, the function \( g \circ f \) would be computable, a contradiction, since \( A \) is assumed to be noncomputable (nonrecursive). In the present case, we pick \( A = K \). To find the computable function \( f : K \to \text{TOTAL} \), we use the s-m-n Theorem. Let \( \theta \) be the function defined below:

for all \( x, y \in \mathbb{N} \),

\[
\theta(x, y) = \begin{cases} 
\varphi_x(x) & \text{if } x \in K, \\
\text{undefined} & \text{if } x \notin K.
\end{cases}
\]

Note that \( \theta \) does not depend on \( y \). The function \( \theta \) is partial computable. Indeed, we have

\[
\theta(x, y) = \varphi_x(x) = \varphi_{\text{univ}}(x, x).
\]

Thus, \( \theta \) has some index \( j \), so that \( \theta = \varphi_j \), and by the s-m-n Theorem, we have

\[
\varphi_{s(j, 1, x)}(y) = \varphi_j(x, y) = \theta(x, y).
\]

Let \( f \) be the computable function defined such that

\[
f(x) = s(j, 1, x)
\]

for all \( x \in \mathbb{N} \). Then we have

\[
\varphi_{f(x)}(y) = \begin{cases} 
\varphi_x(x) & \text{if } x \in K, \\
\text{undefined} & \text{if } x \notin K
\end{cases}
\]

for all \( y \in \mathbb{N} \). Thus, observe that \( \varphi_{f(x)} \) is a total function iff \( x \in K \), that is,

\[
x \in K \iff f(x) \in \text{TOTAL},
\]

where \( f \) is computable. As we explained earlier, this shows that TOTAL is not computable (not recursive).
The above argument can be generalized to yield a result known as Rice's theorem. Let \( \varphi_0, \varphi_1, \ldots \) be any indexing of the partial computable functions, and let \( C \) be any set of partial computable functions. We define the set \( P_C \) as

\[
P_C = \{ x \in \mathbb{N} \mid \varphi_x \in C \}.
\]

We can view \( C \) as a property of some of the partial computable functions. For example

\[ C = \{ \text{all total computable functions} \} \]

We say that \( C \) is nontrivial if \( C \) is neither empty nor the set of all partial computable functions. Equivalently \( C \) is nontrivial iff \( P_C \neq \emptyset \) and \( P_C \neq \mathbb{N} \). We may think of \( P_C \) as the set of programs computing the functions in \( C \).

**Theorem 3.6. (Rice's Theorem, 1953)** For any acceptable indexing \( \varphi_0, \varphi_1, \ldots \) of the partial computable functions, for any set \( C \) of partial computable functions, the set

\[
P_C = \{ x \in \mathbb{N} \mid \varphi_x \in C \}
\]

is not computable (not recursive) unless \( C \) is trivial.

**Proof.** Assume that \( C \) is nontrivial. A set is computable (recursive) iff its complement is computable (recursive) (the proof is trivial). Hence, we may assume that the totally undefined function is not in \( C \), and since \( C \neq \emptyset \), let \( \psi \) be some other function in \( C \). We produce a computable function \( f \) such that

\[
\varphi_f(x)(y) = \begin{cases} 
\psi(y) & \text{if } x \in K, \\
\text{undefined} & \text{if } x \notin K,
\end{cases}
\]

for all \( y \in \mathbb{N} \). We get \( f \) by using the s-m-n Theorem. Let \( \psi = \varphi_i \), and define \( \theta \) as follows:

\[
\theta(x, y) = \varphi_{\text{univ}}(i, y) + (\varphi_{\text{univ}}(x, x) \downarrow \varphi_{\text{univ}}(x, x)),
\]

where \( \downarrow \) is the primitive recursive function monus for truncated subtraction; see Section 1.7. Recall that \( \varphi_{\text{univ}}(x, x) \downarrow \varphi_{\text{univ}}(x, x) \) is defined iff \( \varphi_{\text{univ}}(x, x) \) is defined iff \( x \in K \), and so

\[
\theta(x, y) = \varphi_{\text{univ}}(i, y) = \varphi_i(y) = \psi(y) \quad \text{iff} \quad x \in K
\]

and \( \theta(x, y) \) is undefined otherwise. Clearly \( \theta \) is partial computable, and we let \( \theta = \varphi_j \). By the s-m-n Theorem, we have

\[
\varphi_{s(j, 1, x)}(y) = \varphi_j(x, y) = \theta(x, y)
\]

for all \( x, y \in \mathbb{N} \). Letting \( f \) be the computable function such that

\[
f(x) = s(j, 1, x),
\]
by definition of $\theta$, we get

$$\varphi_{f(x)}(y) = \theta(x, y) = \begin{cases} 
\psi(y) & \text{if } x \in K, \\
\text{undefined} & \text{if } x \notin K.
\end{cases}$$

Thus, $f$ is the desired reduction function. Now we have

$$x \in K \text{ iff } f(x) \in P_C,$$

and thus, the characteristic function $C_K$ of $K$ is equal to $C_P \circ f$, where $C_P$ is the characteristic function of $P_C$. Therefore, $P_C$ is not computable (not recursive), since otherwise, $K$ would be computable, a contradiction. \hfill \square

Rice’s theorem shows that all nontrivial properties of the input/output behavior of programs are undecidable!

It is important to understand that Rice’s theorem says that the set $P_C$ of indices of all partial computable functions equal to some function in a given set $C$ of partial computable functions is not computable if $C$ is nontrivial, not that the set $C$ is not computable if $C$ is nontrivial. The second statement does not make any sense because our machinery only applies to sets of natural numbers (or sets of strings). For example, the set $C = \{ \varphi_{i_0} \}$ consisting of a single partial computable function is nontrivial, and being finite, under the second wrong interpretation it would be computable. But we need to consider the set

$$P_C = \{ n \in \mathbb{N} \mid \varphi_n = \varphi_{i_0} \}$$

of indices of all partial computable functions $\varphi_n$ that are equal to $\varphi_{i_0}$, and by Rice’s theorem, this set is not computable. In other words, it is undecidable whether an arbitrary partial computable function is equal to some fixed partial computable function.

The scenario to apply Rice’s theorem to a class $C$ of partial functions is to show that some partial computable function belongs to $C$ ($C$ is not empty), and that some partial computable function does not belong to $C$ ($C$ is not all the partial computable functions). This demonstrates that $C$ is nontrivial.

In particular, the following properties are undecidable.

**Proposition 3.7.** The following properties of partial computable functions are undecidable.

(a) A partial computable function is a constant function.

(b) Given any integer $y \in \mathbb{N}$, is $y$ in the range of some partial computable function.

(c) Two partial computable functions $\varphi_x$ and $\varphi_y$ are identical. More precisely, the set

$$\{ \langle x, y \rangle \mid \varphi_x = \varphi_y \}$$

is not computable.

(d) A partial computable function $\varphi_x$ is equal to a given partial computable function $\varphi_a$. 
(e) A partial computable function yields output \( z \) on input \( y \), for any given \( y, z \in \mathbb{N} \).

(f) A partial computable function diverges for some input.

(g) A partial computable function diverges for all input.

The above proposition is left as an easy exercise. For example, in (a), we need to exhibit a constant (partial) computable function, such as \( \text{zero}(n) = 0 \), and a nonconstant (partial) computable function, such as the identity function (or \( \text{succ}(n) = n + 1 \)).

A property may be undecidable although it is partially decidable. By partially decidable, we mean that there exists a computable function \( g \) that enumerates the set \( P_C = \{ x \mid \varphi_x \in C \} \). This means that there is a computable function \( g \) whose range is \( P_C \). We say that \( P_C \) is listable, or computably enumerable, or recursively enumerable. Indeed, \( g \) provides a recursive enumeration of \( P_C \), with possible repetitions. Listable sets are the object of the next section.

### 3.4 Listable (Recursively Enumerable) Sets

In this section and the next our focus is on subsets of \( \mathbb{N} \) rather than on numerical functions. Consider the set

\[
A = \{ k \in \mathbb{N} \mid \varphi_k(a) \text{ is defined} \},
\]

where \( a \in \mathbb{N} \) is any fixed natural number. By Rice’s theorem, \( A \) is not computable (not recursive); check this. We claim that \( A \) is the range of a computable function \( g \). For this, we use the \( T \)-predicate introduced in Definition 2.13. Recall that the predicate \( T(i, y, z) \) is defined as follows:

\[
T(i, y, z) \text{ holds iff } i \text{ codes a RAM program, } y \text{ is an input, and } z \text{ codes a halting computation of program } P_i \text{ on input } y.
\]

We produce a function which is actually primitive recursive. First, note that \( A \) is nonempty (why?), and let \( x_0 \) be any index in \( A \). We define \( g \) by primitive recursion as follows:

\[
g(0) = x_0,
\]

\[
g(x + 1) = \begin{cases} 
\Pi_1(x) & \text{if } T(\Pi_1(x), a, \Pi_2(x)), \\
 x_0 & \text{otherwise}.
\end{cases}
\]

Since this type of argument is new, it is helpful to explain informally what \( g \) does. For every input \( x \), the function \( g \) tries finitely many steps of a computation on input \( a \) for some partial computable function \( \varphi_i \) computed by the RAM program \( P_i \). Since we need to consider all pairs \((i, z)\) but we only have one variable \( x \) at our disposal, we use the trick of packing \( i \) and \( z \) into \( x = (i, z) \). Then the index \( i \) of the partial function is given by \( i = \Pi_1(x) \) and the guess for the code of the computation is given by \( z = \Pi_2(x) \). Since \( \Pi_1 \) and \( \Pi_2 \) are projection functions, when \( x \) ranges over \( \mathbb{N} \), both \( i = \Pi_1(x) \) and \( z = \Pi_2(x) \) also range over \( \mathbb{N} \). Thus
every partial function $\varphi_i$ and every code for a computation $z$ will be tried, and whenever $\varphi_i(a)$ is defined, which means that there is a correct guess for the code $z$ of the halting computation of $P_i$ on input $a$, $T(\Pi_1(x), a, \Pi_2(x)) = T(i, a, z)$ is true, and $g(x + 1)$ returns $i$.

Such a process is called a *dovetailing* computation. This type of argument will be used over and over again.

**Definition 3.5.** A subset $X$ of $\mathbb{N}$ is listable, or computably enumerable, or recursively enumerable$^2$ iff either $X = \emptyset$, or $X$ is the range of some total computable function (total recursive function). Similarly, a subset $X$ of $\Sigma^*$ is listable or computably enumerable, or recursively enumerable iff either $X = \emptyset$, or $X$ is the range of some total computable function (total recursive function).

We will often abbreviate computably enumerable as c.e, (and recursively enumerable as r.e.). A computably enumerable set is sometimes called a partially decidable or semidecidable set.

**Remark:** It should be noted that the definition of a listable set (c.e set or r.e. set) given in Definition 3.5 is different from an earlier definition given in terms of acceptance by a Turing machine and it is by no means obvious that these two definitions are equivalent. This equivalence will be proven in Proposition 3.9 ((1) $\iff$ (4)).

The following proposition relates computable sets and listable sets (recursive sets and recursively enumerable sets).

**Proposition 3.8.** A set $A$ is computable (recursive) iff both $A$ and its complement $\overline{A}$ are listable (computably enumerable, recursively enumerable).

**Proof.** Assume that $A$ is computable. Then it is trivial that its complement is also computable. Hence, we only have to show that a computable set is listable. The empty set is listable by definition. Otherwise, let $y \in A$ be any element. Then the function $f$ defined such that

$$f(x) = \begin{cases} x & \text{if } C_A(x) = 1, \\ y & \text{if } C_A(x) = 0, \end{cases}$$

for all $x \in \mathbb{N}$ is computable and has range $A$.

Conversely, assume that both $A$ and $\overline{A}$ are listable. If either $A$ or $\overline{A}$ is empty, then $A$ is computable. Otherwise, let $A = f(\mathbb{N})$ and $\overline{A} = g(\mathbb{N})$, for some computable functions $f$ and $g$. We define the function $C_A$ as follows:

$$C_A(x) = \begin{cases} 1 & \text{if } f(\min \{f(y) = x \lor g(y) = x\}) = x, \\ 0 & \text{otherwise}. \end{cases}$$

The function $C_A$ lists $A$ and $\overline{A}$ in parallel, waiting to see whether $x$ turns up in $A$ or in $\overline{A}$. Note that $x$ must eventually turn up either in $A$ or in $\overline{A}$, so that $C_A$ is a total computable function.$\square$

$^2$Since 1996, the term recursively enumerable has been considered old-fashioned by many researchers, and the terms listable and computably enumerable have been used instead.
Our next goal is to show that the listable (recursively enumerable) sets can be given several equivalent definitions.

**Proposition 3.9.** For any subset $A$ of $\mathbb{N}$, the following properties are equivalent:

1. $A$ is empty or $A$ is the range of a primitive recursive function (Rosser, 1936).
2. $A$ is listable (computably enumerable, recursively enumerable).
3. $A$ is the range of a partial computable function.
4. $A$ is the domain of a partial computable function.

**Proof.** The implication (1) $\Rightarrow$ (2) is trivial, since $A$ is listable iff either it is empty or it is the range of a (total) computable function.

To prove the implication (2) $\Rightarrow$ (3), it suffices to observe that the empty set is the range of the totally undefined function (computed by an infinite loop program), and that a computable function is a partial computable function.

The implication (3) $\Rightarrow$ (4) is shown as follows. Assume that $A$ is the range of $\varphi_i$. Define the function $f$ such that

$$f(x) = \min k[T(i, \Pi_1(k), \Pi_2(k)) \wedge \text{Res}(\Pi_2(k)) = x]$$

for all $x \in \mathbb{N}$. Since $A = \varphi_i(\mathbb{N})$, we have $x \in A$ iff there is some input $y \in \mathbb{N}$ and some computation coded by $z$ such that the RAM program $P_i$ on input $y$ has a halting computation coded by $z$ and produces the output $x$. Using the $T$-predicate, this is equivalent to $T(i, y, z)$ and $\text{Res}(z) = x$. Since we need to search over all pairs $(y, z)$, we pack $y$ and $z$ as $k = \langle y, z \rangle$ so that $y = \Pi_1(k)$ and $z = \Pi_2(k)$, and we search over all $k \in \mathbb{N}$. If the search succeeds, which means that $T(i, y, z)$ and $\text{Res}(z) = x$, we set $f(x) = k = \langle y, z \rangle$, so that $f$ is a function whose domain in the range of $\varphi_i$ (namely $A$). Note that the value $f(x)$ is irrelevant, but it is convenient to pick $k$. Clearly, $f$ is partial computable and has domain $A$.

The implication (4) $\Rightarrow$ (1) is shown as follows. The only nontrivial case is when $A$ is nonempty. Assume that $A$ is the domain of $\varphi_i$. Since $A \neq \emptyset$, there is some $a \in \mathbb{N}$ such that $a \in A$, which means that for some input $y$ the RAM program $P_i$ has a halting computation coded by $z$ on input $a$, so if we pack $y$ and $z$ as $k = \langle y, z \rangle$, the quantity

$$\min k[T(i, \Pi_1(k), \Pi_2(k))] = \min(y, z)[T(i, y, z)]$$

is defined. We can pick $a$ to be

$$a = \Pi_1(\min k[T(i, \Pi_1(k), \Pi_2(k))]).$$

We define the primitive recursive function $f$ as follows:

$$f(0) = a,$$

$$f(x + 1) = \begin{cases} 
\Pi_1(x) & \text{if } T(i, \Pi_1(x), \Pi_2(x)), \\
a & \text{if } \neg T(i, \Pi_1(x), \Pi_2(x)).
\end{cases}$$
Some \( y \in \mathbb{N} \) is in the domain of \( \varphi_i \) (namely \( A \)) iff the RAM program \( P_i \) has a halting computation coded by \( z \) on input \( y \) iff \( T(i, y, z) \) is true. If we pack \( y \) and \( z \) as \( x = \langle y, z \rangle \), then \( T(i, y, z) = T(i, \Pi_1(x), \Pi_2(x)) \), so if we search over all \( x = \langle y, z \rangle \) we search over all \( y \) and all \( z \). Whenever \( T(i, y, z) = T(i, \Pi_1(x), \Pi_2(x)) \) holds, we set \( f(x + 1) = y \) since \( y \in A \), and if \( T(i, y, z) = T(i, \Pi_1(x), \Pi_2(x)) \) is false, we return the default value \( a \in A \). Our search will find all \( y \) such that \( T(i, y, z) = T(i, \Pi_1(x), \Pi_2(x)) \) holds for some \( z \), which means that all \( y \in A \) will be in the range of \( f \). By construction, \( f \) only has values in \( A \). Clearly, \( f \) is primitive recursive.

More intuitive proofs of the implications (3) \( \Rightarrow \) (4) and (4) \( \Rightarrow \) (1) can be given. Assume that \( A \neq \emptyset \) and that \( A = \text{range}(g) \), where \( g \) is a partial computable function. Assume that \( g \) is computed by a RAM program \( P \). To compute \( f(x) \), we start computing the sequence \[ g(0), g(1), \ldots \]
looking for \( x \). If \( x \) turns up as say \( g(n) \), then we output \( n \). Otherwise the computation diverges. Hence, the domain of \( f \) is the range of \( g \).

Assume now that \( A \) is the domain of some partial computable function \( g \), and that \( g \) is computed by some Turing machine \( M \). Since the case where \( A = \emptyset \) is trivial, we may assume that \( A \neq \emptyset \), and let \( n_0 \in A \) be some chosen element in \( A \). We construct another Turing machine performing the following steps: On input \( n \),

(0) Do one step of the computation of \( g(0) \)

\[ \ldots \]

(\( n \)) Do \( n + 1 \) steps of the computation of \( g(0) \)

\[ \text{Do } n \text{ steps of the computation of } g(1) \]

\[ \ldots \]

\[ \text{Do } 2 \text{ steps of the computation of } g(n - 1) \]

\[ \text{Do } 1 \text{ step of the computation of } g(n) \]

During this process, whenever the computation of \( g(m) \) halts for some \( m \leq n \), we output \( m \). Otherwise, we output \( n_0 \).

In this fashion, we will enumerate the domain of \( g \), and since we have constructed a Turing machine that halts for every input, we have a total computable function.

The following proposition can easily be shown using the proof technique of Proposition 3.9.

**Proposition 3.10.** The following facts hold.

1. There is a computable function \( h \) such that

   \[ \text{range}(\varphi_x) = \text{dom}(\varphi_{h(x)}) \quad \text{for all } x \in \mathbb{N}. \]
(2) There is a computable function $k$ such that

$$\text{dom}(\varphi_x) = \text{range}(\varphi_{k(x)})$$

and $\varphi_{k(x)}$ is total computable, for all $x \in \mathbb{N}$ such that $\text{dom}(\varphi_x) \neq \emptyset$.

The proof of Proposition 3.10 is left as an exercise.

Using Proposition 3.9, we can prove that $K$ is a listable set. Indeed, we have $K = \text{dom}(f)$, where

$$f(x) = \varphi_{\text{univ}}(x, x) \quad \text{for all } x \in \mathbb{N}.$$  

The set $K_0 = \{ \langle x, y \rangle | \varphi_x(y) \text{ is defined} \}$ is also a listable set, since $K_0 = \text{dom}(g)$, where

$$g(z) = \varphi_{\text{univ}}(\Pi_1(z), \Pi_2(z)),$$

which is partial computable. It worth recording these facts in the following lemma.

**Proposition 3.11.** The sets $K$ and $K_0$ are listable (c.e., r.e.) sets that are not computable sets (not recursive).

We can now prove that there are sets that are not listable (not c.e., not r.e.).

**Proposition 3.12.** For any indexing of the partial computable functions, the complement $\overline{K}$ of the set

$$K = \{ x \in \mathbb{N} | \varphi_x(x) \text{ is defined} \}$$

is not listable (not computably enumerable, not recursively enumerable).

**Proof.** If $\overline{K}$ was listable, since $K$ is also listable, by Proposition 3.8, the set $K$ would be computable, a contradiction. \qed

The sets $\overline{K}$ and $\overline{K}_0$ are examples of sets that are not listable (not c.e., not r.e.). This shows that the listable (c.e., r.e.) sets are not closed under complementation. However, we leave it as an exercise to prove that the listable (c.e., r.e.) sets are closed under union and intersection.

We will prove later on that TOTAL is not listable (not c.e., not r.e.). This is rather unpleasant. Indeed, this means that there is no way of effectively listing all algorithms (all total computable functions). Hence, in a certain sense, the concept of partial computable function (procedure) is more natural than the concept of a (total) computable function (algorithm).

The next two propositions give other characterizations of the listable (c.e., r.e. sets) and of the computable sets (recursive sets). The proofs are left as an exercise.
Proposition 3.13. The following facts hold.

1. A set $A$ is listable (c.e., r.e.) iff either it is finite or it is the range of an injective computable function.

2. A set $A$ is listable (c.e., r.e.) if either it is empty or it is the range of a monotonic partial computable function.

3. A set $A$ is listable (c.e., r.e.) iff there is a Turing machine $M$ such that, for all $x \in \mathbb{N}$, $M$ halts on $x$ iff $x \in A$.

Proposition 3.14. A set $A$ is computable (recursive) iff either it is finite or it is the range of a strictly increasing computable function.

Another important result relating the concept of partial computable function and that of a listable (c.e., r.e.) set is given below.

Theorem 3.15. For every unary partial function $f$, the following properties are equivalent:

1. $f$ is partial computable.

2. The set 
   \[ \{ \langle x, f(x) \rangle \mid x \in \text{dom}(f) \} \]
   is listable (c.e., r.e.).

Proof. Let $g(x) = \langle x, f(x) \rangle$. Clearly, $g$ is partial computable, and

\[ \text{range}(g) = \{ \langle x, f(x) \rangle \mid x \in \text{dom}(f) \} \]

Conversely, assume that

\[ \text{range}(g) = \{ \langle x, f(x) \rangle \mid x \in \text{dom}(f) \} \]

for some computable function $g$. Then we have

\[ f(x) = \Pi_2(g(\min y[\Pi_1(g(y)) = x])) \quad \text{for all } x \in \mathbb{N}, \]

so that $f$ is partial computable. □

Using our indexing of the partial computable functions and Proposition 3.9, we obtain an indexing of the listable (c.e., r.e.) sets.

Definition 3.6. For any acceptable indexing $\varphi_0, \varphi_1, \ldots$ of the partial computable functions, we define the enumeration $W_0, W_1, \ldots$ of the listable (c.e., r.e.) sets by setting

\[ W_x = \text{dom}(\varphi_x). \]

We now describe a technique for showing that certain sets are listable (c.e., r.e.) but not computable (not recursive), or complements of listable (c.e., r.e.) sets that are not computable (not recursive), or not listable (not c.e., not r.e.), or neither listable (not c.e., not r.e.) nor the complement of a listable (c.e., r.e.) set. This technique is known as reducibility.
3.5 Reducibility and Complete Sets

We already used the notion of reducibility in the proof of Proposition 3.5 to show that TOTAL is not computable (not recursive).

**Definition 3.7.** Let $A$ and $B$ be subsets of $\mathbb{N}$ (or $\Sigma^*$). We say that the set $A$ is many-one reducible to the set $B$ if there is a total computable function (or total recursive function) $f: \mathbb{N} \to \mathbb{N}$ (or $f: \Sigma^* \to \Sigma^*$) such that

$$x \in A \iff f(x) \in B \quad \text{for all } x \in \mathbb{N}.$$  

We write $A \leq B$, and for short, we say that $A$ is reducible to $B$. Sometimes, the notation $A \leq_m B$ is used to stress that this is a many-to-one reduction (that is, $f$ is not necessarily injective).

Intuitively, deciding membership in $B$ is as hard as deciding membership in $A$. This is because any method for deciding membership in $B$ can be converted to a method for deciding membership in $A$ by first applying $f$ to the number (or string) to be tested.

**Remark:** Besides many-to-one reducibility, there is also a notion of one-one reducibility defined as follows: the set $A$ is one-one reducible to the set $B$ if there is a total injective computable function $f: \mathbb{N} \to \mathbb{N}$ such that

$$x \in A \iff f(x) \in B \quad \text{for all } x \in \mathbb{N}.$$  

We write $A \leq_1 B$. Obviously $A \leq_1 B$ implies $A \leq_m B$ so one-one reducibility is a stronger notion. We do not need one-one reducibility for our purposes so we will not discuss it. We refer the interested reader to Rogers [32] (especially Chapter 7) for more on reducibility.

The following simple proposition is left as an exercise to the reader.

**Proposition 3.16.** Let $A, B, C$ be subsets of $\mathbb{N}$ (or $\Sigma^*$). The following properties hold:

1. If $A \leq B$ and $B \leq C$, then $A \leq C$.
2. If $A \leq B$ then $\overline{A} \leq \overline{B}$.
3. If $A \leq B$ and $B$ is listable (c.e., r.e.), then $A$ is listable (c.e., r.e.).
4. If $A \leq B$ and $A$ is not listable (not c.e., not r.e.), then $B$ is not listable (not c.e., not r.e.).
5. If $A \leq B$ and $B$ is computable, then $A$ is computable.
6. If $A \leq B$ and $A$ is not computable, then $B$ is not computable.
Part (4) of Proposition 3.16 is often useful for proving that some set $B$ is not listable. It suffices to reduce some set known to be nonlistable to $B$, for example $\overline{K}$. Similarly, Part (6) of Proposition 3.16 is often useful for proving that some set $B$ is not computable. It suffices to reduce some set known to be noncomputable to $B$, for example $K$.

Observe that $A \leq B$ implies that $\overline{A} \leq \overline{B}$, but not that $\overline{B} \leq \overline{A}$.

Part (3) of Proposition 3.16 may be useful for proving that some set $A$ is listable. It suffices to reduce $A$ to some set known to be listable, for example $K$. Similarly, Part (5) of Proposition 3.16 may be useful for proving that some set $A$ is computable. It suffices to reduce $A$ to some set known to be computable. In practice, it is often easier to prove directly that $A$ is computable by showing that both $A$ and $\overline{A}$ are listable.

Another important concept is the concept of a complete set.

**Definition 3.8.** A listable (c.e., r.e.) set $A$ is complete w.r.t. many-one reducibility iff every listable (c.e., r.e.) set $B$ is reducible to $A$, i.e., $B \leq A$.

For simplicity, we will often say complete for complete w.r.t. many-one reducibility. Intuitively, a complete listable (c.e., r.e.) set is a “hardest” listable (c.e., r.e.) set as far as membership is concerned.

**Theorem 3.17.** The following properties hold:

1. If $A$ is complete, $B$ is listable (c.e, r.e.), and $A \leq B$, then $B$ is complete.
2. $K_0$ is complete.
3. $K_0$ is reducible to $K$. Consequently, $K$ is also complete.

**Proof.**

1. This is left as a simple exercise.
2. Let $W_x$ be any listable set (recall Definition 3.6). Then
   \[ y \in W_x \iff \langle x, y \rangle \in K_0, \]
   and the reduction function is the computable function $f$ such that
   \[ f(y) = \langle x, y \rangle \text{ for all } y \in \mathbb{N}. \]
3. We use the s-m-n Theorem. First, we leave it as an exercise to prove that there is a computable function $f$ such that
   \[ \varphi_f(x)(y) = \begin{cases} 1 & \text{if } \varphi_{\Pi_1(x)}(\Pi_2(x)) \text{ is defined,} \\
                    \text{undefined} & \text{otherwise,} \end{cases} \]
   for all $x, y \in \mathbb{N}$. Then for every $z \in \mathbb{N}$,
   \[ z \in K_0 \iff \varphi_{\Pi_1(z)}(\Pi_2(z)) \text{ is defined,} \]
iff \( \varphi_{f(z)}(y) = 1 \) for all \( y \in \mathbb{N} \). However,

\[
\varphi_{f(z)}(y) = 1 \quad \text{iff} \quad \varphi_{f(z)}(f(z)) = 1,
\]

since \( \varphi_{f(z)} \) is a constant function. This means that

\[
\begin{align*}
z & \in K_0 \quad \text{iff} \quad f(z) \in K,
\end{align*}
\]

and \( f \) is the desired function. \( \square \)

As a corollary of Theorem 3.17, the set \( K \) is also complete.

**Definition 3.9.** Two sets \( A \) and \( B \) have the same *degree of unsolvability* or are *equivalent* iff \( A \leq B \) and \( B \leq A \).

Since \( K \) and \( K_0 \) are both complete, they have the same degree of unsolvability in the set of listable sets.

We will now investigate the reducibility and equivalence of various sets.

Recall that

\[
\text{TOTAL} = \{x \in \mathbb{N} \mid \varphi_x \text{ is total}\}.
\]

We define EMPTY and FINITE, as follows:

\[
\begin{align*}
\text{EMPTY} & = \{x \in \mathbb{N} \mid \varphi_x \text{ is undefined for all input}\}, \\
\text{FINITE} & = \{x \in \mathbb{N} \mid \varphi_x \text{ is defined only for finitely many input}\}.
\end{align*}
\]

Obviously, EMPTY \( \subset \) FINITE, and since

\[
\text{FINITE} = \{x \in \mathbb{N} \mid \varphi_x \text{ has a finite domain}\},
\]

we have

\[
\text{FINITE} = \{x \in \mathbb{N} \mid \varphi_x \text{ has an infinite domain}\},
\]

and thus, TOTAL \( \subset \) FINITE. Since

\[
\text{EMPTY} = \{x \in \mathbb{N} \mid \varphi_x \text{ is undefined for all input}\}
\]

we have

\[
\text{EMPTY} = \{x \in \mathbb{N} \mid \varphi_x \text{ has a finite domain}\},
\]

and \( \text{FINITE} \subseteq \text{EMPTY} \).

**Proposition 3.18.** We have \( K_0 \leq \text{EMPTY} \).

The proof of Proposition 3.18 follows from the proof of Theorem 3.17. We also have the following proposition.
Proposition 3.19. The following properties hold:

(1) **EMPTY** is not listable (not c.e., not r.e.).

(2) **EMPTY** is listable (c.e., r.e.).

(3) \(\overline{K}\) and **EMPTY** are equivalent.

(4) **EMPTY** is complete.

Proof. We prove (1) and (3), leaving (2) and (4) as an exercise (Actually, (2) and (4) follow easily from (3)). First, we show that \(\overline{K} \leq \text{EMPTY}\). By the s-m-n Theorem, there exists a computable function \(f\) such that

\[
\varphi_{f(x)}(y) = \begin{cases} 
\varphi_x(x) & \text{if } \varphi_x(x) \text{ is defined,} \\
\text{undefined} & \text{if } \varphi_x(x) \text{ is undefined,} 
\end{cases}
\]

for all \(x, y \in \mathbb{N}\). Note that for all \(x \in \mathbb{N}\),

\(x \in \overline{K} \iff f(x) \in \text{EMPTY}\),

and thus, \(\overline{K} \leq \text{EMPTY}\). Since \(\overline{K}\) is not listable, **EMPTY** is not listable.

We now prove (3). By the s-m-n Theorem, there is a computable function \(g\) such that

\[
\varphi_{g(x)}(y) = \min z[T(x,\Pi_1(z),\Pi_2(z))], \quad \text{for all } x, y \in \mathbb{N}.
\]

Note that

\(x \in \text{EMPTY} \iff g(x) \in \overline{K}\) for all \(x \in \mathbb{N}\).

Therefore, \(\text{EMPTY} \leq \overline{K}\), and since we just showed that \(\overline{K} \leq \text{EMPTY}\), the sets \(\overline{K}\) and **EMPTY** are equivalent. \(\square\)

Proposition 3.20. The following properties hold:

(1) **TOTAL** and **TOTAL** are not listable (not c.e., not r.e.).

(2) **FINITE** and **FINITE** are not listable (not c.e, not r.e.).

Proof. Checking the proof of Theorem 3.17, we note that \(K_0 \leq \text{TOTAL}\) and \(K_0 \leq \text{FINITE}\). Hence, we get \(\overline{K_0} \leq \text{TOTAL}\) and \(\overline{K_0} \leq \text{FINITE}\), and neither \(\text{TOTAL}\) nor \(\text{FINITE}\) is listable. If **TOTAL** was listable, then there would be a computable function \(f\) such that **TOTAL** = \(\text{range}(f)\). Define \(g\) as follows:

\[g(x) = \varphi_{f(x)}(x) + 1 = \varphi_{\text{univ}}(f(x), x) + 1\]

for all \(x \in \mathbb{N}\). Since \(f\) is total and \(\varphi_{f(x)}\) is total for all \(x \in \mathbb{N}\), the function \(g\) is total computable. Let \(e\) be an index such that

\[g = \varphi_{f(e)}\].
Since $g$ is total, $g(e)$ is defined. Then we have
\[ g(e) = \varphi_{f(e)}(e) + 1 = g(e) + 1, \]
a contradiction. Hence, TOTAL is not listable. Finally, we show that $\text{TOTAL} \leq \text{FINITE}$. This also shows that FINITE is not listable. By the s-m-n Theorem, there is a computable function $f$ such that
\[ \varphi_{f(x)}(y) = \begin{cases} 1 & \text{if } \forall z \leq y(\varphi_x(z) \downarrow), \\ \text{undefined} & \text{otherwise,} \end{cases} \]
for all $x, y \in \mathbb{N}$. It is easily seen that
\[ x \in \text{TOTAL} \iff f(x) \in \text{FINITE} \quad \text{for all } x \in \mathbb{N}. \]

From Proposition 3.20, we have $\text{TOTAL} \leq \text{FINITE}$. It turns out that $\text{FINITE} \leq \text{TOTAL}$, and TOTAL and FINITE are equivalent.

**Proposition 3.21.** The sets TOTAL and FINITE are equivalent.

**Proof.** We show that $\text{FINITE} \leq \text{TOTAL}$. By the s-m-n Theorem, there is a computable function $f$ such that
\[ \varphi_{f(x)}(y) = \begin{cases} 1 & \text{if } \exists z \geq y(\varphi_x(z) \downarrow), \\ \text{undefined} & \text{if } \forall z \geq y(\varphi_x(z) \uparrow), \end{cases} \]
for all $x, y \in \mathbb{N}$. It is easily seen that
\[ x \in \text{FINITE} \iff f(x) \in \text{TOTAL} \quad \text{for all } x \in \mathbb{N}. \]

More advanced topics such that the recursion theorem, the extended Rice Theorem, and creative and productive sets will be discussed in Chapter 5.
Chapter 4

The Lambda-Calculus

The original motivation of Alonzo Church for inventing the λ-calculus was to provide a type-free foundation for mathematics (alternate to set theory) based on higher-order logic and the notion of function in the early 1930’s (1932,1933) This attempt to provide such a foundation for mathematics failed due to a form of Russell’s paradox. Church was clever enough to turn the technical reason for this failure, the existence of fixed-point combinators, into a success, namely to view the λ-calculus as a formalism for defining the notion of computability (1932,1933,1935). The λ-calculus is indeed one of the first computation models, slightly preceding the Turing machine.

Kleene proved in 1936 that all the computable functions (recursive functions) in the sense of Herbrand and Gödel are definable in the λ-calculus, showing that the λ-calculus has universal computing power. In 1937, Turing proved that Turing machines compute the same class of computable functions. (This paper is very hard to read, in part because the definition of a Turing machine is not included in this paper). In short, the λ-calculus and Turing machines have the same computing power. Here we have to be careful. To be precise we should have said that all the total computable functions (total recursive functions) are definable in the λ-calculus. In fact, it is also true that all the partial computable functions (partial recursive functions) are definable in the λ-calculus but this requires more care.

Since the λ-calculus does not have any notion of tape, register, or any other means of storing data, it quite amazing that the λ-calculus has so much computing power.

The λ-calculus is based on three concepts:

1. Application.
2. Abstraction (also called λ-abstraction).
3. β-reduction (and β-conversion).

If $f$ is a function, say the exponential function $f : \mathbb{N} \to \mathbb{N}$ given by $f(n) = 2^n$, and if $n$ a natural number, then the result of applying $f$ to a natural number, say 5, is written as

$$(f5)$$
and is called an application. Here we can agree that \( f \) and \( 5 \) do not have the same type, in the sense that \( f \) is a function and \( 5 \) is a number, so applications such as \((ff)\) or \((55)\) do not make sense, but the \( \lambda \)-calculus is type-free so expressions such as \((ff)\) as allowed. This may seem silly, and even possibly undesirable, but allowing self application turns out to a major reason for the computing power of the \( \lambda \)-calculus.

Given an expression \( M \) containing a variable \( x \), say

\[
M(x) = x^2 + x + 1,
\]

as \( x \) ranges over \( \mathbb{N} \), we obtain the function represented in standard mathematical notation by \( x \mapsto x^2 + x + 1 \). If we supply the input value 5 for \( x \), then the value of the function is \( 5^2 + 5 + 1 = 31 \). Church introduced the notation

\[
\lambda x. (x^2 + x + 1)
\]

for this function. Here, we have an abstraction, in the sense that the static expression \( M(x) \) for \( x \) fixed becomes an “abstract” function denoted \( \lambda x. M \).

It would be pointless to only have the two concepts of application and abstraction. The glue between these two notions is a form of evaluation called \( \beta \)-reduction.\(^1\) Given a \( \lambda \)-abstraction \( \lambda x. M \) and some other term \( N \) (thought of as an argument), we have the “evaluation” rule, we say \( \beta \)-reduction,

\[
(\lambda x. M)N \rightarrow^+_\beta M[x := N],
\]

where \( M[x := N] \) denotes the result of substituting \( N \) for all occurrences of \( x \) in \( M \). For example, if \( M = \lambda x. (x^2 + x + 1) \) and \( N = 2y + 1 \), we have

\[
(\lambda x. (x^2 + x + 1))(2y + 1) \rightarrow^+_\beta (2y + 1)^2 + 2y + 1 + 1.
\]

Observe that \( \beta \)-reduction is a purely formal operation (plugging \( N \) wherever \( x \) occurs in \( M \)), and that the expression \((2y+1)^2+2y+1+1\) is not instantly simplified to \(4y^2+6y+3\). In the \( \lambda \)-calculus, the natural numbers as well as the arithmetic operations + and \( \times \) need to be represented as \( \lambda \)-terms in such a way that they “evaluate” correctly using only \( \beta \)-conversion. In this sense, the \( \lambda \)-calculus is an incredibly low-level programming language. Nevertheless, the \( \lambda \)-calculus is the core of various functional programming languages such as OCaml, ML, Miranda and Haskell, among others.

We now proceed with precise definitions and results. But first we ask the reader not to think of functions as the functions we encounter in analysis or algebra. Instead think of functions as rules for computing (by moving and plugging arguments around), a more combinatory (which does not mean combinatorial) viewpoint.

This chapter relies heavily on the masterly expositions by Barendregt [3, 4]. We also found inspiration from very informative online material by Henk Barendregt, Peter Selinger, and J.R.B. Cockett, whom we thank. Hindley and Seldin [19] and Krivine [22] are also excellent sources (and not as advanced as Barendregt [3]).

\(^1\)Apparently, Church was fond of Greek letters.
4.1 Syntax of the Lambda-Calculus

We begin by defining the lambda-calculus, also called untyped lambda-calculus or pure lambda-calculus, to emphasize that the terms of this calculus are not typed. This formal system consists of

1. A set of terms, called $\lambda$-terms.

2. A notion of reduction, called $\beta$-reduction, which allows a term $M$ to be transformed into another term $N$ in a way that mimics a kind of evaluation.

First we define (pure) $\lambda$-terms. We have a countable set of variables $\{x_0, x_1, \ldots, x_n \ldots\}$ that correspond to the atomic $\lambda$-terms.

**Definition 4.1.** The $\lambda$-terms $M$ are defined inductively as follows:

1. If $x_i$ is a variable, then $x_i$ is a $\lambda$-term.

2. If $M$ and $N$ are $\lambda$-terms, then $(MN)$ is a $\lambda$-term called an application.

3. If $M$ is a $\lambda$-term, and $x$ is a variable, then the expression $(\lambda x. M)$ is a $\lambda$-term called a $\lambda$-abstraction.

Note that the only difference between the $\lambda$-terms of Definition 4.1 and the raw simply-typed $\lambda$-terms of Definition ?? is that in Clause (3), in a $\lambda$-abstraction term $(\lambda x. M)$, the variable $x$ occurs without any type information, whereas in a simply-typed $\lambda$-abstraction term $(\lambda x: \sigma. M)$, the variable $x$ is assigned the type $\sigma$. At this stage this is only a cosmetic difference because raw $\lambda$-terms are not yet assigned types. But there are type-checking rules for assigning types to raw simply-typed $\lambda$-terms that restrict application, so the set of simply-typed $\lambda$-terms that type-check is much more restricted than the set of (untyped) $\lambda$-terms. In particular, no simply-typed $\lambda$-term that type-checks can be a self-application $(MM)$. The fact that self-application is allowed in the untyped $\lambda$-calculus is what gives it its computational power (through fixed-point combinators, see Section 4.5).

**Definition 4.2.** The depth $d(M)$ of a $\lambda$-term $M$ is defined inductively as follows.

1. If $M$ is a variable $x$, then $d(x) = 0$.

2. If $M$ is an application $(M_1M_2)$, then $d(M) = \max\{d(M_1), d(M_2)\} + 1$.

3. If $M$ is a $\lambda$-abstraction $(\lambda x. M_1)$, then $d(M) = d(M_1) + 1$.

It is pretty clear that $\lambda$-terms have representations as (ordered) labeled trees.

**Definition 4.3.** Given a $\lambda$-term $M$, the tree $\text{tree}(M)$ representing $M$ is defined inductively as follows:

1. If $M$ is a variable $x$, then $\text{tree}(M)$ is the one-node tree labeled $x$. 
2. If $M$ is an application ($M_1 M_2$), then $\text{tree}(M)$ is the tree with a binary root node labeled $\lambda x$, and with a left subtree $\text{tree}(M_1)$ and a right subtree $\text{tree}(M_2)$.

3. If $M$ is a $\lambda$-abstraction $\lambda x. M_1$, then $\text{tree}(M)$ is the tree with a unary root node labeled $\lambda x$, and with one subtree $\text{tree}(M_1)$.

Definition 4.3 is illustrated in Figure 4.1.

![Figure 4.1: The tree $\text{tree}(M)$ associated with a pure $\lambda$-term $M$.](image)

Obviously, the depth $d(M)$ of $\lambda$-term is the depth of its tree representation $\text{tree}(M)$. Unfortunately $\lambda$-terms contain a profusion of parentheses so some conventions are commonly used:

1. A term of the form

   $$(\cdots ((FM_1)M_2)\cdots M_n)$$

   is abbreviated (association to the left) as

   $$FM_1\cdots M_n.$$  

2. A term of the form

   $$\lambda x_1. (\lambda x_2. (\cdots (\lambda x_n. M)\cdots))$$

   is abbreviated (association to the right) as

   $$\lambda x_1\cdots x_n. M.$$
4.1. SYNTAX OF THE LAMBDA-CALCULUS

Matching parentheses may be dropped or added for convenience. Here are some examples of \( \lambda \)-terms (and their abbreviation):

\[
\begin{align*}
\lambda x. (yx) & \quad \lambda x. yx \\
((\lambda x. (yx)) z) & \quad (\lambda x. yx) z \\
(((\lambda x. (\lambda y. (yx))) z) w) & \quad (\lambda x. yx) z w.
\end{align*}
\]

Note that \( \lambda x. yx \) is an abbreviation for \((\lambda x. (yx))\), not \((\lambda x. y) x\).

The variables occurring in a \( \lambda \)-term are free of bound.

**Definition 4.4.** For any \( \lambda \)-term \( M \), the set \( FV(M) \) of free variables of \( M \) and the set \( BV(M) \) of bound variables in \( M \) are defined inductively as follows:

1. If \( M = x \) (a variable), then
   \[ FV(x) = \{ x \}, \quad BV(x) = \emptyset. \]

2. If \( M = (M_1 M_2) \), then
   \[ FV(M) = FV(M_1) \cup FV(M_2), \quad BV(M) = BV(M_1) \cup BV(M_2). \]

3. If \( M = (\lambda x. M_1) \), then
   \[ FV(M) = FV(M_1) - \{ x \}, \quad BV(M) = BV(M_1) \cup \{ x \}. \]

If \( x \in FV(M_1) \), we say that the occurrences of the variable \( x \) occur in the scope of \( \lambda \).

A \( \lambda \)-term \( M \) is closed or a combinator if \( FV(M) = \emptyset \), that is, if it has no free variables.

For example

\[ FV((\lambda x. yx) z) = \{ y, z \}, \quad BV((\lambda x. yx) z) = \{ x \}, \]

and

\[ FV((\lambda xy. yx) zw) = \{ z \}, \quad BV((\lambda xy. yx) zw) = \{ x, y \}. \]

Before proceeding with the notion of substitution we must address an issue with bound variables. The point is that bound variables are really place-holders so they can be renamed freely without changing the reduction behavior of the term as long as they do not clash with free variables. For example, the terms \( \lambda x. x(\lambda y. (yx)) \) and \( \lambda x. x(\lambda z. x(zx)) \) should be considered as equivalent. Similarly, the terms \( \lambda x. x(\lambda y. (yx)) \) and \( \lambda w. w(\lambda z. w(zw)) \) should be considered as equivalent.
One way to deal with this issue is to use the tree representation of λ-terms given in Definition 4.3. For every leaf labeled with a bound variable \( x \), we draw a backpointer to an ancestor of \( x \) determined as follows. Given a leaf labeled with a bound variable \( x \), climb up to the closest ancestor labeled \( \lambda x \), and draw a backpointer to this node. Then all bound variables can be erased. An example is shown in Figure 4.2 for the term \( M = \lambda x. x(\lambda y. x(yx)) \).

A clever implementation of the idea of backpointers is the formalism of de Bruijn indices; see Pierce [28] (Chapter 6) or Barendregt [3] (Appendix C).

Church introduced the notion of \( \alpha \)-conversion to deal with this issue. First we need to define substitutions.

A substitution \( \varphi \) is a finite set of pairs \( \varphi = \{(x_1, N_1), \ldots, (x_n, N_n)\} \), where the \( x_i \) are distinct variables and the \( N_i \) are λ-terms. We write

\[
\varphi = [N_1/x_1, \ldots, N_n/x_n] \quad \text{or} \quad \varphi = [x_1 := N_1, \ldots, x_n := N_n].
\]

The second notation indicates more clearly that each term \( N_i \) is substituted for the variable \( x_i \), and it seems to have been almost universally adopted.

Given a substitution \( \varphi = [x_1 := N_1, \ldots, x_n := N_n] \), for any variable \( x_i \), we denote by \( \varphi_{-x_i} \) the new substitution where the pair \( (x_i, N_i) \) is replaced by the pair \( (x_i, x_i) \) (that is, the new substitution leaves \( x_i \) unchanged).

**Definition 4.5.** Given any λ-term \( M \) and any substitution \( \varphi = [x_1 := N_1, \ldots, x_n := N_n] \), we define the λ-term \( M[\varphi] \), the result of applying the substitution \( \varphi \) to \( M \), as follows:
4.1. SYNTAX OF THE LAMBDA-CALCULUS

(1) If $M = y$, with $y \neq x_i$ for $i = 1, \ldots, n$, then $M[\varphi] = y = M$.

(2) If $M = x_i$ for some $i \in \{1, \ldots, n\}$, then $M[\varphi] = N_i$.

(3) If $M = (PQ)$, then $M[\varphi] = (P[\varphi]Q[\varphi])$.

(4) If $M = \lambda x. N$ and $x \neq x_i$ for $i = 1, \ldots, n$, then $M[\varphi] = \lambda x. N[\varphi]$.

(5) If $M = \lambda x. N$ and $x = x_i$ for some $i \in \{1, \ldots, n\}$, then

$M[\varphi] = \lambda x. N[\varphi]_x = x_i$.

The term $M$ is safe for the substitution $\varphi = [x_1 := N_1, \ldots, x_n := N_n]$ if $BV(M) \cap (FV(N_1) \cup \cdots \cup FV(N_n)) = \emptyset$, that is, if the free variables in the substitution do not become bound.

Note that Clause (5) ensures that a substitution only substitutes the terms $N_i$ for the variables $x_i$ free in $M$. Thus if $M$ is a closed term, then for every substitution $\varphi$, we have $M[\varphi] = M$.

There is a problem with the present definition of a substitution in Cases (4) and (5), which is that the result of substituting a term $N_i$ containing the variable $x$ free causes this variable to become bound after the substitution. We say that $\varphi$ is captured. We should only apply a substitution $\varphi$ to a term $M$ if $M$ is safe for $\varphi$. To remedy this problem, Church defined $\alpha$-conversion.

**Definition 4.6.** The binary relation $\rightarrow_\alpha$ on $\lambda$-terms called immediate $\alpha$-conversion$^2$ is the smallest relation satisfying the following properties: for all $\lambda$-terms $M, N, P, Q$:

$\lambda x. M \rightarrow_\alpha \lambda y. M[x := y]$, for all $y \notin FV(M) \cup BV(M)$

if $M \rightarrow_\alpha N$ then $MQ \rightarrow_\alpha NQ$ and $PM \rightarrow_\alpha PN$

if $M \rightarrow_\alpha N$ then $\lambda x. M \rightarrow_\alpha \lambda x. N$.

The least equivalence relation $\equiv_\alpha = (\rightarrow_\alpha \cup \rightarrow_\alpha^{-1})^*$ containing $\rightarrow_\alpha$ (the reflexive and transitive closure of $\rightarrow_\alpha \cup \rightarrow_\alpha^{-1}$) is called $\alpha$-conversion. Here $\rightarrow_\alpha^{-1}$ denotes the converse of the relation $\rightarrow_\alpha$, that is, $M \rightarrow_\alpha^{-1} N$ iff $N \rightarrow_\alpha M$.

For example,

$\lambda f. x. f(f(x)) = \lambda f. \lambda x. f(f(x)) \rightarrow_\alpha \lambda f. \lambda y. f(f(y)) \rightarrow_\alpha \lambda g. \lambda y. g(g(y)) = \lambda g y. g(g(y))$.

Now given a $\lambda$-term $M$ and a substitution $\varphi = [x_1 := N_1, \ldots, x_n := N_n]$, before applying $\varphi$ to $M$ we first perform some $\alpha$-conversion to obtain a term $M' \equiv_\alpha M$ whose set of bound variables $BV(M')$ is disjoint from $FV(N_1) \cup \cdots \cup FV(N_n)$ so that $M'$ is safe for $\varphi$, and the result of the substitution is $M'[, \varphi]$. For example,

$(\lambda x y z. (x y) z) (y z) \equiv_\alpha (\lambda x u v. (x u) v) (y z) \rightarrow_\beta (\lambda u v. (x u) v)[x := y z] = \lambda u v. ((y z) u) v$.

$^2$We told you that Church was fond of Greek letters.
CHAPTER 4. THE LAMBDA-CALCULUS

From now on, we consider two λ-terms \( M \) and \( M' \) such that \( M \equiv \alpha M' \) as identical (to be rigorous, we deal with equivalence classes of terms with respect to \( \alpha \)-conversion). Even the experts are lax about \( \alpha \)-conversion so we happily go along with them. The convention is that bound variables are always renamed to avoid clashes (with free or bound variables).

Note that the representation of λ-terms as trees with back-pointers also ensures that substitutions are safe. However, this requires some extra effort. No matter what, it takes some effort to deal properly with bound variables.

4.2 \( \beta \)-Reduction and \( \beta \)-Conversion; the Church–Rosser Theorem

The computational engine of the λ-calculus is \( \beta \)-reduction.

**Definition 4.7.** The relation \( \rightarrow \beta \), called immediate \( \beta \)-reduction, is the smallest relation satisfying the following properties for all λ-terms \( M, N, P, Q \):

\[
(\lambda x. M)N \rightarrow \beta M[x := N], \quad \text{where } M \text{ is safe for } [x := N]
\]

if \( M \rightarrow \beta N \) then \( MQ \rightarrow \beta NQ \) and \( PM \rightarrow \beta PN \)

if \( M \rightarrow \beta N \) then \( \lambda x. M \rightarrow \beta \lambda x. N \).

The transitive closure of \( \rightarrow \beta \) is denoted by \( \rightarrow^+ \beta \), the reflexive and transitive closure of \( \rightarrow \beta \) is denoted by \( \rightarrow^* \beta \), and we define \( \beta \)-conversion, denoted by \( \leftrightarrow^* \beta \), as the smallest equivalence relation \( \leftrightarrow^* \beta = (\rightarrow \beta \cup \rightarrow^{-1} \beta)^* \) containing \( \rightarrow \beta \). A subterm of the form \( (\lambda x. M)N \) occurring in another term is called a \( \beta \)-redex. A λ-term \( M \) is a \( \beta \)-normal form if there is no λ-term \( N \) such that \( M \rightarrow \beta N \), equivalently if \( M \) contains no \( \beta \)-redex.

For example,

\[
(\lambda xy. x)uv = ((\lambda x. (\lambda y. x)u)v \rightarrow \beta ((\lambda y. x)[x := u])v = (\lambda y. u)v \rightarrow \beta u[y := v] = u
\]

and

\[
((\lambda xy. y)uv = ((\lambda x. (\lambda y. y)u)v \rightarrow \beta ((\lambda y. y)[x := u])v = (\lambda y. y)v \rightarrow \beta y[y := v] = v.
\]

This shows that \( \lambda xy. x \) behaves like the projection onto the first argument and \( \lambda xy. y \) behaves like the projection onto the second. More interestingly, if we let \( \omega = \lambda x. (xx) \), then

\[
\Omega = \omega \omega = (\lambda x. (xx))(\lambda x. (xx)) \rightarrow \beta (xx)[x := \lambda x. (xx)] = \omega \omega = \Omega.
\]

The above example shows that \( \beta \)-reduction sequences may be infinite. This is a curse and a miracle of the λ-calculus!
4.2. \( \beta \)-REDUCTION AND \( \beta \)-CONVERSION; THE CHURCH–ROSSER THEOREM

There are even \( \beta \)-reductions where the evolving term grows in size:

\[
(\lambda x.xxx)(\lambda x.xxx) \rightarrow^+ \beta (\lambda x.xxx)(\lambda x.xxx)(\lambda x.xxx) \\
\rightarrow^+ \beta (\lambda x.xxx)(\lambda x.xxx)(\lambda x.xxx)(\lambda x.xxx) \\
\rightarrow^+ \beta \ldots
\]

In general, a \( \lambda \)-term contains many different \( \beta \)-redex. One then might wonder if there is any sort of relationship between any two terms \( M_1 \) and \( M_2 \) arising through two \( \beta \)-reduction sequences \( M \rightarrow^* \beta M_1 \) and \( M \rightarrow^* \beta M_2 \) starting with the same term \( M \). The answer is given by the following famous theorem.

**Theorem 4.1.** (Church–Rosser Theorem) The following two properties hold:

1. The \( \lambda \)-calculus is **confluent**: for any three \( \lambda \)-terms \( M, M_1, M_2 \), if \( M \rightarrow^* \beta M_1 \) and \( M \rightarrow^* \beta M_2 \), then there is some \( \lambda \)-term \( M_3 \) such that \( M_1 \rightarrow^+ \beta M_3 \) and \( M_2 \rightarrow^+ \beta M_3 \). See Figure 4.3.

![Figure 4.3: The confluence property](image)

2. The \( \lambda \)-calculus has the **Church–Rosser property**: for any two \( \lambda \)-terms \( M_1, M_2 \), if \( M_1 \leftrightarrow^* \beta M_2 \), then there is some \( \lambda \)-term \( M_3 \) such that \( M_1 \rightarrow^* \beta M_3 \) and \( M_2 \rightarrow^* \beta M_3 \). See Figure 4.4.

Furthermore (1) and (2) are equivalent, and if a \( \lambda \)-term \( M \) \( \beta \)-reduces to a \( \beta \)-normal form \( N \), then \( N \) is unique (up to \( \alpha \)-conversion).

**Proof.** I am not aware of any easy proof of Part (1) or Part (2) of Theorem 4.1, but the equivalence of (1) and (2) is easily shown by induction.
Assume that (2) holds. Since $\rightarrow^*_{\beta}$ is contained in $\leftrightarrow^*_{\beta}$, if $M \rightarrow^*_{\beta} M_1$ and $M \rightarrow^*_{\beta} M_2$, then $M_1 \leftrightarrow^*_{\beta} M_2$, and since (2) holds, then there is some $\lambda$-term $M_3$ such that $M_1 \rightarrow^*_{\beta} M_3$ and $M_2 \rightarrow^*_{\beta} M_3$, which is (1).

To prove that (1) implies (2) we need the following observation.

Since $\leftrightarrow^*_{\beta} = (\rightarrow_{\beta} \cup \rightarrow^{-1}_{\beta})^*$, we see immediately that $M_1 \leftrightarrow^*_{\beta} M_2$ iff either

(a) $M_1 = M_2$, or

(b) there is some $M_3$ such that $M_1 \rightarrow_{\beta} M_3$ and $M_3 \leftarrow^*_{\beta} M_2$, or

(c) there is some $M_3$ such that $M_3 \rightarrow_{\beta} M_1$ and $M_3 \leftarrow^*_{\beta} M_2$.

Assume (1). We proceed by induction on the number of steps in $M_1 \leftarrow^*_{\beta} M_2$. If $M_1 \leftarrow^*_{\beta} M_2$, as discussed before, there are three cases.

Case a. Base case, $M_1 = M_2$. Then (2) holds with $M_3 = M_1 = M_2$.

Case b. There is some $M_3$ such that $M_1 \rightarrow_{\beta} M_3$ and $M_3 \leftarrow^*_{\beta} M_2$. Since $M_3 \leftarrow^*_{\beta} M_2$ contains one less step than $M_1 \leftarrow^*_{\beta} M_2$, by the induction hypothesis there is some $M_4$ such that $M_3 \rightarrow^*_{\beta} M_4$ and $M_2 \rightarrow^*_{\beta} M_4$, and then $M_1 \rightarrow_{\beta} M_3 \rightarrow^*_{\beta} M_4$ and $M_2 \rightarrow^*_{\beta} M_4$, proving (2). See Figure 4.5.

Case c. There is some $M_3$ such that $M_3 \rightarrow_{\beta} M_1$ and $M_3 \leftarrow^*_{\beta} M_2$. Since $M_3 \leftarrow^*_{\beta} M_2$ contains one less step than $M_1 \leftarrow^*_{\beta} M_2$, by the induction hypothesis there is some $M_4$ such that $M_3 \rightarrow^*_{\beta} M_4$ and $M_2 \rightarrow^*_{\beta} M_4$. Now $M_3 \rightarrow_{\beta} M_1$ and $M_3 \rightarrow^*_{\beta} M_4$, so by (1) there is some $M_5$ such that $M_1 \rightarrow^*_{\beta} M_5$ and $M_4 \rightarrow^*_{\beta} M_5$. Putting derivations together we get $M_1 \rightarrow^*_{\beta} M_5$ and $M_2 \rightarrow^*_{\beta} M_4 \rightarrow^*_{\beta} M_5$, which proves (2). See Figure 4.6.
Suppose $M \xrightarrow{\beta} N_1$ and $M \xrightarrow{\beta} N_2$ where $N_1$ and $N_2$ are both $\beta$-normal forms. Then by confluence there is some $N$ such that $N_1 \xrightarrow{\beta} N$ and $N_2 \xrightarrow{\beta} N$. Since $N_1$ and $N_2$ are both $\beta$-normal forms, we must have $N_1 = N = N_2$ (up to $\alpha$-conversion).

Barendregt gives an elegant proof of the confluence property in [3] (Chapter 11).

Another immediate corollary of the Church–Rosser theorem is that if $M \xleftarrow{\beta} N$ and if $N$ is a $\beta$-normal form, then in fact $M \xrightarrow{\beta} N$. We leave this fact as an exercise.

This fact will be useful in showing that the recursive functions are computable in the $\lambda$-calculus.
4.3 Some Useful Combinators

In this section we provide some evidence for the expressive power of the $\lambda$-calculus.

First we make a remark about the representation of functions of several variables in the $\lambda$-calculus. The $\lambda$-calculus makes the implicit assumption that a function has a single argument. This is the idea behind application: given a term $M$ viewed as a function and an argument $N$, the term $(MN)$ represents the result of applying $M$ to the argument $N$, except that the actual evaluation is suspended. Evaluation is performed by $\beta$-conversion. To deal with functions of several arguments we use a method known as Currying (after Haskell Curry). In this method, a function of $n$ arguments is viewed as a function of one argument taking a function of $n-1$ arguments as argument. Consider the case of two arguments, the general case being similar. Consider a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. For any fixed $x$, we define the function $F_x : \mathbb{N} \to \mathbb{N}$ given by

$$F_x(y) = f(x, y) \quad y \in \mathbb{N}.$$ 

Using the $\lambda$-notation we can write

$$F_x = \lambda y. f(x, y),$$

and then the function $x \mapsto F_x$, which is a function from $\mathbb{N}$ to the set of functions $[\mathbb{N} \to \mathbb{N}]$ (also denoted $\mathbb{N}^\mathbb{N}$), is denoted by the $\lambda$-term

$$F = \lambda x. F_x = \lambda x. (\lambda y. f(x, y)).$$

And indeed,

$$(FM)N \xrightarrow{\beta} F_M N \xrightarrow{\beta} f(M, N).$$

Remark: Currying is a way to realizing the isomorphism between the sets of functions $[\mathbb{N} \times \mathbb{N} \to \mathbb{N}]$ and $[\mathbb{N} \to [\mathbb{N} \to \mathbb{N}]]$ (or in the standard set-theoretic notation, between $\mathbb{N}^{\mathbb{N} \times \mathbb{N}}$ and $(\mathbb{N}^\mathbb{N})^\mathbb{N}$). Does this remind you of the identity

$$(m^n)^p = m^{np}?$$

It should.

The function space $[\mathbb{N} \to \mathbb{N}]$ is called an exponential. There is a very abstract way to view all this which is to say that we have an instance of a Cartesian closed category (CCC).

**Proposition 4.2.** If $I, K, K_*$, and $S$ are the combinators defined by

$$I = \lambda x. x$$
$$K = \lambda xy. x$$
$$K_* = \lambda xy. y$$
$$S = \lambda xyz. (xz)(yz),$$
then for all $\lambda$-terms $M, N, P$, we have

\[
\begin{align*}
IM & \rightarrow^\beta M \\
KMN & \rightarrow^\beta M \\
KMN & \rightarrow^\beta N \\
SMNP & \rightarrow^\beta (MP)(NP) \\
KI & \rightarrow^\beta K_+ \\
SKK & \rightarrow^\beta I.
\end{align*}
\]

The proof is left as an easy exercise. For example,

\[
SMNP = (\lambda xyz. (xz)(yz))MNP \rightarrow^\beta ((\lambda yz. (xz)(yz))(x := M))NP = (\lambda yz. (Mz)(yz))NP \rightarrow^\beta ((\lambda z. (Mz)(yz))(y := N))P = (\lambda z. (Mz)(Nz))P \rightarrow^\beta ((Mz)(Nz))(z := P) = (MP)(NP).
\]

The need for a conditional construct if then else such that if $T$ then $P$ else $Q$ yields $P$ and if $F$ then $P$ else $Q$ yields $Q$ is indispensable to write nontrivial programs. There is a trick to encode the boolean values $T$ and $F$ in the $\lambda$-calculus to mimick the above behavior of if $B$ then $P$ else $Q$, provided that $B$ is a truth value. Since everything in the $\lambda$-calculus is a function, the booleans values $T$ and $F$ are encoded as $\lambda$-terms. At first, this seems quite odd, but what counts is the behavior of if $B$ then $P$ else $Q$, and it works!

The truth values $T, F$ and the conditional construct if $B$ then $P$ else $Q$ can be encoded in the $\lambda$-calculus as follows.

**Proposition 4.3.** Consider the combinators given by $T = K, F = K_+$, and

\[
\text{if then else} = \lambda b xy. bxy.
\]

Then for all $\lambda$-terms we have

\[
\begin{align*}
\text{if } T \text{ then } P \text{ else } Q & \rightarrow^\beta P \\
\text{if } F \text{ then } P \text{ else } Q & \rightarrow^\beta Q.
\end{align*}
\]

The proof is left as an easy exercise. For example,

\[
\begin{align*}
\text{if } T \text{ then } P \text{ else } Q & = \text{if then else}(TPQ) \\
& = (\lambda b xy. bxy)TPQ \\
& \rightarrow^\beta ((\lambda b xy. bxy)(b := T))PQ = (\lambda y. Txy)PQ \\
& \rightarrow^\beta ((\lambda y. Txy)(x := P))Q = (\lambda y. TPy)Q \\
& \rightarrow^\beta (TPy)(y := Q) = TPQ \\
& = KPQ \rightarrow^\beta P,
\end{align*}
\]
by Proposition 4.2.

The boolean operations \( \land, \lor, \neg \) can be defined in terms of if then else. For example,

\[
\text{And } b_1 b_2 = \text{if } b_1 \text{ then (if } b_2 \text{ then } \text{T else } F \text{) else } F.
\]

**Remark:** If \( B \) is a term different from \( \text{T} \) or \( \text{F} \), then if \( B \) then \( P \) else \( Q \) may not reduce at all, or reduce to something different from \( P \) or \( Q \). The problem is that the conditional statement that we designed only works properly if the input \( B \) is of the correct type, namely a boolean. If we give garbage as input, then we can’t expect a correct result. The \( \lambda \)-calculus being type-free, it is unable to check for the validity of the input. In this sense this is a defect, but it also accounts for its power.

The ability to construct ordered pairs is also crucial.

**Proposition 4.4.** For any two \( \lambda \)-terms \( M \) and \( N \) consider the combinator \( \langle M, N \rangle \) and the combinator \( \pi_1 \) and \( \pi_2 \) given by

\[
\langle M, N \rangle = \lambda z. z MN = \lambda z. \text{if } z \text{ then } M \text{ else } N
\]

\[
\pi_1 = \lambda z. z K
\]

\[
\pi_2 = \lambda z. z K_\ast.
\]

Then

\[
\pi_1(M, N) \to^+ \beta M
\]

\[
\pi_2(M, N) \to^+ \beta N
\]

\[
\langle M, N \rangle \text{T} \to^+ \beta M
\]

\[
\langle M, N \rangle \text{F} \to^+ \beta N.
\]

The proof is left as an easy exercise. For example,

\[
\pi_1(M, N) = (\lambda z. z K)(\lambda z. z MN)
\]

\[
\to^\beta (z K)[z := \lambda z. z MN] = (\lambda z. z MN)K
\]

\[
\to^\beta (z MN)[z := K] = K MN \to^\beta M,
\]

by Proposition 4.2.

In the next section we show how to encode the natural numbers in the \( \lambda \)-calculus and how to compute various arithmetical functions.

### 4.4 Representing the Natural Numbers

Historically the natural numbers were first represented in the \( \lambda \)-calculus by Church in the 1930’s. Later in 1976 Barendregt came up with another representation which is more convenient to show that the recursive functions are \( \lambda \)-definable. We start with Church’s representation.
4.4. REPRESENTING THE NATURAL NUMBERS

First, given any two \( \lambda \)-terms \( F \) and \( M \), for any natural number \( n \in \mathbb{N} \), we define \( F^n(M) \) inductively as follows:

\[
\begin{align*}
F^0(M) &= M \\
F^{n+1}(M) &= F(F^n(M)).
\end{align*}
\]

**Definition 4.8.** (Church Numerals) The Church numerals \( c_0, c_1, c_2, \ldots \) are defined by

\[
c_n = \lambda fx. f^n(x).
\]

So \( c_0 = \lambda fx. x = \text{K} \), \( c_1 = \lambda fx. fx \), \( c_2 = \lambda fx. f(fx) \), etc. The Church numerals are \( \beta \)-normal forms.

Observe that

\[
c_n F z = (\lambda fx. f^n(x))Fz \xrightarrow{\beta} F^n(z).
\]

This shows that \( c_n \) *iterates* \( n \) times the function represented by the term \( F \) on initial input \( z \). This is the trick behind the definition of the Church numerals. This suggests the following definition.

**Definition 4.9.** The iteration combinator \( \text{Iter} \) is given by

\[
\text{Iter} = \lambda nfx. nfx.
\]

Observe that

\[
\text{Iter} c_n F X \xrightarrow{\beta} F^n X,
\]

that is, the result of iterating \( F \) for \( n \) steps starting with the initial term \( X \).

Let us show how some basic functions on the natural numbers can be defined. We begin with the constant function \( Z \) given by \( Z(n) = 0 \) for all \( n \in \mathbb{N} \). We claim that \( Z_c = \lambda x. c_0 \) works. Indeed, we have

\[
Z_c c_n = (\lambda x. c_0)c_n \rightarrow^{\beta} c_0[x := c_n] = c_0
\]

since \( c_0 \) is a closed term.

The successor function \( \text{Succ} \) is given by

\[
\text{Succ}(n) = n + 1.
\]

We claim that

\[
\text{Succ}_c = \lambda nfx. f(nfx)
\]

computes \( \text{Succ} \). Indeed we have

\[
\begin{align*}
\text{Succ}_c c_n &= (\lambda nfx. f(nfx))c_n \\
&\rightarrow^{\beta} (\lambda fx. f(f(nfx)))[n := c_n] = \lambda fx. f(c_nfx) \\
&\rightarrow^{\beta} \lambda fx. f(f^n(x)) \\
&= \lambda fx. f^{n+1}(x) = c_{n+1}.
\end{align*}
\]
CHAPTER 4. THE LAMBDA-CALCULUS

The function \textbf{IsZero} which tests whether a natural number is equal to 0 is defined by the combinator

\[ \text{IsZero}_c = \lambda x. x(\text{K F})\text{T}. \]

The proof that it works is left as an exercise.

Addition and multiplication are a little more tricky to define.

\textbf{Proposition 4.5. (J.B. Rosser)} Define \textbf{Add} and \textbf{Mult} as the combinators given by

\begin{align*}
\text{Add} &= \lambda mnf.x.m(f(nx)) \\
\text{Mult} &= \lambda xyz.x(yz).
\end{align*}

We have

\begin{align*}
\text{Add } c_m c_n &\xrightarrow{\beta} c_{m+n} \\
\text{Mult } c_m c_n &\xrightarrow{\beta} c_{m+n}
\end{align*}

for all \( m, n \in \mathbb{N} \).

\textit{Proof.} We have

\begin{align*}
\text{Add } c_m c_n &= (\lambda mnf.x.m(f(nx)))c_m c_n \\
&\xrightarrow{\beta} (\lambda fx.c_m f (c_n f x)) \\
&\xrightarrow{\beta} \lambda fx.f^m(f^n(x)) \\
&= \lambda fx.f^{m+n}(x) = c_{m+n}.
\end{align*}

For multiplication we need to prove by induction on \( m \) that

\((c_n x)^m(y) \xrightarrow{\beta} x^{m+n}(y)\). (\*)

If \( m = 0 \) then both sides are equal to \( y \).

For the induction step we have

\begin{align*}
(c_n x)^{m+1}(y) &= c_n x((c_n x)^m(y)) \\
&\xrightarrow{\ast} c_n x(x^{m+n}(y)) \quad \text{by induction} \\
&\xrightarrow{\ast} x^n(x^{m+n}(y)) \\
&= x^{n+m+n}(y) = x^{(m+1)n}(y).
\end{align*}

We now have

\begin{align*}
\text{Mult } c_m c_n &= (\lambda xyz.x(yz))c_m c_n \\
&\xrightarrow{\beta} \lambda z.(c_m(c_n z)) \\
&= \lambda z.((\lambda fy. f^m(y))(c_n z)) \\
&\xrightarrow{\beta} \lambda zy.(c_n z)^m(y),
\end{align*}
4.4. REPRESENTING THE NATURAL NUMBERS

and since we proved in (*) that

\[(c_n z)^m(y) \rightarrow^\beta z^{m+n}(y),\]

we get

\[\text{Mult } c_mC_n \rightarrow^\beta \lambda zy. (c_n z)^m(y) \rightarrow^\beta \lambda zy. z^{m+n}(y) = c_{m+n},\]

which completes the proof. \qed

As an exercise the reader should prove that addition and multiplication can also be defined in terms of \texttt{Iter} (see Definition 4.9) by

\[
\begin{align*}
\text{Add} &= \lambda m n. \texttt{Iter} m \texttt{Succ}_c n \\
\text{Mult} &= \lambda m n. \texttt{Iter} m (\texttt{Add} n) c_0.
\end{align*}
\]

The above expressions are close matches to the primitive recursive definitions of addition and multiplication. To check that they work, prove that

\[
\text{Add } c_m c_n \rightarrow^\beta (\texttt{Succ}_c)^m(c_n) \rightarrow^\beta c_{m+n}
\]

and

\[
\text{Mult } c_m c_n \rightarrow^\beta (\texttt{Add} n)^m(c_0) \rightarrow^\beta c_{m+n}.
\]

A version of the exponential function can also be defined. A function that plays an important technical role is the predecessor function \texttt{Pred} defined such that

\[
\begin{align*}
\texttt{Pred}(0) &= 0 \\
\texttt{Pred}(n + 1) &= n.
\end{align*}
\]

It turns out that it is quite tricky to define this function in terms of the Church numerals. Church and his students struggled for a while until Kleene found a solution in his famous 1936 paper. The story goes that Kleene found his solution when he was sitting in the dentist’s chair! The trick is to make use of pairs. Kleene’s solution is

\[
\texttt{Pred}_\pi = \lambda n. \pi_2(\texttt{Iter} n \lambda z. (\texttt{Succ}_c(\pi_1 z), \pi_1 z) \langle c_0, c_0 \rangle).
\]

The reason this works is that we can prove that

\[
(\lambda z. (\texttt{Succ}_c(\pi_1 z), \pi_1 z))^0\langle c_0, c_0 \rangle \rightarrow^\beta \langle c_0, c_0 \rangle,
\]

and by induction that

\[
(\lambda z. (\texttt{Succ}_c(\pi_1 z), \pi_1 z))^{n+1}\langle c_0, c_0 \rangle \rightarrow^\beta \langle c_{n+1}, c_n \rangle.
\]

For the base case \(n = 0\) we get

\[
(\lambda z. (\texttt{Succ}_c(\pi_1 z), \pi_1 z))\langle c_0, c_0 \rangle \rightarrow^\beta \langle c_1, c_0 \rangle.
\]
For the induction step we have

\[(\lambda z. \langle \text{Succ}_c(\pi_1 z), \pi_1 z \rangle)^{n+2} (c_0, c_0) =\]

\[(\lambda z. \langle \text{Succ}_c(\pi_1 z), \pi_1 z \rangle)((\lambda z. \langle \text{Succ}_c(\pi_1 z), \pi_1 z \rangle)^{n+1} (c_0, c_0))\]

\[\xrightarrow{\beta} (\lambda z. \langle \text{Succ}_c(\pi_1 z), \pi_1 z \rangle)(c_{n+1}, c_n) \xrightarrow{\beta} (c_{n+2}, c_{n+1}).\]

Here is another tricky solution due to J. Velmans (according to H. Barendregt):

\[\text{Pred}_c = \lambda xyz. x(\lambda pq. q(py))(Kz)I.\]

We leave it to the reader to verify that it works.

The ability to construct pairs together with the \texttt{Iter} combinator allows the definition of a large class of functions, because \texttt{Iter} is “type-free” in its second and third arguments so it really allows higher-order primitive recursion.

For example, the factorial function defined such that

\[0! = 1\]

\[(n + 1)! = (n + 1)n!\]

can be defined. First we define \(h\) by

\[h = \lambda xn. \text{Mult Succ}_c n x\]

and then

\[\text{fact} = \lambda n. \pi_1 (\text{Iter} n \lambda z. \langle h(\pi_1 z) (\pi_2 z), \text{Succ}_c(\pi_2 z) \rangle (c_1, c_0)).\]

The above term works because

\[(\lambda z. \langle h(\pi_1 z) (\pi_2 z), \text{Succ}_c(\pi_2 z) \rangle)^0 (c_1, c_0) \xrightarrow{\beta} (c_1, c_0) = (c_0!, c_0),\]

and

\[(\lambda z. \langle h(\pi_1 z) (\pi_2 z), \text{Succ}_c(\pi_2 z) \rangle)^{n+1} (c_1, c_0) \xrightarrow{\beta} (c_{(n+1)}!, c_{n+1}) = (c_{(n+1)}!, c_{n+1}).\]

We leave the details as an exercise.

Barendregt came up with another way of representing the natural numbers that makes things easier.

**Definition 4.10.** (Barendregt Numerals) The Barendregt numerals \(b_n\) are defined as follows:

\[b_0 = I = \lambda x. x\]

\[b_{n+1} = \langle F, b_n \rangle.\]

The Barendregt numerals are \(\beta\)-normal forms. Barendregt uses the notation \(\uparrow n \uparrow\) instead of \(b_n\) but this notation is also used for the Church numerals by other authors so we prefer using \(b_n\) (which is consistent with the use of \(c_n\) for the Church numerals). The Barendregt numerals are tuples, which makes operating on them simpler than the Church numerals which encode \(n\) as the composition \(f^n\).
Proposition 4.6. The functions Succ, Pred and IsZero are defined in terms of the Barendregt numerals by the combinators

\[ \text{Succ}_b = \lambda x. (F, x) \]
\[ \text{Pred}_b = \lambda x. (xF) \]
\[ \text{IsZero}_b = \lambda x. (xT), \]

and we have

\[ \text{Succ}_b b_n \xrightarrow{+\beta} b_{n+1} \]
\[ \text{Pred}_b b_0 \xrightarrow{+\beta} b_0 \]
\[ \text{Pred}_b b_{n+1} \xrightarrow{+\beta} b_n \]
\[ \text{IsZero}_b b_0 \xrightarrow{+\beta} T \]
\[ \text{IsZero}_b b_{n+1} \xrightarrow{+\beta} F. \]

The proof is left as an exercise.

Since there is an obvious bijection between the Church combinators and the Barendregt combinators there should be combinators effecting the translations. Indeed we have the following result.

Proposition 4.7. The combinator T given by

\[ T = \lambda x. (x\text{Succ}_b)b_0 \]

has the property that

\[ Tc_n \xrightarrow{+\beta} b_n \quad \text{for all } n \in \mathbb{N}. \]

Proof. We proceed by induction on n. For the base case

\[ Tc_0 = (\lambda x. (x\text{Succ}_b)b_0)c_0 \]
\[ \xrightarrow{+\beta} c_0(\text{Succ}_b)b_0 \]
\[ \xrightarrow{+\beta} b_0. \]

For the induction step,

\[ Tc_n = (\lambda x. (x\text{Succ}_b)b_0)c_n \]
\[ \xrightarrow{+\beta} (c_n\text{Succ}_b)b_0 \]
\[ \xrightarrow{+\beta} \text{Succ}_b^n(b_0). \]

Thus we need to prove that

\[ \text{Succ}_b^n(b_0) \xrightarrow{+\beta} b_n. \quad (\ast) \]
For the base case \( n = 0 \), the left-hand side reduces to \( b_0 \).

For the induction step, we have

\[
\text{Succ}_b^{n+1}(b_0) = \text{Succ}_b(\text{Succ}_b^n(b_0))
\]
\[
= \xrightarrow{\beta} \text{Succ}_b(b_n) \quad \text{by induction}
\]
\[
= \xrightarrow{\beta} b_{n+1},
\]

which concludes the proof. \( \square \)

There is also a combinator defining the inverse map but it is defined recursively and we don’t know how to express recursive definitions in the \( \lambda \)-calculus. This is achieved by using fixed-point combinators.

\section{Fixed-Point Combinators and Recursively Defined Functions}

Fixed-point combinators are the key to the definability of recursive functions in the \( \lambda \)-calculus. We begin with the \( \text{Y} \)-combinator due to Curry.

\begin{proposition} \text{(Curry \text{Y}-combinator)} \end{proposition}

If we define the combinator \( \text{Y} \) as

\[
\text{Y} = \lambda f. (\lambda x. f(xx))(\lambda x. f(xx)),
\]

then for any \( \lambda \)-term \( F \) we have

\[
F(\text{Y}F) \xleftrightarrow{\beta} \text{Y}F.
\]

\begin{proof}
Write \( W = \lambda x. F(xx) \). We have

\[
F(\text{Y}F) = F((\lambda f. (\lambda x. f(xx))(\lambda x. f(xx)))F) \xrightarrow{\beta} F((\lambda x. F(xx))(\lambda x. F(xx))) = F(WW),
\]

and

\[
\text{Y}F = (\lambda f. (\lambda x. f(xx))(\lambda x. f(xx)))F \xrightarrow{\beta} (\lambda x. F(xx))(\lambda x. F(xx)) = (\lambda x. F(xx))W \xrightarrow{\beta} F(WW).
\]

Therefore \( F(\text{Y}F) \xleftrightarrow{\beta} \text{Y}F \), as claimed. \( \square \)

Observe that neither \( F(\text{Y}F) \xrightarrow{\beta} \text{Y}F \) nor \( \text{Y}F \xrightarrow{\beta} F(\text{Y}F) \). This is a slight disadvantage of the Curry \( \text{Y} \)-combinator. Turing came up with another fixed-point combinator that does not have this problem.
Proposition 4.9. (Turing $\Theta$-combinator) If we define the combinator $\Theta$ as

$$\Theta = (\lambda xy. y(xy)) (\lambda xy. y(xy)),$$

then for any $\lambda$-term $F$ we have

$$\Theta F \xrightarrow{\beta} F(\Theta F).$$

Proof. If we write $A = (\lambda xy. y(xy))$, then $\Theta = AA$. We have

$$\Theta F = (AA)F = ((\lambda xy. y(xy))A)F \xrightarrow{\beta} (\lambda y. y(AAy))F \xrightarrow{\beta} F(AAF) = F(\Theta F),$$

as claimed.

Now we show how to use the fixed-point combinators to represent recursively-defined functions in the $\lambda$-calculus. For example, there is a combinator $G$ such that

$$GX \xrightarrow{\beta} X(XG) \quad \text{for all } X.$$

Informally the idea is to consider the “functional” $F = \lambda gx. x(xg)$, and to find a fixed-point of this functional. Pick

$$G = \Theta\lambda gx. x(xg) = \Theta F.$$

Since by Proposition 4.9 we have $G = \Theta F \xrightarrow{\beta} F(\Theta F) = FG$, and we also have

$$FG = (\lambda gx. x(xg))G \xrightarrow{\beta} \lambda x. x(xG),$$

so $G \xrightarrow{+} \lambda x. x(xG)$, which implies

$$GX \xrightarrow{+} (\lambda x. x(xG))X \xrightarrow{\beta} X(XG).$$

In general, if we want to define a function $G$ recursively such that

$$GX \xrightarrow{+} M(X, G)$$

where $M(X, G)$ is $\lambda$-term containing recursive occurrences of $G$, we let $F = \lambda gx. M(x, g)$ and

$$G = \Theta F.$$

Then we have

$$G \xrightarrow{+} FG = (\lambda gx. M(x, g))G \xrightarrow{\beta} \lambda x. M(x, g)[g := G] = \lambda x. M(x, G),$$

so

$$GX \xrightarrow{+} (\lambda x. M(x, G))X \xrightarrow{\beta} M(x, G)[x := X] = M(X, G),$$
as desired.

As another example, here is how the factorial function can be defined (using the Church numerals). Let

$$ F = \lambda g n. \text{if } \text{IsZero}_c n \text{ then } c_1 \text{ else } \text{Mult} n g (\text{Pred}_c n). $$

Then the term $G = \Theta F$ defines factorial function. The verification of the above fact is left as an exercise.

As usual with recursive definitions there is no guarantee that the function that we obtain terminates for all input. For example, if we consider

$$ F = \lambda g n. \text{if } \text{IsZero}_c n \text{ then } c_1 \text{ else } \text{Mult} n g (\text{Succ}_c n) $$

then for $n \geq 1$ the reduction behavior is

$$ Gc_n \xrightarrow{\beta} \text{Mult} c_n Gc_{n+1}, $$

which does not terminate.

We leave it as an exercise to show that the inverse of the function $T$ mapping the Church numerals to the Barendregt numerals is given by the combinator

$$ T^{-1} = \Theta (\lambda f x. \text{if } \text{IsZero}_b x \text{ then } c_0 \text{ else } \text{Succ}_c (f (\text{Pred}_b x)). $$

It is remarkable that the $\lambda$-calculus allows the implementation of arbitrary recursion without a stack, just using $\lambda$-terms are the data-structure and $\beta$-reduction. This does not mean that this evaluation mechanism is efficient but this is another story (as well as evaluation strategies, which have to do with parameter-passing strategies, call-by-name, call-by-value).

Now we have all the ingredients to show that all the (total) computable functions are definable in the $\lambda$-calculus.

### 4.6 $\lambda$-Definability of the Computable Functions

Let us begin by reviewing the definition of the computable functions (recursive functions) (à la Herbrand–Gödel–Kleene). For our purposes it suffices to consider functions (partial or total) $f : \mathbb{N}^n \to \mathbb{N}$ as opposed to the more general case of functions $f : (\Sigma^*)^n \to \Sigma^*$ defined on strings.

**Definition 4.11.** The base functions are the functions $Z, S, P^n_i$ defined as follows:

1. The constant zero function $Z$ such that
   $$ Z(n) = 0, \quad \text{for all } n \in \mathbb{N}. $$

2. The successor function $S$ such that
   $$ S(n) = n + 1, \quad \text{for all } n \in \mathbb{N}. $$
(3) For every \( n \geq 1 \) and every \( i \) with \( 1 \leq i \leq n \), the projection function \( P_i^n \) such that
\[
P_i^n(x_1, \ldots, x_n) = x_i, \quad x_1, \ldots, x_n \in \mathbb{N}.
\]

Next comes (extended) composition.

**Definition 4.12.** Given any partial or total function \( g: \mathbb{N}^m \to \mathbb{N} \) \((m \geq 1)\) and any \( m \) partial or total functions \( h_i: \mathbb{N}^n \to \mathbb{N} \) \((n \geq 1)\), the composition of \( g \) and \( h_1, \ldots, h_m \), denoted \( g \circ (h_1, \ldots, h_m) \), is the partial or total function function \( f: \mathbb{N}^n \to \mathbb{N} \) given by
\[
f(x_1, \ldots, x_n) = g(h_1(x_1, \ldots, x_n), \ldots, h_m(x_1, \ldots, x_n)), \quad x_1, \ldots, x_n \in \mathbb{N}.
\]

If \( g \) or any of the \( h_i \) are partial functions, then \( f(x_1, \ldots, x_n) \) is defined if and only if all \( h_i(x_1, \ldots, x_n) \) are defined and \( g(h_1(x_1, \ldots, x_n), \ldots, h_m(x_1, \ldots, x_n)) \) is defined.

Note that even if \( g \) “ignores” one of its arguments, say the \( i \)th one, \( g(h_1(x_1, \ldots, x_n), \ldots, h_m(x_1, \ldots, x_n)) \) is undefined if \( h_i(x_1, \ldots, x_n) \) is undefined.

**Definition 4.13.** Given any partial or total functions \( g: \mathbb{N}^m \to \mathbb{N} \) and \( h: \mathbb{N}^{m+2} \to \mathbb{N} \) \((m \geq 1)\), the partial or total function function \( f: \mathbb{N}^{m+1} \to \mathbb{N} \) is defined by primitive recursion from \( g \) and \( h \) if \( f \) is given by:
\[
f(0, x_1, \ldots, x_m) = g(x_1, \ldots, x_m)
f(n+1, x_1, \ldots, x_m) = h(f(n, x_1, \ldots, x_m), n, x_1, \ldots, x_m)
\]
for all \( n, x_1, \ldots, x_m \in \mathbb{N} \). If \( m = 0 \), then \( g \) is some fixed natural number and we have
\[
f(0) = g
\]
\[
f(n+1) = h(f(n), n).
\]

It can be shown that if \( g \) and \( h \) are total functions, then so if \( f \).

Note that the second clause of the definition of primitive recursion is
\[
f(n+1, x_1, \ldots, x_m) = h(f(n, x_1, \ldots, x_m), n, x_1, \ldots, x_m)
\]
but in an earlier definition it was
\[
f(n+1, x_1, \ldots, x_m) = h(n, f(n, x_1, \ldots, x_m), x_1, \ldots, x_m), \quad (*_1)
\]
with the first two arguments of \( h \) permuted. Since
\[
h \circ (P_2^{m+2}, P_1^{m+2}, P_3^{m+2}, \ldots, P_{m+2}^{m+2})(n, f(n, x_1, \ldots, x_m), x_1, \ldots, x_m)
\]
\[
= h(f(n, x_1, \ldots, x_m), n, x_1, \ldots, x_m)
\]
and
\[
\begin{align*}
  h \circ (P^m_2, P^m_1, P^m_3, \ldots, P^m_{m+2})(f(n, x_1, \ldots, x_m), n, x_1, \ldots, x_m) \\
  = h(n, f(n, x_1, \ldots, x_m), x_1, \ldots, x_m),
\end{align*}
\]
the two definitions are equivalent. In this section we chose version \((\ast_1)\) because it matches
the treatment in Barendregt [3] and will make it easier for the reader to follow Barendregt
[3] if they wish.

The last operation is minimization (sometimes called minimalization).

**Definition 4.14.** Given any partial or total function \(g : \mathbb{N}^{m+1} \rightarrow \mathbb{N} (m \geq 0)\), the partial or
total function function \(f : \mathbb{N}^m \rightarrow \mathbb{N}\) is defined as follows: for all \(x_1, \ldots, x_m \in \mathbb{N}\),
\[
f(x_1, \ldots, x_m) = \text{the least } n \in \mathbb{N} \text{ such that } g(n, x_1, \ldots, x_m) = 0,
\]
and undefined if there is no \(n\) such that \(g(n, x_1, \ldots, x_m) = 0\). We say that \(f\) is defined by
minimization from \(g\), and we write
\[
f(x_1, \ldots, x_m) = \mu x [g(x, x_1, \ldots, x_m) = 0].
\]
For short, we write \(f = \mu g\).

Even if \(g\) is a total function, \(f\) may be undefined for some (or all) of its inputs.

**Definition 4.15.** (Herbrand–Gödel–Kleene) The set of partial computable (or partial recursive)
functions is the smallest set of partial functions (defined on \(\mathbb{N}^n\) for some \(n \geq 1\)) which
contains the base functions and is closed under

1. Composition.
2. Primitive recursion.

The set of computable (or recursive) functions is the subset of partial computable functions
that are total functions (that is, defined for all input).

We proved earlier the Kleene normal form, which says that every partial computable
function \(f : \mathbb{N}^m \rightarrow \mathbb{N}\) is computable as
\[
f = g \circ \mu h,
\]
for some primitive recursive functions \(g : \mathbb{N} \rightarrow \mathbb{N}\) and \(h : \mathbb{N}^{m+1} \rightarrow \mathbb{N}\). The significance of this
result is that \(f\) is built up from total functions using composition and primitive recursion,
and only a single minimization is needed at the end.

Before stating our main theorem, we need to define what it means for a (numerical)
function in the \(\lambda\)-calculus. This requires some care to handle partial functions.

Since there are combinators for translating Church numerals to Barendregt numerals and
vice-versa, it does not matter which numerals we pick. We pick the Church numerals because
primitive recursion is definable without using a fixed-point combinator.
Definition 4.16. A function (partial or total) \( f : \mathbb{N}^n \to \mathbb{N} \) is \( \lambda \)-definable if for all \( m_1, \ldots, m_n \in \mathbb{N} \), there is a combinator (a closed \( \lambda \)-term) \( F \) with the following properties:

1. The value \( f(m_1, \ldots, m_n) \) is defined if and only if \( Fc_{m_1} \cdots c_{m_n} \) reduces to a \( \beta \)-normal form (necessarily unique by the Church–Rosser theorem).

2. If \( f(m_1, \ldots, m_n) \) is defined, then \( Fc_{m_1} \cdots c_{m_n} \) reduces to \( \beta \)-normal form.

In view of the Church–Rosser theorem (Theorem 4.1) and the fact that \( c_{f(m_1, \ldots, m_n)} \) is a \( \beta \)-normal form, we can replace

\[
Fc_{m_1} \cdots c_{m_n} \xrightarrow{\ast} Fc_{f(m_1, \ldots, m_n)}
\]

by

\[
Fc_{m_1} \cdots c_{m_n} \xrightarrow{\ast} c_{f(m_1, \ldots, m_n)}
\]

Note that the termination behavior of \( f \) on inputs \( m_1, \ldots, m_n \) has to match the reduction behavior of \( Fc_{m_1} \cdots c_{m_n} \), namely \( f(m_1, \ldots, m_n) \) is undefined if no reduction sequence from \( Fc_{m_1} \cdots c_{m_n} \) reaches a \( \beta \)-normal form. Condition (2) ensures that if \( f(m_1, \ldots, m_n) \) is defined, then the correct value \( c_{f(m_1, \ldots, m_n)} \) is computed by some reduction sequence from \( Fc_{m_1} \cdots c_{m_n} \).

We only care about total functions then we require that \( Fc_{m_1} \cdots c_{m_n} \) reduces to a \( \beta \)-normal form for all \( m_1, \ldots, m_n \) and (2). A stronger and more elegant version of \( \lambda \)-definability that better captures when a function is undefined for some input is considered in Section 4.7.

We have the following remarkable theorems.

**Theorem 4.10.** If a total function \( f : \mathbb{N}^n \to \mathbb{N} \) is \( \lambda \)-definable, then it is (total) computable. If a partial function \( f : \mathbb{N}^n \to \mathbb{N} \) is \( \lambda \)-definable, then it is partial computable.

Although Theorem 4.10 is intuitively obvious since computation by \( \beta \)-reduction sequences are “clearly” computable, a detailed proof is long and very tedious. One has to define primitive recursive functions to mimick \( \beta \)-conversion, etc. Most books sweep this issue under the rug. Barendregt observes that the “\( \lambda \)-calculus is recursively axiomatized,” which implies that the graph of the function being defined is recursively enumerable, but no details are provided; see Barendregt [3] (Chapter 6, Theorem 6.3.13). Kleene (1936) provides a detailed and very tedious proof. This is an amazing paper, but very hard to read. If the reader is not content she/he should work out the details over many long lonely evenings.

**Theorem 4.11.** (Kleene, 1936) If a (total) function \( f : \mathbb{N}^n \to \mathbb{N} \) is computable, then it is \( \lambda \)-definable. If a (partial) function \( f : \mathbb{N}^n \to \mathbb{N} \) is partial computable, then it is \( \lambda \)-definable.

**Proof.** First we assume all functions to be total. There are several steps.

**Step 1.** The base functions are \( \lambda \)-definable.
We already showed that $Z_c$ computes $Z$ and that $Succ_c$ computes $S$. Observe that $U^n_i$ given by
\[ U^n_i = \lambda x_1 \cdots x_n \cdot x_i \]
computes $P^n_i$.

**Step 2.** Closure under composition.
If $g$ is $\lambda$-defined by the combinator $G$ and $h_1, \ldots, h_m$ are $\lambda$-defined by the combinators $H_1, \ldots, H_m$, then $g \circ (h_1, \ldots, h_m)$ is $\lambda$-defined by
\[ F = \lambda x_1 \cdots x_n \cdot G(H_1 x_1 \cdots x_n) \cdots (H_m x_1 \cdots x_n). \]
Since the functions are total, there is no problem.

**Step 3.** Closure under primitive recursion.
We could use a fixed-point combinator but the combinator $\text{Iter}$ and pairing do the job. If $f$ is defined by primitive recursion from $g$ and $h$, and if $G$ $\lambda$-defines $g$ and $H$ $\lambda$-defines $h$, then $f$ is $\lambda$-defined by
\[ F = \lambda n x_1 \cdots x_m \cdot \text{Iter}_n \lambda z \cdot \langle H \pi_1 \pi_2 z x_1 \cdots x_m, Succ_c(\pi_2 z) \rangle \langle G x_1 \cdots x_m, c_0 \rangle. \]
The reason $F$ works is that we can prove by induction that
\[ (\lambda z \cdot \langle H \pi_1 \pi_2 z c_{n_1} \cdots c_{n_m}, Succ_c(\pi_2 z) \rangle)^n \langle G c_{n_1} \cdots c_{n_m}, c_0 \rangle \rightarrow^\beta \langle c_{f(n,n_1,\ldots,n_m)}, c_n \rangle. \]
For the base case $n = 0$,
\[ (\lambda z \cdot \langle H \pi_1 \pi_2 z c_{n_1} \cdots c_{n_m}, Succ_c(\pi_2 z) \rangle)^0 \langle G c_{n_1} \cdots c_{n_m}, c_0 \rangle \rightarrow^\beta \langle G c_{n_1} \cdots c_{n_m}, c_0 \rangle = \langle c_{g(n_1,\ldots,n_m)}, c_0 \rangle = \langle c_{f(0,n_1,\ldots,n_m)}, c_0 \rangle. \]
For the induction step,
\[ (\lambda z \cdot \langle H \pi_1 \pi_2 z c_{n_1} \cdots c_{n_m}, Succ_c(\pi_2 z) \rangle)^{n+1} \langle G c_{n_1} \cdots c_{n_m}, c_0 \rangle \rightarrow^\beta \langle H c_{f(n_1,\ldots,n_m)} c_{n_1} \cdots c_{n_m}, Succ_c c_n \rangle \rightarrow^\beta \langle c_{h(f(n_1,\ldots,n_m),n_{n_1,\ldots,n_m})}, c_{n+1} \rangle = \langle c_{f(n+1,n_1,\ldots,n_m)}, c_{n+1} \rangle. \]
Since the functions are total, there is no problem.
We can also show that primitive recursion can be achieved using a fixed-point combinator. Define the combinators $J$ and $F$ by
\[ J = \lambda f x_1 \cdots x_m \cdot \text{if \text{IsZero}_c x then } G x_1 \cdots x_m \text{ else } H(f(Pred_c x) x_1 \cdots x_m)(Pred_c x) x_1 \cdots x_m, \]
and

\[ F = \Theta J. \]

Then \( F \) \( \lambda \)-defines \( f \), and since the functions are total, there is no problem. This method must be used if we use the Barendregt numerals.

**Step 4. Closure under minimization.**

Suppose \( f \) is total and defined by minimization from \( g \) and that \( g \) is \( \lambda \)-defined by \( G \).

Define the combinators \( J \) and \( F \) by

\[
J = \lambda f x_1 \ldots x_m. \text{if } \text{IsZero}_c G x_1 \ldots x_m \text{ then } x \text{ else } f(\text{Succ}_c x) x_1 \ldots x_m
\]

and

\[ F = \Theta J. \]

It is not hard to check that

\[
F c_1 c_2 \ldots c_n \xrightarrow{\beta} \begin{cases} c_n & \text{if } g(n, n_1, \ldots, n_m) = 0 \\ F c_{n+1} c_1 \ldots c_n & \text{otherwise,} \end{cases}
\]

and we can use this to prove that \( F \) \( \lambda \)-defines \( f \). Since we assumed that \( f \) is total, some least \( n \) will be found. We leave the details as an exercise.

This finishes the proof that every total computable function is \( \lambda \)-definable.

To prove the result for the partial computable functions we appeal to the Kleene normal form: every partial computable function \( f: \mathbb{N}^m \to \mathbb{N} \) is computable as

\[ f = g \circ \mu h, \]

for some primitive recursive functions \( g: \mathbb{N} \to \mathbb{N} \) and \( h: \mathbb{N}^{m+1} \to \mathbb{N} \). Then our previous proof yields combinators \( G \) and \( H \) that \( \lambda \)-define \( g \) and \( h \). The minimization of \( h \) may fail but since \( g \) is a total function of a single argument, \( f(n_1, \ldots, n_m) \) is defined iff \( g(\mu_n[h(n, n_1, \ldots, n_m) = 0]) \) is defined so it should be clear that \( F \) computes \( f \), but the reader may want to provide a rigorous argument. A detailed proof is given in Hindley and Seldin [19] (Chapter 4, Theorem 4.18).

Combining Theorem 4.10 and Theorem 4.11 we have established the remarkable result that the set of \( \lambda \)-definable total functions is exactly the set of (total) computable functions, and similarly for partial functions. So the \( \lambda \)-calculus has universal computing power.

**Remark:** With some work, it is possible to show that lists can be represented in the \( \lambda \)-calculus. Since a Turing machine tape can be viewed as a list, it should be possible (but very tedious) to simulate a Turing machine in the \( \lambda \)-calculus. This simulation should be somewhat analogous to the proof that a Turing machine computes a computable function (defined à la Herbrand–Gödel–Kleene).

Since the \( \lambda \)-calculus has the same power as Turing machines we should expect some undecidability results analogous to the undecidability of the halting problem or Rice’s theorem. We state the following analog of Rice’s theorem without proof. It is a corollary of a theorem known as the Scott–Curry theorem.
**Theorem 4.12.** (D. Scott) Let $A$ be any nonempty set of $\lambda$-terms not equal to the set of all $\lambda$-terms. If $A$ is closed under $\beta$-reduction, then it is not computable (not recursive).

Theorem 4.12 is proven in Barendregt [3] (Chapter 6, Theorem 6.6.2) and Barendregt [4].

As a corollary of Theorem 4.12 it is undecidable whether a $\lambda$-term has a $\beta$-normal form, a result originally proved by Church. This is an analog of the undecidability of the halting problem, but it seems more spectacular because the syntax of $\lambda$-terms is really very simple. The problem is that $\beta$-reduction is very powerful and elusive.

### 4.7 Definability of Functions in Typed Lambda-Calculi

In the pure $\lambda$-calculus, some $\lambda$-terms have no $\beta$-normal form, and worse, it is undecidable whether a $\lambda$-term has a $\beta$-normal form. In contrast, by Theorem 4.12, every raw $\lambda$-term that type-checks in the simply-typed $\lambda$-calculus has a $\beta$-normal form. Thus it is natural to ask whether the natural numbers are definable in the simply-typed $\lambda$-calculus because if the answer is positive, then the numerical functions definable in the simply-typed $\lambda$-calculus are guaranteed to be total.

This indeed possible. If we pick any base type $\sigma$, then we can define typed Church numerals $c_n$ as terms of type $\text{Nat}_\sigma = (\sigma \to \sigma) \to (\sigma \to \sigma)$, by

$$c_n = \lambda f : (\sigma \to \sigma). \lambda x : \sigma. f^n(x).$$

The notion of $\lambda$-definable function is defined just as before. Then we can define $\text{Add}$ and $\text{Mult}$ as terms of type $\text{Nat}_\sigma \to (\text{Nat}_\sigma \to \text{Nat}_\sigma)$ essentially as before, but surprise, not much more is definable. Among other things, strong typing of terms restricts the iterator combinator too much. It was shown by Schwichtenberg and Statman that the numerical functions definable in the simply-typed $\lambda$-calculus are the extended polynomials; see Statman [35] and Troelstra and Schwichtenberg [37]. The extended polynomials are the smallest class of numerical functions closed under composition containing

1. The constant functions 0 and 1.
2. The projections.
3. Addition and multiplication.
4. The function $\text{IsZero}_c$.

Is there a way to get a larger class of total functions?

There are indeed various ways of doing this. One method is to add the natural numbers and the booleans as data types to the simply-typed $\lambda$-calculus, and to also add product types, an iterator combinator, and some new reduction rules. This way we obtain a system equivalent to Gödel’s system $T$. A large class of numerical total functions containing the
primitive recursive functions is definable in this system; see Girard–Lafond–Taylor [16]. Although theoretically interesting, this is not a practical system.

Another wilder method is to allow more general types to the simply-typed λ-calculus, the so-called second-order types or polymorphic types. In addition to base types, we allow type variables (often denoted $X, Y, \ldots$) ranging over simple types and new types of the form $\forall X. \sigma$. For example, $\forall X. (X \to X)$ is such a new type, and so is $\forall X. (X \to ((X \to X) \to X))$.

Actually, the second-order types that we just defined are special cases of the QBF (quantified boolean formulae) arising in complexity theory restricted to implication and universal quantifiers; see Section 11.3. Remarkably, the other connectives $\land, \lor, \neg$ and $\exists$ are definable in terms of $\to$ (as a logical connective, $\Rightarrow$) and $\forall$; see Troelstra and Schwichtenberg [37] (Chapter 11).

Remark: The type $\textbf{Nat} = \forall X. (X \to ((X \to X) \to X))$.

can be chosen to represent the type of the natural numbers. The type of the natural numbers can also be chosen to be $\forall X. (X \to X) \to (X \to X)$.

This makes essentially no difference but the first choice has some technical advantages.

There is also a new form of type abstraction, $\Lambda X. M$, and of type application, $M\sigma$, where $M$ is a λ-term and $\sigma$ is a type. There are two new typing rules:

$\Gamma \vdash M : \sigma$

$\Gamma \vdash (\Lambda X. M) : \forall X. \sigma$ (type abstraction)

provided that $X$ does not occur free in any of the types in $\Gamma$, and

$\Gamma \vdash M : \forall X. \sigma$

$\Gamma \vdash (M\tau) : \sigma[X := \tau]$ (type application)

where $\tau$ is any type (and no capture of variable takes place).

From the point of view where types are viewed as propositions and λ-terms are viewed as proofs, type abstraction is an introduction rule and type application is an elimination rule, both for the second-order quantifier $\forall$.

We also have a new reduction rule

$(\Lambda X. M)\sigma \to_{\beta\forall} M[X := \sigma]$ that corresponds to a new form of redundancy in proofs having to do with a $\forall$-elimination immediately following a $\forall$-introduction. Here in the substitution $M[X := \tau]$, all free occurrences of $X$ in $M$ and the types in $M$ are replaced by $\tau$. For example,

$(\Lambda X. \lambda f : (X \to X). \lambda x : X. \lambda g : \forall Y. (Y \to Y). gX f x)[X := \tau]$

$= \lambda f : (\tau \to \tau). \lambda x : \tau. \lambda g : \forall Y. (Y \to Y). g\tau x f$.

---

3Barendregt and others used Greek variables to denote type variables but we find this confusing.
CHAPTER 4. THE LAMBDA-CALCULUS

For technical details, see Gallier [13].

This new typed \(\lambda\)-calculus is the \textit{second-order polymorphic lambda calculus}. It was invented by Girard (1972) who named it \textit{system F}; see Girard [17, 18], and it is denoted \(\lambda_2\) by Barendregt. From the point of view of logic, Girard’s system is a proof system for \textit{intuitionistic second-order propositional logic}. We define \(\rightarrow_{\lambda_2}^+\) and \(\rightarrow_{\lambda_2}^*\) as the relations

\[
\rightarrow_{\lambda_2}^+ = (\rightarrow_{\beta} \cup \rightarrow_{\beta\forall})^+
\]
\[
\rightarrow_{\lambda_2}^* = (\rightarrow_{\beta} \cup \rightarrow_{\beta\forall})^*
\]

A variant of system F was also introduced independently by John Reynolds (1974) but for very different reasons.

The intuition behind terms of type \(\forall X.\sigma\) is that a term \(M\) of type \(\forall X.\sigma\) is a sort of \textit{generic function} such that for any type \(\tau\), the function \(M\tau\) is a \textit{specialized version} of type \(\sigma[X := \tau]\) of \(M\). For example, \(M\) could be the function that appends an element to a list, and for specific types such as the natural numbers \(\text{Nat}\), strings \(\text{String}\), trees \(\text{Tree}\), etc., the functions \(M\text{Nat}, M\text{String}, M\text{Tree}\), are the specialized versions of \(M\) to lists of elements having the specific data types \(\text{Nat}, \text{String}, \text{Tree}\).

For example, if \(\sigma\) is any type, we have the closed term

\[
A_{\sigma} = \lambda x: \sigma. \lambda f: (\sigma \rightarrow \sigma). fx,
\]

of type \(\sigma \rightarrow ((\sigma \rightarrow \sigma) \rightarrow \sigma)\), such that for every term \(F\) of type \(\sigma \rightarrow \sigma\) and every term \(a\) of type \(\sigma\),

\[
A_{\sigma} Fa \rightarrow_{\lambda_2}^+ Fa.
\]

Since \(A_{\sigma}\) has the same behavior for all types \(\sigma\), it is natural to define the generic function \(A\) given by

\[
A = \Lambda X. \lambda x: X. \lambda f: (X \rightarrow X). fx,
\]

which has type \(\forall X. (X \rightarrow ((X \rightarrow X) \rightarrow X))\), and then \(A\sigma\) has the same behavior as \(A_{\sigma}\). We will see shortly that \(A\) is the Church numeral \(c_1\) in \(\lambda_2\).

Remarkably, system F is \textit{strongly normalizing}, which means that every \(\lambda\)-term typeable in system F has a \(\beta\)-normal form. The proof of this theorem is hard and was one of Girard’s accomplishments in his dissertation, Girard [18]. The Church–Rosser property also holds for system F. The proof technique used to prove that system F is strongly normalizing is thoroughly analyzed in Gallier [13].

We stated earlier that deciding whether a simple type \(\sigma\) is provable, that is, whether there is a closed \(\lambda\)-term \(M\) that type-checks in the simply-typed \(\lambda\)-calculus such that the judgement \(\vdash M: \sigma\) is provable is a hard problem. Indeed Statman proved that this problem is P-space complete; see Statman [34] and Section 11.4.

It is natural so ask whether it is decidable whether given any second-order type \(\sigma\), there is a closed \(\lambda\)-term \(M\) that type-checks in system F such that the judgement \(\vdash M: \sigma\) is provable (if \(\sigma\) is viewed as a second-order logical formula, the problem is to decide whether \(\sigma\)
is provable). Surprisingly the answer is no; this problem (called \textit{inhabitation}) is undecidable. This result was proven by L"ob around 1976, see Barendregt [4].

This undecidability result is troubling and at first glance seems paradoxical. Indeed, viewed as a logical formula, a second-order type $\sigma$ is a QBF (a quantified boolean formula), and if we assign the truth values $F$ and $T$ to the boolean variables in it, we can decide whether such a proposition is valid in exponential time and polynomial space (in fact, we will see that later QBF validity is P-space complete). This seems in contradiction with the fact that provability is undecidable.

But the proof system corresponding to system F is an \textit{intuitionistic} proof system, so there are (non-quantified) propositions that are valid in the truth-value semantics but not provable in intuitionistic propositional logic. The set of second-order propositions provable in intuitionistic second-order logic is a \textit{proper} subset of the set of valid QBF (under the truth-value semantics), and it is \textit{not computable}. So there is no paradox after all.

Going back to the issue of computability of numerical functions, a version of the Church numerals can be defined as

$$c_n = \Lambda X. \lambda x: X. \lambda f: (X \to X). f^n(x).$$

Observe that $c_n$ has type $\text{Nat}$. Inspired by the definition of $\text{Succ}$ given in Section 4.4, we can define the successor function on the natural numbers as

$$\text{Succ} = \lambda n: \text{Nat}. \Lambda X. \lambda x: X. \lambda f: (X \to X). f(nXxf).$$

Note how $n$, which is of type $\text{Nat} = \forall X. (X \to ((X \to X) \to X))$, is applied to the type variable $X$ in order to become a term $nX$ of type $X \to ((X \to X) \to X)$, so that $nXxf$ has type $X$, thus $f(nXxf)$ also has type $X$.

For every type $\sigma$, every term $F$ of type $\sigma \to \sigma$ and every term $a$ of type $\sigma$, we have

$$c_n\sigma aF = (\Lambda X. \lambda x: X. \lambda f: (X \to X). f^n(x))\sigma aF$$

$$\lambda \to \lambda_2 (\lambda x: \sigma. \lambda f: (\sigma \to \sigma). f^n(x))aF$$

$$\lambda \to \lambda_2 F^n(a);$$

that is,

$$c_n\sigma aF \lambda \to \lambda_2 F^n(a).$$

So $c_n\sigma$ iterates $F$ $n$ times starting with $a$. As a consequence,

$$\text{Succ} c_n = (\lambda n: \text{Nat}. \Lambda X. \lambda x: X. \lambda f: (X \to X). f(nXxf)) c_n$$

$$\lambda \to \lambda_2 \Lambda X. \lambda x: X. \lambda f: (X \to X). f(c_nXxf)$$

$$\lambda \to \lambda_2 \Lambda X. \lambda x: X. \lambda f: (X \to X). f(f^n(x))$$

$$= \Lambda X. \lambda x: X. \lambda f: (X \to X). f^{n+1}(x) = c_{n+1}. $$
We can also define addition of natural numbers as
\[ \text{Add} = \lambda m : \text{Nat}. \lambda n : \text{Nat}. \Lambda X. \lambda x : X. \lambda f : (X \to X). (mX f(nX xf))f. \]

Note how \( m \) and \( n \), which are of type \( \text{Nat} = \forall X. (X \to ((X \to X) \to X)) \), are applied to the type variable \( X \) in order to become terms \( mX \) and \( nX \) of type \( X \to (X \to X) \to X \), so that \( nX xf \) has type \( X \), thus \( f(nX xf) \) also has type \( X \), and \( mX f(nX xf) \) has type \( (X \to X) \to X \), and finally \( (mX f(nX xf))f \) has type \( X \).

Many of the constructions that can be performed in the pure \( \lambda \)-calculus can be mimicked in system F, which explains its expressive power. For example, for any two second-order types \( \sigma \) and \( \tau \), we can define a pairing function \( \langle -, - \rangle \) (to be very precise, \( \langle -, - \rangle_{\sigma, \tau} \)) given by
\[ \langle -, - \rangle = \lambda u : \sigma. \lambda v : \tau. \Lambda X. \lambda f : \sigma \to (\tau \to X). fuv, \]
of type \( \sigma \to (\tau \to (\forall X. ((\sigma \to (\tau \to X)) \to X))) \). Given any term \( M \) of type \( \sigma \) and any term \( N \) of type \( \tau \), we have
\[ \langle -, - \rangle_{\sigma, \tau} MN \xrightarrow{\Delta 2} \Lambda X. \lambda f : \sigma \to (\tau \to X). fMN. \]
Thus we define \( \langle M, N \rangle \) as
\[ \langle M, N \rangle = \Lambda X. \lambda f : \sigma \to (\tau \to X). fMN, \]
and the type
\[ \forall X. ((\sigma \to (\tau \to X)) \to X) \]
of \( \langle M, N \rangle \) is denoted by \( \sigma \times \tau \). As a logical formula it is equivalent to \( \sigma \land \tau \), which means that if we view \( \sigma \) and \( \tau \) as (second-order) propositions, then
\[ \sigma \land \tau \equiv \forall X. ((\sigma \to (\tau \to X)) \to X) \]
is provable intuitionistically. This is a special case of the result that we mentioned earlier: the connectives \( \land, \lor, \neg \) and \( \exists \) are definable in terms of \( \to \) (as a logical connective, \( \Rightarrow \)) and \( \forall \).

**Proposition 4.13.** The connectives \( \land, \lor, \neg, \perp \) and \( \exists \) are definable in terms of \( \to \) and \( \forall \), which means that the following equivalences are provable intuitionistically, where \( X \) is not free in \( \sigma \) or \( \tau \):
\[
\begin{align*}
\sigma \land \tau & \equiv \forall X. ((\sigma \to (\tau \to X)) \to X) \\
\sigma \lor \tau & \equiv \forall X. ((\sigma \to X) \to ((\tau \to X) \to X)) \\
\perp & \equiv \forall X. X \\
\neg \sigma & \equiv \sigma \to \forall X. X \\
\exists Y. \sigma & \equiv \forall X. (((\forall Y. (\sigma \to X)) \to X). \]

4.7. DEFINABILITY OF FUNCTIONS IN TYPED LAMBDA-CALCULI

We leave the proof as an exercise, or see Troelstra and Schwichtenberg [37] (Chapter 11).

Remark: The rule of type application implies that \( \bot \rightarrow \sigma \) is intuitionistically provable for all propositions (types) \( \sigma \). So in second-order logic there is no difference between minimal and intuitionistic logic.

We also have two projections \( \pi_1 \) and \( \pi_2 \) (to be very precise \( \pi_1^{\sigma \times \tau} \) and \( \pi_2^{\sigma \times \tau} \)) given by

\[
\pi_1 = \lambda g: \sigma \times \tau. g\sigma(\lambda x: \sigma. \lambda y: \tau. x) \\
\pi_2 = \lambda g: \sigma \times \tau. g\tau(\lambda x: \sigma. \lambda y: \tau. y).
\]

It is easy to check that \( \pi_1 \) has type \( (\sigma \times \tau) \rightarrow \sigma \) and that \( \pi_2 \) has type \( (\sigma \times \tau) \rightarrow \tau \). The reader should check that for any \( M \) of type \( \sigma \) and any \( N \) of type \( \tau \) we have

\[
\pi_1(M, N) \overset{+}{\rightarrow}_{\lambda^2} M \text{ and } \pi_2(M, N) \overset{+}{\rightarrow}_{\lambda^2} N.
\]

For example, we have

\[
\pi_1(M, N) = (\lambda g: \sigma \times \tau. g\sigma(\lambda x: \sigma. \lambda y: \tau. x))(\Lambda X. \lambda f: \sigma \rightarrow (\tau \rightarrow X). fMN)
\]

\[
\overset{+}{\rightarrow}_{\lambda^2} (\Lambda X. \lambda f: \sigma \rightarrow (\tau \rightarrow X). fMN)\sigma(\lambda x: \sigma. \lambda y: \tau. x)
\]

\[
\overset{+}{\rightarrow}_{\lambda^2} (\lambda f: \sigma \rightarrow (\tau \rightarrow \sigma). fMN)(\lambda x: \sigma. \lambda y: \tau. x)
\]

\[
\overset{+}{\rightarrow}_{\lambda^2} (\lambda x: \sigma. \lambda y: \tau. x)MN
\]

\[
\overset{+}{\rightarrow}_{\lambda^2} (\lambda y: \tau. M)N
\]

\[
\overset{+}{\rightarrow}_{\lambda^2} M.
\]

The booleans can be defined as

\( T = \Lambda X. \lambda x: X. \lambda y: X. x \)

\( F = \Lambda X. \lambda x: X. \lambda y: X. y, \)

both of type \( \text{Bool} = \forall X. (X \rightarrow (X \rightarrow X)) \). We also define if then else as

\( \text{if then else} = \Lambda X. \lambda z: \text{Bool}. zX \)

of type \( \forall X. \text{Bool} \rightarrow (X \rightarrow (X \rightarrow X)) \). It is easy that for any type \( \sigma \) and any two terms \( M \) and \( N \) of type \( \sigma \) we have

\[
(\text{if } T \text{ then } M \text{ else } N)\sigma \overset{+}{\rightarrow}_{\lambda^2} M
\]

\[
(\text{if } F \text{ then } M \text{ else } N)\sigma \overset{+}{\rightarrow}_{\lambda^2} N,
\]
where we write \((\text{if } T \text{ then } M \text{ else } N) \sigma\) instead of \((\text{if then else}) \sigma T M N\) (and similarly for the other term). For example, we have

\[(\text{if } T \text{ then } M \text{ else } N) \sigma = (\Lambda X. \lambda z : \text{Bool}. z X) \sigma T M N\]

\[\overset{\lambda_2}{\rightarrow} (\lambda z : \text{Bool}. z \sigma) T M N\]

\[\overset{\lambda_2}{\rightarrow} (T \sigma) M N\]

\[= ((\Lambda X. \lambda x : X. \lambda y : X. x) \sigma) M N\]

\[\overset{\lambda_2}{\rightarrow} (\lambda x : \sigma. \lambda y : \sigma. x) M N\]

\[\overset{\lambda_2}{\rightarrow} M.\]

Lists, trees, and other inductively data structures are also representable in system F; see Girard–Lafond–Taylor [16].

We can also define an iterator \(\text{Iter}\) given by

\[\text{Iter} = \Lambda X. \lambda u : X. \lambda f : (X \rightarrow X). \lambda z : \text{Nat}. z X u f\]

of type \(\forall X. (X \rightarrow ((X \rightarrow X) \rightarrow (\text{Nat} \rightarrow X)))\). The idea is that given \(f\) of type \(\sigma \rightarrow \sigma\) and \(u\) of type \(\sigma\), the term \(\text{Iter} \sigma u f c_n\) iterates \(f\) \(n\) times over the input \(u\). It is easy to show that for any term \(t\) of type \(\text{Nat}\) we have

\[\text{Iter} \sigma u f c_0 \overset{\lambda_2}{\rightarrow} u\]

\[\text{Iter} \sigma u f (\text{Succ}_c t) \overset{\lambda_2}{\rightarrow} f(\text{Iter} \sigma u f t),\]

and that

\[\text{Iter} \sigma u f c_n \overset{\lambda_2}{\rightarrow} f^n(u).\]

Then mimicking what we did in the pure \(\lambda\)-calculus, we can show that the primitive recursive functions are \(\lambda\)-definable in system \(F\). Actually, higher-order primitive recursion is definable. So, for example, Ackermann’s function is definable.

Remarkably, the class of numerical functions definable in system \(F\) is a class of (total) computable functions much bigger than the class of primitive recursive functions. This class of functions was characterized by Girard as the functions that are provably-recursive in a formalization of arithmetic known as \textit{intuitionistic second-order arithmetic}; see Girard [18], Troelstra and Schwichtenberg [37] and Girard–Lafond–Taylor [16]. It can also be shown (using a diagonal argument) that there are (total) computable functions not definable in system \(F\).

From a theoretical point of view, every (total) function that we will ever want to compute is definable in system \(F\). However, from a practical point of view, programming in system \(F\) is very tedious and usually leads to very inefficient programs. Nevertheless polymorphism is an interesting paradigm which had made its way in certain programming languages.

Type systems even more powerful than system \(F\) have been designed, the ultimate system being the \textit{calculus of constructions} due to Huet and Coquand, but these topics are beyond the scope of these notes.
4.8. HEAD NORMAL-FORMS AND THE PARTIAL COMPUTABLE FUNCTIONS 121

One last comment has to do with the use of the simply-typed \( \lambda \)-calculus as the core of a programming language. In the early 1970’s Dana Scott defined a system named LCF based on the the simply-typed \( \lambda \)-calculus and obtained by adding the natural numbers and the booleans as data types, product types, and a fixed-point operator. Robin Milner then extended LCF, and as a by-product, defined a programming language known as ML, which is the ancestor of most functional programming languages. A masterful and thorough exposition of type theory and its use in programming language design is given in Pierce [28].

We now revisit the problem of defining the partial computable functions.

4.8 Head Normal-Forms and the Partial Computable Functions

One defect of the proof of Theorem 4.11 in the case where a computable function is partial is the use of the Kleene normal form. The difficulty has to do with composition. Given a partial computable function \( g \) \( \lambda \)-defined by a closed term \( G \) and a partial computable function \( h \) \( \lambda \)-defined by a closed term \( H \) (for simplicity we assume that both \( g \) and \( h \) have a single argument), it would be nice if the composition \( h \circ g \) was represented by \( \lambda x. H(Gx) \). This is true if both \( g \) and \( h \) are total, but false if either \( g \) or \( h \) is partial as shown by the following example from Barendregt [3] (Chapter 2, §2). If \( g \) is the function undefined everywhere and \( h \) is the constant function 0, then \( g \) is \( \lambda \)-defined by \( G = K\Omega \) and \( h \) is \( \lambda \)-defined by \( H = Kc_0 \), with \( \Omega = (\lambda x.(xx))(\lambda x.(xx)) \). We have

\[
\lambda x. H(Gx) = \lambda x. Kc_0(K\Omega x) \xrightarrow{+} \lambda x. Kc_0\Omega \xrightarrow{+} \lambda x. c_0,
\]

but \( h \circ g = g \) is the function undefined everywhere, and \( \lambda x. c_0 \) represents the total function \( h \), so \( \lambda x. H(Gx) \) does not \( \lambda \)-define \( h \circ g \).

It turns out that the \( \lambda \)-definability of the partial computable functions can be obtained in a more elegant fashion without having recourse to the Kleene normal form by capturing the fact that a function is undefined for some input is a more subtle way. The key notion is the notion of head normal form, which is more general than the notion of \( \beta \)-normal form. As a consequence, there are fewer \( \lambda \)-terms having no head normal form than \( \lambda \)-terms having no \( \beta \)-normal form, and we capture a stronger form of divergence.

Recall that a \( \lambda \)-term is either a variable \( x \), or an application \( (MN) \), or a \( \lambda \)-abstraction \( (\lambda x.M) \). We can sharpen this characterization as follows.

**Proposition 4.14.** The following properties hold:

1. Every application term \( M \) is of the form

\[
M = (N_1N_2\cdots N_{n-1})N_n, \quad n \geq 2,
\]

where \( N_1 \) is not an application term.
(2) Every abstraction term \(M\) is of the form
\[ M = \lambda x_1 \cdots x_n. N, \quad n \geq 1, \]
where \(N\) is not an abstraction term.

(3) Every \(\lambda\)-term \(M\) is of one of the following two forms:
\[ M = \lambda x_1 \cdots x_n. xM_1 \cdots M_m, \quad m, n \geq 0 \]  \hspace{1cm} (a)
\[ M = \lambda x_1 \cdots x_n. (\lambda x. M_0)M_1 \cdots M_m, \quad m \geq 1, n \geq 0, \]  \hspace{1cm} (b)
where \(x\) is a variable.

Proof. (1) Suppose that \(M\) is an application \(M = M_1M_2\). We proceed by induction on the depth of \(M_1\). For the base case \(M_1\) must be variables and we are done. For the induction step, if \(M_1\) is a \(\lambda\)-abstraction, we are done. If \(M_1\) is an application, then by the induction hypothesis it is of the form
\[ M_1 = (N_1N_2 \cdots N_{n-1})N_n, \quad n \geq 2, \]
where \(N_1\) is not an application term, and then
\[ M = M_1M_2 = ((N_1N_2 \cdots N_{n-1})N_n)M_2 \quad n \geq 2, \]
where \(N_1\) is not an application term.

The proof of (2) is similar.

(3) We proceed by induction on the depth of \(M\). If \(M\) is a variable, then we are in Case (a) with \(m = n = 0\).

If \(M\) is an application, then by (1) it is of the form \(M = N_1N_2 \cdots N_p\) with \(N_1\) not an application term. This means that either \(N_1\) is a variable, in which case we are in Case (a) with \(n = 0\), or \(N_1\) is an abstraction, in which case we are in Case (b) also with \(n = 0\).

If \(M\) is an abstraction \(\lambda x. N\), then by the induction hypothesis \(N\) is of the form (a) or (b), and by adding one more binder \(\lambda x\) in front of these expressions we preserve the shape of (a) and (b) by increasing \(n\) by 1.

The terms, \(I, K, K_s, S\), the Church numerals \(c_n\), if then else, \((M, N)\), \(\pi_1, \pi_2\), \(\text{Iter}, \text{Succ}_e\), \(\text{Add}\) and \(\text{Mult}\) as in Proposition 4.5, are \(\lambda\)-terms of type (a). However, \(\text{Pred}_K\), \(\Omega = (\lambda x. (xx))(\lambda x. (xx))\), \(Y\) (the Curry \(Y\)-combinator), \(\Theta\) (the Turing \(\Theta\)-combinator) are of type (b).

Proposition 4.14 motivates the following definition.

**Definition 4.17.** A \(\lambda\)-term \(M\) is a head normal form (for short \(hnf\)) if it is of the form (a), namely
\[ M = \lambda x_1 \cdots x_n. xM_1 \cdots M_m, \quad m, n \geq 0, \]
where \(x\) is a variable called the head variable.
A \( \lambda \)-term \( M \) has a head normal form if there is some head normal form \( N \) such that \( M \xrightarrow{\ast} \beta N \).

In a term \( M \) of the form (b),

\[
M = \lambda x_1 \cdots x_n. (\lambda x. M_0)M_1 \cdots M_m, \quad m \geq 1, n \geq 0,
\]

the subterm \((\lambda x. M_0)M_1\) is called the head redex of \( M \).

In addition to the terms of type (a) that we listed after Proposition 4.14, the term \( \lambda x. x \Omega \) is a head normal form. It is the head normal form of the term \( \lambda x. (Ix)\Omega \), which has no \( \beta \)-normal form.

Not every term has a head normal form. For example, the term

\[
\Omega = (\lambda x. (xx))(\lambda x. (xx))
\]

has no head normal form. Every \( \beta \)-normal form must be a head normal form, but the converse is false as we saw with

\[
M = \lambda x. x \Omega,
\]

which is a head normal form but has no \( \beta \)-normal form.

Note that a head redex of a term is a leftmost redex, but not conversely, as shown by the term \( \lambda x. x((\lambda y. y)x) \).

A term may have more than one head normal form but here is a way of obtaining a head normal form (if there is one) in a systematic fashion.

**Definition 4.18.** The relation \( \rightarrow_h \), called one-step head reduction, is defined as follows: For any two terms \( M \) and \( N \), if \( M \) contains a head redex \((\lambda x. M_0)M_1\), which means that \( M \) is of the form

\[
M = \lambda x_1 \cdots x_n. (\lambda x. M_0)M_1 \cdots M_m, \quad m \geq 1, n \geq 0,
\]

then \( M \xrightarrow{\rightarrow_h} N \) with

\[
N = \lambda x_1 \cdots x_n. (M_0[x := M_1])M_2 \cdots M_m.
\]

We denote by \( \xrightarrow{+}_h \) the transitive closure of \( \rightarrow_h \) and by \( \xrightarrow{*}_h \) the reflexive and transitive closure of \( \rightarrow_h \).

Given a term \( M \) containing a head redex, the head reduction sequence of \( M \) is the uniquely determined sequence of one-step head reductions

\[
M = M_0 \rightarrow_h M_1 \rightarrow_h \cdots \rightarrow_h M_n \rightarrow_h \cdots.
\]

If the head reduction sequence reaches a term \( M_n \) which is a head normal form we say that the sequence terminates, and otherwise we say that \( M \) has an infinite head reduction.

The following result is shown in Barendregt [3] (Chapter 8, \S 3).
CHAPTER 4. THE LAMBDA-CALCULUS

Theorem 4.15. (Wadsworth) A \( \lambda \)-term \( M \) has a head normal form if and only if the head reduction sequence terminates.

In some intuitive sense, a \( \lambda \)-term \( M \) that does not have any head normal form has a strong divergence behavior with respect to \( \beta \)-reduction.

Remark: There is a notion more general than the notion of head normal form which comes up in functional languages (for example, Haskell). A \( \lambda \)-term \( M \) is a weak head normal form if it of one of the two forms
\[ \lambda x. N \quad \text{or} \quad y N_1 \cdots N_m \]
where \( y \) is a variable. These are exactly the terms that do not have a redex of the form \((\lambda x. M_0) M_1 N_1 \cdots N_m\). Every head normal form is a weak head normal form, but there are many more weak head normal forms than there are head normal forms since a term of the form \( \lambda x. N \) where \( N \) is arbitrary is a weak head normal form, but not a head normal form unless \( N \) is of the form \( \lambda x_1 \cdots x_n. xM_1 \cdots M_m \), with \( m, n \geq 0 \).

Reducing to a weak head normal form is a lazy evaluation strategy.

There is also another useful notion which turns out to be equivalent to having a head normal form.

Definition 4.19. A closed \( \lambda \)-term \( M \) is solvable if there are closed terms \( N_1, \ldots, N_n \) such that
\[ MN_1 \cdots N_n \overset{*}{\rightarrow}_\beta \textbf{I}. \]
A \( \lambda \)-term \( M \) with free variables \( x_1, \ldots, x_m \) is solvable if the closed term \( \lambda x_1 \cdots x_m. M \) is solvable. A term is unsolvable if it is not solvable.

The following result is shown in Barendregt [3] (Chapter 8, \S 3).

Theorem 4.16. (Wadsworth) A \( \lambda \)-term \( M \) has a head normal form if and only if it is solvable.

Actually, the proof that having a head normal form implies solvable is not hard. We are now ready to revise the notion of \( \lambda \)-definability of numerical functions.

Definition 4.20. A function (partial or total) \( f : \mathbb{N}^n \rightarrow \mathbb{N} \) is strongly \( \lambda \)-definable if for all \( m_1, \ldots, m_n \in \mathbb{N} \), there is a combinator (a closed \( \lambda \)-term) \( F \) with the following properties:

1. If the value \( f(m_1, \ldots, m_n) \) is defined, then \( Fc_{m_1} \cdots c_{m_n} \) reduces to the \( \beta \)-normal form \( c_{f(m_1, \ldots, m_n)} \).
2. If \( f(m_1, \ldots, m_n) \) is undefined, then \( Fc_{m_1} \cdots c_{m_n} \) has no head normal form, or equivalently, is unsolvable.
Observe that in Case (2), when the value \( f(m_1, \ldots, m_n) \) is undefined, the divergence behavior of \( Fc_{m_1} \cdots c_{m_n} \) is stronger than in Definition 4.16. Not only \( Fc_{m_1} \cdots c_{m_n} \) has no \( \beta \)-normal form, but actually it has no head normal form.

The following result is proven in Barendregt [3] (Chapter 8, §4). The proof does not use the Kleene normal form. Instead, it makes clever use of the term KII. Another proof is given in Krivine [22] (Chapter II).

**Theorem 4.17.** Every partial or total computable function is strongly \( \lambda \)-definable. Conversely, every strongly \( \lambda \)-definable function is partial computable.

Making sure that a composition \( g \circ (h_1, \ldots, h_m) \) is defined for some input \( x_1, \ldots, x_n \) iff all the \( h_i(x_1, \ldots, x_n) \) and \( g(h_1(x_1, \ldots, x_n), \ldots, h_m(x_1, \ldots, x_n)) \) are defined is tricky. The term KII comes to the rescue! If \( g \) is strongly \( \lambda \)-definable by \( G \) and the \( h_i \) are strongly \( \lambda \)-definable by \( H_i \), then it can be shown that the combinator \( F \) given by

\[
F = \lambda x_1 \cdots x_n. (H_1 x_1 \cdots x_n \text{KII}) \cdots (H_m x_1 \cdots x_n \text{KII})(G(H_1 x_1 \cdots x_n) \cdots (G(H_m x_1 \cdots x_n))
\]

strongly \( \lambda \)-defines \( F \); see Barendregt [3] (Chapter 8, Lemma 8.4.6).
Chapter 5

Recursion Theory; More Advanced Topics

This chapter is devoted to three advanced topics of recursion theory:

(1) The recursion theorem.

(2) The extended Rice theorem.

(3) Creative and productive sets and their use in proving a strong version of Gödel’s first incompleteness theorem.

The recursion theorem is a deep result and an important technical tool in recursion theory.

The extended Rice theorem gives a characterization of the sets of partial computable functions that are listable in terms of extensions of partial computable functions with finite domains.

Productive and creative sets arise when dealing with truth and provability in arithmetic. The “royal road” to Gödel’s first incompleteness theorem is to first prove that for any proof system for arithmetic that only proves true statements (and is rich enough), the set of true sentences of arithmetic is productive. Productive sets are not listable in a strong sense, so we deduce that it is impossible to axiomatize the set of true sentences of arithmetic in a computable manner. The set of provable sentences of arithmetic is creative, which implies that it is impossible to decide whether a sentence of arithmetic is provable. This also implies that there are true sentences $F$ such that neither $F$ nor $\neg F$ are provable.

5.1 The Recursion Theorem

The recursion theorem, due to Kleene, is a fundamental result in recursion theory. Let $f$ be a total computable function. Then it turns out that there is some $n$ such that

$$\varphi_n = \varphi_{f(n)}.$$
To understand why such a mysterious result is interesting, consider the recursive definition of the factorial function $fact(n) = n!$ given by

$$
\begin{align*}
    fact(0) &= 1 \\
    fact(n + 1) &= (n + 1)fact(n).
\end{align*}
$$

The trick is to define the partial computable function $g$ (defined on $\mathbb{N}^2$) given by

$$
\begin{align*}
    g(m, 0) &= 1 \\
    g(m, n + 1) &= (n + 1)\varphi_m(n)
\end{align*}
$$

for all $m, n \in \mathbb{N}$. By the s-m-n Theorem, there is a computable function $f$ such that

$$
g(m, n) = \varphi_{f(m)}(n) \quad \text{for all } m, n \in \mathbb{N}.
$$

Then the equations above become

$$
\begin{align*}
    \varphi_{f(m)}(0) &= 1 \\
    \varphi_{f(m)}(n + 1) &= (n + 1)\varphi_m(n).
\end{align*}
$$

Since $f$ is (total) recursive, there is some $m_0$ such that $\varphi_{m_0} = \varphi_{f(m_0)}$, and for $m_0$ we get

$$
\begin{align*}
    \varphi_{m_0}(0) &= 1 \\
    \varphi_{m_0}(n + 1) &= (n + 1)\varphi_{m_0}(n),
\end{align*}
$$

so the partial recursive function $\varphi_{m_0}$ satisfies the recursive definition of factorial, which means that it is a fixed point of the recursive equations defining factorial. Since factorial is a total function, $\varphi_{m_0} = fact$, that is, factorial is a total computable function.

More generally, if a function $h$ (over $\mathbb{N}^k$) is defined in terms of recursive equations of the form

$$
h(z_1, \ldots, z_k) = t(h(y_1, \ldots, y_k))
$$

where $y_1, \ldots, y_k, z_1, \ldots, z_k$ are expressions in some variables $x_1, \ldots, x_k$ ranging over $\mathbb{N}$ and where $t$ is an expression containing recursive occurrences of $h$, if we can show that the equations

$$
g(m, z_1, \ldots, z_k) = t(\varphi_m(y_1, \ldots, y_k))
$$

define a partial computable function $g$, then we can use the above trick to put them in the form

$$
\varphi_{f(m)}(z_1, \ldots, z_k) = t(\varphi_m(y_1, \ldots, y_k)).
$$

for some computable function $f$. Such a formalism is described in detail in Chapter XI of Kleene I.M [21]. By the recursion theorem, there is some $m_0$ such that $\varphi_{m_0} = \varphi_{f(m_0)}$, so $\varphi_{m_0}$ satisfies the recursive equations

$$
\varphi_{m_0}(z_1, \ldots, z_k) = t(\varphi_{m_0}(y_1, \ldots, y_k)).
$$
and \( \varphi_{m_0} \) is a fixed point of these recursive equations. If we can show that \( \varphi_{m_0} \) is total, then we found the fixed point of this set of recursive equations and \( h = \varphi_{m_0} \) is a total computable function. If \( \varphi_{m_0} \) is a partial function, it is still a fixed point. However in general there is more than one fixed point and we don’t which one \( \varphi_{m_0} \) is (it could be the partial function undefined everywhere).

**Theorem 5.1.** *(Recursion Theorem, Version 1)* Let \( \varphi_0, \varphi_1, \ldots \) be any acceptable indexing of the partial computable functions. For every total computable function \( f \), there is some \( n \) such that

\[
\varphi_n = \varphi_f(n).
\]

**Proof.** Consider the function \( \theta \) defined such that

\[
\theta(x, y) = \varphi_{\text{univ}}(\varphi_{\text{univ}}(x, x), y) \quad \text{for all } x, y \in \mathbb{N}.
\]

The function \( \theta \) is partial computable, and there is some index \( j \) such that \( \varphi_j = \theta \). By the s-m-n Theorem, there is a computable function \( g \) such that

\[
\varphi_g(x)(y) = \theta(x, y).
\]

Consider the function \( f \circ g \). Since it is computable, there is some index \( m \) such that \( \varphi_m = f \circ g \). Let

\[
n = g(m).
\]

Since \( \varphi_m \) is total, \( \varphi_m(m) \) is defined, and we have

\[
\varphi_n(y) = \varphi_g(m)(y) = \theta(m, y) = \varphi_{\text{univ}}(\varphi_{\text{univ}}(m, m), y) = \varphi_{\varphi_{\text{univ}}(m, m)}(y)
\]

\[
= \varphi_{\varphi_m}(m)(y) = \varphi_{f \circ g(m)}(y) = \varphi_{f(g(m))}(y) = \varphi_{f(n)}(y),
\]

for all \( y \in \mathbb{N} \). Therefore, \( \varphi_n = \varphi_{f(n)} \), as desired. \( \square \)

The recursion theorem can be strengthened as follows.

**Theorem 5.2.** *(Recursion Theorem, Version 2)* Let \( \varphi_0, \varphi_1, \ldots \) be any acceptable indexing of the partial computable functions. There is a total computable function \( h \) such that for all \( x \in \mathbb{N} \), if \( \varphi_x \) is total, then

\[
\varphi_{\varphi_x(h(x))} = \varphi_{h(x)}.
\]

**Proof.** The computable function \( g \) obtained in the proof of Theorem 5.1 satisfies the condition

\[
\varphi_g(x) = \varphi_{\varphi_x(x)},
\]

and it has some index \( i \) such that \( \varphi_i = g \). Recall that \( c \) is a computable composition function such that

\[
\varphi_c(x, y) = \varphi_x \circ \varphi_y.
\]

It is easily verified that the function \( h \) defined such that

\[
h(x) = g(c(x, i)) \quad \text{for all } x \in \mathbb{N}
\]

does the job. \( \square \)
A third version of the recursion Theorem is given below.

**Theorem 5.3.** (Recursion Theorem, Version 3) For all \( n \geq 1 \), there is a total computable function \( h \) of \( n + 1 \) arguments, such that for all \( x \in \mathbb{N} \), if \( \varphi_x \) is a total computable function of \( n + 1 \) arguments, then

\[
\varphi_{\varphi_x}(h(x,x_1,\ldots,x_n),x_1,\ldots,x_n) = \varphi_x(h(x_1,\ldots,x_n)),
\]

for all \( x_1,\ldots,x_n \in \mathbb{N} \).

**Proof.** Let \( \theta \) be the function defined such that

\[
\theta(x,x_1,\ldots,x_n,y) = \varphi_{\varphi_x}(x,x_1,\ldots,x_n)(y) = \varphi_{\text{univ}}(\varphi_{\text{univ}}(x,x_1,\ldots,x_n),y)
\]

for all \( x, x_1, \ldots, x_n, y \in \mathbb{N} \). By the s-m-n Theorem, there is a computable function \( g \) such that

\[
\varphi_{\varphi_x}(x_1,\ldots,x_n) = \varphi_{\varphi_x}(x_1,\ldots,x_n).
\]

It is easily shown that there is a computable function \( c \) such that

\[
\varphi_{\varphi_c}(x_1,\ldots,x_n) = \varphi_c(\varphi_{\varphi_c}(x_1,\ldots,x_n),x_1,\ldots,x_n)
\]

for any two partial computable functions \( \varphi_i \) and \( \varphi_j \) (viewed as functions of \( n + 1 \) arguments) and all \( x, x_1, \ldots, x_n \in \mathbb{N} \). Let \( \varphi_i = g \), and define \( h \) such that

\[
h(x,x_1,\ldots,x_n) = g(c(x,i),x_1,\ldots,x_n),
\]

for all \( x, x_1, \ldots, x_n \in \mathbb{N} \). We have

\[
\varphi_{h(x_1,\ldots,x_n)} = \varphi_{g(c(x,i),x_1,\ldots,x_n)} = \varphi_{\varphi_c(x_1,\ldots,x_n)},
\]

and using the fact that \( \varphi_i = g \),

\[
\varphi_{\varphi_c(x_1,\ldots,x_n)} = \varphi_{\varphi_x(\varphi_i(c(x,i),x_1,\ldots,x_n),x_1,\ldots,x_n)},
\]

\[
= \varphi_{\varphi_x(g(c(x,i),x_1,\ldots,x_n),x_1,\ldots,x_n)},
\]

\[
= \varphi_{\varphi_x(h(x,x_1,\ldots,x_n),x_1,\ldots,x_n)}.
\]

As a first application of the recursion theorem, we can show that there is an index \( n \) such that \( \varphi_n \) is the constant function with output \( n \). Loosely speaking, \( \varphi_n \) prints its own name. Let \( f \) be the computable function such that

\[
f(x,y) = x
\]

for all \( x,y \in \mathbb{N} \). By the s-m-n Theorem, there is a computable function \( g \) such that

\[
\varphi_{g(x)}(y) = f(x,y) = x
\]
5.1. THE RECURSION THEOREM

for all $x, y \in \mathbb{N}$. By the Theorem 5.1, there is some $n$ such that

$$\varphi_{g(n)} = \varphi_n,$$

the constant function with value $n$.

As a second application, we get a very short proof of Rice’s theorem. Let $C$ be such that $P_C \neq \emptyset$ and $P_C \neq \mathbb{N}$, and let $j \in P_C$ and $k \in \mathbb{N} - P_C$. Define the function $f$ as follows:

$$f(x) = \begin{cases} j & \text{if } x \notin P_C, \\ k & \text{if } x \in P_C, \end{cases}$$

If $P_C$ is computable, then $f$ is computable. By the recursion theorem (Theorem 5.1), there is some $n$ such that

$$\varphi_{f(n)} = \varphi_n.$$

But then we have

$$n \in P_C \iff f(n) \notin P_C$$

by definition of $f$, and thus,

$$\varphi_{f(n)} \neq \varphi_n,$$

a contradiction. Hence, $P_C$ is not computable.

As a third application, we prove the following proposition.

**Proposition 5.4.** Let $C$ be a set of partial computable functions and let

$$A = \{x \in \mathbb{N} | \varphi_x \in C\}.$$

The set $A$ is not reducible to its complement $\overline{A}$.

**Proof.** Assume that $A \leq \overline{A}$. Then there is a computable function $f$ such that

$$x \in A \iff f(x) \in \overline{A}$$

for all $x \in \mathbb{N}$. By the recursion theorem, there is some $n$ such that

$$\varphi_{f(n)} = \varphi_n.$$

But then,

$$\varphi_n \in C \iff n \in A \iff f(n) \in \overline{A} \iff \varphi_{f(n)} \in \overline{C},$$

contradicting the fact that

$$\varphi_{f(n)} = \varphi_n.$$

\[\Box\]
The recursion theorem can also be used to show that functions defined by recursive
definitions other than primitive recursion are partial computable, as we discussed a the
beginning of this section. This is the case for the function known as Ackermann’s function
discussed in Section 1.9 and defined recursively as follows:

\begin{align*}
f(0, y) &= y + 1, \\
f(x + 1, 0) &= f(x, 1), \\
f(x + 1, y + 1) &= f(x, f(x + 1, y)).
\end{align*}

It can be shown that this function is not primitive recursive. Intuitively, it outgrows all
primitive recursive functions. However, \( f \) is computable, but this is not so obvious. We can
use the recursion theorem to prove that \( f \) is computable. Using the technique described at
the beginning of this section consider the following definition by cases:

\begin{align*}
g(n, 0, y) &= y + 1, \\
g(n, x + 1, 0) &= \varphi_{univ}(n, x, 1), \\
g(n, x + 1, y + 1) &= \varphi_{univ}(n, x, \varphi_{univ}(n, x + 1, y)).
\end{align*}

Clearly, \( g \) is partial computable. By the s-m-n Theorem, there is a computable function \( h \)
such that

\( \varphi_{h(n)}(x, y) = g(n, x, y). \)

The equations defining \( g \) yield

\begin{align*}
\varphi_{h(n)}(0, y) &= y + 1, \\
\varphi_{h(n)}(x + 1, 0) &= \varphi_{n}(x, 1), \\
\varphi_{h(n)}(x + 1, y + 1) &= \varphi_{n}(x, \varphi_{n}(x + 1, y)).
\end{align*}

By the recursion theorem, there is an \( m \) such that

\( \varphi_{h(m)} = \varphi_{m}. \)

Therefore, the partial computable function \( \varphi_{m}(x, y) \) satisfies the equations

\begin{align*}
\varphi_{m}(0, y) &= y + 1, \\
\varphi_{m}(x + 1, 0) &= \varphi_{m}(x, 1), \\
\varphi_{m}(x + 1, y + 1) &= \varphi_{m}(x, \varphi_{m}(x + 1, y))
\end{align*}

defining Ackermann’s function. We showed in Section 1.9 that \( \varphi_{m}(x, y) \) is a total function,
and thus, \( f = \varphi_{m} \) and Ackermann’s function is a total computable function.

Hence, the recursion theorem justifies the use of certain recursive definitions. However,
note that there are some recursive definitions that are only satisfied by the completely un-
defined function.

In the next section, we prove the extended Rice theorem.
5.2 Extended Rice Theorem

The extended Rice theorem characterizes the sets of partial computable functions $C$ such that $P_C$ is listable (c.e., r.e.). First, we need to discuss a way of indexing the partial computable functions that have a finite domain. Using the uniform projection function $\Pi$ (see Definition 2.3), we define the primitive recursive function $F$ such that

$$F(x, y) = \Pi(y + 1, \Pi_1(x) + 1, \Pi_2(x)).$$

We also define the sequence of partial functions $P_0, P_1, \ldots$ as follows:

$$P_x(y) = \begin{cases} F(x, y) - 1 & \text{if } 0 < F(x, y) \text{ and } y < \Pi_1(x) + 1, \\ \text{undefined} & \text{otherwise}. \end{cases}$$

**Proposition 5.5.** Every $P_x$ is a partial computable function with finite domain, and every partial computable function with finite domain is equal to some $P_x$.

The proof is left as an exercise. The easy part of the extended Rice theorem is the following lemma. Recall that given any two partial functions $f : A \rightarrow B$ and $g : A \rightarrow B$, we say that $g$ extends $f$ iff $f \subseteq g$, which means that $g(x)$ is defined whenever $f(x)$ is defined, and if so, $g(x) = f(x)$.

**Proposition 5.6.** Let $C$ be a set of partial computable functions. If there is a listable (c.e., r.e.) set $A$ such that $\varphi_x \in C$ iff there is some $y \in A$ such that $\varphi_x$ extends $P_y$, then $P_C = \{ x \mid \varphi_x \in C \}$ is listable (c.e., r.e.).

**Proof.** Proposition 5.6 can be restated as

$$P_C = \{ x \mid \exists y \in A, P_y \subseteq \varphi_x \}$$

is listable. If $A$ is empty, so is $P_C$, and $P_C$ is listable. Otherwise, let $f$ be a computable function such that $A = \text{range}(f)$.

Let $\psi$ be the following partial computable function:

$$\psi(z) = \begin{cases} \Pi_1(z) & \text{if } P_f(\Pi_2(z)) \subseteq \varphi_{\Pi_1(z)}, \\ \text{undefined} & \text{otherwise}. \end{cases}$$

It is clear that

$$P_C = \text{range}(\psi).$$

To see that $\psi$ is partial computable, write $\psi(z)$ as follows:

$$\psi(z) = \begin{cases} \Pi_1(z) & \text{if } \forall w \leq \Pi_1(f(\Pi_2(z))) \\
[F(f(\Pi_2(z)), w) > 0 \Rightarrow \varphi_{\Pi_1(z)}(w) = F(f(\Pi_2(z)), w) - 1], \\
\text{undefined} & \text{otherwise}. \end{cases}$$

This completes the proof. ☐
To establish the converse of Proposition 5.6, we need two propositions.

**Proposition 5.7.** If $P_C$ is listable (c.e., r.e.) and $\varphi \in C$, then there is some $P_y \subseteq \varphi$ such that $P_y \in C$.

*Proof.* Assume that $P_C$ is listable and that $\varphi \in C$. By an s-m-n construction, there is a computable function $g$ such that

$$
\varphi_g(x)(y) = \begin{cases} 
\varphi(y) & \text{if } \forall z \leq y [\neg T(x, x, z)], \\
\text{undefined} & \text{if } \exists z \leq y [T(x, x, z)],
\end{cases}
$$

for all $x, y \in \mathbb{N}$. Observe that if $x \in K$, then $\varphi_g(x)$ is a finite subfunction of $\varphi$, and if $x \in \overline{K}$, then $\varphi_g(x) = \varphi$. Assume that no finite subfunction of $\varphi$ is in $C$. Then

$$
x \in \overline{K} \iff g(x) \in P_C
$$

for all $x \in \mathbb{N}$, that is, $\overline{K} \leq P_C$. Since $P_C$ is listable, $\overline{K}$ would also be listable, a contradiction. \qed

As a corollary of Proposition 5.7, we note that TOTAL is not listable.

**Proposition 5.8.** If $P_C$ is listable (c.e., r.e.), $\varphi \in C$, and $\varphi \subseteq \psi$, where $\psi$ is a partial computable function, then $\psi \in C$.

*Proof.* Assume that $P_C$ is listable. We claim that there is a computable function $h$ such that

$$
\varphi_h(x)(y) = \begin{cases} 
\psi(y) & \text{if } x \in K, \\
\varphi(y) & \text{if } x \in \overline{K},
\end{cases}
$$

for all $x, y \in \mathbb{N}$. Assume that $\psi \notin C$. Then

$$
x \in \overline{K} \iff h(x) \in P_C
$$

for all $x \in \mathbb{N}$, that is, $\overline{K} \leq P_C$, a contradiction, since $P_C$ is listable. Therefore, $\psi \in C$. To find the function $h$ we proceed as follows: Let $\varphi = \varphi_j$ and define $\Theta$ such that

$$
\Theta(x, y, z) = \begin{cases} 
\varphi(y) & \text{if } T(j, y, z) \land \neg T(x, y, w), \text{ for } 0 \leq w < z \\
\psi(y) & \text{if } T(x, x, z) \land \neg T(j, y, w), \text{ for } 0 \leq w < z \\
\text{undefined} & \text{otherwise}.
\end{cases}
$$

Observe that if $x = y = j$, then $\Theta(j, j, z)$ is multiply defined, but since $\psi$ extends $\varphi$, we get the same value $\psi(y) = \varphi(y)$, so $\Theta$ is a well defined partial function. Clearly, for all $(m, n) \in \mathbb{N}^2$, there is at most one $z \in \mathbb{N}$ so that $\Theta(x, y, z)$ is defined, so the function $\sigma$ defined by

$$
\sigma(x, y) = \begin{cases} 
z & \text{if } (x, y, z) \in \text{dom}(\Theta) \\
\text{undefined} & \text{otherwise}
\end{cases}
$$
is a partial computable function. Finally, let
\[ \theta(x, y) = \Theta(x, y, \sigma(x, y)), \]
a partial computable function. It is easy to check that
\[ \theta(x, y) = \begin{cases} \psi(y) & \text{if } x \in K, \\ \varphi(y) & \text{if } x \in \overline{K}, \end{cases} \]
for all \( x, y \in \mathbb{N} \). By the s-m-n Theorem, there is a computable function \( h \) such that
\[ \varphi_{h(x)}(y) = \theta(x, y) \]
for all \( x, y \in \mathbb{N} \).

Observe that Proposition 5.8 yields a new proof that \( \overline{\text{TOTAL}} \) is not listable (not c.e., not r.e.). Finally we can prove the extended Rice theorem.

**Theorem 5.9.** *(Extended Rice Theorem)* The set \( P_C \) is listable (c.e., r.e.) iff there is a listable (c.e., r.e) set \( A \) such that
\[ \varphi_x \in C \iff \exists y \in A (P_y \subseteq \varphi_x). \]

**Proof.** Let \( P_C = \text{dom}(\varphi_i) \). Using the s-m-n Theorem, there is a computable function \( k \) such that
\[ \varphi_{k(y)} = P_y \text{ for all } y \in \mathbb{N}. \]
Define the listable set \( A \) such that
\[ A = \text{dom}(\varphi_i \circ k). \]
Then
\[ y \in A \iff \varphi_i(k(y)) \downarrow \iff P_y \in C. \]
Next, using Proposition 5.7 and Proposition 5.8, it is easy to see that
\[ \varphi_x \in C \iff \exists y \in A (P_y \subseteq \varphi_x). \]
Indeed, if \( \varphi_x \in C \), by Proposition 5.7, there is a finite subfunction \( P_y \subseteq \varphi_x \) such that \( P_y \in C \), but
\[ P_y \in C \iff y \in A, \]
as desired. On the other hand, if
\[ P_y \subseteq \varphi_x \]
for some \( y \in A \), then
\[ P_y \in C, \]
and by Proposition 5.8, since \( \varphi_x \) extends \( P_y \), we get
\[ \varphi_x \in C. \]
5.3 Creative and Productive Sets; Incompleteness in Arithmetic

In this section, we discuss some special sets that have important applications in logic: creative and productive sets. These notions were introduced by Post and Dekker (1944, 1955). The concepts to be described are illustrated by the following situation. Assume that

\[ W_x \subseteq K \]

for some \( x \in \mathbb{N} \) (recall that \( W_x \) was introduced in Definition 3.6). We claim that

\[ x \in K - W_x. \]

Indeed, if \( x \in W_x \), then \( \varphi_x(x) \) is defined, and by definition of \( K \), we get \( x \notin K \), a contradiction. Therefore, \( \varphi_x(x) \) must be undefined, that is,

\[ x \in K - W_x. \]

The above situation can be generalized as follows.

**Definition 5.1.** A set \( A \subseteq \mathbb{N} \) is **productive** iff there is a total computable function \( f \) such that for every listable set \( W_x \),

\[
\text{if } W_x \subseteq A \text{ then } f(x) \in A - W_x \text{ for all } x \in \mathbb{N}.
\]

The function \( f \) is called the **productive function of** \( A \). A set \( A \) is **creative** if it is listable (c.e., r.e.) and if its complement \( \overline{A} \) is productive.

As we just showed, \( K \) is creative and \( \overline{K} \) is productive. It is also easy to see that TOTAL is productive. But TOTAL is worse than \( K \), because by Proposition 3.20, TOTAL is not listable.

The following facts are immediate consequences of the definition.

1. A productive set is not listable (not c.e., not r.e.), since \( A \neq W_x \) for all listable sets \( W_x \) (the image of the productive function \( f \) is a subset of \( A - W_x \), which can't be empty since \( f \) is total).

2. A creative set is not computable (not recursive).

Productiveness is a technical way of saying that a nonlistable set \( A \) is not listable in a rather strong and constructive sense. Indeed, there is a computable function \( f \) such that no matter how we attempt to approximate \( A \) with a listable set \( W_x \subseteq A \), then \( f(x) \) is an element in \( A \) not in \( W_x \).
5.3. CREATING AND PRODUCTIVE SETS; INCOMPLETENESS

Remark: In Rogers [32] (Chapter 7, Section 3), the definition of a productive set only requires the productive function \( f \) to be *partial computable*. However, it is proven in Theorem XI of Rogers that this weaker requirement is equivalent to the stronger requirement of Definition 5.1.

Creative and productive sets arise in logic. The set of theorems of a logical theory is often creative. For example, the set of theorems in Peano’s arithmetic is creative, and the set of true sentences of Peano’s arithmetic is productive. This yields incompleteness results. We will return to this topic at the end of this section.

**Proposition 5.10.** If a set \( A \) is productive, then it has an infinite listable (c.e., r.e.) subset.

**Proof.** We first give an informal proof. Let \( f \) be the computable productive function of \( A \). We define a computable function \( g \) as follows: Let \( x_0 \) be an index for the empty set, and let

\[
g(0) = f(x_0).
\]

Assuming that \( \{g(0), g(1), \ldots, g(y)\} \) is known, let \( x_{y+1} \) be an index for this finite set, and let

\[
g(y + 1) = f(x_{y+1}).
\]

Since \( W_{x_{y+1}} \subseteq A \), we have \( f(x_{y+1}) \in A \).

For the formal proof, following Rogers [32] (Chapter 7, Section 7, Theorem X), we use the following facts whose proof is left as an exercise:

1. There is a computable function \( u \) such that

\[
W_{u(x,y)} = W_x \cup W_y.
\]

2. There is a computable function \( t \) such that

\[
W_{t(x)} = \{x\}.
\]

Letting \( x_0 \) be an index for the empty set, we define the function \( h \) as follows:

\[
h(0) = x_0,
\]

\[
h(y + 1) = u(t(f(y)), h(y)).
\]

We define \( g \) such that

\[
g = f \circ h.
\]

It is easily seen that \( g \) does the job. \( \square \)
Another important property of productive sets is the following.

**Proposition 5.11.** If a set $A$ is productive, then $\overline{K} \leq A$.

*Proof.* Let $f$ be a productive function for $A$. Using the s-m-n Theorem, we can find a computable function $h$ such that

$$W_{h(y,x)} = \begin{cases} \{f(y)\} & \text{if } x \in K, \\ \emptyset & \text{if } x \in \overline{K}. \end{cases}$$

The above can be restated as follows:

$$\varphi_{h(y,x)}(z) = \begin{cases} 1 & \text{if } x \in K \text{ and } z = f(y), \\ \text{undefined} & \text{if } x \in \overline{K}, \end{cases}$$

for all $x, y, z \in \mathbb{N}$. By the third version of the recursion theorem (Theorem 5.3), there is a computable function $g$ such that

$$W_{g(x)} = W_{h(g(x), x)} \quad \text{for all } x \in \mathbb{N}.$$ 

Let

$$k = f \circ g.$$

We claim that

$$x \in \overline{K} \text{ iff } k(x) \in A \quad \text{for all } x \in \mathbb{N}.$$ 

Substituting $g(x)$ for $y$ in the equation for $W_{h(y,x)}$ and using the fact that $W_{g(x)} = W_{h(g(x),x)}$ and $k(x) = f(g(x))$, we get

$$W_{g(x)} = \begin{cases} \{f(g(x))\} = \{k(x)\} & \text{if } x \in K, \\ \emptyset & \text{if } x \in \overline{K}. \end{cases}$$

Because $f$ is a productive function for $A$, if $x \in \overline{K}$, then $W_{g(x)} = \emptyset \subseteq A$, so $k(x) = f(g(x)) \in A$. Conversely, assume that $k(x) = f(g(x)) \in A$. If $x \in K$, then $W_{g(x)} = \{f(g(x))\}$, so $W_{g(x)} \subseteq A$, and since $f$ is a productive function for $A$, we have $f(g(x)) \in A - W_{g(x)} = A - \{f(g(x))\}$, a contradiction. Therefore, $x \notin \overline{K}$ and the reduction is achieved. Thus, $\overline{K} \leq A$. \ \ \ \Box$

Using Part (1) of Proposition 5.12 stated next we obtain the converse of Proposition 5.11. Thus a set $A$ is productive iff $\overline{K} \leq A$. This fact is recorded in the next proposition.

The following results can also be shown.

**Proposition 5.12.** The following facts hold.

1. If $A$ is productive and $A \leq B$, then $B$ is productive.
2. $A$ is creative iff $A$ is complete.
(3) A is creative iff \( A \) is equivalent to \( K \).

(4) A is productive iff \( K \leq A \).

Part (1) is easy to prove; see Rogers [32] (Chapter 7, Theorem V(b)). Part (2) is proven in Rogers [32] (Chapter 11, Corollary V). Part (3) follows from Part (2) since \( K \) is complete. Part (4) follows from Proposition 5.11 and Part (1).

We conclude with a discussion of the significance of the notions of productive and creative sets to logic. A more detailed discussion can be found in Rogers [32] (Chapter 7, Section 8). In Section ?? we discussed Peano arithmetic and the reader is invited to review it. It is convenient to add a countable set of constants \( 0, 1, 2, \ldots \), denoting the natural numbers to the language of arithmetic, and the new axioms

\[ S^n(0) = n, \quad n \in \mathbb{N}. \]

By a now fairly routine process (using a pairing function and an extended pairing function), it is possible to assign a Gödel number \( \#(A) \) to every first-order sentence \( A \) in the language of arithmetic; see Enderton [9] (Chapter III) or Kleene I.M. [21] (Chapter X). With some labor, it is possible to construct a formula \( F_x \) with one free variable \( x \) having the following property:

\[ n \in K \text{ iff } (F_n \text{ is true in } \mathbb{N}) \]
\[ n \notin K \text{ iff } (F_n \text{ is false in } \mathbb{N}) \text{ iff } (\neg F_x \text{ is true in } \mathbb{N}). \]

One should not underestimate the technical difficulty of this task. One of Gödel’s most original steps in proving his first incompleteness theorem was to define a variant of the formula \( F_x \). Later on, simpler proofs were given, but they are still very technical. The brave reader should attempt to solve Exercises 7.64 and 7.65 in Rogers [32].

Observe that the sentences \( F_n \) are special kinds of sentences of arithmetic but of course there are many more sentences of arithmetic. The following “basic lemma” from Rogers [32] (Chapter 7, Section 8) is easily shown.

**Proposition 5.13.** For any two subsets \( S \) and \( T \) of \( \mathbb{N} \), if \( T \) is listable and if \( S \cap T \) is productive, then \( S \) is productive. In particular, if \( T \) is computable and if \( S \cap T \) is productive, then \( S \) is productive.

With a slight abuse of notation, we say that a set \( T \) is sentences of arithmetic is computable (resp. listable) iff the set of Gödel numbers \( \#(A) \) of sentences \( A \) in \( T \) is computable (resp. listable). Then the following remarkable (historically shocking) facts hold.

**Theorem 5.14.** (Unaxiomatizability of arithmetic) The following facts hold.

(1) The set of sentences of arithmetic true in \( \mathbb{N} \) is a productive set. Consequently, the set of true sentences is not listable.
(2) The set of sentences of arithmetic false in \( \mathbb{N} \) is a productive set. Consequently, the set of false sentences is not listable.

Proof sketch. (1) It is easy to show that the set \( \{ \neg F_x \mid x \in \mathbb{N} \} \) is computable. Since
\[
\{ n \in \mathbb{N} \mid \neg F_n \text{ is true in } \mathbb{N} \} = \overline{K}
\]
is productive and
\[
\{ A \mid A \text{ is true in } \mathbb{N} \} \cap \{ \neg F_x \mid x \in \mathbb{N} \} = \{ \neg F_x \mid \neg F_x \text{ is true in } \mathbb{N} \}
\]
\[
= \{ \neg F_x \mid x \in \overline{K} \},
\]
by Proposition 5.13, the set \( \{ A \mid A \text{ is true in } \mathbb{N} \} \) is also productive.

(2) It is also easy to show that the set \( \{ F_x \mid x \in \mathbb{N} \} \) is computable. Since
\[
\{ n \in \mathbb{N} \mid F_n \text{ is false in } \mathbb{N} \} = \overline{K}
\]
is productive and
\[
\{ A \mid A \text{ is false in } \mathbb{N} \} \cap \{ F_x \mid x \in \mathbb{N} \} = \{ F_x \mid F_x \text{ is false in } \mathbb{N} \}
\]
\[
= \{ F_x \mid x \in \overline{K} \},
\]
by Proposition 5.13, the set \( \{ A \mid A \text{ is false in } \mathbb{N} \} \) is also productive. \( \square \)

Definition 5.2. A proof system for arithmetic is axiomatizable if the the set of provable sentences is listable.

Since the set of provable sentences of an axiomatizable proof system is listable, Theorem 5.14 annihilates any hope of finding an axiomatization of arithmetic. Theorem 5.14 also shows that it is impossible to decide effectively (algorithmically) whether a sentence of arithmetic is true. In fact the set of true sentences of arithmetic is not even listable.

If we consider proof systems for arithmetic, such as Peano arithmetic, then creative sets show up.

Definition 5.3. A proof system for arithmetic is sound if every provable sentence is true (in \( \mathbb{N} \)). A proof system is consistent if there is no sentence \( A \) such that both \( A \) and \( \neg A \) are provable.

Clearly, a sound proof system is consistent.

Assume that a proof system for arithmetic is sound and strong enough so that the formula \( F_x \) with the free variable \( x \) introduced just before Proposition 5.13 has the following properties:
\[
n \in K \text{ iff } (F_n \text{ is provable})
\]
\[
n \notin K \text{ iff } (F_n \text{ is not provable}).
\]
Peano arithmetic is such a proof system. Then we have the following theorem.
Theorem 5.15. (Undecidability of provability in arithmetic) Consider any axiomatizable proof system for arithmetic satisfying the hypotheses stated before the statement of the theorem. The following facts hold.

(1) The set of unprovable sentences of arithmetic is a productive set. Consequently, the set of unprovable sentences is not listable.

(2) The set of provable sentences of arithmetic is a creative set. Consequently, the set of provable sentences is not computable.

As a corollary of Theorem 5.15, there is no algorithm to decide whether a sentence of arithmetic is provable or not. But things are worse. Because the set of unprovable sentences of arithmetic is productive, there is a recursive function $f$, which for any attempt to find a listable subset $W$ of the nonprovable sentences of arithmetic, produces another nonprovable sentence not in $W$.

Theorem 5.15 also implies Gödel’s first incompleteness theorem. Indeed, it is immediately seen that the set $\{F_x \mid \neg F_x \text{ is provable}\}$ is listable (because $\{\neg F_x \mid x \in \mathbb{N}\}$ is computable and $\{A \mid A \text{ is provable}\}$ is listable). But since our proof system is assumed to be sound, $\neg F_x$ provable implies that $F_x$ is not provable, so by

$$n \notin K \iff (F_n \text{ is not provable}),$$

we have

$$\{x \in \mathbb{N} \mid \neg F_x \text{ is provable}\} \subseteq \{x \in \mathbb{N} \mid F_x \text{ is not provable}\} = \overline{K}.$$ 

Since $\overline{K}$ is productive and $\{x \in \mathbb{N} \mid \neg F_x \text{ is provable}\}$ is listable, we have

$$W_y = \{x \in \mathbb{N} \mid \neg F_x \text{ is provable}\}$$

for some $y$, and if $f$ is the productive function associated with $\overline{K}$, then for $x_0 = f(y)$ we have

$$F_{x_0} \in \{F_x \mid F_x \text{ is not provable}\} - \{\neg F_x \mid \neg F_x \text{ is provable}\},$$

that is, both $F_{x_0}$ and $\neg F_{x_0}$ are not provable. Furthermore, since

$$n \notin K \iff (F_n \text{ is not provable})$$

and

$$n \notin K \iff (F_n \text{ is false in } \mathbb{N})$$

we see that $F_{x_0}$ is false in $\mathbb{N}$, and so $\neg F_{x_0}$ is true in $\mathbb{N}$. In summary, we proved the following result.

Theorem 5.16. (Incompleteness in arithmetic (Gödel 1931)) Consider any axiomatizable proof system for arithmetic satisfying the hypotheses stated earlier. Then there exists a sentence $F$ of arithmetic ($F = \neg F_{x_0}$) such that neither $F$ nor $\neg F$ are provable. Furthermore, $F$ is true in $\mathbb{N}$. 
Theorem 5.15 holds under the weaker assumption that the proof system is consistent (as opposed to sound), and that there is a formula $G$ with one free variable $x$ such that

$$ n \in K \text{ iff } (G_n \text{ is provable}). $$

The formula $G$ is due to Rosser. The incompleteness theorem (Theorem 5.16) also holds under the weaker assumption of consistency, except for the statement that $F$ is true.

To summarize informally the above negative results:

1. No (effective) axiomatization of mathematics can exactly capture all true statements of arithmetic.

2. From any (effective) axiomatization which yields only true statements of arithmetic, a new true statement can be found not provable in that axiomatization.

Fact (2) is what inspired Post to use the term creative for the type of sets arising in Definition 5.1. Indeed, one has to be creative to capture truth in arithmetic.

Another (relatively painless) way to prove incompleteness results in arithmetic is to use Diophantine definability; see Section 6.4.
Chapter 6

Listable Sets and Diophantine Sets; Hilbert’s Tenth Problem

6.1 Diophantine Equations and Hilbert’s Tenth Problem

There is a deep and a priori unexpected connection between the theory of computable and listable sets and the solutions of polynomial equations involving polynomials in several variables with integer coefficients. These are polynomials in \( n \geq 1 \) variables \( x_1, \ldots, x_n \) which are finite sums of *monomials* of the form

\[
a x_1^{k_1} \cdots x_n^{k_n},
\]

where \( k_1, \ldots, k_n \in \mathbb{N} \) are nonnegative integers, and \( a \in \mathbb{Z} \) is an integer (possibly negative). The natural number \( k_1 + \cdots + k_n \) is called the *degree* of the monomial \( a x_1^{k_1} \cdots x_n^{k_n} \).

For example, if \( n = 3 \), then

1. 5, \(-7\), are monomials of degree 0.
2. 3\( x_1 \), \(-2x_2\), are monomials of degree 1.
3. \( x_1 x_2, 2x_1^2, 3x_1 x_3, -5x_2^3 \), are monomials of degree 2.
4. \( x_1 x_2 x_3, x_1^2 x_3, -x_3^3 \), are monomials of degree 3.
5. \( x_1^4, -x_1^2 x_3^2, x_1 x_2^2 x_3 \), are monomials of degree 4.

It is convenient to introduce multi-indices, where an *n-dimensional multi-index* is an \( n \)-tuple \( \alpha = (k_1, \ldots, k_n) \) with \( n \geq 1 \) and \( k_i \in \mathbb{N} \). Let \( |\alpha| = k_1 + \cdots + k_n \). Then we can write

\[
x^\alpha = x_1^{k_1} \cdots x_n^{k_n}.
\]

For example, for \( n = 3 \),

\[
x^{(1,2,1)} = x_1 x_2^2 x_3, \quad x^{(0,2,2)} = x_2^2 x_3^2.
\]
Definition 6.1. A *polynomial* $P(x_1, \ldots, x_n)$ in the variables $x_1, \ldots, x_n$ with integer coefficients is a finite sum of monomials of the form

$$P(x_1, \ldots, x_n) = \sum_\alpha a_\alpha x^\alpha,$$

where the $\alpha$'s are $n$-dimensional multi-indices, and with $a_\alpha \in \mathbb{Z}$. The maximum of the degrees $|\alpha|$ of the monomials $a_\alpha x^\alpha$ is called the *total degree* of the polynomial $P(x_1, \ldots, x_n)$.

The set of all such polynomials is denoted by $\mathbb{Z}[x_1, \ldots, x_n]$.

Sometimes, we write $P$ instead of $P(x_1, \ldots, x_n)$. We also use variables $x, y, z$ etc. instead of $x_1, x_2, x_3, \ldots$.

For example, $2x - 3y - 1$ is a polynomial of total degree 1, $x^2 + y^2 - z^2$ is a polynomial of total degree 2, and $x^3 + y^3 + z^3 - 29$ is a polynomial of total degree 3.

Mathematicians have been interested for a long time in the problem of solving equations of the form

$$P(x_1, \ldots, x_n) = 0,$$

with $P \in \mathbb{Z}[x_1, \ldots, x_n]$, seeking only *integer solutions* for $x_1, \ldots, x_n$.

Diophantus of Alexandria, a Greek mathematician of the 3rd century, was one of the first to investigate such equations. For this reason, seeking integer solutions of polynomials in $\mathbb{Z}[x_1, \ldots, x_n]$ is referred to as solving *Diophantine equations*.

This problem is not as simple as it looks. The equation

$$2x - 3y - 1 = 0$$

obviously has the solution $x = 2, y = 1$, and more generally $x = -1 + 3a, y = -1 + 2a$, for any integer $a \in \mathbb{Z}$.

The equation

$$x^2 + y^2 - z^2 = 0$$

has the solution $x = 3, y = 4, z = 5$, since $3^2 + 4^2 = 9 + 16 = 25 = 5^2$. More generally, the reader should check that

$$x = t^2 - 1, \; y = 2t, \; z = t^2 + 1$$

is a solution for all $t \in \mathbb{Z}$.

The equation

$$x^3 + y^3 + z^3 - 29 = 0$$

has the solution $x = 3, y = 1, z = 1$.

What about the equation

$$x^3 + y^3 + z^3 - 30 = 0?$$
Amazingly, the only known integer solution is

\[(x, y, z) = (-283059965, -2218888517, 2220422932),\]
discovered in 1999 by E. Pine, K. Yarbrough, W. Tarrant, and M. Beck, following an approach suggested by N. Elkies.

And what about solutions of the equation

\[x^3 + y^3 + z^3 - 33 = 0\]?

Well, nobody knows whether this equation is solvable in integers!

In 1900, at the International Congress of Mathematicians held in Paris, the famous mathematician David Hilbert presented a list of ten open mathematical problems. Soon after, Hilbert published a list of 23 problems. The tenth problem is this:

**Hilbert’s tenth problem (H10)**

Find an algorithm that solves the following problem:

Given as input a polynomial \( P \in \mathbb{Z}[x_1, \ldots, x_n] \) with integer coefficients, return YES or NO, according to whether there exist integers \( a_1, \ldots, a_n \in \mathbb{Z} \) so that \( P(a_1, \ldots, a_n) = 0 \); that is, the Diophantine equation \( P(x_1, \ldots, x_n) = 0 \) has a solution.

It is important to note that at the time Hilbert proposed his tenth problem, a rigorous mathematical definition of the notion of algorithm did not exist. In fact, the machinery needed to even define the notion of algorithm did not exist. It is only around 1930 that precise definitions of the notion of computability due to Turing, Church, and Kleene, were formulated, and soon after shown to be all equivalent.

So to be precise, the above statement of Hilbert’s tenth should say: find a RAM program (or equivalently a Turing machine) that solves the following problem: ...

In 1970, the following somewhat surprising resolution of Hilbert’s tenth problem was reached:

**Theorem** (Davis-Putnam-Robinson-Matiyasevich)

*Hilbert’s tenth problem is undecidable; that is, there is no algorithm for solving Hilbert’s tenth problem.*

In 1962, Davis, Putnam and Robinson had shown that if a fact known as *Julia Robinson hypothesis* could be proven, then Hilbert’s tenth problem would be undecidable. At the time, the Julia Robinson hypothesis seemed implausible to many, so it was a surprise when in 1970 Matiyasevich found a set satisfying the Julia Robinson hypothesis, thus completing the proof of the undecidability of Hilbert’s tenth problem. It is also a bit startling that Matiyasevich’ set involves the Fibonacci numbers.

A detailed account of the history of the proof of the undecidability of Hilbert’s tenth problem can be found in Martin Davis’ classical paper Davis [6].
Even though Hilbert’s tenth problem turned out to have a negative solution, the knowledge gained in developing the methods to prove this result is very significant. What was revealed is that polynomials have considerable expressive powers. This is what we discuss in the next section.

6.2 Diophantine Sets and Listable Sets

We begin by showing that if we can prove that the version of Hilbert’s tenth problem with solutions restricted to belong to \( \mathbb{N} \) is undecidable, then Hilbert’s tenth problem (with solutions in \( \mathbb{Z} \)) is undecidable.

**Proposition 6.1.** If we had an algorithm for solving Hilbert’s tenth problem (with solutions in \( \mathbb{Z} \)), then we would have an algorithm for solving Hilbert’s tenth problem with solutions restricted to belong to \( \mathbb{N} \) (that is, nonnegative integers).

**Proof.** The above statement is not at all obvious, although its proof is short with the help of some number theory. Indeed, by a theorem of Lagrange (Lagrange’s four square theorem), every natural number \( m \) can be represented as the sum of four squares,

\[
m = a_0^2 + a_1^2 + a_2^2 + a_3^2, \quad a_0, a_1, a_2, a_3 \in \mathbb{Z}.
\]

We reduce Hilbert’s tenth problem restricted to solutions in \( \mathbb{N} \) to Hilbert’s tenth problem (with solutions in \( \mathbb{Z} \)). Given a Diophantine equation \( P(x_1, \ldots, x_n) = 0 \), we can form the polynomial

\[
Q = P(u_1^2 + v_1^2 + y_1^2 + z_1^2, \ldots, u_n^2 + v_n^2 + y_n^2 + z_n^2)
\]

in the \( 4n \) variables \( u_i, v_i, y_i, z_i \) (1 \( \leq i \leq n \)) obtained by replacing \( x_i \) by \( u_i^2 + v_i^2 + y_i^2 + z_i^2 \) for \( i = 1, \ldots, n \). If \( Q = 0 \) has a solution \( (p_1, q_1, r_1, s_1, \ldots, p_n, q_n, r_n, s_n) \) with \( p_i, q_i, r_i, s_i \in \mathbb{Z} \), then if we set \( a_i = p_i^2 + q_i^2 + r_i^2 + s_i^2 \), obviously \( P(a_1, \ldots, a_n) = 0 \) with \( a_i \in \mathbb{N} \). Conversely, if \( P(a_1, \ldots, a_n) = 0 \) with \( a_i \in \mathbb{N} \), then by Lagrange’s theorem there exist some \( p_i, q_i, r_i, s_i \in \mathbb{Z} \) (in fact \( \mathbb{N} \)) such that \( a_i = p_i^2 + q_i^2 + r_i^2 + s_i^2 \) for \( i = 1, \ldots, n \), and the equation \( Q = 0 \) has the solution \( (p_1, q_1, r_1, s_1, \ldots, p_n, q_n, r_n, s_n) \) with \( p_i, q_i, r_i, s_i \in \mathbb{Z} \). Therefore \( Q = 0 \) has a solution \( (p_1, q_1, r_1, s_1, \ldots, p_n, q_n, r_n, s_n) \) with \( p_i, q_i, r_i, s_i \in \mathbb{Z} \) iff \( P = 0 \) has a solution \( (a_1, \ldots, a_n) \) with \( a_i \in \mathbb{N} \). If we had an algorithm to decide whether \( Q \) has a solution with its components in \( \mathbb{Z} \), then we would have an algorithm to decide whether \( P = 0 \) has a solution with its components in \( \mathbb{N} \).

As consequence, the contrapositive of Proposition 6.1 shows that if the version of Hilbert’s tenth problem restricted to solutions in \( \mathbb{N} \) is undecidable, so is Hilbert’s original problem (with solutions in \( \mathbb{Z} \)).

In fact, the Davis-Putnam-Robinson-Matiyasevich theorem establishes the undecidability of the version of Hilbert’s tenth problem restricted to solutions in \( \mathbb{N} \). From now on, we restrict our attention to this version of Hilbert’s tenth problem.
A key idea is to use Diophantine equations with parameters, to \textit{define} sets of numbers. For example, consider the polynomial

\[ P_1(a, y, z) = (y + 2)(z + 2) - a. \]

For \( a \in \mathbb{N} \) fixed, the equation \((y + 2)(z + 2) - a = 0\), equivalently

\[ a = (y + 2)(z + 2), \]

has a solution with \( y, z \in \mathbb{N} \) iff \( a \) is composite.

If we now consider the polynomial

\[ P_2(a, y, z) = y(2z + 3) - a, \]

for \( a \in \mathbb{N} \) fixed, the equation \( y(2z + 3) - a = 0 \), equivalently

\[ a = y(2z + 3), \]

has a solution with \( y, z \in \mathbb{N} \) iff \( a \) is not a power of 2.

For a slightly more complicated example, consider the polynomial

\[ P_3(a, y) = 3y + 1 - a^2. \]

We leave it as an exercise to show that the natural numbers \( a \) that satisfy the equation \( 3y + 1 - a^2 = 0 \), equivalently

\[ a^2 = 3y + 1, \]

or \((a - 1)(a + 1) = 3y\), are of the form \( a = 3k + 1 \) or \( a = 3k + 2 \), for any \( k \in \mathbb{N} \).

In the first case, if we let \( S_1 \) be the set of composite natural numbers, then we can write

\[ S_1 = \{ a \in \mathbb{N} \mid (\exists y, z)((y + 2)(z + 2) - a = 0) \}, \]

where it is understood that the existentially quantified variables \( y, z \) take their values in \( \mathbb{N} \).

In the second case, if we let \( S_2 \) be the set of natural numbers that are not powers of 2, then we can write

\[ S_2 = \{ a \in \mathbb{N} \mid (\exists y, z)(y(2z + 3) - a = 0) \}. \]

In the third case, if we let \( S_3 \) be the set of natural numbers that are congruent to 1 or 2 modulo 3, then we can write

\[ S_3 = \{ a \in \mathbb{N} \mid (\exists y)(3y + 1 - a^2 = 0) \}. \]

A more explicit Diophantine definition for \( S_3 \) is

\[ S_3 = \{ a \in \mathbb{N} \mid (\exists y)((a - 3y - 1)(a - 3y - 2) = 0) \}. \]

The natural generalization is as follows.
Definition 6.2. A set \( S \subseteq \mathbb{N} \) of natural numbers is Diophantine (or Diophantine definable) if there is a polynomial \( P(a, x_1, \ldots, x_n) \in \mathbb{Z}[a, x_1, \ldots, x_n] \), with \( n \geq 0 \), such that
\[
S = \{ a \in \mathbb{N} \mid (\exists x_1, \ldots, x_n)(P(a, x_1, \ldots, x_n) = 0) \},
\]
where it is understood that the existentially quantified variables \( x_1, \ldots, x_n \) take their values in \( \mathbb{N} \). More generally, a relation \( R \subseteq \mathbb{N}^m \) is Diophantine (\( m \geq 2 \)) if there is a polynomial \( P(a_1, \ldots, a_m, x_1, \ldots, x_n) \in \mathbb{Z}[a_1, \ldots, a_m, x_1, \ldots, x_n] \), with \( n \geq 0 \), such that
\[
R = \{(a_1, \ldots, a_m) \in \mathbb{N}^m \mid (\exists x_1, \ldots, x_n)(P(a_1, \ldots, a_m, x_1, \ldots, x_n) = 0) \},
\]
where it is understood that the existentially quantified variables \( x_1, \ldots, x_n \) take their values in \( \mathbb{N} \).

For example, the strict order relation \( a_1 < a_2 \) is defined as follows:
\[
a_1 < a_2 \quad \text{iff} \quad (\exists x)(a_1 + 1 + x - a_2 = 0),
\]
and the divisibility relation \( a_1 \mid a_2 \) (\( a_1 \) divides \( a_2 \)) is defined as follows:
\[
a_1 \mid a_2 \quad \text{iff} \quad (\exists x)(a_1 x - a_2 = 0).
\]

What about the ternary relation \( R \subseteq \mathbb{N}^3 \) given by
\[(a_1, a_2, a_3) \in R \quad \text{iff} \quad a_1 \mid a_2 \quad \text{and} \quad a_1 < a_3^2 \]

At first glance it is not obvious how to “convert” a conjunction of Diophantine definitions into a single Diophantine definition, but we can do this using the following trick: given any finite number of Diophantine equations in the variables \( x_1, \ldots, x_n \),
\[
P_1 = 0, \ P_2 = 0, \ldots, \ P_m = 0,
\]
observe that (*) has a solution \( (a_1, \ldots, a_n) \), which means that \( P_i(a_1, \ldots, a_n) = 0 \) for \( i = 1, \ldots, m \), iff the single equation
\[
P_1^2 + P_2^2 + \cdots + P_m^2 = 0 \quad (**)\]
also has the solution \( (a_1, \ldots, a_n) \). This is because, since \( P_1^2, P_2^2, \ldots, P_m^2 \) are all nonnegative, their sum is equal to zero iff they are all equal to zero, that is \( P_i^2 = 0 \) for \( i = 1 \ldots, m \), which is equivalent to \( P_i = 0 \) for \( i = 1 \ldots, m \).

Using this trick, we see that
\[(a_1, a_2, a_3) \in R \quad \text{iff} \quad (\exists u, v)((a_1 u - a_2)^2 + (a_1 + 1 + v - a_3)^2 = 0).
\]

We can also define the notion of Diophantine function.

\(^1\)We have to allow \( n = 0 \). Otherwise singleton sets would not be Diophantine.
Definition 6.3. A function $f : \mathbb{N}^n \rightarrow \mathbb{N}$ is Diophantine iff its graph $\{(a_1, \ldots, a_n, b) \subseteq \mathbb{N}^{n+1} | b = f(a_1, \ldots, a_n)\}$ is Diophantine.

For example, the pairing function $J$ and the projection functions $K, L$ due to Cantor introduced in Section 2.1 are Diophantine, since

- $z = J(x, y)$ iff $(x + y - 1)(x + y) + 2x - 2z = 0$
- $x = K(z)$ iff $(\exists y)((x + y - 1)(x + y) + 2x - 2z = 0)$
- $y = L(z)$ iff $(\exists x)((x + y - 1)(x + y) + 2x - 2z = 0)$.

How extensive is the family of Diophantine sets? The remarkable fact proved by Davis-Putnam-Robinson-Matiyasevich is that they coincide with the listable sets (the recursively enumerable sets). This is a highly nontrivial result.

The easy direction is the following result.

Proposition 6.2. Every Diophantine set is listable (recursively enumerable).

Proof sketch. Suppose $S$ is given as

$$S = \{ a \in \mathbb{N} | (\exists x_1, \ldots, x_n)(P(a, x_1, \ldots, x_n) = 0) \},$$

Using the extended pairing function $\langle x_1, \ldots, x_n \rangle$ of Section 2.1, we enumerate all $n$-tuples $(x_1, \ldots, x_n) \in \mathbb{N}^n$, and during this process we compute $P(a, x_1, \ldots, x_n)$. If $P(a, x_1, \ldots, x_n)$ is zero, then we output $a$, else we go on. This way, $S$ is the range of a computable function, and it is listable. \qed

It is also easy to see that every Diophantine function is partial computable. The main theorem of the theory of Diophantine sets is the following deep result.

Theorem 6.3. (Davis-Putnam-Robinson-Matiyasevich, 1970) Every listable subset of $\mathbb{N}$ is Diophantine. Every partial computable function is Diophantine.

Theorem 6.3 is often referred to as the DPRM theorem. A complete proof of Theorem 6.3 is provided in Davis [6]. As noted by Davis, although the proof is certainly long and nontrivial, it only uses elementary facts of number theory, nothing more sophisticated than the Chinese remainder theorem. Nevertheless, the proof is a tour de force.

One of the most difficult steps is to show that the exponential function $h(n, k) = n^k$ is Diophantine. This is done using the Pell equation. According to Martin Davis, the proof given in Davis [6] uses a combination of ideas from Matiyasevich and Julia Robinson. Matiyasevich’s proof used the Fibonacci numbers.

Using some results from the theory of computation it is now easy to deduce that Hilbert’s tenth problem is undecidable. To achieve this, recall that there are listable sets that are not
computable. For example, it is shown in Lemma 3.11 that $K = \{x \in \mathbb{N} \mid \varphi_x(x) \text{ is defined}\}$ is listable but not computable. Since $K$ is listable, by Theorem 6.3, it is defined by some Diophantine equation

$$P(a, x_1, \ldots, x_n) = 0,$$

which means that

$$K = \{a \in \mathbb{N} \mid (\exists x_1, \ldots, x_n)(P(a, x_1, \ldots, x_n) = 0)\}.$$

We have the following strong form of the undecidability of Hilbert’s tenth problem, in the sense that it shows that Hilbert’s tenth problem is already undecidable for a fixed Diophantine equation in one parameter.

**Theorem 6.4.** There is no algorithm which takes as input the polynomial $P(a, x_1, \ldots, x_n)$ defining $K$ and any natural number $a \in \mathbb{N}$ and decides whether

$$P(a, x_1, \ldots, x_n) = 0.$$

Consequently, Hilbert’s tenth problem is undecidable.

**Proof.** If there was such an algorithm, then $K$ would be decidable, a contradiction.

Any algorithm for solving Hilbert’s tenth problem could be used to decide whether or not $P(a, x_1, \ldots, x_n) = 0$, but we just showed that there is no such algorithm. \(\square\)

It is an open problem whether Hilbert’s tenth problem is undecidable if we allow rational solutions (that is, $x_1, \ldots, x_n \in \mathbb{Q}$).

Alexandra Shlapentokh proved that various extensions of Hilbert’s tenth problem are undecidable. These results deal with some algebraic number theory beyond the scope of these notes. Incidentally, Alexandra was an undergraduate at Penn, and she worked on a logic project for me (finding a Gentzen system for a subset of temporal logic).

Having now settled once and for all the undecidability of Hilbert’s tenth problem, we now briefly explore some interesting consequences of Theorem 6.3.

### 6.3 Some Applications of the DPRM Theorem

The first application of the DPRM theorem is a particularly striking way of defining the listable subsets of $\mathbb{N}$ as the nonnegative ranges of polynomials with integer coefficients. This result is due to Hilary Putnam.

**Theorem 6.5.** For every listable subset $S$ of $\mathbb{N}$, there is some polynomial $Q(x, x_1, \ldots, x_n)$ with integer coefficients such that

$$S = \{Q(a, b_1, \ldots, b_n) \mid Q(a, b_1, \ldots, b_n) \in \mathbb{N}, a, b_1, \ldots, b_n \in \mathbb{N}\}.$$
Proof. By the DPRM theorem (Theorem 6.3), there is some polynomial \( P(x, x_1, \ldots, x_n) \) with integer coefficients such that
\[
S = \{ a \in \mathbb{N} | (\exists x_1, \ldots, x_n)(P(a, x_1, \ldots, x_n) = 0) \}.
\]
Let \( Q(x, x_1, \ldots, x_n) \) be given by
\[
Q(x, x_1, \ldots, x_n) = (x + 1)(1 - P^2(x, x_1, \ldots, x_n)) - 1.
\]
We claim that \( Q \) satisfies the statement of the theorem. If \( a \in S \), then \( P(a, b_1, \ldots, b_n) = 0 \) for some \( b_1, \ldots, b_n \in \mathbb{N} \), so
\[
Q(a, b_1, \ldots, b_n) = (a + 1)(1 - 0) - 1 = a.
\]
This shows that all \( a \in S \) show up the the nonnegative range of \( Q \). Conversely, assume that \( Q(a, b_1, \ldots, b_n) \geq 0 \) for some \( a, b_1, \ldots, b_n \in \mathbb{N} \). Then by definition of \( Q \) we must have
\[
(a + 1)(1 - P^2(a, b_1, \ldots, b_n)) - 1 \geq 0,
\]
that is,
\[
(a + 1)(1 - P^2(a, b_1, \ldots, b_n)) \geq 1,
\]
and since \( a \in \mathbb{N} \), this implies that \( P^2(a, b_1, \ldots, b_n) < 1 \), but since \( P \) is a polynomial with integer coefficients and \( a, b_1, \ldots, b_n \in \mathbb{N} \), the expression \( P^2(a, b_1, \ldots, b_n) \) must be a nonnegative integer, so we must have
\[
P(a, b_1, \ldots, b_n) = 0,
\]
which shows that \( a \in S \).

Remark: It should be noted that in general, the polynomials \( Q \) arising in Theorem 6.5 may take on negative integer values, and to obtain all listable sets, we must restrict ourself to their nonnegative range.

As an example, the set \( S_3 \) of natural numbers that are congruent to 1 or 2 modulo 3 is given by
\[
S_3 = \{ a \in \mathbb{N} | (\exists y)(3y + 1 - a^2 = 0) \}.
\]
so by Theorem 6.5, \( S_3 \) is the nonnegative range of the polynomial
\[
Q(x, y) = (x + 1)(1 - (3y + 1 - x^2)^2)) - 1
\]
\[
= -(x + 1)((3y - x^2)^2 + 2(3y - x^2)) - 1
\]
\[
= (x + 1)(x^2 - 3y)(2 - (x^2 - 3y)) - 1.
\]
Observe that \( Q(x, y) \) takes on negative values. For example, \( Q(0, 0) = -1 \). Also, in order for \( Q(x, y) \) to be nonnegative, \( (x^2 - 3y)(2 - (x^2 - 3y)) \) must be positive, but this can only happen if \( x^2 - 3y = 1 \), that is, \( x^2 = 3y + 1 \), which is the original equation defining \( S_3 \).
There is no miracle. The nonnegativity of \( Q(x, x_1, \ldots, x_n) \) must subsume the solvability of the equation \( P(x, x_1, \ldots, x_n) = 0 \).

A particularly interesting listable set is the set of primes. By Theorem 6.5, in theory, the set of primes is the positive range of some polynomial with integer coefficients.

Remarkably, some explicit polynomials have been found. This is a nontrivial task. In particular, the process involves showing that the exponential function is definable, which was the stumbling block to the completion of the DPRM theorem for many years.

To give the reader an idea of how the proof begins, observe by the Bezout identity, if \( p = s + 1 \) and \( q = s! \), then we can assert that \( p \) and \( q \) are relatively prime (\( \gcd(p, q) = 1 \)) as the fact that the Diophantine equation

\[
ap - bq = 1
\]

is satisfied for some \( a, b \in \mathbb{N} \). Then it is not hard to see that \( p \in \mathbb{N} \) is prime iff the following set of equations has a solution for \( a, b, s, r, q \in \mathbb{N} \):

\[
p = s + 1
\]
\[
p = r + 2
\]
\[
q = s!
\]
\[
ap - bq = 1.
\]

The problem with the above is that the equation \( q = s! \) is not Diophantine. The next step is to show that the factorial function is Diophantine, and this involves a lot of work. One way to proceed is to show that the above system is equivalent to a system allowing the use of the exponential function. The final step is to show that the exponential function can be eliminated in favor of polynomial equations.

We refer the interested reader to the remarkable expository paper by Davis, Matiyasevich and Robinson [7] for details. Here is a polynomial of total degree 25 in 26 variables (due to J. Jones, D. Sato, H. Wada, D. Wiens) which produces the primes as its positive range:

\[
(k + 2) \left[ 1 - ([wz + h + j - q]^2 + [(gk + 2g + k + 1)(h + j) + h - z]^2 + [16(k + 1)^3(k + 2)(n + 1)^2 + 1 - f]^2 + [2n + p + q + z - e]^2 + [e^3(e + 2)(a + 1)^2 + 1 - o]^2 + [(a^2 - 1)y^2 + 1 - x]^2 + [16r^2y^4(a^2 - 1) + 1 - u]^2 + ((a + u^2(u^2 - a))^2 - 1)(n + 4dy)^2 + 1 - (x + cu)^2]^2 + [(a^2 - 1)t^2 + 1 - m]^2 + [ai + k + 1 - l - i]^2 + [n + l + v - y]^2 + [p + l(a - n - 1) + b(2an + 2a - n^2 - 2n - 2) - m]^2 + [q + y(a - p - 1) + s(2ap + 2a - p^2 - 2p - 2) - x]^2 + [z + pl(a - p) + t(2ap - p^2 - 1) - pm]^2 \right].
\]
Around 2004, Nachi Gupta, an undergraduate student at Penn, and I tried to produce the prime 2 as one of the values of the positive range of the above polynomial. It turns out that this leads to values of the variables that are so large that we never succeeded!

Other interesting applications of the DPRM theorem are the re-statements of famous open problems, such as the Riemann hypothesis, as the unsolvability of certain Diophantine equations. For all this, see Davis, Matiyasevich and Robinson [7]. One may also obtain a nice variant of Gödel’s incompleteness theorem.

### 6.4 Gödel's Incompleteness Theorem

Gödel published his famous incompleteness theorem in 1931. At the time, his result rocked the mathematical world, and certainly the community of logicians.

In order to understand why his result had such impact one needs to step back in time. In the late 1800’s, Hilbert had advanced the thesis that it should be possible to completely formalize mathematics in such a way that every true statement should be provable “mechanically.” In modern terminology, Hilbert believed that one could design a theorem prover that should be complete. His quest is known as Hilbert’s program. In order to achieve his goal, Hilbert was led to investigate the notion of proof, and with some collaborators including Ackerman, Hilbert developed a significant amount of what is known as proof theory. When the young Gödel announced his incompleteness theorem, Hilbert’s program came to an abrupt halt. Even the quest for a complete proof system for arithmetic was impossible.

It should be noted that when Gödel proved his incompleteness theorem, computability theory basically did not exist, so Gödel had to start from scratch. His proof is really a tour de force. Gödel’s theorem also triggered extensive research on the notion of computability and undecidability between 1931 and 1936, the major players being Church, Gödel himself, Herbrand, Kleene, Rosser, Turing, and Post.

In this section we will give a (deceptively) short proof that relies on the DPRM and the existence of universal functions. The proof is short because the hard work lies in the proof of the DPRM!

The first step is to translate the fact that there is a universal partial computable function $\varphi_{\text{univ}}$ (see Proposition 2.6), such that for all $x, y \in \mathbb{N}$, if $\varphi_x$ is the $x$th partial computable function, then

$$\varphi_x(y) = \varphi_{\text{univ}}(x, y).$$

Also recall from Definition 3.6 that for any acceptable indexing of the partial computable functions, the listable (c.e. r.e.) sets $W_x$ are given by

$$W_x = \text{dom}(\varphi_x), \quad x \in \mathbb{N}.$$ 

Since $\varphi_{\text{univ}}$ is a partial computable function, it can be converted into a Diophantine equation so that we have the following result.
Theorem 6.6. (Universal Equation Theorem) There is a Diophantine equation
\[ U(m, a, x_1, \ldots, x_\nu) = 0 \]
such that for every listable (c.e., r.e.) set \( W_m \) \( (m \in \mathbb{N}) \) we have
\[ a \in W_m \iff (\exists x_1, \ldots, x_\nu)(U(m, a, x_1, \ldots, x_\nu) = 0). \]

Proof. We have
\[ W_m = \{ a \in \mathbb{N} \mid (\exists x_1)(\varphi_{\text{univ}}(m, a) = x_1) \}, \]
and since \( \varphi_{\text{univ}} \) is partial computable, by the DPRM (Theorem 6.3), there is Diophantine
polynomial \( U(m, a, x_1, \ldots, x_\nu) \) such that
\[ x_1 = \varphi_{\text{univ}}(m, a) \iff (\exists x_2, \ldots, x_\nu)(U(m, a, x_1, \ldots, x_\nu) = 0), \]
and so
\[ W_m = \{ a \in \mathbb{N} \mid (\exists x_1, \ldots, x_\nu)(U(m, a, x_1, \ldots, x_\nu) = 0) \}, \]
as claimed. \( \square \)

The Diophantine equation \( U(m, a, x_1, \ldots, x_\nu) = 0 \) is called a \textit{universal Diophantine equation}.
It is customary to denote \( U(m, a, x_1, \ldots, x_\nu) \) by \( P_m(a, x_1, \ldots, x_\nu) \).

Gödel’s incompleteness theorem applies to sets of logical (first-order) formulae of arithmetic built from the mathematical symbols \( 0, S, +, \cdot, < \) and the logical connectives \( \land, \lor, \neg, \Rightarrow, =, \forall, \exists \). Recall that logical equivalence, \( \equiv \), is defined by
\[ P \equiv Q \iff (P \Rightarrow Q) \land (Q \Rightarrow P). \]
The term
\[ S(S(\ldots S(0) \ldots)) \]
under the notation \( S^n(0) \), and represents the natural number \( n \).

For example,
\[ \exists x(S(S(0))) < (S(S(0)) + x), \]
\[ \exists x \exists y \exists z((x > 0) \land (y > 0) \land (z > 0) \land ((x \cdot x + y \cdot y) = z \cdot z)), \]
and
\[ \forall x \forall y \forall z((x > 0) \land (y > 0) \land (z > 0) \Rightarrow \neg((x \cdot x \cdot x \cdot x + y \cdot y \cdot y \cdot y) = z \cdot z \cdot z \cdot z)) \]
are formulae in the language of arithmetic. All three are true. The first formula is satisfied
by \( x = S(S(0)) \), the second by \( x = S^3(0), y = S^4(0) \) and \( z = S^5(0) \) (since \( 3^3 + 4^2 = 9 + 16 = 25 = 5^2 \)),
and the third formula asserts a special case of Fermat’s famous theorem: for every \( n \geq 3 \),
the equation \( x^n + y^n = z^n \) has no solution with \( x, y, z \in \mathbb{N} \) and \( x > 0, y > 0, z > 0 \).
The third formula corresponds to \( n = 4 \). Even for this case, the proof is hard.

To be completely rigorous we should explain precisely what is a formal proof. Roughly
speaking, a proof system consists of axioms and inference rule. A proof is a certain kind of
6.4. Gödel’s Incompleteness Theorem

A tree whose nodes are labeled with formulae, and this tree is constructed in such a way that for every node some inference rule is applied. Proof systems are discussed in Chapter ?? and in more detail in Chapter ?? The reader is invited to review this material. Such proof systems are also presented in Gallier [15, 14].

Given a polynomial $P(x_1, \ldots, x_m)$ in $\mathbb{Z}[x_1, \ldots, x_m]$, we need a way to “prove” that some natural numbers $n_1, \ldots, n_m \in \mathbb{N}$ are a solution of the Diophantine equation

$$P(x_1, \ldots, x_m) = 0,$$

which means that we need to have enough formulae of arithmetic to allow us to simplify the expression $P(n_1, \ldots, n_m)$ and check whether or not it is equal to zero.

For example, if $P(x, y) = 2x - 3y - 1$, we have the solution $x = 2$ and $y = 1$. What we do is to group all monomials with positive signs, $2x$, and all monomials with negative signs, $3y + 1$, plug in the values for $x$ and $y$, simplify using the arithmetic tables for $+$ and $\cdot$, and then compare the results. If they are equal, then we proved that the equation has a solution.

In our language, $x = S^2(0)$, $2x = S^2(0) \cdot x$, and $y = S^1(0)$, $3y + 1 = S^3(0) \cdot y + S(0)$. We need to simplify the expressions

$$2x = S^2(0) \cdot S^2(0) \quad \text{and} \quad 3y + 1 = S^3(0) \cdot S(0) + S(0).$$

Using the formulae

$$S^m(0) + S^n(0) = S^{m+n}(0)$$
$$S^m(0) \cdot S^n(0) = S^{mn}(0)$$
$$S^m(0) < S^n(0) \quad \text{iff} \quad m < n,$$

with $m, n \in \mathbb{N}$, we simplify $S^2(0) \cdot S^2(0)$ to $S^4(0)$, $S^3(0) \cdot S(0) + S(0)$ to $S^4(0)$, and we see that the results are equal.

In general, given a polynomial $P(x_1, \ldots, x_m)$ in $\mathbb{Z}[x_1, \ldots, x_m]$, we write it as

$$P(x_1, \ldots, x_m) = P_{\text{pos}}(x_1, \ldots, x_m) - P_{\text{neg}}(x_1, \ldots, x_m),$$

where $P_{\text{pos}}(x_1, \ldots, x_m)$ consists of the monomials with positive coefficients, and $-P_{\text{neg}}(x_1, \ldots, x_m)$ consists of the monomials with negative coefficients. Next we plug in $S^{n_1}(0), \ldots, S^{n_m}(0)$ in $P_{\text{pos}}(x_1, \ldots, x_m)$, and evaluate using the formulae for the addition and multiplication tables obtaining a term of the form $S^p(0)$. Similarly, we plug in $S^{n_1}(0), \ldots, S^{n_m}(0)$ in $P_{\text{neg}}(x_1, \ldots, x_m)$, and evaluate using the formulae for the addition and multiplication tables obtaining a term of the form $S^q(0)$. Then we test in parallel (using the above formulae) whether

$$S^p(0) = S^q(0), \quad \text{or} \quad S^p(0) < S^q(0), \quad \text{or} \quad S^q(0) < S^p(0).$$

Only one outcome is possible, and we obtain a proof that either $P(n_1, \ldots, n_m) = 0$ or $P(n_1, \ldots, n_m) \neq 0$. 
A more economical way that does use not an infinite number of formulae expressing the addition and multiplication tables is to use various axiomatizations of arithmetic.

One axiomatization known as Robinson arithmetic (R. M. Robinson (1950)) consists of the following seven axioms:

\[
\begin{align*}
\forall x & \neg (S(x) = 0) \\
\forall x \forall y ((S(x) = S(y)) \Rightarrow (x = y)) \\
\forall y ((y = 0) \lor \exists x (S(x) = y)) \\
\forall x (x + 0 = x) \\
\forall x \forall y (x + S(y) = S(x + y)) \\
\forall x (x \cdot 0 = 0) \\
\forall x \forall y (x \cdot S(y) = x \cdot y + x).
\end{align*}
\]

Peano arithmetic is obtained from Robinson arithmetic by adding a rule schema expressing induction:

\[
[\varphi(0) \land \forall n (\varphi(n) \Rightarrow \varphi(n + 1))] \Rightarrow \forall m \varphi(m),
\]

where \(\varphi(x)\) is any (first-order) formula of arithmetic. To deal with \(<\), we also have the axiom

\[
\forall x \forall y (x < y \equiv \exists z (S(z) + x = y)).
\]

It is easy to prove that the formulae

\[
\begin{align*}
S^m(0) + S^n(0) &= S^{m+n}(0) \\
S^m(0) \cdot S^n(0) &= S^{mn}(0) \\
S^m(0) &< S^n(0) \text{ iff } m < n,
\end{align*}
\]

are provable in Robinson arithmetic, and thus in Peano arithmetic (with \(m, n \in \mathbb{N}\)).

Gödel’s incompleteness applies to sets \(\mathcal{A}\) of formulae of arithmetic that are “nice” and strong enough. A set \(\mathcal{A}\) of formulae is nice if it is listable and consistent (see Definition 5.3), which means that it is impossible to prove \(\varphi\) and \(\neg \varphi\) from \(\mathcal{A}\) for some formula \(\varphi\). In other words, \(\mathcal{A}\) is free of contradictions.

Since the axioms of Peano arithmetic are obviously true statements about \(\mathbb{N}\) and since the induction principle holds for \(\mathbb{N}\), the set of all formulae provable in Robinson arithmetic and in Peano arithmetic is consistent.

As in Section 5.3, it is possible to assign a Gödel number \(\#(A)\) to every first-order sentence \(A\) in the language of arithmetic; see Enderton [9] (Chapter III) or Kleene I.M. [21] (Chapter X). With a slight abuse of notation, we say that a set \(T\) is sentences of arithmetic is computable (resp. listable) iff the set of Gödel numbers \(\#(A)\) of sentences \(A\) in \(T\) is computable (resp. listable). It can be shown that the set of all formulae provable in Robinson arithmetic and in Peano arithmetic are listable.

Here is a rather strong version of Gödel’s incompleteness from Davis, Matiyasevich and Robinson [7].
Theorem 6.7. (Gödel’s Incompleteness Theorem) Let $A$ be a set of formulae of arithmetic satisfying the following properties:

(a) The set $A$ is consistent.

(b) The set $A$ is listable (c.e., r.e.)

(c) The set $A$ is strong enough to prove all formulae

\[ S^m(0) + S^n(0) = S^{m+n}(0) \]
\[ S^m(0) \cdot S^n(0) = S^{mn}(0) \]
\[ S^m(0) < S^n(0) \quad \text{iff} \quad m < n, \]

for all $m, n \in \mathbb{N}$.

Then we can construct a Diophantine equation $F(x_1, \ldots, x_\nu) = 0$ corresponding to $A$ such that $F(x_1, \ldots, x_\nu) = 0$ has no solution with $x_1, \ldots, x_\nu \in \mathbb{N}$ but the formula

\[ \neg(\exists x_1, \ldots, x_\nu)(F(x_1, \ldots, x_\nu) = 0) \]  

is not provable from $A$. In other words, there is a true statement of arithmetic not provable from $A$; that is, $A$ is incomplete.

Proof. Define the subset $A \subseteq \mathbb{N}$ as follows:

\[ A = \{ a \in \mathbb{N} | \neg(\exists x_1, \ldots, x_\nu)(P_a(a, x_1, \ldots, x_\nu) = 0) \text{ is provable from } A \}, \]

where $P_m(a, x_1, \ldots, x_\nu)$ is defined just after Theorem 6.6. Because by (b) $A$ is listable, it is easy to see (because the set of formulae provable from a listable set is listable) that $A$ is listable, so by the DPRM $A$ is Diophantine, and by Theorem 6.6, there is some $k \in \mathbb{N}$ such that

\[ A = W_k = \{ a \in \mathbb{N} | (\exists x_1, \ldots, x_\nu)(P_k(a, x_1, \ldots, x_\nu) = 0) \} \]

The trick is now to see whether $k \notin W_k$ or not.

We claim that $k \notin W_k$.

We proceed by contradiction. Assume that $k \in W_k$. This means that

\[ (\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0), \]  

(\dagger_1)

and since $A = W_k$, by (**)\text{, that}

\[ \neg(\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0) \text{ is provable from } A. \]  

(\dagger_2)

By (\dagger_1) and (c), since the equation $P_k(k, x_1, \ldots, x_\nu) = 0$ has a solution, we can prove the formula

\[ (\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0) \]
from $\mathcal{A}$. By $\dagger_2$, the formula $\neg(\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0)$ is provable from $\mathcal{A}$, but since $(\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0)$ is also provable from $\mathcal{A}$, this contradicts the fact that $\mathcal{A}$ is consistent (which is hypothesis (a)).

Therefore we must have $k \not\in W_k$. This means that $P_k(k, x_1, \ldots, x_\nu) = 0$ has no solution with $x_1, \ldots, x_\nu \in \mathbb{N}$, and since $A = W_k$, the formula

\[ \neg(\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0) \]

is not provable from $\mathcal{A}$. 

**Remark:** Going back to the proof of Theorem 5.16, observe that $A$ plays the role of $\{F_x | \neg F_x \text{ is provable}\}$, that $k$ plays the role of $x_0$, and that the fact that

\[ \neg(\exists x_1, \ldots, x_\nu)(P_k(k, x_1, \ldots, x_\nu) = 0) \]

is not provable from $\mathcal{A}$ corresponds to $\neg F_{x_0}$ being true.

As a corollary of Theorem 6.7, since the theorems provable in Robinson arithmetic satisfy (a), (b), (c), we deduce that there are true theorems of arithmetic not provable in Robinson arithmetic; in short, Robinson arithmetic is incomplete. Since Robinson arithmetic does not have induction axioms, this shows that induction is not the culprit behind incompleteness. Since Peano arithmetic is an extension (consistent) of Robinson arithmetic, it is also incomplete. This is Gödel’s original incompleteness theorem, but Gödel had to develop from scratch the tools needed to prove his result, so his proof is very different (and a tour de force).

But the situation is even more dramatic. Adding a true unprovable statement to a set $\mathcal{A}$ satisfying (a), (b), (c) preserves properties (a), (b), (c), so there is no escape from incompleteness (unless perhaps we allow unreasonable sets of formulae violating (b)). The reader should compare this situation with the results given by Theorem 5.14 and Theorem 5.15.

Gödel’s incompleteness theorem is a negative result, in the sense that it shows that there is no hope of obtaining proof systems capable of proving all true statements for various mathematical theories such as arithmetic. We can also view Gödel’s incompleteness theorem positively as evidence that mathematicians will never be replaced by computers! There is always room for creativity.

The true but unprovable formulae arising in Gödel’s incompleteness theorem are rather contrived and by no means “natural.” For many years after Gödel’s proof was published logicians looked for natural incompleteness phenomena. In the early 1980’s such results were found, starting with a result of Kirby and Paris. Harvey Friedman then found more spectacular instances of natural incompleteness, one of which involves a finite miniaturization of Kruskal’s tree theorem. The proof of such results uses some deep methods of proof theory involving a tool known as ordinal notations. A survey of such results can be found in Gallier [10].
Chapter 7

The Post Correspondence Problem; Applications to Undecidability Results

7.1 The Post Correspondence Problem

The Post correspondence problem (due to Emil Post) is another undecidable problem that turns out to be a very helpful tool for proving problems in logic or in formal language theory to be undecidable.

**Definition 7.1.** Let \( \Sigma \) be an alphabet with at least two letters. An instance of the *Post Correspondence problem* (for short, PCP) is given by two nonempty sequences \( U = (u_1, \ldots, u_m) \) and \( V = (v_1, \ldots, v_m) \) of strings \( u_i, v_i \in \Sigma^* \). Equivalently, an instance of the PCP is a sequence of pairs \( (u_1, v_1), \ldots, (u_m, v_m) \).

The problem is to find whether there is a (finite) sequence \( (i_1, \ldots, i_p) \), with \( i_j \in \{1, \ldots, m\} \) for \( j = 1, \ldots, p \), so that
\[
 u_{i_1} u_{i_2} \cdots u_{i_p} = v_{i_1} v_{i_2} \cdots v_{i_p}.
\]

**Example 7.1.** Consider the following problem:

\[
(abab, ababaa), (aaabbb, bb), (aab, baab), (ba, baa), (ab, ba), (aa, a).
\]

There is a solution for the string 123456:

\[
abab\ aaabbb\ aab\ ba\ ab\ ab\ aa = ababaa\ bb\ baab\ baa\ ba\ ba\ a.
\]

If you are not convinced that this is a hard problem, try solving the following instance of the PCP:

\[
\{(aab, a), (ab, abb), (ab, bab), (ba, aab)\}.
\]

The shortest solution is a sequence of length 66.

We are beginning to suspect that this is a hard problem. Indeed, it is undecidable!
Theorem 7.1. (Emil Post, 1946) The Post correspondence problem is undecidable, provided that the alphabet $\Sigma$ has at least two symbols.

There are several ways of proving Theorem 7.1, but the strategy is more or less the same: reduce the halting problem to the PCP, by encoding sequences of ID’s as partial solutions of the PCP. In Machtey and Young [25] (Section 2.6), the undecidability of the PCP is shown by demonstrating how to simulate the computation of a Turing machine as a sequence of ID’s. We give a proof involving special kinds of RAM programs (called Post machines in Manna [26]), which is an adaptation of a proof due to Dana Scott presented in Manna [26] (Section 1.5.4, Theorem 1.8).

Proof. The first step of the proof is to show that a RAM program with $p \geq 2$ registers can be simulated by a RAM program using a single register. The main idea of the simulation is that by using the instructions add, tail, and jmp, it is possible to perform cyclic permutations on the string held by a register.

First we can also assume that RAM programs only uses instructions of the form

\begin{align*}
(1_j) & \quad N \quad \text{add}_j \quad X \\
(2) & \quad N \quad \text{tail} \quad X \\
(6,a) & \quad N \quad X \quad \text{jmp}_j \quad N1a \\
(6,b) & \quad N \quad X \quad \text{jmp}_j \quad N1b \\
(7) & \quad N \quad \text{continue}
\end{align*}

We can simulate $p \geq 2$ registers with a single register, by encoding the contents $r_1, \ldots, r_p$ of the $p$ registers as the string

$$r_1#r_2# \cdots #r_p,$$

using a single marker #. For instance, if $p = 2$, the effect of the instruction add$_b$ on register R1 is achieved as follows: Assuming that the initial contents are

$$aab#baba$$

using cyclic permutations (also inserting or deleting # whenever necessary), we get

$$aab#baba$$

$$ab#baba#a$$

$$b#baba#aa$$

$$baba#aab$$

$$baba#aabb \quad \text{(add b on the right)}$$

$$aba#aabb#b$$

$$ba#aabb#ba$$

$$a#aabb#bab$$

$$aabb#baba$$
Similarly, the effect of the instruction \texttt{tail} on register $R2$ is achieved as follows

\begin{align*}
\texttt{aab}\#\texttt{baba} \\
\texttt{ab}\#\texttt{baba}\#\texttt{a} \\
\texttt{b}\#\texttt{baba}\#\texttt{aa} \\
\texttt{baba}\#\texttt{aab} \\
\texttt{aba}\#\texttt{aab} \quad \text{(delete the leftmost letter)} \\
\texttt{ba}\#\texttt{aab}\#\texttt{a} \\
\texttt{a}\#\texttt{aab}\#\texttt{ab} \\
\texttt{aab}\#\texttt{aba}
\end{align*}

Since the halting problem for RAM programs is undecidable and since every RAM program can be simulated by another RAM with a single register, the halting problem for RAM programs with a single register is undecidable.

The second step of the proof is to reduce the halting problem for RAM programs with one register to the PCP (over an alphabet with at least two symbols).

Recall that $\Sigma = \{a_1, \ldots, a_k\}$. First, it is easily shown that every RAM program $P$ (with a single register $X$) is equivalent to a RAM program $P'$ such that all instructions are labeled with distinct line numbers, and such that there is only one occurrence of the instruction \texttt{continue} at the end of the program.

In order to obtain a reasonably simple reduction of the halting problem for RAM programs with a single register to the PCP, we modify the jump instructions as follows: the new instruction

\begin{equation*}
\texttt{Jmp } N_1, \ldots, N_k, N_{k+1}
\end{equation*}

tests whether $\text{head}(X) = a_j$, with $1 \leq j \leq k$. Since there is a single register $X$, it is omitted in the instruction \texttt{Jmp}. If $\text{head}(X) = a_j$, then jump to the instruction labeled $N_j$ \textit{and} perform the \texttt{tail} instruction so that $X = \text{tail}(X)$, otherwise if $X = \epsilon$ (which implies that $j = k + 1$), jump to the instruction labeled $N_{k+1}$. The instruction \texttt{tail} is eliminated. We leave it as an exercise to show how to simulate the new instruction

\begin{equation*}
\texttt{Jmp } N_1, \ldots, N_k, N_{k+1}
\end{equation*}

using the instructions \texttt{tail}, \texttt{jmp}_j \texttt{Na} and \texttt{jmp}_j \texttt{Nb}, and vice-versa. From now on we will use the second version using the instructions

\begin{equation*}
\texttt{Jmp } N_1, \ldots, N_k, N_{k+1}.
\end{equation*}

For the purpose of deciding whether a RAM program terminates, we may assume without loss of generality that we deal with programs that clear the register $X$ when they halt. In
fact, by adding an extra symbol # to the alphabet (which now has \( k + 1 \) symbols), we may also assume that in every instruction

\[
\text{Jmp } N_1, \ldots, N_{k+1}, N_{k+2},
\]

\( N_{k+2} \) is the line number of the last instruction in the RAM program, which must be a \texttt{continue}. This implies that the program clears the register \( X \) before it halts. We can execute the instruction \texttt{add}_{k+1} X \textit{at the very beginning of the program and perform an add}_{k+1} X \textit{after each tail instruction to make sure that in the new program the register X always has # as its rightmost symbol. When the original program performs an instruction}

\[
\text{Jmp } N_1, \ldots, N_{k+1}, N_{k+2}
\]

with \( X = \epsilon \), the new program performs the instruction

\[
\text{Jmp } N_1, \ldots, N_{k+2}, N_1.
\]

Since \( X \) is never empty during execution of the new program, the line number \( N_1 \) is irrelevant. Finally, when the original program halts, the new program clears the register \( X \) and then jumps to the last \texttt{continue}. We leave the details as an exercise.

From now on, we assume that \( \Sigma = \{a_1, \ldots, a_k, \#\} \). Given a RAM program \( P \) satisfying all the restrictions described above, we construct an instance of the PCP as follows. Assume that \( P \) has \( q \) lines numbered \( N_1, \ldots, N_q \). The alphabet \( \Delta \) of the PCP is \( \Delta = \Sigma \cup \{\ast, N_0, N_1, \ldots, N_q\} \). Indeed, the construction requires one more line number \( N_0 \) to force a solution of the PCP to start with some specified pair.

The lists \( U \) and \( V \) are constructed so that given any nonempty input \( x = x_1 \cdots x_m \) (with \( x_i \in \Sigma \)), the only possible \( U \)-lists \( u \) and \( V \)-lists \( v \) that could lead to a solution are of the form

\[
u = N_0 w_0 \ast N_1 w_1 \ast \cdots \ast N_{i_{n-1}} w_{n-1} \ast N_{i_n}
\]

and

\[
v = N_0 w_0 \ast N_1 w_1 \ast \cdots \ast N_{i_{n-1}} w_{n-1} \ast N_{i_n} w_n \ast N_{i_{n+1}} \ast
\]

where each \( w_i \) is of the form

\[
w_i = \ast w_{i,1} \ast \cdots \ast w_{i,n_i}
\]

or \( w_i = \epsilon \),

with

\[
w_0 = \ast x_1 \ast x_2 \ast \cdots \ast x_m,
\]

where \( w_{i,j} \in \Sigma, 1 \leq j \leq n_i, 1 \leq i \leq n \).

The sequence \( N_1, \ldots, N_{i_{n+1}} \) is the sequence of line numbers of the instructions executed by the RAM program \( P \) after \( n \) steps, started on input \( x \), and \( w_j \) is the value of the (single) register \( X \) just after executing the \( j \)th step, \textit{i.e.}, the instruction at line number \( N_{i_j} \). We make sure that the \( V \)-list is always ahead of the \( U \)-list by one instruction.
The lists $U$ and $V$ are defined according to the following rules. Rather than defining $U$ and $V$ explicitly, we define the pairs $(u_i, v_i)$, where $u_i \in U$ and $v_i \in V$.

To get started: we have the initial pair
\[(N_0, N_0 \ast x_1 \ast x_2 \ast \cdots \ast x_m \ast N_1 \ast).\]

To simulate an instruction $N_i \text{ add}_j X$, create the pair
\[(\ast N_i, a_j \ast N_{i+1} \ast), \quad \text{for all } a \in \Sigma.\]

To simulate an instruction of the form $N_i \text{ Jmp } N_1, \ldots, N_{k+1}, N_{k+2}$, create the pairs
\[(\ast N_i \ast a_j, N_j \ast), \quad 1 \leq j \leq k + 1,\]
and
\[(\ast N_i \ast N_q, N_q).\]

To build up the register contents, we need pairs
\[(\ast a, a \ast), \quad \text{for all } a \in \Sigma.\]

Note that we used the alphabet $\Delta = \Sigma \cup \{\ast, N_0, N_1, \ldots, N_q\}$, which uses more than 2 symbols in general. Let us finish our reduction for an instance of the PCP over the alphabet $\Delta$. Then after this construction is finished we will explain how to convert the instance of the PCP that we obtained to an instance of the PCP over a two-symbol alphabet.

The pairs of the PCP are designed so that the only possible $U$-lists $u$ and $V$-lists $v$ that could lead to a solution are of the form
\[u = N_0 w_0 \ast N_1 w_1 \ast \cdots \ast N_{i_{n-1}} w_{n-1} \ast N_{i_n}\]
and
\[v = N_0 w_0 \ast N_1 w_1 \ast \cdots \ast N_{i_{n-1}} w_{n-1} \ast N_{i_n} w_n \ast N_{i_{n+1}} \ast,\]
where each $w_i$ is of the form
\[w_i = \ast w_{i,1} \ast \cdots \ast w_{i,n_i} \quad \text{or} \quad w_i = \epsilon,\]
with
\[w_0 = \ast x_1 \ast x_2 \ast \cdots \ast x_m,\]
where $w_{i,j} \in \Sigma$, $1 \leq j \leq n_i$, $1 \leq i \leq n$, and where $v$ is an encoding of $n$ steps of the computation of the RAM program $P$ on input $x = x_1 \cdots x_m$, and $u$ lags behind $v$ by one step.

For example, let us see how the $U$-list and the $V$-list are updated, assuming that $N_{in}$ is the following instruction:

\[ N_{in} \quad \text{add}_b \quad X \]

Just after execution of the instruction at line number $N_{in}$, we have

\[ u = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} \]

and

\[ v = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} w_{n-1} * N_{in} * . \]

Since $w_{n-1} = * w_{n-1,1} * \cdots * w_{n-1,n-1}$, using the pairs

\[ (* w_{n-1,1}, w_{n-1,1} *), (* w_{n-1,2}, w_{n-1,2} *), \ldots, (* w_{n-1,n-1}, w_{n-1,n-1} *), \]

we get

\[ u = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} w_{n-1} \]

and

\[ v = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} w_{n-1} * N_{in} w_{n-1} * . \]

Next we use the pair

\[ (* N_{in}, b * N_{in+1} *) \]

simulating $\text{add}_b$, and we get

\[ u = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} w_{n-1} * N_{in} \]

and

\[ v = N_0 w_0 * N_1 w_1 * \cdots * N_{in-1} w_{n-1} * N_{in} w_{n-1} * b * N_{in+1} * . \]

Observe that the only chance for getting a solution of the PCP is to start with the pairs involving $N_0$. It is easy to see that the PCP constructed from $P$ has a solution iff $P$ halts on input $x$. However, the halting problem for RAM’s with a single register is undecidable, and thus, the PCP over the alphabet $\Delta$ is also undecidable.

It remains to show that we can recode the instance of the PCP that we obtained over the alphabet $\Delta = \Sigma \cup \{*, N_0, N_1, \ldots, N_q\}$ as an instance of the PCP over the alphabet $\{a_1, *\}$. To achieve this, we recode each symbol $a_i$ in $\Sigma$ as $*a_1^i$ (with $a_{k+1} = \#$) and each $N_j$ as $*a_1^{k+j+2}$. This way, we are only using the alphabet $\Delta = \{a_1, *\}$. We need the second character $\ast$, whose purpose is to avoid trivial solutions of the form

\[(u, u)\].
This could happen if we had used pairs \((a, a)\) to build up the register. Then we substitute \(a_i^1\) for \(a_i\) and \(a_i^{k+j+2}\) for \(N_j\) in the pairs that we created. Observe that the pairs \((a, a)^*\) become pairs involving longer strings. It is easy to see that the original PCP over \(\Delta\) has a solution iff the new PCP over \(\{a_1, *\}\) has a solution, so the PCP over two-letter alphabet is undecidable.

In the next two sections we present some undecidability results for context-free grammars and context-free languages.

### 7.2 Some Undecidability Results for CFG’s

**Theorem 7.2.** It is undecidable whether a context-free grammar is ambiguous.

**Proof.** We reduce the PCP to the ambiguity problem for CFG’s. Given any instance \(U = (u_1, \ldots, u_m)\) and \(V = (v_1, \ldots, v_m)\) of the PCP, let \(c_1, \ldots, c_m\) be \(m\) new symbols, and consider the following languages:

\[
L_U = \left\{ u_{i_1} \cdots u_{i_p} c_{i_p} \cdots c_{i_1} \mid 1 \leq i_j \leq m, \\
1 \leq j \leq p, p \geq 1 \right\},
\]

\[
L_V = \left\{ v_{i_1} \cdots v_{i_p} c_{i_p} \cdots c_{i_1} \mid 1 \leq i_j \leq m, \\
1 \leq j \leq p, p \geq 1 \right\},
\]

and \(L_{U,V} = L_U \cup L_V\).

We can easily construct a CFG, \(G_{U,V}\), generating \(L_{U,V}\). The productions are:

\[
S \rightarrow S_U \\
S \rightarrow S_V \\
S_U \rightarrow u_i S_U c_i \\
S_U \rightarrow u_i c_i \\
S_V \rightarrow v_i S_V c_i \\
S_V \rightarrow v_i c_i.
\]

It is easily seen that the PCP for \((U, V)\) has a solution iff \(L_U \cap L_V \neq \emptyset\) iff \(G\) is ambiguous.

**Remark:** As a corollary, we also obtain the following result: it is undecidable for arbitrary context-free grammars \(G_1\) and \(G_2\) whether \(L(G_1) \cap L(G_2) = \emptyset\) (see also Theorem 7.4).

Recall that the computations of a Turing Machine, \(M\), can be described in terms of instantaneous descriptions, \(upav\).
We can encode computations

\[ ID_0 \vdash ID_1 \vdash \cdots \vdash ID_n \]

halting in a proper ID, as the language, \( L_M \), consisting all of strings

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k} \# w_{2k+1}^R, \]

or

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k-2} \# w_{2k-1}^R \# w_{2k}, \]

where \( k \geq 0, w_0 \) is a starting ID, \( w_i \vdash w_{i+1} \) for all \( i \) with \( 0 \leq i < 2k + 1 \) and \( w_{2k+1} \) is proper halting ID in the first case, \( 0 \leq i < 2k \) and \( w_{2k} \) is proper halting ID in the second case.

The language \( L_M \) turns out to be the intersection of two context-free languages \( L_M^0 \) and \( L_M^1 \) defined as follows:

1. The strings in \( L_M^0 \) are of the form

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k} \# w_{2k+1}^R \]

or

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k-2} \# w_{2k-1}^R \# w_{2k}, \]

where \( w_{2i} \vdash w_{2i+1} \) for all \( i \geq 0 \), and \( w_{2k} \) is a proper halting ID in the second case.

2. The strings in \( L_M^1 \) are of the form

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k} \# w_{2k+1}^R \]

or

\[ w_0 \# w_1^R \# w_2 \# w_3^R \# \cdots \# w_{2k-2} \# w_{2k-1}^R \# w_{2k}, \]

where \( w_{2i+1} \vdash w_{2i+2} \) for all \( i \geq 0 \), \( w_0 \) is a starting ID, and \( w_{2k+1} \) is a proper halting ID in the first case.

**Theorem 7.3.** Given any Turing machine \( M \), the languages \( L_M^0 \) and \( L_M^1 \) are context-free, and \( L_M = L_M^0 \cap L_M^1 \).

**Proof.** We can construct PDA’s accepting \( L_M^0 \) and \( L_M^1 \). It is easily checked that \( L_M = L_M^0 \cap L_M^1 \). \( \square \)

As a corollary, we obtain the following undecidability result:

**Theorem 7.4.** It is undecidable for arbitrary context-free grammars \( G_1 \) and \( G_2 \) whether \( L(G_1) \cap L(G_2) = \emptyset \).
Proof. We can reduce the problem of deciding whether a partial recursive function is undefined everywhere to the above problem. By Rice’s theorem, the first problem is undecidable.

However, this problem is equivalent to deciding whether a Turing machine never halts in a proper ID. By Theorem 7.3, the languages $L^0_M$ and $L^1_M$ are context-free. Thus, we can construct context-free grammars $G_1$ and $G_2$ so that $L^0_M = L(G_1)$ and $L^1_M = L(G_2)$. Then $M$ never halts in a proper ID iff $L_M = \emptyset$ iff (by Theorem 7.3), $L_M = L(G_1) \cap L(G_2) = \emptyset$. \qed

Given a Turing machine $M$, the language $L_M$ is defined over the alphabet $\Delta = \Gamma \cup Q \cup \{$. The following fact is also useful to prove undecidability:

**Theorem 7.5.** Given any Turing machine $M$, the language $\Delta^* - L_M$ is context-free.

*Proof.* One can easily check that the conditions for not belonging to $L_M$ can be checked by a PDA. \qed

As a corollary, we obtain:

**Theorem 7.6.** Given any context-free grammar, $G = (V, \Sigma, P, S)$, it is undecidable whether $L(G) = \Sigma^*$.

*Proof.* We can reduce the problem of deciding whether a Turing machine never halts in a proper ID to the above problem.

Indeed, given $M$, by Theorem 7.5, the language $\Delta^* - L_M$ is context-free. Thus, there is a CFG, $G$, so that $L(G) = \Delta^* - L_M$. However, $M$ never halts in a proper ID iff $L_M = \emptyset$ iff $L(G) = \Delta^*$. \qed

As a consequence, we also obtain the following:

**Theorem 7.7.** Given any two context-free grammar, $G_1$ and $G_2$, and any regular language, $R$, the following facts hold:

1. $L(G_1) = L(G_2)$ is undecidable.
2. $L(G_1) \subseteq L(G_2)$ is undecidable.
3. $L(G_1) = R$ is undecidable.
4. $R \subseteq L(G_2)$ is undecidable.

In contrast to (4), the property $L(G_1) \subseteq R$ is decidable!
7.3 More Undecidable Properties of Languages; Greibach’s Theorem

We discuss a nice theorem of S. Greibach, which is a sort of version of Rice’s theorem for families of languages.

Let $\mathcal{L}$ be a countable family of languages. We assume that there is a coding function $c: \mathcal{L} \rightarrow \mathbb{N}$ and that this function can be extended to code the regular languages (all alphabets are subsets of some given countably infinite set).

We also assume that $\mathcal{L}$ is effectively closed under union, and concatenation with the regular languages.

This means that given any two languages $L_1$ and $L_2$ in $\mathcal{L}$, we have $L_1 \cup L_2 \in \mathcal{L}$, and $c(L_1 \cup L_2)$ is given by a recursive function of $c(L_1)$ and $c(L_2)$, and that for every regular language $R$, we have $L_1R \in \mathcal{L}$, $RL_1 \in \mathcal{L}$, and $c(RL_1)$ and $c(L_1R)$ are recursive functions of $c(R)$ and $c(L_1)$.

Given any language, $L \subseteq \Sigma^*$, and any string, $w \in \Sigma^*$, we define $L/w$ by

$$L/w = \{u \in \Sigma^* \mid uw \in L\}.$$ 

**Theorem 7.8.** (Greibach) Let $\mathcal{L}$ be a countable family of languages that is effectively closed under union and concatenation with the regular languages, and assume that the problem $L = \Sigma^*$ is undecidable for $L \in \mathcal{L}$ and any given sufficiently large alphabet $\Sigma$. Let $P$ be any nontrivial property of languages that is true for the regular languages, so that if $P(L)$ holds for any $L \in \mathcal{L}$, then $P(L/a)$ also holds for any letter $a$. Then $P$ is undecidable for $\mathcal{L}$.

**Proof.** Since $P$ is nontrivial for $\mathcal{L}$, there is some $L_0 \in \mathcal{L}$ so that $P(L_0)$ is false.

Let $\Sigma$ be large enough, so that $L_0 \subseteq \Sigma^*$, and the problem $L = \Sigma^*$ is undecidable for $L \in \mathcal{L}$.

We show that given any $L \in \mathcal{L}$, with $L \subseteq \Sigma^*$, we can construct a language $L_1 \in \mathcal{L}$, so that $L = \Sigma^*$ iff $P(L_1)$ holds. Thus, the problem $L = \Sigma^*$ for $L \in \mathcal{L}$ reduces to property $P$ for $\mathcal{L}$, and since for $\Sigma$ big enough, the first problem is undecidable, so is the second.

For any $L \in \mathcal{L}$, with $L \subseteq \Sigma^*$, let

$$L_1 = L_0 \# \Sigma^* \cup \Sigma^* \# L.$$ 

Since $\mathcal{L}$ is effectively closed under union and concatenation with the regular languages, we have $L_1 \in \mathcal{L}$.

If $L = \Sigma^*$, then $L_1 = \Sigma^* \# \Sigma^*$, a regular language, and thus, $P(L_1)$ holds, since $P$ holds for the regular languages.

Conversely, we would like to prove that if $L \neq \Sigma^*$, then $P(L_1)$ is false.
Since \( L \neq \Sigma^* \), there is some \( w \notin L \). But then,
\[
L_1/\#w = L_0.
\]
Since \( P \) is preserved under quotient by a single letter, by a trivial induction, if \( P(L_1) \) holds, then \( P(L_0) \) also holds. However, \( P(L_0) \) is false, so \( P(L_1) \) must be false.

Thus, we proved that \( L = \Sigma^* \) iff \( P(L_1) \) holds, as claimed. \( \Box \)

Greibach’s theorem can be used to show that it is undecidable whether a context-free grammar generates a regular language.

It can also be used to show that it is undecidable whether a context-free language is inherently ambiguous.

### 7.4 Undecidability of Validity in First-Order Logic

The PCP can also be used to give a quick proof of Church’s famous result stating that validity in first-order logic is undecidable. Here we are considering first-order formulae as defined in Section ???. Given a first-order language \( \mathcal{L} \) consisting of constant symbols \( c \), function symbols \( f \), and predicate symbols \( P \), a \textit{first-order structure} \( \mathcal{M} \) consists of a nonempty domain \( M \), of an assignment of some element of \( c_M \in M \) to every constant symbol \( c \), of a function \( f_M: M^n \rightarrow M \) to every \( n \)-ary function symbol \( f \), and to a boolean-valued function \( P_M: M^m \rightarrow \{T,F\} \) to any \( m \)-ary predicate symbol \( P \).

Then given any assignment \( \rho: X \rightarrow M \) to the first-order variables \( x_i \in X \), we can define recursively the truth value \( \varphi_M[\rho] \) of every first-order formula \( \varphi \). If \( \varphi \) is a sentence, which means that \( \varphi \) has no free variables, then the truth value \( \varphi_M[\rho] \) is independent of \( \rho \), so we simply write \( \varphi_M \). Details can be found in Gallier [15], Enderton [9], or Shoenfeld [33]. The formula \( \varphi \) is \textit{valid in} \( \mathcal{M} \) if \( \varphi_M[\rho] = T \) for all \( \rho \). We also say that \( \mathcal{M} \) is a \textit{model} of \( \varphi \) and we write
\[
\mathcal{M} \models \varphi.
\]

The formula \( \varphi \) is \textit{valid} (or \textit{universally valid}) if it is valid in every first-order structure \( \mathcal{M} \); this denoted by
\[
\models \varphi.
\]

The \textit{validity problem} in first-order logic is to decide whether there is algorithm to decide whether any first-order formula is valid.

**Theorem 7.9.** (Church, 1936) The validity problem for first-order logic is undecidable.

**Proof.** The following proof due to R. Floyd is given in Manna [26] (Section 2.16). The proof consists in reducing the PCP over the alphabet \{0, 1\} to the validity problem. Given an instance \( S = (U, V) \) of the PCP, we construct a first-order sentence \( \Phi_S \) (using a computable
function) such that \( S \) has a solution if and only if \( \Phi_S \) is valid. Since the PCP is undecidable, so is the validity problem for first-order logic.

For this construction, we need a constant symbol \( a \), two unary function symbols \( f_0 \) and \( f_1 \), and a binary predicate symbol \( P \). We denote the term

\[ f_{s_p}(\cdots(f_{s_2}(f_{s_1}(x)))\cdots) \]

as \( f_{s_1s_2\cdots s_p} \), where \( s_i \in \{0, 1\} \). Suppose \( S \) is the set of pairs

\[ S = \{(u_1, v_1), \ldots, (u_m, v_m)\} \]

The key ingredient is the sentence

\[ \Phi_S \equiv \left( \bigwedge_{i=1}^{m} P(f_{u_i}(a), f_{v_i}(a)) \land \forall x \forall y \left( P(x, y) \Rightarrow \bigwedge_{i=1}^{m} P(f_{u_i}(x), f_{v_i}(x)) \right) \right) \Rightarrow \exists z P(z, z) \]

We claim that the PCP \( S \) has a solution iff \( \Phi_S \) is valid.

**Step 1.** We prove that if \( \Phi_S \) is valid, then the PCP has a solution. Consider the first-order structure \( M \) with domain \( \{0, 1\}^* \), with \( a_M = \epsilon \), \( (f_0)_M \) is concatenation on the right with 0 \( ((f_0)_M(x) = x0) \), \( (f_1)_M \) is concatenation on the right with 1 \( ((f_1)_M(x) = x1) \), and

\[ P(x, y) = T \iff x = u_{i_1}u_{i_2}\cdots u_{i_n}, \quad y = v_{i_1}v_{i_2}\cdots v_{i_n}, \]

for some nonempty sequence \( i_1, i_2, \ldots, i_n \) with \( 1 \leq i_j \leq m \).

Since \( \Phi_S \) is valid, it must be valid in \( M \), but then we see immediately that both

\[ \bigwedge_{i=1}^{m} P(f_{u_i}(a), f_{v_i}(a)) \]

and

\[ \forall x \forall y \left( P(x, y) \Rightarrow \bigwedge_{i=1}^{m} P(f_{u_i}(x), f_{v_i}(x)) \right) \]

are valid in \( M \), thus

\[ \exists z P(z, z) \]

is also valid in \( M \). This means that there is some nonempty sequence \( i_1, i_2, \ldots, i_n \) with \( 1 \leq i_j \leq m \) such that

\[ z = u_{i_1}u_{i_2}\cdots u_{i_n} = v_{i_1}v_{i_2}\cdots v_{i_n}, \]

and so we have a solution of the PCP.

**Step 2.** We prove that if the PCP has a solution, then \( \Phi_S \) is valid. Let \( i_1, i_2, \ldots, i_n \) be a nonempty sequence with \( 1 \leq i_j \leq m \) such that

\[ u_{i_1}u_{i_2}\cdots u_{i_n} = v_{i_1}v_{i_2}\cdots v_{i_n}, \]
which means that \(i_1, i_2, \ldots, i_n\) is a solution of the PCP \(S\). We prove that for every first-order structure \(M\), if \[
\bigwedge_{i=1}^{m} P(f_{u_i}(a), f_{v_i}(a))
\]
and
\[
\forall x \forall y \left( P(x, y) \Rightarrow \bigwedge_{i=1}^{m} P(f_{u_i}(x), f_{v_i}(x)) \right)
\]
are valid in \(M\), then
\[
\exists z P(z, z)
\]
is also valid in \(M\). But then \(\Phi_S\) is valid in every first-order structure \(M\), and thus it is valid.

To finish the proof, assume that \(M\) is any first-order structure such that
\[
\bigwedge_{i=1}^{m} P(f_{u_i}(a), f_{v_i}(a)) \quad (*)_1
\]
and
\[
\forall x \forall y \left( P(x, y) \Rightarrow \bigwedge_{i=1}^{m} P(f_{u_i}(x), f_{v_i}(x)) \right) \quad (*)_2
\]
are valid in \(M\). Using \((*)_1\), by repeated application on \((*)_2\), we deduce that
\[
P(f_{u_1 u_2 \cdots u_n}(a), f_{v_1 v_2 \cdots v_n}(a)),
\]
is valid in \(M\). For example, since \((u_{i_1}, v_{i_1})\) is a pair in the PCP instance, by \((*)_1\) the proposition \(P(f_{u_1}(a), f_{v_1}(a))\) holds, so by \((*)_2\) with \(x = f_{u_1}(a)\) and \(v = f_{v_1}(a)\), we get the implication
\[
P(f_{u_1}(a), f_{v_1}(a)) \Rightarrow \bigwedge_{i=1}^{m} P(f_{u_i}(f_{u_1}(a)), f_{v_i}(f_{v_1}(a))),
\]
and since \(P(f_{u_1}(a), f_{v_1}(a))\) holds, we deduce that \(\bigwedge_{i=1}^{m} P(f_{u_i}(f_{u_1}(a)), f_{v_i}(f_{v_1}(a)))\) holds, and consequently \(P(f_{u_2}(f_{u_1}(a)), f_{v_2}(f_{v_1}(a))) = P(f_{u_1 u_2}(a), f_{v_1 v_2}(a))\) holds.

Since by hypothesis
\[
u_{i_1} u_{i_2} \cdots u_{i_n} = v_{i_1} v_{i_2} \cdots v_{i_n},
\]
we deduce that \(\exists z P(z, z)\) is valid in \(M\), and so \(\Phi_S\) is valid in \(M\), as claimed.

There are other ways of proving Church’s theorem. Among other sources, see Shoenfield [33] (Section 6.8) and Machtey and Young [25] (Chapter 4, theorem 4.3.6). These proofs are rather long and involve complicated arguments. Floyd’s proof has the virtue of being quite short and transparent, if we accept the undecidability of the PCP.

Lewis shows the stronger result than even with a single unary function symbol \(f\), one constant \(a\), and one binary predicate symbol \(P\), the validity problem is undecidable; see
Lewis [23] (Chapter IIC). Lewis’ proof is a very clever reduction of a tiling problem. Lewis’ book also contains an extensive classification of undecidable classes of first-order sentences. On the positive side, Dreben and Goldfarb [8] contains a very complete study of classes of first-order sentences for which the validity problem is decidable.
Chapter 8

Computational Complexity; \( \mathcal{P} \) and \( \mathcal{NP} \)

8.1 The Class \( \mathcal{P} \)

In the previous two chapters, we clarified what it means for a problem to be decidable or undecidable. This chapter is heavily inspired by Lewis and Papadimitriou’s excellent treatment [24].

In principle, if a problem is decidable, then there is an algorithm (i.e., a procedure that halts for every input) that decides every instance of the problem.

However, from a practical point of view, knowing that a problem is decidable may be useless, if the number of steps (time complexity) required by the algorithm is excessive, for example, exponential in the size of the input, or worse.

For instance, consider the traveling salesman problem, which can be formulated as follows:

We have a set \( \{c_1, \ldots, c_n\} \) of cities, and an \( n \times n \) matrix \( D = (d_{ij}) \) of nonnegative integers, the distance matrix, where \( d_{ij} \) denotes the distance between \( c_i \) and \( c_j \), which means that \( d_{ii} = 0 \) and \( d_{ij} = d_{ji} \) for all \( i \neq j \).

The problem is to find a shortest tour of the cities, that is, a permutation \( \pi \) of \( \{1, \ldots, n\} \) so that the cost

\[
C(\pi) = d_{\pi(1)\pi(2)} + d_{\pi(2)\pi(3)} + \cdots + d_{\pi(n-1)\pi(n)} + d_{\pi(n)\pi(1)}
\]

is as small as possible (minimal).

One way to solve the problem is to consider all possible tours, i.e., \( n! \) permutations. Actually, since the starting point is irrelevant, we need only consider \((n-1)!\) tours, but this still grows very fast. For example, when \( n = 40 \), it turns out that \( 39! \) exceeds \( 10^{45} \), a huge number.
Consider the $4 \times 4$ symmetric matrix given by
\[
D = \begin{pmatrix}
0 & 2 & 1 & 1 \\
2 & 0 & 1 & 1 \\
1 & 1 & 0 & 3 \\
1 & 1 & 3 & 0
\end{pmatrix},
\]
and the budget $B = 4$. The tour specified by the permutation
\[
\pi = \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 4 & 2 & 3
\end{pmatrix}
\]
has cost 4, since
\[
c(\pi) = d_{\pi(1)\pi(2)} + d_{\pi(2)\pi(3)} + d_{\pi(3)\pi(4)} + d_{\pi(4)\pi(1)}
= d_{14} + d_{42} + d_{23} + d_{31}
= 1 + 1 + 1 + 1 = 4.
\]
The cities in this tour are traversed in the order
\[(1, 4, 2, 3, 1).\]

Remark: The permutation $\pi$ shown above is described in Cauchy’s two-line notation,
\[
\pi = \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 4 & 2 & 3
\end{pmatrix},
\]
where every element in the second row is the image of the element immediately above it in the first row: thus
\[\pi(1) = 1, \pi(2) = 4, \pi(3) = 2, \pi(4) = 3.\]

Thus, to capture the essence of practically feasible algorithms, we must limit our computational devices to run only for a number of steps that is bounded by a polynomial in the length of the input.

We are led to the definition of polynomially bounded computational models.

**Definition 8.1.** A deterministic Turing machine $M$ is said to be polynomially bounded if there is a polynomial $p(X)$ so that the following holds: for every input $x \in \Sigma^*$, there is no ID $ID_n$ so that
\[ID_0 \vdash ID_1 \vdash \cdots \vdash ID_{n-1} \vdash ID_n, \quad \text{with} \quad n > p(|x|),\]
where $ID_0 = q_0x$ is the starting ID.

A language $L \subseteq \Sigma^*$ is polynomially decidable if there is a polynomially bounded Turing machine that accepts $L$. The family of all polynomially decidable languages is denoted by $\mathcal{P}$. 
Remark: Even though Definition 8.1 is formulated for Turing machines, it can also be formulated for other models, such as RAM programs. The reason is that the conversion of a Turing machine into a RAM program (and vice versa) produces a program (or a machine) whose size is polynomial in the original device.

The following proposition, although trivial, is useful:

**Proposition 8.1.** The class $\mathcal{P}$ is closed under complementation.

Of course, many languages do not belong to $\mathcal{P}$. One way to obtain such languages is to use a diagonal argument. But there are also many natural languages that are not in $\mathcal{P}$, although this may be very hard to prove for some of these languages.

Let us consider a few more problems in order to get a better feeling for the family $\mathcal{P}$.

### 8.2 Directed Graphs, Paths

Recall that a directed graph, $G$, is a pair $G = (V, E)$, where $E \subseteq V \times V$. Every $u \in V$ is called a node (or vertex) and a pair $(u, v) \in E$ is called an edge of $G$.

We will restrict ourselves to simple graphs, that is, graphs without edges of the form $(u, u)$; equivalently, $G = (V, E)$ is a simple graph if whenever $(u, v) \in E$, then $u \neq v$.

Given any two nodes $u, v \in V$, a path from $u$ to $v$ is any sequence of $n + 1$ edges $(n \geq 0)$

$$(u, v_1), (v_1, v_2), \ldots, (v_n, v).$$

(If $n = 0$, a path from $u$ to $v$ is simply a single edge, $(u, v)$.)

A graph $G$ is strongly connected if for every pair $(u, v) \in V \times V$, there is a path from $u$ to $v$. A closed path, or cycle, is a path from some node $u$ to itself.

We will restrict out attention to finite graphs, i.e. graphs $(V, E)$ where $V$ is a finite set.

**Definition 8.2.** Given a graph $G$, an Eulerian cycle is a cycle in $G$ that passes through all the nodes (possibly more than once) and every edge of $G$ exactly once. A Hamiltonian cycle is a cycle that passes through all the nodes exactly once (note, some edges may not be traversed at all).

**Eulerian Cycle Problem:** Given a graph $G$, is there an Eulerian cycle in $G$?

**Hamiltonian Cycle Problem:** Given a graph $G$, is there an Hamiltonian cycle in $G$?
8.3 Eulerian Cycles

The following graph is a directed graph version of the Königsberg bridge problem, solved by Euler in 1736.

The nodes $A, B, C, D$ correspond to four areas of land in Königsberg and the edges to the seven bridges joining these areas of land.

![Figure 8.1: A directed graph modeling the Königsberg bridge problem.](image)

The problem is to find a closed path that crosses every bridge exactly once and returns to the starting point.

In fact, the problem is unsolvable, as shown by Euler, because some nodes do not have the same number of incoming and outgoing edges (in the undirected version of the problem, some nodes do not have an even degree.)

It may come as a surprise that the Eulerian Cycle Problem does have a polynomial time algorithm, but that so far, not such algorithm is known for the Hamiltonian Cycle Problem. The reason why the Eulerian Cycle Problem is decidable in polynomial time is the following theorem due to Euler:

**Theorem 8.2.** A graph $G = (V, E)$ has an Eulerian cycle iff the following properties hold:

1. The graph $G$ is strongly connected.
2. Every node has the same number of incoming and outgoing edges.

Proving that properties (1) and (2) hold if $G$ has an Eulerian cycle is fairly easy. The converse is harder, but not that bad (try!).
Theorem 8.2 shows that it is necessary to check whether a graph is strongly connected. This can be done by computing the transitive closure of $E$, which can be done in polynomial time (in fact, $O(n^3)$).

Checking property (2) can clearly be done in polynomial time. Thus, the Eulerian cycle problem is in $\mathcal{P}$. Unfortunately, no theorem analogous to Theorem 8.2 is known for Hamiltonian cycles.

### 8.4 Hamiltonian Cycles

A game invented by Sir William Hamilton in 1859 uses a regular solid dodecahedron whose twenty vertices are labeled with the names of famous cities.

The player is challenged to “travel around the world” by finding a closed cycle along the edges of the dodecahedron which passes through every city exactly once (this is the undirected version of the Hamiltonian cycle problem).

In graphical terms, assuming an orientation of the edges between cities, the graph $D$ shown in Figure 8.2 is a plane projection of a regular dodecahedron and we want to know if there is a Hamiltonian cycle in this directed graph.

![Figure 8.2: A tour “around the world.”](image-url)
Finding a Hamiltonian cycle in this graph does not appear to be so easy!
A solution is shown in Figure 8.3 below.

Figure 8.3: A Hamiltonian cycle in $D$.

**Remark:** We talked about problems being decidable in polynomial time. Obviously, this is equivalent to deciding some property of a certain class of objects, for example, finite graphs.

Our framework requires that we first encode these classes of objects as strings (or numbers), since $\mathcal{P}$ consists of languages.

Thus, when we say that a property is decidable in polynomial time, we are really talking about the encoding of this property as a language. Thus, we have to be careful about these encodings, but it is rare that encodings cause problems.

### 8.5 Propositional Logic and Satisfiability

We define the syntax and the semantics of propositions in conjunctive normal form (CNF).
8.5. PROPOSITIONAL LOGIC AND SATISFIABILITY

The syntax has to do with the legal form of propositions in CNF. Such propositions are interpreted as truth functions, by assigning truth values to their variables.

We begin by defining propositions in CNF. Such propositions are constructed from a countable set, $PV$, of propositional (or boolean) variables, say

$$PV = \{x_1, x_2, \ldots\},$$

using the connectives $\land$ (and), $\lor$ (or) and $\neg$ (negation).

**Definition 8.3.** We define a literal (or atomic proposition), $L$, as $L = x$ or $L = \neg x$, also denoted by $\bar{x}$, where $x \in PV$.

A clause, $C$, is a disjunction of pairwise distinct literals,

$$C = (L_1 \lor L_2 \lor \cdots \lor L_m).$$

Thus, a clause may also be viewed as a nonempty set

$$C = \{L_1, L_2, \ldots, L_m\}.$$

We also have a special clause, the empty clause, denoted $\bot$ or $\emptyset$ (or $\{\}$). It corresponds to the truth value false.

A proposition in CNF, or boolean formula, $P$, is a conjunction of pairwise distinct clauses

$$P = C_1 \land C_2 \land \cdots \land C_n.$$

Thus, a boolean formula may also be viewed as a nonempty set

$$P = \{C_1, \ldots, C_n\},$$

but this time, the comma is interpreted as conjunction. We also allow the proposition $\bot$, and sometimes the proposition $\top$ (corresponding to the truth value true).

For example, here is a boolean formula:

$$P = \{(x_1 \lor x_2 \lor x_3), (\bar{x}_1 \lor x_2), (\bar{x}_2 \lor x_3), (\bar{x}_3 \lor x_1), (\bar{x}_1 \lor \bar{x}_2 \lor \bar{x}_3)\}.$$

In order to interpret boolean formulae, we use truth assignments.

**Definition 8.4.** We let $BOOL = \{F, T\}$, the set of truth values, where $F$ stands for false and $T$ stands for true. A truth assignment (or valuation), $v$, is any function $v: PV \to BOOL$.

**Example 8.1.** The function $v_F: PV \to BOOL$ given by

$$v_F(x_i) = F \quad \text{for all } i \geq 1$$
is a truth assignment, and so is the function \( v_T : \mathbf{PV} \to \text{BOOL} \) given by
\[
v_T(x_i) = T \quad \text{for all } i \geq 1.\]
The function \( v : \mathbf{PV} \to \text{BOOL} \) given by
\[
\begin{align*}
v(x_1) &= T \\
v(x_2) &= F \\
v(x_3) &= T \\
v(x_i) &= T \quad \text{for all } i \geq 4
\end{align*}
\]
is also a truth assignment.

**Definition 8.5.** Given a truth assignment \( v : \mathbf{PV} \to \text{BOOL} \), we define the truth value \( \hat{v}(X) \) of a literal, clause, and boolean formula, \( X \), using the following recursive definition:

1. \( \hat{v}(\bot) = F \), \( \hat{v}(\top) = T \).
2. \( \hat{v}(x) = v(x) \), if \( x \in \mathbf{PV} \).
3. \( \hat{v}(x) = \overline{v(x)} \), if \( x \in \mathbf{PV} \), where \( \overline{v(x)} = F \) if \( v(x) = T \) and \( \overline{v(x)} = T \) if \( v(x) = F \).
4. \( \hat{v}(C) = F \) if \( C \) is a clause and \( \hat{v}(L_i) = F \) for all literals \( L_i \) in \( C \), otherwise \( T \).
5. \( \hat{v}(P) = T \) if \( P \) is a boolean formula and \( \hat{v}(C_j) = T \) for all clauses \( C_j \) in \( P \), otherwise \( F \).

Since a boolean formula \( P \) only contains a finite number of variables, say \( \{x_{i_1}, \ldots, x_{i_n}\} \), one should expect that its truth value \( \hat{v}(P) \) depends only on the truth values assigned by the truth assignment \( v \) to the variables in the set \( \{x_{i_1}, \ldots, x_{i_n}\} \), and this is indeed the case. The following proposition is easily shown by induction on the depth of \( P \) (viewed as a tree).

**Proposition 8.3.** Let \( P \) be a boolean formula containing the set of variables \( \{x_{i_1}, \ldots, x_{i_n}\} \). If \( v_1 : \mathbf{PV} \to \text{BOOL} \) and \( v_2 : \mathbf{PV} \to \text{BOOL} \) are any truth assignments agreeing on the set of variables \( \{x_{i_1}, \ldots, x_{i_n}\} \), which means that
\[
v_1(x_{i_j}) = v_2(x_{i_j}) \quad \text{for } j = 1, \ldots, n,
\]
then \( \hat{v}_1(P) = \hat{v}_2(P) \).

In view of Proposition 8.3, given any boolean formula \( P \), we only need to specify the values of a truth assignment \( v \) for the variables occurring on \( P \).

**Example 8.2.** Given the boolean formula
\[
P = \{ (x_1 \lor x_2 \lor x_3), (\overline{x_1} \lor x_2), (\overline{x_2} \lor x_3), (\overline{x_3} \lor x_1), (\overline{x_1} \lor \overline{x_2} \lor \overline{x_3}) \},
\]
we only need to specify \(v(x_1), v(x_2), v(x_3)\). Thus there are \(2^3 = 8\) distinct truth assignments:

\[
\begin{array}{ccc}
F, F, F & T, F, F \\
F, F, T & T, F, T \\
F, T, F & T, T, F \\
F, T, T & T, T, T.
\end{array}
\]

In general, there are \(2^n\) distinct truth assignments to \(n\) distinct variables.

**Example 8.3.** Here is an example showing the evaluation of the truth value \(\hat{v}(P)\) for the boolean formula

\[
P = (x_1 \lor x_2 \lor x_3) \land (\overline{x_1} \lor x_2) \land (\overline{x_2} \lor x_3) \land (\overline{x_3} \lor x_1) \land (\overline{x_1} \lor \overline{x_2} \lor x_3)
\]

\[
= \{(x_1 \lor x_2 \lor x_3), (\overline{x_1} \lor x_2), (\overline{x_2} \lor x_3), (\overline{x_3} \lor x_1), (\overline{x_1} \lor \overline{x_2} \lor x_3)\},
\]

and the truth assignment

\[
v(x_1) = T, \quad v(x_2) = F, \quad v(x_3) = F.
\]

For the literals, we have

\[
\hat{v}(x_1) = T, \quad \hat{v}(x_2) = F, \quad \hat{v}(x_3) = F, \quad \hat{v}(\overline{x_1}) = F, \quad \hat{v}(\overline{x_2}) = T, \quad \hat{v}(\overline{x_3}) = T,
\]

for the clauses

\[
\hat{v}(x_1 \lor x_2 \lor x_3) = \hat{v}(x_1) \lor \hat{v}(x_2) \lor \hat{v}(x_3) = T \lor F \lor F = T,
\]

\[
\hat{v}(\overline{x_1} \lor x_2) = \hat{v}(\overline{x_1}) \lor \hat{v}(x_2) = F \lor F = F,
\]

\[
\hat{v}(\overline{x_2} \lor x_3) = \hat{v}(\overline{x_2}) \lor \hat{v}(x_3) = T \lor F = T,
\]

\[
\hat{v}(\overline{x_3} \lor x_1) = \hat{v}(\overline{x_3}) \lor \hat{v}(x_1) = T \lor T = T,
\]

\[
\hat{v}(\overline{x_1} \lor \overline{x_2} \lor x_3) = \hat{v}(\overline{x_1}) \lor \hat{v}(\overline{x_2}) \lor \hat{v}(x_3) = F \lor T \lor T = T,
\]

and for the conjunction of the clauses,

\[
\hat{v}(P) = \hat{v}(x_1 \lor x_2 \lor x_3) \land \hat{v}(\overline{x_1} \lor x_2) \land \hat{v}(\overline{x_2} \lor x_3) \land \hat{v}(\overline{x_3} \lor x_1) \land \hat{v}(\overline{x_1} \lor \overline{x_2} \lor x_3) = T \land F \land T \land T \land T = F.
\]

Therefore, \(\hat{v}(P) = F\).

**Definition 8.6.** We say that a truth assignment \(v\) satisfies a boolean formula \(P\), if \(\hat{v}(P) = T\). In this case, we also write

\[
v \models P.
\]

A boolean formula \(P\) is satisfiable if \(v \models P\) for some truth assignment \(v\), otherwise, it is unsatisfiable. A boolean formula \(P\) is valid (or a tautology) if \(v \models P\) for all truth assignments \(v\), in which case we write

\[
\models P.
\]
One should check that the boolean formula

\[ P = \{ (x_1 \lor x_2 \lor x_3), (\overline{x_1} \lor x_2), (\overline{x_2} \lor x_3), (\overline{x_3} \lor x_1), (\overline{x_1} \lor \overline{x_2} \lor \overline{x_3}) \} \]

is unsatisfiable.

One may think that it is easy to test whether a proposition is satisfiable or not. Try it, it is not that easy!

As a matter of fact, the **satisfiability problem**, testing whether a boolean formula is satisfiable, also denoted SAT, is not known to be in \( \mathcal{P} \). Moreover, it is an \( \mathcal{NP} \)-complete problem (see Section 8.6). Most people believe that the satisfiability problem is not in \( \mathcal{P} \), but a proof still eludes us!

Before we explain what is the class \( \mathcal{NP} \), we state the following result.

**Proposition 8.4.** The satisfiability problem for clauses containing at most two literals (2-satisfiability, or 2-SAT) is solvable in polynomial time.

**Proof sketch.** The first step consists in observing that if every clause in \( P \) contains at most two literals, then we can reduce the problem to testing satisfiability when every clause has exactly two literals.

Indeed, if \( P \) contains some clause \((x)\), then any valuation satisfying \( P \) must make \( x \) true. Then all clauses containing \( x \) will be true, and we can delete them, whereas we can delete \( \overline{x} \) from every clause containing it, since \( \overline{x} \) is false.

Similarly, if \( P \) contains some clause \((\overline{x})\), then any valuation satisfying \( P \) must make \( x \) false. Then all clauses containing \( \overline{x} \) will be true, and we can delete them, whereas we can delete \( x \) from every clause containing it.

Thus in a finite number of steps, either all the clauses were satisfied and \( P \) is satisfiable, or we get the empty clause and \( P \) is unsatisfiable, or we get a set of clauses with exactly two literals. The number of steps is clearly linear in the number of literals in \( P \). Here are some examples illustrating the three possible outcomes.

(1) Consider the conjunction of clauses

\[ P_1 = (x_1 \lor \overline{x_2}) \land (x_2) \land (x_1 \lor x_3) \land (x_2 \lor \overline{x_3}). \]

We must set \( x_2 \) to \( T \), so \((x_1 \lor \overline{x_2})\) becomes \((x_1)\) and \((x_2 \lor \overline{x_3})\) becomes \( T \) and can be deleted. We now have

\[ P = (x_1) \land (x_1 \lor x_3) \land (\overline{x_3}). \]

We must set \( x_1 \) to \( T \), so \((x_1 \lor x_3)\) becomes \( T \) and can be deleted. We must also set \( x_3 \) to \( F \), so \((\overline{x_3})\) becomes \( T \) and all the clauses are satisfied.
Consider the conjunction of clauses

\[ P_2 = (x_1) \land (x_3) \land (\overline{x_1} \lor x_2) \land (\overline{x_2} \lor \overline{x_3}). \]

We must set \( x_1 \) to \( T \), so \((\overline{x_1} \lor x_2)\) becomes \((x_2)\). We now have

\[ (x_3) \land (x_2) \land (\overline{x_2} \lor \overline{x_3}). \]

We must set \( x_3 \) to \( T \), so \((\overline{x_2} \lor \overline{x_3})\) becomes \((\neg x_2)\). We now have

\[ (x_2) \land (\overline{x_2}). \]

We must set \( x_2 \) to \( T \), so \((\overline{x_2})\) becomes the empty clause, which means that \( P_2 \) is unsatisfiable.

For the second step, we construct a directed graph from \( P \). The purpose of this graph is to propagate truth. The nodes of this graph are the literals in \( P \), and edges are defined as follows:

1. For every clause \((\overline{x} \lor y)\), there is an edge from \( x \) to \( y \) and an edge from \( \overline{y} \) to \( \overline{x} \).
2. For every clause \((x \lor y)\), there is an edge from \( \overline{x} \) to \( y \) and an edge from \( \overline{y} \) to \( x \).
3. For every clause \((\overline{x} \lor \overline{y})\), there is an edge from \( x \) to \( \overline{y} \) and an edge from \( y \) to \( \overline{x} \).

Then it can be shown that \( P \) is unsatisfiable iff there is some \( x \) so that there is a cycle containing \( x \) and \( \overline{x} \). As a consequence, 2-satisfiability is in \( \mathcal{P} \). \( \square \)

**Example 8.4.** Consider the following conjunction of clauses:

\[ P = (x_1 \lor \overline{x_2}) \land (x_1 \lor x_2) \land (x_2 \lor \overline{x_3}). \]

It is satisfied by any valuation \( v \) such that \( v(x_1) = T \), and if \( v(x_2) = F \) then \( v(x_3) = F \). The construction of the graph associated with \( P \) is shown in Figure 8.4.

**8.6 The Class \( \mathcal{NP} \), Polynomial Reducibility, \( \mathcal{NP} \)-Completeness**

One will observe that the hard part in trying to solve either the Hamiltonian cycle problem or the satisfiability problem, SAT, is to **find** a solution, but that **checking** that a candidate solution is indeed a solution can be done easily in polynomial time.

This is the essence of problems that can be solved **nondeterministically** in polynomial time: a solution can be guessed and then checked in polynomial time.
Definition 8.7. A nondeterministic Turing machine \( M \) is said to be **polynomially bounded** if there is a polynomial \( p(X) \) so that the following holds: For every input \( x \in \Sigma^* \), there is no \( ID = ID_n \) so that

\[
ID_0 \vdash ID_1 \vdash^* ID_{n-1} \vdash ID_n, \quad \text{with} \quad n > p(|x|),
\]

where \( ID_0 = q_0x \) is the starting ID.

A language \( L \subseteq \Sigma^* \) is **nondeterministic polynomially decidable** if there is a polynomially bounded nondeterministic Turing machine that accepts \( L \). The family of all nondeterministic polynomially decidable languages is denoted by \( \mathcal{NP} \).

Of course, we have the inclusion

\[
\mathcal{P} \subseteq \mathcal{NP},
\]

but whether or not we have equality is one of the most famous open problems of theoretical computer science and mathematics.

In fact, the question \( \mathcal{P} \neq \mathcal{NP} \) is one of the open problems listed by the CLAY Institute, together with the Poincaré conjecture and the Riemann hypothesis, among other problems, and for which one million dollar is offered as a reward! Actually the Poincaré conjecture...
was settled by G. Perelman in 2006, but he rejected receiving the prize in 2010! He also declined the Fields Medal which was awarded to him in 2006.

It is easy to check that SAT is in $\mathcal{NP}$, and so is the Hamiltonian cycle problem.

As we saw in recursion theory, where we introduced the notion of many-one reducibility, in order to compare the “degree of difficulty” of problems, it is useful to introduce the notion of reducibility and the notion of a complete set.

**Definition 8.8.** A function $f : \Sigma^* \rightarrow \Sigma^*$ is **polynomial-time computable** if there is a polynomial $p(X)$ so that the following holds: there is a deterministic Turing machine $M$ computing it so that for every input $x \in \Sigma^*$, there is no ID $ID_n$ so that

$$ID_0 \vdash ID_1 \vdash \ldots \vdash ID_{n-1} \vdash ID_n,$$

with $n > p(|x|)$,

where $ID_0 = q_0x$ is the starting ID.

Given two languages $L_1, L_2 \subseteq \Sigma^*$, a **polynomial-time reduction from $L_1$ to $L_2$** is a polynomial-time computable function $f : \Sigma^* \rightarrow \Sigma^*$ so that for all $u \in \Sigma^*$,

$$u \in L_1 \iff f(u) \in L_2.$$ 

The notation $L_1 \leq_P L_2$ is often used to denote the fact that there is polynomial-time reduction from $L_1$ to $L_2$. Sometimes, the notation $L_1 \leq_m^P L_2$ is used to stress that this is a many-to-one reduction (that is, $f$ is not necessarily injective). This type of reduction is also known as a **Karp reduction**.

A polynomial reduction $f : \Sigma^* \rightarrow \Sigma^*$ from a language $L_1$ to a language $L_2$ is a method that converts in polynomial time every string $u \in \Sigma^*$ (viewed as an instance of a problem $A$ encoded by language $L_1$) to a string $f(u) \in \Sigma^*$ (viewed as an instance of a problem $B$ encoded by language $L_2$) in such way that membership in $L_1$, that is $u \in L_1$, is equivalent to membership in $L_2$, that is $f(u) \in L_2$.

As a consequence, if we have a procedure to decide membership in $L_2$ (to solve every instance of problem $B$), then we have a procedure for solving membership in $L_1$ (to solve every instance of problem $A$), since given any $u \in L_1$, we can first apply $f$ to $u$ to produce $f(u)$, and then apply our procedure to decide whether $f(u) \in L_2$; the defining property of $f$ says that this is equivalent to deciding whether $u \in L_1$. Furthermore, if the procedure for deciding membership in $L_2$ runs deterministically in polynomial time, since $f$ runs deterministically in polynomial time, so does the procedure for deciding membership in $L_1$, and similarly if the procedure for deciding membership in $L_2$ runs non deterministically in polynomial time.

For the above reason, we see that membership in $L_2$ can be considered at least as hard as membership in $L_1$, since any method for deciding membership in $L_2$ yields a method for deciding membership in $L_1$. Thus, if we view $L_1$ an encoding a problem $A$ and $L_2$ as encoding a problem $B$, then $B$ is at least as hard as $A$.

The following version of Proposition 3.16 for polynomial-time reducibility is easy to prove.
Proposition 8.5. Let \( A, B, C \) be subsets of \( \mathbb{N} \) (or \( \Sigma^* \)). The following properties hold:

1. If \( A \leq_P B \) and \( B \leq_P C \), then \( A \leq_P C \).
2. If \( A \leq_P B \) then \( \overline{A} \leq_P \overline{B} \).
3. If \( A \leq_P B \) and \( B \in \mathcal{NP} \), then \( A \in \mathcal{NP} \).
4. If \( A \leq_P B \) and \( A \notin \mathcal{NP} \), then \( B \notin \mathcal{NP} \).
5. If \( A \leq_P B \) and \( B \in \mathcal{P} \), then \( A \in \mathcal{P} \).
6. If \( A \leq_P B \) and \( A \notin \mathcal{P} \), then \( B \notin \mathcal{P} \).

Intuitively, we see that if \( L_1 \) is a hard problem and \( L_1 \) can be reduced to \( L_2 \) in polynomial time, then \( L_2 \) is also a hard problem.

For example, one can construct a polynomial reduction from the Hamiltonian cycle problem to the satisfiability problem \( \text{SAT} \). Given a directed graph \( G = (V, E) \) with \( n \) nodes, say \( V = \{1, \ldots, n\} \), we need to construct in polynomial time a set \( F = \tau(G) \) of clauses such that \( G \) has a Hamiltonian cycle iff \( \tau(G) \) is satisfiable. We need to describe a permutation of the nodes that forms a Hamiltonian cycle. For this we introduce \( n^2 \) boolean variables \( x_{ij} \), with the intended interpretation that \( x_{ij} \) is true iff node \( i \) is the \( j \)th node in a Hamiltonian cycle.

To express that at least one node must appear as the \( j \)th node in a Hamiltonian cycle, we have the \( n \) clauses

\[
(x_{1j} \lor x_{2j} \lor \cdots \lor x_{nj}), \quad 1 \leq j \leq n. \tag{1}
\]

The conjunction of these clauses is satisfied iff for every \( j = 1, \ldots, n \) there is some node \( i \) which is the \( j \)th node in the cycle. These \( n \) clauses can be produced in time \( O(n^2) \).

To express that only one node appears in the cycle, we have the clauses

\[
(x_{ij} \lor x_{kj}), \quad 1 \leq i, j, k \leq n, \quad i \neq k. \tag{2}
\]

Since \( (x_{ij} \lor x_{kj}) \) is equivalent to \( (x_{ij} \land x_{kj}) \), each such clause asserts that no two distinct nodes may appear as the \( j \)th node in the cycle. Let \( S_1 \) be the set of all clauses of type (1) or (2). These \( n^3 \) clauses can be produced in time \( O(n^3) \).

The conjunction of the clauses in \( S_1 \) assert that exactly one node appear at the \( j \)th node in the Hamiltonian cycle. We still need to assert that each node \( i \) appears exactly once in the cycle. For this, we have the clauses

\[
(x_{i1} \lor x_{i2} \lor \cdots \lor x_{in}), \quad 1 \leq i \leq n, \tag{3}
\]

and

\[
(x_{ij} \lor x_{ik}), \quad 1 \leq i, j, k \leq n, \quad j \neq k. \tag{4}
\]

Let \( S_2 \) be the set of all clauses of type (3) or (4). These \( n^3 \) clauses can be produced in time \( O(n^3) \).
The conjunction of the clauses in \( S_1 \cup S_2 \) asserts that the \( x_{ij} \) represents a bijection of \( \{1, 2, \ldots, n\} \), in the sense that for any truth assignment \( v \) satisfying all these clauses, \( i \mapsto j \) if \( v(x_{ij}) = T \) defines a bijection of \( \{1, 2, \ldots, n\} \).

It remains to assert that this permutation of the nodes is a Hamiltonian cycle, which means that if \( x_{ij} \) and \( x_{kj+1} \) are both true then there must be an edge \((i, k)\). By contrapositive, this equivalent to saying that if \((i, k)\) is not an edge of \( G \), then \((x_{ij} \land x_{kj+1})\) is true, which as a clause is equivalent to \((\overline{x_{ij}} \lor x_{kj+1})\).

Therefore, for all \((i, k)\) such that \((i, k) \notin E \) (with \( i, k \in \{1, 2, \ldots, n\} \)), we have the clauses

\[
(x_{ij} \lor x_{kj+1 (\text{mod } n)}), \quad j = 1, \ldots, n. \tag{5}
\]

Let \( S_3 \) be the set of clauses of type (5). These \( n \) clauses can be produced in time \( O(n^2) \).

The conjunction of all the clauses in \( S_1 \cup S_2 \cup S_3 \) is the boolean formula \( F = \tau(G) \). It can be produced in time \( O(n^3) \).

We leave it as an exercise to prove that \( G \) has a Hamiltonian cycle iff \( F = \tau(G) \) is satisfiable.

**Example 8.5.** Here is an example of a graph with four nodes and four edges shown in Figure 8.5. The Hamiltonian circuit is \((x_1, x_3, x_1, x_2)\).

![Figure 8.5: A directed graph with a Hamiltonian cycle](image)

It is also possible to construct a reduction of the satisfiability problem to the Hamiltonian cycle problem but this is harder. It is easier to construct this reduction in two steps by
introducing an intermediate problem, the exact cover problem, and to provide a polynomial reduction from the satisfiability problem to the exact cover problem, and a polynomial reduction from the exact cover problem to the Hamiltonian cycle problem. These reductions are carried out in Section 9.2.

The above construction of a set \( F = \tau(G) \) of clauses from a graph \( G \) asserting that \( G \) has a Hamiltonian cycle iff \( F \) is satisfiable illustrates the expressive power of propositional logic.

Remarkably, every language in \( \mathcal{NP} \) can be reduced to SAT. Thus, SAT is a hardest problem in \( \mathcal{NP} \) (since it is in \( \mathcal{NP} \)).

Definition 8.9. A language \( L \) is \( \mathcal{NP} \)-hard if there is a polynomial reduction from every language \( L_1 \in \mathcal{NP} \) to \( L \). A language \( L \) is \( \mathcal{NP} \)-complete if \( L \in \mathcal{NP} \) and \( L \) is \( \mathcal{NP} \)-hard.

Thus, an \( \mathcal{NP} \)-hard language is as hard to decide as any language in \( \mathcal{NP} \).

Remark: There are \( \mathcal{NP} \)-hard languages that do not belong to \( \mathcal{NP} \). Such problems are really difficult. Two standard examples are \( K_0 \) and \( K \), which encode the halting problem. Since \( K_0 \) and \( K \) are not computable, they can’t be in \( \mathcal{NP} \). Furthermore, since every language \( L \) in \( \mathcal{NP} \) is accepted nondeterministically in polynomial time \( p(X) \), for some polynomial \( p(X) \), for every input \( w \) we can try all computations of length at most \( p(|w|) \) (there can be exponentially many, but only a finite number), so every language in \( \mathcal{NP} \) is computable. Finally, it is shown in Theorem 3.17 that \( K_0 \) and \( K \) are complete with respect to many-one reducibility, so in particular they are \( \mathcal{NP} \)-hard. An example of a computable \( \mathcal{NP} \)-hard language not in \( \mathcal{NP} \) will be described after Theorem 8.7.

The importance of \( \mathcal{NP} \)-complete problems stems from the following theorem which follows immediately from Proposition 8.5.

Theorem 8.6. Let \( L \) be an \( \mathcal{NP} \)-complete language. Then \( \mathcal{P} = \mathcal{NP} \) iff \( L \in \mathcal{P} \).

There are analogies between \( \mathcal{P} \) and the class of computable sets, and \( \mathcal{NP} \) and the class of listable sets, but there are also important differences. One major difference is that the family of computable sets is properly contained in the family of listable sets, but it is an open problem whether \( \mathcal{P} \) is properly contained in \( \mathcal{NP} \). We also know that a set \( L \) is computable iff both \( L \) and \( \overline{L} \) are listable, but it is also an open problem whether if both \( L \in \mathcal{NP} \) and \( \overline{L} \in \mathcal{NP} \), then \( L \in \mathcal{P} \). This suggests defining

\[
\text{co}\mathcal{NP} = \{ \overline{L} \mid L \in \mathcal{NP} \},
\]

that is, \( \text{co}\mathcal{NP} \) consists of all complements of languages in \( \mathcal{NP} \). Since \( \mathcal{P} \subseteq \mathcal{NP} \) and \( \mathcal{P} \) is closed under complementation,

\[
\mathcal{P} \subseteq \text{co}\mathcal{NP},
\]

and thus

\[
\mathcal{P} \subseteq \mathcal{NP} \cap \text{co}\mathcal{NP},
\]
but nobody knows whether the inclusion is proper. There are problems in \( \mathcal{NP} \cap \text{co}\mathcal{NP} \) not known to be in \( \mathcal{P} \); see Section 9.3. It is unknown whether \( \mathcal{NP} \) is closed under complementation, that is, nobody knows whether \( \mathcal{NP} = \text{co}\mathcal{NP} \). This is considered unlikely. We will come back to \( \text{co}\mathcal{NP} \) in Section 9.3.

Next we prove a famous theorem of Steve Cook and Leonid Levin (proven independently): SAT is \( \mathcal{NP} \)-complete.

## 8.7 The Bounded Tiling Problem is \( \mathcal{NP} \)-Complete

Instead of showing directly that SAT is \( \mathcal{NP} \)-complete, which is rather complicated, we proceed in two steps, as suggested by Lewis and Papadimitriou.

1. First, we define a tiling problem adapted from H. Wang (1961) by Harry Lewis, and we prove that it is \( \mathcal{NP} \)-complete.

2. We show that the tiling problem can be reduced to SAT.

We are given a finite set \( \mathcal{T} = \{t_1, \ldots, t_p\} \) of tile patterns, for short, tiles. We assume that these tiles are unit squares. Copies of these tile patterns may be used to tile a rectangle of predetermined size \( 2s \times s \) \((s > 1)\). However, there are constraints on the way that these tiles may be adjacent horizontally and vertically.

The horizontal constraints are given by a relation \( H \subseteq \mathcal{T} \times \mathcal{T} \), and the vertical constraints are given by a relation \( V \subseteq \mathcal{T} \times \mathcal{T} \).

Thus, a tiling system is a triple \( \mathcal{T} = (\mathcal{T}, V, H) \) with \( V \) and \( H \) as above.

The bottom row of the rectangle of tiles is specified before the tiling process begins.

**Example 8.6.** For example, consider the following tile patterns:

\[
\begin{array}{cccc}
  a & c & & \\
  c & a & & \\
  b & d & & \\
  c & d & & \\
  c & d & & \\
  d & e & & \\
  e & e & & \\
  e & e & & \\
  e & e & & \\
  e & e & & \\
  e & e & & \\
\end{array}
\]

The horizontal and the vertical constraints are that the letters on adjacent edges match (blank edges do not match).
For $s = 3$, given the bottom row

$$
\begin{array}{cccccccc}
a & b & c & d & c & d & e & e \\
c & c & d & d & e & e & e & e \\
\end{array}
$$

we have the tiling shown below:

$$
\begin{array}{cccccccccccc}
c & d & e & e & e & e & e & e & e \\
b & c & d & d & d & e & e & e & e \\
a & b & c & d & e & e & e & e & e \\
a & b & c & d & e & e & e & e & e \\
\end{array}
$$

Formally, the problem is then as follows:

**The Bounded Tiling Problem**

Given any tiling system $(\mathcal{T}, V, H)$, any integer $s > 1$, and any initial row of tiles $\sigma_0$ (of length $2s$)

$$
\sigma_0: \{1, 2, \ldots, s, s + 1, \ldots, 2s\} \rightarrow \mathcal{T},
$$

find a $2s \times s$-tiling $\sigma$ extending $\sigma_0$, i.e., a function

$$
\sigma: \{1, 2, \ldots, s, s + 1, \ldots, 2s\} \times \{1, \ldots, s\} \rightarrow \mathcal{T}
$$

so that

1. $\sigma(m, 1) = \sigma_0(m)$, for all $m$ with $1 \leq m \leq 2s$.
2. $(\sigma(m, n), \sigma(m + 1, n)) \in H$, for all $m$ with $1 \leq m \leq 2s - 1$, and all $n$, with $1 \leq n \leq s$.
3. $(\sigma(m, n), \sigma(m, n + 1)) \in V$, for all $m$ with $1 \leq m \leq 2s$, and all $n$, with $1 \leq n \leq s - 1$.

Formally, an *instance of the tiling problem* is a triple $((\mathcal{T}, V, H), \hat{s}, \sigma_0)$, where $(\mathcal{T}, V, H)$ is a tiling system, $\hat{s}$ is the string representation of the number $s \geq 2$, in binary and $\sigma_0$ is an initial row of tiles (the bottom row).

For example, if $s = 1025$ (as a decimal number), then its binary representation is $\hat{s} = 10000000001$. The length of $\hat{s}$ is $\log_2 s + 1$. 
Recall that the input must be a string. This is why the number \( s \) is represented by a string in binary. If we only included a single tile \( \sigma_0 \) in position \((s + 1, 1)\), then the length of the input \(((\mathcal{T}, V, H), \hat{s}, \sigma_0)\) would be \( \log_2 s + 1 + C + 1 = \log_2 s + C + 2 \) for some constant \( C \) corresponding to the length of the string encoding \((\mathcal{T}, V, H)\).

However, the rectangular grid has size \( 2s^2 \), which is exponential in the length \( \log_2 s + C + 2 \) of the input \(((\mathcal{T}, V, H), \hat{s}, \sigma_0)\). Thus, it is impossible to check in polynomial time that a proposed solution is a tiling.

However, if we include in the input the bottom row \( \sigma_0 \) of length \( 2s \), then the length of input is \( \log_2 s + 1 + C + 2s = \log_2 s + C + 2s + 1 \) and the size \( 2s^2 \) or the grid is indeed polynomial in the size of the input.

**Theorem 8.7.** The tiling problem defined earlier is \( \mathcal{NP} \)-complete.

**Proof.** Let \( L \subseteq \Sigma^* \) be any language in \( \mathcal{NP} \) and let \( u \) be any string in \( \Sigma^* \). Assume that \( L \) is accepted in polynomial time bounded by \( p(|u|) \).

We show how to construct an instance of the tiling problem, \(((\mathcal{T}, V, H)_L, \hat{s}, \sigma_0)\), where \( s = p(|u|) + 2 \), and where the bottom row encodes the starting ID, so that \( u \in L \) iff the tiling problem \(((\mathcal{T}, V, H)_L, \hat{s}, \sigma_0)\) has a solution.

First, note that the problem is indeed in \( \mathcal{NP} \), since we have to guess a rectangle of size \( 2s^2 \), and that checking that a tiling is legal can indeed be done in \( O(s^2) \), where \( s \) is bounded by the size of the input \(((\mathcal{T}, V, H), \hat{s}, \sigma_0)\), since the input contains the bottom row of \( 2s \) symbols (this is the reason for including the bottom row of \( 2s \) tiles in the input!).

The idea behind the definition of the tiles is that, in a solution of the tiling problem, the labels on the horizontal edges between two adjacent rows represent a legal ID, \( upav \). In a given row, the labels on vertical edges of adjacent tiles keep track of the change of state and direction.

Let \( \Gamma \) be the tape alphabet of the TM, \( M \). As before, we assume that \( M \) signals that it accepts \( u \) by halting with the output 1 (true).

From \( M \), we create the following tiles:

1. For every \( a \in \Gamma \), tiles

   \[
   \begin{array}{c}
   a \\
   a
   \end{array}
   \]

2. For every \( a \in \Gamma \), the bottom row uses tiles

   \[
   \begin{array}{c}
   a \\
   q_0, a
   \end{array}
   \]

   where \( q_0 \) is the start state.
(3) For every instruction \((p, a, b, R, q) \in \delta\), for every \(c \in \Gamma\), tiles
\[
\begin{array}{c|c}
q, R & q, c \\
p, a & c
\end{array}
\]

(4) For every instruction \((p, a, b, L, q) \in \delta\), for every \(c \in \Gamma\), tiles
\[
\begin{array}{c|c}
q, L & b \\
c & q, L
\end{array}
\]

(5) For every halting state, \(p\), tiles
\[
\begin{array}{c|c}
p, 1 \\
p, 1
\end{array}
\]

The purpose of tiles of type (5) is to fill the \(2s \times s\) rectangle iff \(M\) accepts \(u\). Since \(s = p(|u|) + 2\) and the machine runs for at most \(p(|u|)\) steps, the \(2s \times s\) rectangle can be tiled iff \(u \in L\).

The vertical and the horizontal constraints are that adjacent edges have the same label (or no label).

If \(u = u_1 \cdots u_k\), the initial bottom row \(\sigma_0\), of length \(2s\), is
\[
\begin{array}{c|c|c|c|c|c}
B & \cdots & q_0, u_1 & \cdots & u_k & B
\end{array}
\]

where the tile labeled \(q_0, u_1\) is in position \(s + 1\).

The example below illustrates the construction:

**Example 8.7.**
We claim that \( u = u_1 \cdots u_k \) is accepted by \( M \) iff the tiling problem just constructed has a solution.

The upper horizontal edge of the first (bottom) row of tiles represents the starting configuration \( B^s q_0 u B^{s-|u|} \). By induction, we see that after \( i \) \((i \leq p(|u|) = s - 2)\) steps the upper horizontal edge of the \((i + 1)\)th row of tiles represents the current ID upav reached by the Turing machine; see Example 8.7. Since the machine runs for at most \( p(|u|) \) steps and since \( s = p(|u|) + 2 \), when the computation stops, at most the lowest \( p(|u|) + 1 = s - 1 \) rows of the the \( 2s \times s \) rectangle have been tiled. Assume the machine \( M \) stops after \( r \leq s - 2 \) steps. Then the lowest \( r + 1 \) rows have been tiled, and since no further instruction can be executed (since the machine entered a halting state), the remaining \( s - r - 1 \) rows can be filled iff tiles of type (5) can be used iff the machine stopped in an ID containing a pair \( pq \) where \( p \) is a halting state. Therefore, the machine \( M \) accepts \( u \) iff the \( 2s \times s \) rectangle can be tiled. \( \Box \)

Remark:

(1) The problem becomes harder if we only specify a **single** tile \( \sigma_0 \) as input, instead of a row of length \( 2s \). If \( s \) is specified in binary (or any other base, but not in tally notation), then the \( 2s^2 \) grid has size exponential in the length \( \log_2 s + C + 2 \) of the input \( ((T, V, H), \hat{s}, \sigma_0) \), and this tiling problem is actually \( \text{NEXPTIME} \)-complete! The class \( \text{NEXPTIME} \) is the family of languages that can be accepted by a nondeterministic Turing machine that runs in time bounded by \( 2^{p(|x|)} \), for every \( x \), where \( p \) is a polynomial; see the remark after Definition 9.5. By the time hierarchy theorem (Cook, Seiferas, Fischer, Meyer, Zak), it is known that \( \text{NP} \) is properly contained in \( \text{NEXPTIME} \); see Papadimitriou [27] (Chapters 7 and 20) and Arora and Barak [2] (Chapter 3, Section 3.2). Then the tiling problem with a single tile as input is a computable \( \text{NP} \)-hard problem not in \( \text{NP} \).

(2) If we relax the finiteness condition and require that the entire upper half-plane be tiled, i.e., for every \( s > 1 \), there is a solution to the \( 2s \times s \)-tiling problem, then the problem is undecidable.

In 1972, Richard Karp published a list of twenty one \( \text{NP} \)-complete problems.

### 8.8 The Cook–Levin Theorem: SAT is \( \text{NP} \)-Complete

We finally prove the Cook-Levin theorem.

**Theorem 8.8.** (Cook, 1971, Levin, 1973) The satisfiability problem SAT is \( \text{NP} \)-complete.

**Proof.** We reduce the tiling problem to SAT. Given a tiling problem, \( ((T, V, H), \hat{s}, \sigma_0) \), we introduce boolean variables

\[ x_{mnt}, \]

for all \( m \) with \( 1 \leq m \leq 2s \), all \( n \) with \( 1 \leq n \leq s \), and all tiles \( t \in T \).
The intuition is that $x_{mnt} = T$ iff tile $t$ occurs in some tiling $\sigma$ so that $\sigma(m, n) = t$.

We define the following clauses:

(1) For all $m, n$ in the correct range, as above,

$$(x_{mnt_1} \lor x_{mnt_2} \lor \cdots \lor x_{mnt_p}),$$

for all $p$ tiles in $T$.

This clause states that every position in $\sigma$ is tiled.

(2) For any two distinct tiles $t \neq t' \in T$, for all $m, n$ in the correct range, as above,

$$(\overline{x_{mnt}} \lor \overline{x_{mnt'}}).$$

This clause states that a position may not be occupied by more than one tile.

(3) For every pair of tiles $(t, t') \in T \times T - H$, for all $m$ with $1 \leq m \leq 2s - 1$, and all $n$, with $1 \leq n \leq s$,

$$(\overline{x_{mnt}} \lor \overline{x_{m+1nt'}}).$$

This clause enforces the horizontal adjacency constraints.

(4) For every pair of tiles $(t, t') \in T \times T - V$, for all $m$ with $1 \leq m \leq 2s$, and all $n$, with $1 \leq n \leq s - 1$,

$$(\overline{x_{mnt}} \lor \overline{x_{m+1nt+1}}).$$

This clause enforces the vertical adjacency constraints.

(5) For all $m$ with $1 \leq m \leq 2s$,

$$(x_{m1\sigma_0(m)}).$$

This clause states that the bottom row is correctly tiled with $\sigma_0$.

It is easily checked that the tiling problem has a solution iff the conjunction of the clauses just defined is satisfiable. Thus, SAT is $NP$-complete.

We sharpen Theorem 8.8 to prove that 3-SAT is also $NP$-complete. This is the satisfiability problem for clauses containing at most three literals.

We know that we can’t go further and retain $NP$-completeness, since 2-SAT is in $P$.

**Theorem 8.9.** *(Cook, 1971)* The satisfiability problem 3-SAT is $NP$-complete.

**Proof.** We have to break “long clauses”

$$C = (L_1 \lor \cdots \lor L_k),$$

i.e., clauses containing $k \geq 4$ literals, into clauses with at most three literals, in such a way that satisfiability is preserved.
Example 8.8. For example, consider the following clause with \( k = 6 \) literals:

\[
C = (L_1 \lor L_2 \lor L_3 \lor L_4 \lor L_5 \lor L_6).
\]

We create 3 new boolean variables \( y_1, y_2, y_3 \), and the 4 clauses

\[
(L_1 \lor L_2 \lor y_1), (\overline{y_1} \lor L_3 \lor y_2), (\overline{y_2} \lor L_4 \lor y_3), (\overline{y_3} \lor L_5 \lor L_6).
\]

Let \( C' \) be the conjunction of these clauses. We claim that \( C \) is satisfiable iff \( C' \) is.

Assume that \( C' \) is satisfiable but \( C \) is not. If so, in any truth assignment \( v \), \( v(L_i) = \text{F} \), for \( i = 1, 2, \ldots, 6 \). To satisfy the first clause, we must have \( v(y_1) = \text{T} \). Then to satisfy the second clause, we must have \( v(y_2) = \text{T} \), and similarly satisfy the third clause, we must have \( v(y_3) = \text{T} \). However, since \( v(L_5) = \text{F} \) and \( v(L_6) = \text{F} \), the only way to satisfy the fourth clause is to have \( v(y_3) = \text{F} \), contradicting that \( v(y_3) = \text{T} \). Thus, \( C \) is indeed satisfiable.

Let us now assume that \( C \) is satisfiable. This means that there is a smallest index \( i \) such that \( L_i \) is satisfied.

Say \( i = 1 \), so \( v(L_1) = \text{T} \). Then if we let \( v(y_1) = v(y_2) = v(y_3) = \text{F} \), we see that \( C' \) is satisfied.

Say \( i = 2 \), so \( v(L_1) = \text{F} \) and \( v(L_2) = \text{T} \). Again if we let \( v(y_1) = v(y_2) = v(y_3) = \text{F} \), we see that \( C' \) is satisfied.

Say \( i = 3 \), so \( v(L_1) = \text{F} \), \( v(L_2) = \text{F} \), and \( v(L_3) = \text{T} \). If we let \( v(y_1) = \text{T} \) and \( v(y_2) = v(y_3) = \text{F} \), we see that \( C' \) is satisfied.

Say \( i = 4 \), so \( v(L_1) = \text{F} \), \( v(L_2) = \text{F} \), \( v(L_3) = \text{F} \), and \( v(L_4) = \text{T} \). If we let \( v(y_1) = \text{T} \), \( v(y_2) = \text{T} \) and \( v(y_3) = \text{F} \), we see that \( C' \) is satisfied.

Say \( i = 5 \), so \( v(L_1) = \text{F} \), \( v(L_2) = \text{F} \), \( v(L_3) = \text{F} \), \( v(L_4) = \text{F} \), and \( v(L_5) = \text{T} \). If we let \( v(y_1) = \text{T} \), \( v(y_2) = \text{T} \) and \( v(y_3) = \text{T} \), we see that \( C' \) is satisfied.

Say \( i = 6 \), so \( v(L_1) = \text{F} \), \( v(L_2) = \text{F} \), \( v(L_3) = \text{F} \), \( v(L_4) = \text{F} \), \( v(L_5) = \text{F} \), and \( v(L_6) = \text{T} \). Again, if we let \( v(y_1) = \text{T} \), \( v(y_2) = \text{T} \) and \( v(y_3) = \text{T} \), we see that \( C' \) is satisfied.

Therefore if \( C \) is satisfied, then \( C' \) is satisfied in all cases.

In general, for every long clause, create \( k - 3 \) new boolean variables \( y_1, \ldots, y_{k-3} \), and the \( k - 2 \) clauses

\[
(L_1 \lor L_2 \lor y_1), (\overline{y_1} \lor L_3 \lor y_2), (\overline{y_2} \lor L_4 \lor y_3), \ldots,
\]

\[
(\overline{y}_{k-4} \lor L_{k-2} \lor y_{k-3}), (\overline{y}_{k-3} \lor L_{k-1} \lor L_k).
\]

Let \( C' \) be the conjunction of these clauses. We claim that \( C \) is satisfiable iff \( C' \) is.

Assume that \( C' \) is satisfiable, but that \( C \) is not. Then for every truth assignment \( v \), we have \( v(L_i) = \text{F} \), for \( i = 1, \ldots, k \).
However, \( C' \) is satisfied by some \( v \), and the only way this can happen is that \( v(y_1) = T \), to satisfy the first clause. Then \( v(\overline{y_1}) = F \), and we must have \( v(y_2) = T \), to satisfy the second clause.

By induction, we must have \( v(y_{k-3}) = T \), to satisfy the next to the last clause. However, the last clause is now false, a contradiction.

Thus, if \( C' \) is satisfiable, then so is \( C \).

Conversely, assume that \( C \) is satisfiable. If so, there is some truth assignment, \( v \), so that \( v(C) = T \), and thus, there is a smallest index \( i \), with \( 1 \leq i \leq k \), so that \( v(L_i) = T \) (and so, \( v(L_j) = F \) for all \( j < i \)).

Let \( v' \) be the assignment extending \( v \) defined so that
\[
v'(y_j) = \begin{cases} F & \text{if } \max\{1, i-1\} \leq j \leq k-3, \\ T & \text{otherwise.} \end{cases}
\]
and \( v'(y_j) = T \), otherwise.

It is easily checked that \( v'(C') = T \). \( \square \)

Another version of 3-SAT can be considered, in which every clause has exactly three literals. We will call this the problem exact 3-SAT.

**Theorem 8.10.** (Cook, 1971) The satisfiability problem for exact 3-SAT is \( \mathcal{NP} \)-complete.

**Proof.** A clause of the form \((L)\) is satisfiable iff the following four clauses are satisfiable:
\[
(L \lor u \lor v), (L \lor \overline{u} \lor v), (L \lor u \lor \overline{v}), (L \lor \overline{u} \lor \overline{v})
\]
where \( u, v \) are new variables. A clause of the form \((L_1 \lor L_2)\) is satisfiable iff the following two clauses are satisfiable:
\[
(L_1 \lor L_2 \lor u), (L_1 \lor L_2 \lor \overline{u}).
\]
Thus, we have a reduction of 3-SAT to exact 3-SAT. \( \square \)

We now make some remarks about the conversion of propositions to CNF and about the satisfiability of arbitrary propositions.

### 8.9 Satisfiability of Arbitrary Propositions and CNF

We begin by discussing how to convert an arbitrary proposition to conjunctive normal form. In general, given a proposition \( A \), a proposition \( A' \) in CNF equivalent to \( A \) may have an exponential length in the size of \( A \). Thus it is not obvious that the satisfiability problem for arbitrary propositions belongs to \( \mathcal{NP} \). However, using new variables, there is an algorithm to convert a proposition \( A \) to another proposition \( A' \) (containing the new variables) whose length is polynomial in the length of \( A \) and such that \( A \) is satisfiable iff \( A' \) is satisfiable.
Thus the satisfiability problem for arbitrary propositions belongs to \( \mathcal{NP} \). We also briefly discuss the issue of uniqueness of the CNF. In short, it is not unique!

Recall the definition of arbitrary propositions.

**Definition 8.10.** The set of propositions (over the connectives \( \lor \), \( \land \), and \( \neg \)) is defined inductively as follows:

1. Every propositional letter, \( x \in \mathbf{PV} \), is a proposition (an atomic proposition).
2. If \( A \) is a proposition, then \( \neg A \) is a proposition.
3. If \( A \) and \( B \) are propositions, then \( (A \lor B) \) is a proposition.
4. If \( A \) and \( B \) are propositions, then \( (A \land B) \) is a proposition.

Two propositions \( A \) and \( B \) are equivalent, denoted \( A \equiv B \), if
\[
v \models A \iff v \models B
\]
for all truth assignments, \( v \). It is easy to show that \( A \equiv B \) iff the proposition
\[
(\neg A \lor B) \land (\neg B \lor A)
\]
is valid.

**Definition 8.11.** A proposition \( P \) is in conjunctive normal form (CNF) if it is a conjunction \( P = C_1 \land \cdots \land C_n \) of propositions \( C_j \) which are disjunctions of literals (a literal is either a variable \( x \) or the negation \( \neg x \) (also denoted \( \overline{x} \)) of a variable \( x \)).

A proposition \( P \) is in disjunctive normal form (DNF) if it is a disjunction \( P = D_1 \lor \cdots \lor D_n \) of propositions \( D_j \) which are conjunctions of literals.

There are propositions such that any equivalent proposition in CNF has size exponential in terms of the original proposition.

**Example 8.9.** Here is such an example:
\[
A = (x_1 \land x_2) \lor (x_3 \land x_4) \lor \cdots \lor (x_{2n-1} \land x_{2n}).
\]
Observe that it is in DNF. We will prove a little later that any CNF for \( A \) contains \( 2^n \) occurrences of variables.

**Proposition 8.11.** Every proposition \( A \) is equivalent to a proposition \( A' \) in CNF.

There are several ways of proving Proposition 8.11. One method is algebraic, and consists in using the algebraic laws of boolean algebra. First one may convert a proposition to negation normal form, or \( \text{nnf} \).
Definition 8.12. A proposition is in negation normal form or nnf if all occurrences of $\neg$ only appear in front of propositional variables, but not in front of compound propositions.

Any proposition can be converted to an equivalent one in nnf by using the de Morgan laws:

\[
\neg(A \lor B) \equiv (\neg A \land \neg B)
\]
\[
\neg(A \land B) \equiv (\neg A \lor \neg B)
\]
\[
\neg\neg A \equiv A.
\]

Observe that if $A$ has $n$ connectives, then the equivalent formula $A'$ in nnf has at most $2n - 1$ connectives. Then a proposition in nnf can be converted to CNF.

A nice method to convert a proposition in nnf to CNF is to construct a tree whose nodes are labeled with sets of propositions using the following (Gentzen-style) rules:

\[
\frac{P, \Delta \quad Q, \Delta}{(P \land Q), \Delta}
\]

and

\[
\frac{P, Q, \Delta}{(P \lor Q), \Delta}
\]

where $\Delta$ stands for any set of propositions (even empty), and the comma stands for union. Thus, it is assumed that $(P \land Q) \notin \Delta$ in the first case, and that $(P \lor Q) \notin \Delta$ in the second case.

Since we interpret a set, $\Gamma$, of propositions as a disjunction, a valuation, $v$, satisfies $\Gamma$ iff it satisfies some proposition in $\Gamma$.

Observe that a valuation $v$ satisfies the conclusion of a rule iff it satisfies both premises in the first case, and the single premise in the second case. Using these rules, we can build a finite tree whose leaves are labeled with sets of literals.

By the above observation, a valuation $v$ satisfies the proposition labeling the root of the tree iff it satisfies all the propositions labeling the leaves of the tree.

But then, a CNF for the original proposition $A$ (in nnf, at the root of the tree) is the conjunction of the clauses appearing as the leaves of the tree. We may exclude the clauses that are tautologies, and we may discover in the process that $A$ is a tautology (when all leaves are tautologies).

Example 8.10. An illustration of the above method to convert the proposition

\[
A = (x_1 \land y_1) \lor (x_2 \land y_2)
\]

is shown below:
We obtain the CNF
\[ B = (x_1 \lor x_2) \land (x_1 \lor y_2) \land (y_1 \lor x_2) \land (y_1 \lor x_2). \]

**Remark:** Rules for dealing for \( \neg \) can also be created. In this case, we work with pairs of sets of propositions,
\[ \Gamma \to \Delta, \]
where, the propositions in \( \Gamma \) are interpreted conjunctively, and the propositions in \( \Delta \) are interpreted disjunctively. We obtain a sound and complete proof system for propositional logic (a “Gentzen-style” proof system, see *Logic for Computer Science*, Gallier [15]).

Going back to our “bad” proposition \( A \) from Example 8.9, by induction, we see that any tree for \( A \) has \( 2^n \) leaves.

However, the following result holds.

**Proposition 8.12.** For any proposition \( A \), we can construct in polynomial time a formula \( A' \) in CNF, so that \( A \) is satisfiable iff \( A' \) is satisfiable, by creating new variables.

**Sketch of proof.** First we convert \( A \) to nnf, which yields a proposition at most twice as long. Then we proceed recursively. For a conjunction \( C \land D \), we apply recursively the procedure to \( C \) and \( D \). The trick is that for a disjunction \( C \lor D \), first we apply recursively the procedure to \( C \) and \( D \) obtain
\[ (C_1 \land \cdots \land C_m) \lor (D_1 \land \cdots \land D_n) \]
where the \( C_i \)'s and the \( D_j \)'s are clauses. Then we create
\[ (C_1 \lor y) \land \cdots \land (C_m \lor y) \land (D_1 \lor \bar{y}) \land \cdots \land (D_n \lor \bar{y}), \]
where \( y \) is a new variable.

It can be shown that the number of new variables required is at most quadratic in the size of \( A \). For details on this construction see Hopcroft, Motwani and Ullman [20] (Section 10.3.3), but beware that the proof on page 455 contains a mistake. Repair the mistake.

**Example 8.11.** Consider the proposition
\[ A = (x_1 \land \neg x_2) \lor ((\neg x_1 \land x_2) \lor (x_2 \lor x_3)). \]
First, since \( x_1 \) and \( \neg x_2 \) are clauses, we get
\[ A_1 = x_1 \land \neg x_2. \]
Since \( \neg x_1, x_2 \) and \( x_2 \lor x_3 \) are clauses, from \((\neg x_1 \land x_2) \lor (x_2 \lor x_3)\) we construct

\[
A_2 = (\neg x_1 \lor y_1) \land (x_2 \lor y_1) \land (x_2 \lor x_3 \lor \neg y_1).
\]

Next, since \( A_1 \) and \( A_2 \) are conjunctions of clauses, we construct

\[
A' = (x_1 \lor y_2) \land (\neg x_2 \lor y_2) \land (\neg x_1 \lor y_1 \lor \neg y_2) \land (x_2 \lor y_1 \lor \neg y_2) \land (x_2 \lor x_3 \lor \neg y_1 \lor \neg y_2),
\]

a conjunction of clauses which is satisfiable iff \( A \) is satisfiable.

**Warning:** In general, the proposition \( A' \) is *not* equivalent to the proposition \( A \).

The question of uniqueness of the CNF is a bit tricky. For example, the proposition

\[
A = (u \land (x \lor y)) \lor (\neg u \land (x \lor y))
\]

has

\[
A_1 = (u \lor x \lor y) \land (\neg u \lor x \lor y),
A_2 = (u \lor \neg u) \land (x \lor y),
A_3 = x \lor y,
\]

as equivalent propositions in CNF!

We can get a *unique* CNF equivalent to a given proposition if we do the following:

1. Let \( \text{Var}(A) = \{x_1, \ldots, x_m\} \) be the set of variables occurring in \( A \).

2. Define a *maxterm* w.r.t. \( \text{Var}(A) \) as any disjunction of \( m \) pairwise distinct literals formed from \( \text{Var}(A) \), and not containing both some variable \( x_i \) and its negation \( \neg x_i \).

3. Then it can be shown that for any proposition \( A \) that is not a tautology, there is a *unique* proposition in CNF *equivalent* to \( A \), whose clauses consist of maxterms formed from \( \text{Var}(A) \).

The above definition can yield strange results. For instance, the CNF of any unsatisfiable proposition with \( m \) distinct variables is the conjunction of all of its \( 2^m \) maxterms! The above notion does not cope well with minimality.

For example, according to the above, the CNF of

\[
A = (u \land (x \lor y)) \lor (\neg u \land (x \lor y))
\]

should be

\[
A_1 = (u \lor x \lor y) \land (\neg u \lor x \lor y).
\]
Chapter 9

Some $\mathcal{NP}$-Complete Problems

9.1 Statements of the Problems

In this chapter we will show that certain classical algorithmic problems are $\mathcal{NP}$-complete. This chapter is heavily inspired by Lewis and Papadimitriou’s excellent treatment [24]. In order to study the complexity of these problems in terms of resource (time or space) bounded Turing machines (or RAM programs), it is crucial to be able to encode instances of a problem $P$ as strings in a language $L_P$. Then an instance of a problem $P$ is solvable iff the corresponding string belongs to the language $L_P$. This implies that our problems must have a yes–no answer, which is not always the usual formulation of optimization problems where what is required is to find some optimal solution, that is, a solution minimizing or maximizing some objective (cost) function $F$. For example the standard formulation of the traveling salesman problem asks for a tour (of the cities) of minimal cost.

Fortunately, there is a trick to reformulate an optimization problem as a yes–no answer problem, which is to explicitly incorporate a budget (or cost) term $B$ into the problem, and instead of asking whether some objective function $F$ has a minimum or a maximum $w$, we ask whether there is a solution $w$ such that $F(w) \leq B$ in the case of a minimum solution, or $F(w) \geq B$ in the case of a maximum solution.

If we are looking for a minimum of $F$, we try to guess the minimum value $B$ of $F$ and then we solve the problem of finding $w$ such that $F(w) \leq B$. If our guess for $B$ is too small, then we fail. In this case, we try again with a larger value of $B$. Otherwise, if $B$ was not too small we find some $w$ such that $F(w) \leq B$, but $w$ may not correspond to a minimum of $F$, so we try again with a smaller value of $B$, and so on. This yields an approximation method to find a minimum of $F$.

Similarly, if we are looking for a maximum of $F$, we try to guess the maximum value $B$ of $F$ and then we solve the problem of finding $w$ such that $F(w) \geq B$. If our guess for $B$ is too large, then we fail. In this case, we try again with a smaller value of $B$. Otherwise, if $B$ was not too large we find some $w$ such that $F(w) \geq B$, but $w$ may not correspond to a maximum of $F$, so we try again with a greater value of $B$, and so on. This yields an
approximation method to find a maximum of $F$.

We will see several examples of this technique in Problems 5–8 listed below.

The problems that will consider are

(1) Exact Cover
(2) Hamiltonian Cycle for directed graphs
(3) Hamiltonian Cycle for undirected graphs
(4) The Traveling Salesman Problem
(5) Independent Set
(6) Clique
(7) Node Cover
(8) Knapsack, also called subset sum
(9) Inequivalence of $*$-free Regular Expressions
(10) The 0-1-integer programming problem

We begin by describing each of these problems.

(1) Exact Cover

We are given a finite nonempty set $U = \{u_1, \ldots, u_n\}$ (the universe), and a family $\mathcal{F} = \{S_1, \ldots, S_m\}$ of $m \geq 1$ nonempty subsets of $U$. The question is whether there is an exact cover, that is, a subfamily $\mathcal{C} \subseteq \mathcal{F}$ of subsets in $\mathcal{F}$ such that the sets in $\mathcal{C}$ are disjoint and their union is equal to $U$.

For example, let $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, and let $\mathcal{F}$ be the family

$$\mathcal{F} = \{\{u_1, u_3\}, \{u_2, u_3, u_6\}, \{u_1, u_5\}, \{u_2, u_3, u_4\}, \{u_5, u_6\}, \{u_2, u_4\}\}.$$

The subfamily

$$\mathcal{C} = \{\{u_1, u_3\}, \{u_5, u_6\}, \{u_2, u_4\}\}$$

is an exact cover.

It is easy to see that Exact Cover is in $\mathcal{NP}$. To prove that it is $\mathcal{NP}$-complete, we will reduce the Satisfiability Problem to it. This means that we provide a method running in polynomial time that converts every instance of the Satisfiability Problem to an instance of Exact Cover, such that the first problem has a solution iff the converted problem has a solution.
(2) **Hamiltonian Cycle (for Directed Graphs)**

Recall that a directed graph $G$ is a pair $G = (V, E)$, where $E \subseteq V \times V$. Elements of $V$ are called nodes (or vertices). A pair $(u, v) \in E$ is called an edge of $G$. We will restrict ourselves to simple graphs, that is, graphs without edges of the form $(u, u)$; equivalently, $G = (V, E)$ is a simple graph if whenever $(u, v) \in E$, then $u \neq v$.

Given any two nodes $u, v \in V$, a path from $u$ to $v$ is any sequence of $n + 1$ edges ($n \geq 0$)

$$(u, v_1), (v_1, v_2), \ldots, (v_n, v).$$

(If $n = 0$, a path from $u$ to $v$ is simply a single edge, $(u, v)$.)

A directed graph $G$ is strongly connected if for every pair $(u, v) \in V \times V$, there is a path from $u$ to $v$. A closed path, or cycle, is a path from some node $u$ to itself. We will restrict our attention to finite graphs, i.e. graphs $(V, E)$ where $V$ is a finite set.

**Definition 9.1.** Given a directed graph $G$, a Hamiltonian cycle is a cycle that passes through all the nodes exactly once (note, some edges may not be traversed at all).

**Hamiltonian Cycle Problem (for Directed Graphs):** Given a directed graph $G$, is there an Hamiltonian cycle in $G$?

Is there a Hamiltonian cycle in the directed graph $D$ shown in Figure 9.1?

Finding a Hamiltonian cycle in this graph does not appear to be so easy! A solution is shown in Figure 9.2 below.

It is easy to see that **Hamiltonian Cycle (for Directed Graphs)** is in $\mathcal{NP}$. To prove that it is $\mathcal{NP}$-complete, we will reduce **Exact Cover** to it. This means that we provide a method running in polynomial time that converts every instance of **Exact Cover** to an instance of **Hamiltonian Cycle (for Directed Graphs)** such that the first problem has a solution iff the converted problem has a solution. This is perhaps the hardest reduction.
Figure 9.1: A tour “around the world.”

Figure 9.2: A Hamiltonian cycle in $D$. 
(3) **Hamiltonian Cycle (for Undirected Graphs)**

Recall that an undirected graph $G$ is a pair $G = (V, E)$, where $E$ is a set of subsets $\{u, v\}$ of $V$ consisting of exactly two distinct elements. Elements of $V$ are called nodes (or vertices). A pair $\{u, v\} \in E$ is called an edge of $G$.

Given any two nodes $u, v \in V$, a path from $u$ to $v$ is any sequence of $n$ nodes $(n \geq 2)$

$$u = u_1, u_2, \ldots, u_n = v$$

such that $\{u_i, u_{i+1}\} \in E$ for $i = 1, \ldots, n - 1$. (If $n = 2$, a path from $u$ to $v$ is simply a single edge, $\{u, v\}$.)

An undirected graph $G$ is connected if for every pair $(u, v) \in V \times V$, there is a path from $u$ to $v$. A closed path, or cycle, is a path from some node $u$ to itself.

**Definition 9.2.** Given an undirected graph $G$, a Hamiltonian cycle is a cycle that passes through all the nodes exactly once (note, some edges may not be traversed at all).

**Hamiltonian Cycle Problem (for Undirected Graphs):** Given an undirected graph $G$, is there an Hamiltonian cycle in $G$?

An instance of this problem is obtained by changing every directed edge in the directed graph of Figure 9.1 to an undirected edge. The directed Hamiltonian cycle given in Figure 9.1 is also an undirected Hamiltonian cycle of the undirected graph of Figure 9.3.

We see immediately that Hamiltonian Cycle (for Undirected Graphs) is in $NP$. To prove that it is $NP$-complete, we will reduce Hamiltonian Cycle (for Directed Graphs) to it. This means that we provide a method running in polynomial time that converts every instance of Hamiltonian Cycle (for Directed Graphs) to an instance of Hamiltonian Cycle (for Undirected Graphs) such that the first problem has a solution iff the converted problem has a solution. This is an easy reduction.

(4) **Traveling Salesman Problem**

We are given a set $\{c_1, c_2, \ldots, c_n\}$ of $n \geq 2$ cities, and an $n \times n$ matrix $D = (d_{ij})$ of nonnegative integers, where $d_{ij}$ is the distance (or cost) of traveling from city $c_i$ to city $c_j$. We assume that $d_{ii} = 0$ and $d_{ij} = d_{ji}$ for all $i, j$, so that the matrix $D$ is symmetric and has zero diagonal.

**Traveling Salesman Problem:** Given some $n \times n$ matrix $D = (d_{ij})$ as above and some integer $B \geq 0$ (the budget of the traveling salesman), find a permutation $\pi$ of $\{1, 2, \ldots, n\}$ such that

$$c(\pi) = d_{\pi(1)\pi(2)} + d_{\pi(2)\pi(3)} + \cdots + d_{\pi(n-1)\pi(n)} + d_{\pi(n)\pi(1)} \leq B.$$
The quantity \( c(\pi) \) is the cost of the trip specified by \( \pi \). The Traveling Salesman Problem has been stated in terms of a budget so that it has a yes or no answer, which allows us to convert it into a language. A minimal solution corresponds to the smallest feasible value of \( B \).

**Example 9.1.** Consider the \( 4 \times 4 \) symmetric matrix given by

\[
D = \begin{pmatrix}
0 & 2 & 1 & 1 \\
2 & 0 & 1 & 1 \\
1 & 1 & 0 & 3 \\
1 & 1 & 3 & 0
\end{pmatrix},
\]

and the budget \( B = 4 \). The tour specified by the permutation

\[
\pi = \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 4 & 2 & 3
\end{pmatrix}
\]

has cost 4, since

\[
c(\pi) = d_{\pi(1)\pi(2)} + d_{\pi(2)\pi(3)} + d_{\pi(3)\pi(4)} + d_{\pi(4)\pi(1)}
= d_{14} + d_{42} + d_{23} + d_{31}
= 1 + 1 + 1 + 1 = 4.
\]
The cities in this tour are traversed in the order

\[(1, 4, 2, 3, 1)\].

It is clear that the **Traveling Salesman Problem** is in \(\mathcal{NP}\). To show that it is \(\mathcal{NP}\)-complete, we reduce the **Hamiltonian Cycle Problem (Undirected Graphs)** to it. This means that we provide a method running in polynomial time that converts every instance of **Hamiltonian Cycle Problem (Undirected Graphs)** to an instance of the **Traveling Salesman Problem** such that the first problem has a solution iff the converted problem has a solution.

(5) **Independent Set**

The problem is this: Given an undirected graph \(G = (V, E)\) and an integer \(K \geq 2\), is there a set \(C\) of nodes with \(|C| \geq K\) such that for all \(v_i, v_j \in C\), there is no edge \(\{v_i, v_j\} \in E\)?

A maximal independent set with 3 nodes is shown in Figure 9.4. A maximal solution corresponds to the largest feasible value of \(K\). The problem **Independent Set** is obviously in \(\mathcal{NP}\). To show that it is \(\mathcal{NP}\)-complete, we reduce **Exact 3-Satisfiability** to it. This means that we provide a method running in polynomial time that converts every instance of **Exact 3-Satisfiability** to an instance of **Independent Set** such that the first problem has a solution iff the converted problem has a solution.
(6) **Clique**

The problem is this: Given an undirected graph $G = (V, E)$ and an integer $K \geq 2$, is there a set $C$ of nodes with $|C| \geq K$ such that for all $v_i, v_j \in C$, there is some edge $\{v_i, v_j\} \in E$? Equivalently, does $G$ contain a complete subgraph with at least $K$ nodes?

A maximal clique with 4 nodes is shown in Figure 9.5. A maximal solution corresponds to the largest feasible value of $K$. The problem **Clique** is obviously in $\mathcal{NP}$. To show that it is $\mathcal{NP}$-complete, we reduce **Independent Set** to it. This means that we provide a method running in polynomial time that converts every instance of **Independent Set** to an instance of **Clique** such that the first problem has a solution iff the converted problem has a solution.

![Figure 9.5: A maximal Clique in a graph.](image)

(7) **Node Cover**

The problem is this: Given an undirected graph $G = (V, E)$ and an integer $B \geq 2$, is there a set $C$ of nodes with $|C| \leq B$ such that $C$ covers all edges in $G$, which means that for every edge $\{v_i, v_j\} \in E$, either $v_i \in C$ or $v_j \in C$?

A minimal node cover with 6 nodes is shown in Figure 9.6. A minimal solution corresponds to the smallest feasible value of $B$. The problem **Node Cover** is obviously in $\mathcal{NP}$. To show that it is $\mathcal{NP}$-complete, we reduce **Independent Set** to it. This means that we provide a method running in polynomial time that converts every instance of
**Independent Set** to an instance of **Node Cover** such that the first problem has a solution iff the converted problem has a solution.

The Node Cover problem has the following interesting interpretation: think of the nodes of the graph as rooms of a museum (or art gallery *etc.*), and each edge as a straight corridor that joins two rooms. Then Node Cover may be useful in assigning as few as possible guards to the rooms, so that all corridors can be seen by a guard.

(8) **Knapsack** (also called **Subset sum**)

The problem is this: Given a finite nonempty set $S = \{a_1, a_2, \ldots, a_n\}$ of nonnegative integers, and some integer $K \geq 0$, all represented in binary, is there a nonempty subset $I \subseteq \{1, 2, \ldots, n\}$ such that

$$\sum_{i \in I} a_i = K?$$

A “concrete” realization of this problem is that of a hiker who is trying to fill her/his backpack to its maximum capacity with items of varying weights or values.

It is easy to see that the **Knapsack** Problem is in $\mathcal{NP}$. To show that it is $\mathcal{NP}$-complete, we reduce **Exact Cover** to it. This means that we provide a method running in polynomial time that converts every instance of **Exact Cover** to an instance of **Knapsack** Problem such that the first problem has a solution iff the converted problem has a solution.
Remark: The 0-1 Knapsack Problem is defined as the following problem. Given a set of \( n \) items, numbered from 1 to \( n \), each with a weight \( w_i \in \mathbb{N} \) and a value \( v_i \in \mathbb{N} \), given a maximum capacity \( W \in \mathbb{N} \) and a budget \( B \in \mathbb{N} \), is there a set of \( n \) variables \( x_1, \ldots, x_n \) with \( x_i \in \{0, 1\} \) such that

\[
\sum_{i=1}^{n} x_i v_i \geq B,
\]

\[
\sum_{i=1}^{n} x_i w_i \leq W.
\]

Informally, the problem is to pick items to include in the knapsack so that the sum of the values exceeds a given minimum \( B \) (the goal is to maximize this sum), and the sum of the weights is less than or equal to the capacity \( W \) of the knapsack. A maximal solution corresponds to the largest feasible value of \( B \).

The Knapsack Problem as we defined it (which is how Lewis and Papadimitriou define it) is the special case where \( v_i = w_i = 1 \) for \( i = 1, \ldots, n \) and \( W = B \). For this reason, it is also called the Subset Sum Problem. Clearly, the Knapsack (Subset Sum) Problem reduces to the 0-1 Knapsack Problem, and thus the 0-1 Knapsack Problem is also NP-complete.

(9) Inequivalence of \( \ast \)-free Regular Expressions

Recall that the problem of deciding the equivalence \( R_1 \cong R_2 \) of two regular expressions \( R_1 \) and \( R_2 \) is the problem of deciding whether \( R_1 \) and \( R_2 \) define the same language, that is, \( \mathcal{L}[R_1] = \mathcal{L}[R_2] \). Is this problem in \( \mathcal{NP} \)?

In order to show that the equivalence problem for regular expressions is in \( \mathcal{NP} \) we would have to be able to somehow check in polynomial time that two expressions define the same language, but this is still an open problem.

What might be easier is to decide whether two regular expressions \( R_1 \) and \( R_2 \) are inequivalent. For this, we just have to find a string \( w \) such that either \( w \in \mathcal{L}[R_1] - \mathcal{L}[R_2] \) or \( w \in \mathcal{L}[R_2] - \mathcal{L}[R_1] \). The problem is that if we can guess such a string \( w \), we still have to check in polynomial time that \( w \in (\mathcal{L}[R_1] - \mathcal{L}[R_2]) \cup (\mathcal{L}[R_2] - \mathcal{L}[R_1]) \), and this implies that there is a bound on the length of \( w \) which is polynomial in the sizes of \( R_1 \) and \( R_2 \). Again, this is an open problem.

To obtain a problem in \( \mathcal{NP} \) we have to consider a restricted type of regular expressions, and it turns out that \( \ast \)-free regular expressions are the right candidate. A \( \ast \)-free regular expression is a regular expression which is built up from the atomic expressions using only \( + \) and \( \cdot \), but not \( \ast \). For example,

\[
R = ((a + b)aa(a + b) + aba(a + b)b)
\]
9.1. STATEMENTS OF THE PROBLEMS

is such an expression.

It is easy to see that if $R$ is a $\ast$-free regular expression, then for every string $w \in \mathcal{L}[R]$ we have $|w| \leq |R|$. In particular, $\mathcal{L}[R]$ is finite. The above observation shows that if $R_1$ and $R_2$ are $\ast$-free and if there is a string $w \in (\mathcal{L}[R_1] - \mathcal{L}[R_2]) \cup (\mathcal{L}[R_2] - \mathcal{L}[R_1])$, then $|w| \leq |R_1| + |R_2|$, so we can indeed check this in polynomial time. It follows that the inequivalence problem for $\ast$-free regular expressions is in \textit{NP}. To show that it is \textit{NP}-complete, we reduce the \textbf{Satisfiability Problem} to it. This means that we provide a method running in polynomial time that converts every instance of \textbf{Satisfiability Problem} to an instance of \textbf{Inequivalence of Regular Expressions} such that the first problem has a solution iff the converted problem has a solution.

Observe that both problems of \textbf{Inequivalence of Regular Expressions} and \textbf{Equivalence of Regular Expressions} are as hard as \textbf{Inequivalence of $\ast$-free Regular Expressions}, since if we could solve the first two problems in polynomial time, then we could solve \textbf{Inequivalence of $\ast$-free Regular Expressions} in polynomial time, but since this problem is \textit{NP}-complete, we would have $\mathcal{P} = \mathcal{NP}$. This is very unlikely, so the complexity of \textbf{Equivalence of Regular Expressions} remains open.

(10) \textbf{0-1 integer programming problem}

Let $A$ be any $p \times q$ matrix with integer coefficients and let $b \in \mathbb{Z}^p$ be any vector with integer coefficients. The \textbf{0-1 integer programming problem} is to find whether a system of $p$ linear equations in $q$ variables

$$a_{11}x_1 + \cdots + a_{1q}x_q = b_1$$
$$\vdots \hspace{1cm} \vdots$$
$$a_{i1}x_1 + \cdots + a_{iq}x_q = b_i$$
$$\vdots \hspace{1cm} \vdots$$
$$a_{p1}x_1 + \cdots + a_{pq}x_q = b_p$$

with $a_{ij}, b_i \in \mathbb{Z}$ has any solution $x \in \{0, 1\}^q$, that is, with $x_i \in \{0, 1\}$. In matrix form, if we let

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1q} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pq} \end{pmatrix}, \hspace{1cm} b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}, \hspace{1cm} x = \begin{pmatrix} x_1 \\ \vdots \\ x_q \end{pmatrix},$$

then we write the above system as

$$Ax = b.$$
Example 9.2. Is there a solution \( x = (x_1, x_2, x_3, x_4, x_5, x_6) \) of the linear system

\[
\begin{pmatrix}
  1 & -2 & 1 & 3 & -1 & 4 \\
  2 & 2 & -1 & 0 & 1 & -1 \\
 -1 & 1 & 2 & 3 & -2 & 3 \\
  3 & 1 & -1 & 2 & -1 & 4 \\
  0 & 1 & -1 & 1 & 1 & 2 \\
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
  x_6 \\
\end{pmatrix}
= \begin{pmatrix}
  9 \\
  0 \\
  7 \\
  8 \\
  2 \\
\end{pmatrix}
\]

with \( x_i \in \{0, 1\} \)?

Indeed, \( x = (1, 0, 1, 1, 0, 1) \) is a solution.

It is immediate that 0-1 integer programming problem is in \( \mathcal{NP} \). To prove that it is \( \mathcal{NP} \)-complete we reduce the bounded tiling problem to it. This means that we provide a method running in polynomial time that converts every instance of the bounded tiling problem to an instance of the 0-1 integer programming problem such that the first problem has a solution iff the converted problem has a solution.

9.2 Proofs of \( \mathcal{NP} \)-Completeness

(1) Exact Cover

To prove that Exact Cover is \( \mathcal{NP} \)-complete, we reduce the Satisfiability Problem to it:

Satisfiability Problem \( \leq_P \) Exact Cover

Given a set \( F = \{C_1, \ldots, C_\ell\} \) of \( \ell \) clauses constructed from \( n \) propositional variables \( x_1, \ldots, x_n \), we must construct in polynomial time (in the sum of the lengths of the clauses) an instance \( \tau(F) = (U, F) \) of Exact Cover such that \( F \) is satisfiable iff \( \tau(F) \) has a solution.

Example 9.3. If

\[
F = \{C_1 = (x_1 \lor \overline{x}_2), C_2 = (\overline{x}_1 \lor x_2 \lor x_3), C_3 = (x_2), C_4 = (\overline{x}_2 \lor \overline{x}_3)\},
\]

then the universe \( U \) is given by

\[
U = \{x_1, x_2, x_3, C_1, C_2, C_3, C_4, p_{11}, p_{12}, p_{21}, p_{22}, p_{23}, p_{31}, p_{41}, p_{42}\}.
\]
and the family $F$ consists of the subsets

\[
\{p_{11}\}, \{p_{12}\}, \{p_{21}\}, \{p_{22}\}, \{p_{23}\}, \{p_{31}\}, \{p_{41}\}, \{p_{42}\}
\]

\[
T_{1,F} = \{x_1, p_{11}\}
\]

\[
T_{1,T} = \{x_1, p_{21}\}
\]

\[
T_{2,F} = \{x_2, p_{22}, p_{31}\}
\]

\[
T_{2,T} = \{x_2, p_{12}, p_{41}\}
\]

\[
T_{3,F} = \{x_3, p_{23}\}
\]

\[
T_{3,T} = \{x_3, p_{42}\}
\]

\[
\{C_1, p_{11}\}, \{C_1, p_{12}\}, \{C_2, p_{21}\}, \{C_2, p_{22}\}, \{C_2, p_{23}\},
\]

\[
\{C_3, p_{31}\}, \{C_4, p_{41}\}, \{C_4, p_{42}\}.
\]

The above construction is illustrated in Figure 9.7.

Figure 9.7: Construction of an exact cover from the set of clauses in Example 9.3.

It is easy to check that the set $C$ consisting of the following subsets is an exact cover:

\[
T_{1,T} = \{x_1, p_{21}\}, T_{2,T} = \{x_2, p_{12}, p_{41}\}, T_{3,F} = \{x_3, p_{23}\},
\]

\[
\{C_1, p_{11}\}, \{C_2, p_{22}\}, \{C_3, p_{31}\}, \{C_4, p_{42}\}.
\]
The general method to construct \((U, \mathcal{F})\) from \(F = \{C_1, \ldots, C_\ell\}\) proceeds as follows. The size \(n\) of the input is the sum of the lengths of the clauses \(C_i\) as strings. Say

\[ C_j = (L_{j1} \lor \cdots \lor L_{jm_j}) \]

is the \(j\)th clause in \(F\), where \(L_{jk}\) denotes the \(k\)th literal in \(C_j\) and \(m_j \geq 1\). The universe of \(\tau(F)\) is the set

\[ U = \{x_i \mid 1 \leq i \leq n\} \cup \{C_j \mid 1 \leq j \leq \ell\} \cup \{p_{jk} \mid 1 \leq j \leq \ell, \ 1 \leq k \leq m_j\} \]

where in the third set \(p_{jk}\) corresponds to the \(k\)th literal in \(C_j\). The universe \(U\) can be constructed in time \(O(n^2)\).

The following subsets are included in \(\mathcal{F}\):

(a) There is a set \(\{p_{jk}\}\) for every \(p_{jk}\).

(b) For every boolean variable \(x_i\), the following two sets are in \(\mathcal{F}\):

\[ T_{i, \mathbf{T}} = \{x_i\} \cup \{p_{jk} \mid L_{jk} = \overline{x_i}\} \]

which contains \(x_i\) and all negative occurrences of \(x_i\); and

\[ T_{i, \mathbf{F}} = \{x_i\} \cup \{p_{jk} \mid L_{jk} = x_i\} \]

which contains \(x_i\) and all its positive occurrences. Note carefully that \(T_{i, \mathbf{T}}\) involves negative occurrences of \(x_i\) whereas \(T_{i, \mathbf{F}}\) involves positive occurrences of \(x_i\).

(c) For every clause \(C_j\), the \(m_j\) sets \(\{C_j, p_{jk}\}\) are in \(\mathcal{F}\).

The subsets in (a), (b), (c) can be constructed in time \(O(n^3)\). It remains to prove that \(F\) is satisfiable iff \(\tau(F)\) has a solution. We claim that if \(v\) is a truth assignment that satisfies \(F\), then we can make an exact cover \(\mathcal{C}\) as follows:

For each \(x_i\), we put the subset \(T_{i, \mathbf{T}}\) in \(\mathcal{C}\) iff \(v(x_i) = \mathbf{T}\), else we put the subset \(T_{i, \mathbf{F}}\) in \(\mathcal{C}\) iff \(v(x_i) = \mathbf{F}\). Also, for every clause \(C_j\), we put some subset \(\{C_j, p_{jk}\}\) in \(\mathcal{C}\) for a literal \(L_{jk}\) which is made true by \(v\). By construction of \(T_{i, \mathbf{T}}\) and \(T_{i, \mathbf{F}}\), this \(p_{jk}\) is not in any set in \(\mathcal{C}\) selected so far. Since by hypothesis \(F\) is satisfiable, such a literal exists for every clause. Having covered all \(x_i\) and \(C_j\), we put a set \(\{p_{jk}\}\) in \(\mathcal{C}\) for every remaining \(p_{jk}\) which has not yet been covered by the sets already in \(\mathcal{C}\).

Going back to Example 9.3, the truth assignment \(v(x_1) = \mathbf{T}, v(x_2) = \mathbf{T}, v(x_3) = \mathbf{F}\) satisfies \(F\), so we put

\[ T_{1, \mathbf{T}} = \{x_1, p_{21}\}, T_{2, \mathbf{T}} = \{x_2, p_{12}, p_{41}\}, T_{3, \mathbf{F}} = \{x_3, p_{23}\}, \]

\(\{C_1, p_{11}\}, \{C_2, p_{22}\}, \{C_3, p_{31}\}, \{C_4, p_{42}\}\)
9.2. PROOFS OF \( \mathcal{NP} \)-COMPLETENESS

in \( \mathcal{C} \).

We leave as an exercise to check that the above procedure works.

Conversely, if \( \mathcal{C} \) is an exact cover of \( \tau(\mathcal{F}) \), we define a truth assignment as follows:

For every \( x_i \), if \( T_{i,\mathbf{T}} \) is in \( \mathcal{C} \), then we set \( v(x_i) = \mathbf{T} \), else if \( T_{i,\mathbf{F}} \) is in \( \mathcal{C} \), then we set \( v(x_i) = \mathbf{F} \). We leave it as an exercise to check that this procedure works.

Example 9.4. Given the exact cover

\[
T_{1,\mathbf{T}} = \{x_1, p_{21}\}, T_{2,\mathbf{T}} = \{x_2, p_{12}, p_{41}\}, T_{3,\mathbf{F}} = \{x_3, p_{23}\},
\{C_1, p_{11}\}, \{C_2, p_{22}\}, \{C_3, p_{31}\}, \{C_4, p_{42}\},
\]
we get the satisfying assignment \( v(x_1) = \mathbf{T}, v(x_2) = \mathbf{T}, v(x_3) = \mathbf{F} \).

If we now consider the proposition is CNF given by

\[
F_2 = \{C_1 = (x_1 \lor \overline{x_2}), C_2 = (\overline{x_1} \lor x_2 \lor x_3), C_3 = (x_2), C_4 = (\overline{x_2} \lor \overline{x_3} \lor x_4)\}
\]

where we have added the boolean variable \( x_4 \) to clause \( C_4 \), then \( U \) also contains \( x_4 \) and \( p_{43} \) so we need to add the following subsets to \( \mathcal{F} \):

\[
T_{4,\mathbf{F}} = \{x_4, p_{43}\}, T_{4,\mathbf{T}} = \{x_4\}, \{C_4, p_{43}\}, \{p_{43}\}.
\]

The truth assignment \( v(x_1) = \mathbf{T}, v(x_2) = \mathbf{T}, v(x_3) = \mathbf{F}, v(x_4) = \mathbf{T} \) satisfies \( F_2 \), so an exact cover \( \mathcal{C} \) is

\[
T_{1,\mathbf{T}} = \{x_1, p_{21}\}, T_{2,\mathbf{T}} = \{x_2, p_{12}, p_{41}\}, T_{3,\mathbf{F}} = \{x_3, p_{23}\}, T_{4,\mathbf{T}} = \{x_4\},
\{C_1, p_{11}\}, \{C_2, p_{22}\}, \{C_3, p_{31}\}, \{C_4, p_{42}\}, \{p_{43}\}.
\]

The above construction is illustrated in Figure 9.8.

Observe that this time, because the truth assignment \( v \) makes both literals corresponding to \( p_{42} \) and \( p_{43} \) true and since we picked \( p_{42} \) to form the subset \( \{C_4, p_{42}\} \), we need to add the singleton \( \{p_{43}\} \) to \( \mathcal{C} \) to cover all elements of \( U \).

(2) Hamiltonian Cycle (for Directed Graphs)

To prove that Hamiltonian Cycle (for Directed Graphs) is \( \mathcal{NP} \)-complete, we will reduce Exact Cover to it:

Exact Cover \( \leq_p \) Hamiltonian Cycle (for Directed Graphs)

We need to find an algorithm working in polynomial time that converts an instance \((U, \mathcal{F})\) of Exact Cover to a directed graph \( G = \tau(U, \mathcal{F}) \) such that \( G \) has a Hamiltonian cycle iff \((U, \mathcal{F})\) has an exact cover. The size \( n \) of the input \((U, \mathcal{F})\) is \(|U| + |\mathcal{F}|\).
The construction of the graph $G$ uses a trick involving a small subgraph $Gad$ with 7 (distinct) nodes known as a gadget shown in Figure 9.9.

The crucial property of the graph $Gad$ is that if $Gad$ is a subgraph of a bigger graph $G$ in such a way that no edge of $G$ is incident to any of the nodes $u, v, w$ unless it is one of the eight edges of $Gad$ incident to the nodes $u, v, w$, then for any Hamiltonian cycle in $G$, either the path $(a, u), (u, v), (v, w), (w, b)$ is traversed or the path $(c, w), (w, v), (v, u), (u, d)$ is traversed, but not both.

The reader should convince herself/himself that indeed, any Hamiltonian cycle that does not traverse either the subpath $(a, u), (u, v), (v, w), (w, b)$ from $a$ to $b$ or the subpath $(c, w), (w, v), (v, u), (u, d)$ from $c$ to $d$ will not traverse one of the nodes $u, v, w$. Also, the fact that node $v$ is traversed exactly once forces only one of the two paths to be traversed but not both. The reader should also convince herself/himself that a smaller graph does not guarantee the desired property.

It is convenient to use the simplified notation with a special type of edge labeled with the exclusive or sign $\oplus$ between the “edges” between $a$ and $b$ and between $d$ and $c$, as shown in Figure 9.10.
Whenever such a figure occurs, the actual graph is obtained by substituting a copy of the graph \( \text{Gad} \) (the four nodes \( a, b, c, d \) must be distinct). This abbreviating device can be extended to the situation where we build gadgets between a given pair \( (a, b) \) and several other pairs \( (c_1, d_1), \ldots, (c_m, d_m) \), all nodes being distinct, as illustrated in Figure 9.11.

Either all three edges \( (c_1, d_1), (c_2, d_2), (c_3, d_3) \) are traversed or the edge \( (a, b) \) is traversed, and these possibilities are mutually exclusive.

The graph \( G = \tau(U, \mathcal{F}) \) where \( U = \{u_1, \ldots, u_n\} \) (with \( n \geq 1 \)) and \( \mathcal{F} = \{S_1, \ldots, S_m\} \) (with \( m \geq 1 \)) is constructed as follows:

The graph \( G \) has \( m + n + 2 \) nodes \( \{u_0, u_1, \ldots, u_n, S_0, S_1, \ldots, S_m\} \). Note that we have added two extra nodes \( u_0 \) and \( S_0 \). For \( i = 1, \ldots, m \), there are two edges \( (S_{i-1}, S_i)_1 \) and \( (S_{i-1}, S_i)_2 \) from \( S_{i-1} \) to \( S_i \). For \( j = 1, \ldots, n \), from \( u_{j-1} \) to \( u_j \), there are as many edges as there are sets \( S_i \in \mathcal{F} \) containing the element \( u_j \). We can think of each edge between \( u_{j-1} \) and \( u_j \) as an occurrence of \( u_j \) in a uniquely determined set \( S_i \in \mathcal{F} \); we denote this edge by \( (u_{j-1}, u_j)_i \). We also have an edge from \( u_n \) to \( S_0 \) and an edge from \( S_m \) to \( u_0 \), thus “closing the cycle.”

What we have constructed so far is not a legal graph since it may have many parallel
edges, but are going to turn it into a legal graph by pairing edges between the \(u\)'s and edges between the \(S\)'s. Indeed, since each edge \((u_{j-1}, u_j)\), between \(u_{j-1}\) and \(u_j\) corresponds to an occurrence of \(u\) in some uniquely determined set \(S_i \in \mathcal{F}\) (that is, \(u_j \in S_i\)), we put an exclusive-or edge between the edge \((u_{j-1}, u_j)\) and the edge \((S_{i-1}, S_i)\) between \(S_{i-1}\) and \(S_i\), which we call the long edge. The other edge \((S_{i-1}, S_i)\) between \(S_{i-1}\) and \(S_i\) (not paired with any other edge) is called the short edge. Effectively, we put a copy of the gadget graph \(Gad\) with \(a = u_{j-1}, b = u_j, c = S_{i-1}, d = S_i\) for any pair \((u_j, S_i)\) such that \(u_j \in S_i\). The resulting object is indeed a directed graph with no parallel edges. The graph \(G\) can be constructed from \((U, \mathcal{F})\) in time \(O(n^2)\).

**Example 9.5.** The above construction is illustrated in Figure 9.12 for the instance of the exact cover problem given by

\[
U = \{u_1, u_2, u_3, u_4\}, \quad \mathcal{F} = \{S_1 = \{u_3, u_4\}, S_2 = \{u_2, u_3, u_4\}, S_3 = \{u_1, u_2\}\}.
\]

It remains to prove that \((U, \mathcal{F})\) has an exact cover iff the graph \(G = \tau(U, \mathcal{F})\) has a Hamiltonian cycle. First, assume that \(G\) has a Hamiltonian cycle. If so, for every \(j\) some unique “edge” \((u_{j-1}, u_j)\) is traversed once (since every \(u_j\) is traversed once), and by the exclusive-or nature of the gadget graphs, the corresponding long edge \((S_{i-1}, S_i)\) can’t be traversed, which means that the short edge \((S_{i-1}, S_i)\) is traversed. Consequently, if \(C\) consists of those subsets \(S_i\) such that the short edge \((S_{i-1}, S_i)\) is traversed, then \(C\) consists of pairwise disjoint subsets whose union is \(U\), namely \(C\) is an exact cover.

In our example, there is a Hamiltonian where the blue edges are traversed between the \(S\) nodes, and the red edges are traversed between the \(u\) nodes, namely

- short \((S_0, S_1)\), long \((S_1, S_2)\), short \((S_2, S_3)\), \((S_3, u_0)\),
- \((u_0, u_1)_3\), \((u_1, u_2)_3\), \((u_2, u_3)_1\), \((u_3, u_4)_1\), \((u_4, S_0)\).
The subsets corresponding to the short \((S_i - 1, S_i)\) edges are \(S_1\) and \(S_3\), and indeed \(C = \{S_1, S_3\}\) is an exact cover.

Note that the exclusive-or property of the gadgets implies the following: since the edge \((u_0, u_1)_3\) must be chosen to obtain a Hamiltonian, the long edge \((S_2, S_3)\) can’t be chosen, so the edge \((u_1, u_2)_3\) must be chosen, but then the edge \((u_1, u_2)_2\) is not chosen so the long edge \((S_1, S_2)\) must be chosen, so the edges \((u_2, u_3)_2\) and \((u_3, u_4)_2\) can’t be chosen, and thus edges \((u_2, u_3)_1\) and \((u_3, u_4)_1\) must be chosen.

Conversely, if \(C\) is an exact cover for \((U, F)\), then consider the path in \(G\) obtained by traversing each short edge \((S_i - 1, S_i)_1\) for which \(S_i \in C\), each edge \((u_{j-1}, u_j)_i\) such that \(u_j \in S_i\), which means that this edge is connected by a \(\oplus\)-sign to the long edge \((S_i - 1, S_i)_2\) (by construction, for each \(u_j\) there is a unique such \(S_i\)), and the edges \((u_n, S_0)\) and \((S_m, u_0)\), then we obtain a Hamiltonian cycle. Observe that the long edges are the inside edges joining the \(S_i\).
In our example, the exact cover $\mathcal{C} = \{S_1, S_3\}$ yields the Hamiltonian

$$\text{short } (S_0, S_1), \text{ long } (S_1, S_2), \text{ short } (S_2, S_3), (S_3, u_0), (u_0, u_1), (u_1, u_2), (u_2, u_3), (u_3, u_4), (u_4, S_0)$$

that we encountered earlier.

(3) Hamiltonian Cycle (for Undirected Graphs)

To show that Hamiltonian Cycle (for Undirected Graphs) is $\mathcal{NP}$-complete we reduce Hamiltonian Cycle (for Directed Graphs) to it:

**Hamiltonian Cycle (for Directed Graphs) $\leq^P$ Hamiltonian Cycle (for Undirected Graphs)**

Given any directed graph $G = (V, E)$ we need to construct in polynomial time an undirected graph $\tau(G) = G' = (V', E')$ such that $G$ has a (directed) Hamiltonian cycle iff $G'$ has a (undirected) Hamiltonian cycle. This is easy. We make three distinct copies $v_0, v_1, v_2$ of every node $v \in V$ which we put in $V'$, and for every edge $(u, v) \in E$ we create five edges $\{u_0, u_1\}, \{u_1, u_2\}, \{u_2, v_0\}, \{v_0, v_1\}, \{v_1, v_2\}$ which we put in $E'$, as illustrated in the diagram shown in Figure 9.13.

![Figure 9.13: Conversion of a directed graph into an undirected graph.](image)

If the size $n$ of the input is $|V| + |E|$, then $G'$ is constructed in time $O(n)$. The crucial point about the graph $G'$ is that although there may be several edges adjacent to a node $u_0$ or a node $u_2$, the only way to reach $u_1$ from $u_0$ is through the edge $\{u_0, u_1\}$ and the only way to reach $u_1$ from $u_2$ is through the edge $\{u_1, u_2\}$.

Suppose there is a Hamiltonian cycle in $G'$. If this cycle arrives at a node $u_0$ from the node $u_1$, then by the above remark, the previous node in the cycle must be $u_2$. Then the predecessor of $u_2$ in the cycle must be a node $v_0$ such that there is an edge $\{u_2, v_0\}$ in $G'$ arising from an edge $(u, v)$ in $G$. The nodes in the cycle in $G'$ are traversed in the order $(v_0, u_2, u_1, u_0)$ where $v_0$ and $u_2$ are traversed in the opposite order in which they occur as the endpoints of the edge $(u, v)$ in $G$. If so, consider the reverse of our Hamiltonian cycle in $G'$, which is also a Hamiltonian cycle since $G'$ is unoriented. In this cycle, we go from $u_0$ to $u_1$, then to $u_2$, and finally to $v_0$. In $G$, we traverse the edge from $u$ to $v$. In order for the cycle in $G'$ to be Hamiltonian, we must continue...
by visiting \( v_1 \) and \( v_2 \), since otherwise \( v_1 \) is never traversed. Now the next node \( w_0 \) in the Hamiltonian cycle in \( G' \) corresponds to an edge \((v, w)\) in \( G \), and by repeating our reasoning we see that our Hamiltonian cycle in \( G' \) determines a Hamiltonian cycle in \( G \). We leave it as an easy exercise to check that a Hamiltonian cycle in \( G \) yields a Hamiltonian cycle in \( G' \). The process of expanding a directed graph into an undirected graph and the inverse process are illustrated in Figure 9.14 and Figure 9.15.

![Figure 9.14: Expanding the directed graph into an undirected graph.](image)

(4) **Traveling Salesman Problem**

To show that the **Traveling Salesman Problem** is \( \mathcal{NP} \)-complete, we reduce the **Hamiltonian Cycle Problem (Undirected Graphs)** to it:

**Hamiltonian Cycle Problem (Undirected Graphs) \( \leq_p \) Traveling Salesman Problem**

This is a fairly easy reduction.

Given an undirected graph \( G = (V, E) \), we construct an instance \( \tau(G) = (D, B) \) of the Traveling Salesman Problem so that \( G \) has a Hamiltonian cycle iff the traveling salesman problem has a solution. If we let \( n = |V| \), we have \( n \) cities and the matrix
Figure 9.15: Collapsing the undirected graph onto a directed graph.

\[ D = (d_{ij}) \] is defined as follows:

\[
d_{ij} = \begin{cases} 
0 & \text{if } i = j \\
1 & \text{if } \{v_i, v_j\} \in E \\
2 & \text{otherwise.}
\end{cases}
\]

We also set the budget \( B \) as \( B = n \). The construction of \((D, B)\) from \( G \) can be done in time \( O(n^2) \).

Any tour of the cities has cost equal to \( n \) plus the number of pairs \( (v_i, v_j) \) such that \( i \neq j \) and \( \{v_i, v_j\} \) is not an edge of \( G \). It follows that a tour of cost \( n \) exists iff there are no pairs \( (v_i, v_j) \) of the second kind iff the tour is a Hamiltonian cycle.

The reduction from Hamiltonian Cycle Problem (Undirected Graphs) to the Traveling Salesman Problem is quite simple, but a direct reduction of say Satisfiability to the Traveling Salesman Problem is hard. By breaking this reduction into several steps made it simpler to achieve.

(5) Independent Set
To show that **Independent Set** is \( \mathcal{NP} \)-complete, we reduce **Exact 3-Satisfiability** to it:

**Exact 3-Satisfiability \( \leq_P \) Independent Set**

Recall that in **Exact 3-Satisfiability** every clause \( C_i \) has exactly three literals \( L_{i1}, L_{i2}, L_{i3} \).

Given a set \( F = \{C_1, \ldots, C_m\} \) of \( m \geq 2 \) such clauses, we construct in polynomial time an undirected graph \( G = (V, E) \) such that \( F \) is satisfiable iff \( G \) has an independent set \( C \) with at least \( K = m \) nodes.

For every \( i \) (\( 1 \leq i \leq m \)), we have three nodes \( c_{i1}, c_{i2}, c_{i3} \) corresponding to the three literals \( L_{i1}, L_{i2}, L_{i3} \) in clause \( C_i \), so there are \( 3m \) nodes in \( V \). The “core” of \( G \) consists of \( m \) triangles, one for each set \( \{c_{i1}, c_{i2}, c_{i3}\} \). We also have an edge \( \{c_{ik}, c_{j\ell}\} \) iff \( L_{ik} \) and \( L_{j\ell} \) are complementary literals. If the size \( n \) of the input is the sum of the lengths of the clauses, then the construction of \( G \) can be done in time \( O(n^2) \).

**Example 9.6.** Let \( F \) be the set of clauses

\[
F = \{C_1 = (x_1 \lor x_2 \lor x_3), C_2 = (\overline{x_1} \lor x_2 \lor x_3), C_3 = (\overline{x_1} \lor \overline{x_2} \lor x_3), C_4 = (x_1 \lor x_2 \lor x_3)\}.
\]

The graph \( G \) associated with \( F \) is shown in Figure 9.16.

![Figure 9.16: The graph constructed from the clauses of Example 9.6.](image)

It remains to show that the construction works. Since any three nodes in a triangle are connected, an independent set \( C \) can have at most one node per triangle and thus has at most \( m \) nodes. Since the budget is \( K = m \), we may assume that there is an independent set with \( m \) nodes. Define a (partial) truth assignment by

\[
v(x_i) = \begin{cases} 
T & \text{if } L_{jk} = x_i \text{ and } c_{jk} \in C \\
F & \text{if } L_{jk} = \overline{x_i} \text{ and } c_{jk} \in C.
\end{cases}
\]
Since the non-triangle edges in $G$ link nodes corresponding to complementary literals and nodes in $C$ are not connected, our truth assignment does not assign clashing truth values to the variables $x_i$. Not all variables may receive a truth value, in which case we assign an arbitrary truth value to the unassigned variables. This yields a satisfying assignment for $F$.

In Example 9.6, the set $C = \{c_{11}, c_{22}, c_{32}, c_{41}\}$ corresponding to the nodes shown in red in Figure 9.16 form an independent set, and they induce the partial truth assignment $v(x_1) = T, v(x_2) = F$. The variable $x_3$ can be assigned an arbitrary value, say $v(x_3) = F$, and $v$ is indeed a satisfying truth assignment for $F$.

Conversely, if $v$ is a truth assignment for $F$, then we obtain an independent set $C$ of size $m$ by picking for each clause $C_i$ a node $c_{ik}$ corresponding to a literal $L_{ik}$ whose value under $v$ is $T$.

(6) **Clique**

To show that **Clique** is $\mathcal{NP}$-complete, we reduce **Independent Set** to it:

**Independent Set $\leq_P$ Clique**

The key to the reduction is the notion of the complement of an undirected graph $G = (V, E)$. The complement $G^c = (V, E^c)$ of the graph $G = (V, E)$ is the graph with the same set of nodes $V$ as $G$ but there is an edge $\{u, v\}$ (with $u \neq v$) in $E^c$ iff $\{u, v\} \notin E$. Then it is not hard to check that there is a bijection between maximum independent sets in $G$ and maximum cliques in $G^c$. The reduction consists in constructing from a graph $G$ its complement $G^c$, and then $G$ has an independent set iff $G^c$ has a clique. Obviously, the reduction can be done in linear time.

This construction is illustrated in Figure 9.17, where a maximum independent set in the graph $G$ is shown in blue and a maximum clique in the graph $G^c$ is shown in red.

![Figure 9.17: A graph (left) and its complement (right).](image-url)
(7) **Node Cover**

To show that **Node Cover** is \( \mathcal{NP} \)-complete, we reduce **Independent Set** to it:

**Independent Set \( \leq_p \) Node Cover**

This time the crucial observation is that if \( N \) is an independent set in \( G \), then the complement \( C = V - N \) of \( N \) in \( V \) is a node cover in \( G \). Thus there is an independent set of size at least \( K \) iff there is a node cover of size at most \( n - K \) where \( n = |V| \) is the number of nodes in \( V \). The reduction leaves the graph unchanged and replaces \( K \) by \( n - K \). Obviously, the reduction can be done in linear time. An example is shown in Figure 9.18 where an independent set is shown in blue and a node cover is shown in red.

![Figure 9.18: An independent set (left) and a node cover (right).](image)

(8) **Knapsack** (also called **Subset sum**)

To show that **Knapsack** is \( \mathcal{NP} \)-complete, we reduce **Exact Cover** to it:

**Exact Cover \( \leq_p \) Knapsack**

Given an instance \((U, \mathcal{F})\) of set cover with \( U = \{u_1, \ldots, u_n\} \) and \( \mathcal{F} = \{S_1, \ldots, S_m\} \), a family of subsets of \( U \), we need to produce in polynomial time an instance \( \tau(U, \mathcal{F}) \) of the Knapsack Problem consisting of \( k \) nonnegative integers \( a_1, \ldots, a_k \) and another integer \( K > 0 \) such that there is a subset \( I \subseteq \{1, \ldots, k\} \) such that \( \sum_{i \in I} a_i = K \) iff there is an exact cover of \( U \) using subsets in \( \mathcal{F} \).

The trick here is the relationship between *set union* and *integer addition*.

**Example 9.7.** Consider the exact cover problem given by \( U = \{u_1, u_2, u_3, u_4\} \) and

\[ \mathcal{F} = \{S_1 = \{u_3, u_4\}, S_2 = \{u_2, u_3, u_4\}, S_3 = \{u_1, u_2\}\}. \]
We can represent each subset $S_j$ by a binary string $a_j$ of length 4, where the $i$th bit from the left is 1 iff $u_i \in S_j$, and 0 otherwise. In our example

$$
a_1 = 0011 \\
a_2 = 0111 \\
a_3 = 1100.
$$

Then the trick is that some family $C$ of subsets $S_j$ is an exact cover if the sum of the corresponding numbers $a_j$ adds up to $1111 = 2^4 - 1 = K$. For example,

$$C = \{ S_1 = \{u_3, u_4\}, \ S_3 = \{u_1, u_2\} \}$$

is an exact cover and

$$a_1 + a_3 = 0011 + 1100 = 1111.$$

Unfortunately, there is a problem with this encoding which has to do with the fact that addition may involve carry. For example, assuming four subsets and the universe $U = \{u_1, \ldots, u_6\}$,

$$11 + 13 + 15 + 24 = 63,$$

in binary

$$001011 + 001101 + 001111 + 011000 = 111111,$$

but if we convert these binary strings to the corresponding subsets we get the subsets

$$S_1 = \{u_3, u_5, u_6\} \\
S_2 = \{u_3, u_4, u_6\} \\
S_3 = \{u_3, u_4, u_5, u_6\} \\
S_4 = \{u_2, u_3\},$$

which are not disjoint and do not cover $U$.

The fix is surprisingly simple: use base $m$ (where $m$ is the number of subsets in $F$) instead of base 2.

**Example 9.8.** Consider the exact cover problem given by $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ and $F$ given by

$$S_1 = \{u_3, u_5, u_6\} \\
S_2 = \{u_3, u_4, u_6\} \\
S_3 = \{u_3, u_4, u_5, u_6\} \\
S_4 = \{u_2, u_3\}, \\
S_5 = \{u_1, u_2, u_4\}.$$
In base $m = 5$, the numbers corresponding to $S_1, \ldots, S_5$ are

\[
a_1 = 001011 \\
a_2 = 001101 \\
a_3 = 001111 \\
a_4 = 011000 \\
a_5 = 110100.
\]

This time,

\[
a_1 + a_2 + a_3 + a_4 = 001011 + 001101 + 001111 + 011000 = 014223 \neq 111111,
\]

so $\{S_1, S_2, S_3, S_4\}$ is not a solution. However

\[
a_1 + a_5 = 001011 + 110100 = 111111,
\]

and $C = \{S_1, S_5\}$ is an exact cover.

Thus, given an instance $(U, \mathcal{F})$ of Exact Cover where $U = \{u_1, \ldots, u_n\}$ and $\mathcal{F} = \{S_1, \ldots, S_m\}$ the reduction to Knapsack consists in forming the $m$ numbers $a_1, \ldots, a_m$ (each of $n$ bits) encoding the subsets $S_j$, namely $a_{ji} = 1$ iff $u_i \in S_j$, else 0, and to let $K = 1 + m^2 + \cdots + m^{n-1}$, which is represented in base $m$ by the string $11\cdots11$. In testing whether $\sum_{i \in I} a_i = K$ for some subset $I \subseteq \{1, \ldots, m\}$, we use arithmetic in base $m$.

If a candidate solution $C$ involves at most $m - 1$ subsets, then since the corresponding numbers are added in base $m$, a carry can never happen. If the candidate solution involves all $m$ subsets, then $a_1 + \cdots + a_m = K$ iff $\mathcal{F}$ is a partition of $U$, since otherwise some bit in the result of adding up these $m$ numbers in base $m$ is not equal to 1, even if a carry occurs. Since the number $K$ is written in binary, it takes time $O(mn)$ to produce $((a_1, \ldots, a_m), K)$ from $(U, \mathcal{F})$.

(9) Inequivalence of $*$-free Regular Expressions

To show that Inequivalence of $*$-free Regular Expressions is $\mathcal{NP}$-complete, we reduce the Satisfiability Problem to it:

Satisfiability Problem $\leq_P$ Inequivalence of $*$-free Regular Expressions

We already argued that Inequivalence of $*$-free Regular Expressions is in $\mathcal{NP}$ because if $R$ is a $*$-free regular expression, then for every string $w \in \mathcal{L}[R]$ we have $|w| \leq |R|$. The above observation shows that if $R_1$ and $R_2$ are $*$-free and if there is a string $w \in (\mathcal{L}[R_1] - \mathcal{L}[R_2]) \cup (\mathcal{L}[R_2] - \mathcal{L}[R_1])$, then $|w| \leq |R_1| + |R_2|$, so we can indeed
check this in polynomial time. It follows that the inequivalence problem for $\ast$-free regular expressions is in $\mathcal{NP}$.

We reduce the \textbf{Satisfiability Problem} to the \textbf{Inequivalence of $\ast$-free Regular Expressions} as follows. For any set of clauses $P = C_1 \land \cdots \land C_p$, if the propositional variables occurring in $P$ are $x_1, \ldots, x_n$, we produce two $\ast$-free regular expressions $R$, $S$ over $\Sigma = \{0, 1\}$, such that $P$ is satisfiable iff $L_R \neq L_S$. The expression $S$ is actually

$$S = (0 + 1)(0 + 1)\cdots(0 + 1).$$

The expression $R$ is of the form

$$R = R_1 + \cdots + R_p,$$

where $R_i$ is constructed from the clause $C_i$ in such a way that $L_{R_i}$ corresponds precisely to the set of truth assignments that falsify $C_i$; see below.

Given any clause $C_i$, let $R_i$ be the $\ast$-free regular expression defined such that, if $x_j$ and $\overline{x_j}$ both belong to $C_i$ (for some $j$), then $R_i = \emptyset$, else

$$R_i = R_i^1 \cdot R_i^2 \cdots R_i^n,$$

where $R_i^j$ is defined by

$$R_i^j = \begin{cases} 
0 & \text{if } x_j \text{ is a literal of } C_i \\
1 & \text{if } \overline{x_j} \text{ is a literal of } C_i \\
(0 + 1) & \text{if } x_j \text{ does not occur in } C_i.
\end{cases}$$

The construction of $R$ from $P$ takes linear time.

\textbf{Example 9.9.} If we apply the above conversion to the clauses of Example 9.3, namely

$$F = \{ C_1 = (x_1 \lor \overline{x_2}), C_2 = (\overline{x_1} \lor x_2 \lor x_3), C_3 = (x_2), C_4 = (\overline{x_2} \lor x_3) \},$$

we get

$$R_1 = 0 \cdot 1 \cdot (0 + 1), \quad R_2 = 1 \cdot 0 \cdot 0, \quad R_3 = (0 + 1) \cdot 0 \cdot (0 + 1), \quad R_4 = (0 + 1) \cdot 1 \cdot 1.$$

Clearly, all truth assignments that falsify $C_i$ must assign $\text{F}$ to $x_j$ if $x_j \in C_i$ or assign $\text{T}$ to $x_j$ if $\overline{x_j} \in C_i$. Therefore, $L_R$ corresponds to the set of truth assignments that falsify $C_i$ (where 1 stands for $\text{T}$ and 0 stands for $\text{F}$) and thus, if we let

$$R = R_1 + \cdots + R_p,$$

then $L_R$ corresponds to the set of truth assignments that falsify $P = C_1 \land \cdots \land C_p$. Since $L_S = \{0, 1\}^n$ (all binary strings of length $n$), we conclude that $L_R \neq L_S$ iff $P$ is satisfiable. Therefore, we have reduced the \textbf{Satisfiability Problem} to our problem and the reduction clearly runs in polynomial time. This proves that the problem of deciding whether $L_R \neq L_S$, for any two $\ast$-free regular expressions $R$ and $S$ is $\mathcal{NP}$-complete.
9.3 Succinct Certificates, coNP, and EXP

(10) 0-1 integer programming problem

It is easy to check that the problem is in \( \mathcal{NP} \).

To prove that the is \( \mathcal{NP} \)-complete we reduce the bounded-tiling problem to it:

**bounded-tiling problem \( \leq_p \) 0-1 integer programming problem**

Given a tiling problem, \(((T, V, H), \hat{s}, \sigma_0)\), we create a 0-1-valued variable \( x_{mnt} \), such that \( x_{mnt} = 1 \) iff tile \( t \) occurs in position \((m, n)\) in some tiling. Write equations or inequalities expressing that a tiling exists and then use “slack variables” to convert inequalities to equations. For example, to express the fact that every position is tiled by a single tile, use the equation

\[
\sum_{t \in T} x_{mnt} = 1, 
\]

for all \( m, n \) with \( 1 \leq m \leq 2s \) and \( 1 \leq n \leq s \). We leave the rest as an exercise.

9.3 Succinct Certificates, coNP, and EXP

All the problems considered in Section 9.1 share a common feature, which is that for each problem, a solution is produced nondeterministically (an exact cover, a directed Hamiltonian cycle, a tour of cities, an independent set, a node cover, a clique etc.), and then this candidate solution is checked deterministically and in polynomial time. The candidate solution is a string called a certificate (or witness).

It turns out that membership on \( \mathcal{NP} \) can be defined in terms of certificates. To be a certificate, a string must satisfy two conditions:

1. It must be *polynomially succinct*, which means that its length is at most a polynomial in the length of the input.
2. It must be *checkable* in polynomial time.

All “yes” inputs to a problem in \( \mathcal{NP} \) must have at least one certificate, while all “no” inputs must have none.

The notion of certificate can be formalized using the notion of a polynomially balanced language.

**Definition 9.3.** Let \( \Sigma \) be an alphabet, and let “;” be a symbol not in \( \Sigma \). A language \( L' \subseteq \Sigma^*; \Sigma^* \) is said to be *polynomially balanced* if there exists a polynomial \( p(X) \) such that for all \( x, y \in \Sigma^* \), if \( x; y \in L' \) then \( |y| \leq p(|x|) \).
Suppose $L'$ is a polynomially balanced language and that $L' \in \mathcal{P}$. Then we can consider the language

$$L = \{ x \in \Sigma^* \mid (\exists y \in \Sigma^*)(x; y \in L') \}.$$

The intuition is that for each $x \in L$, the set

$$\{ y \in \Sigma^* \mid x; y \in L' \}$$

is the set of certificates of $x$. For every $x \in L$, a Turing machine can nondeterministically guess one of its certificates $y$, and then use the deterministic Turing machine for $L'$ to check in polynomial time that $x; y \in L'$. Note that, by definition, strings not in $L$ have no certificate. It follows that $L \in \mathcal{NP}$.

Conversely, if $L \in \mathcal{NP}$ and the alphabet $\Sigma$ has at least two symbols, we can encode the paths in the computation tree for every input $x \in L$, and we obtain a polynomially balanced language $L' \subseteq \Sigma^*; \Sigma^*$ with $L'$ in $\mathcal{P}$ such that

$$L = \{ x \in \Sigma^* \mid (\exists y \in \Sigma^*)(x; y \in L') \}.$$

The details of this construction are left as an exercise. In summary, we obtain the following theorem.

**Theorem 9.1.** Let $L \subseteq \Sigma^*$ be a language over an alphabet $\Sigma$ with at least two symbols, and let ";" be a symbol not in $\Sigma$. Then $L \in \mathcal{NP}$ iff there is a polynomially balanced language $L' \subseteq \Sigma^*; \Sigma^*$ with $L'$ in $\mathcal{P}$ such that

$$L = \{ x \in \Sigma^* \mid (\exists y \in \Sigma^*)(x; y \in L') \}.$$

A striking illustration of the notion of succinct certificate is illustrated by the set of composite integers, namely those natural numbers $n \in \mathbb{N}$ that can be written as the product $pq$ of two numbers $p, q \geq 2$ with $p, q \in \mathbb{N}$. For example, the number

$$4, 294, 967, 297$$

is a composite!

This is far from obvious, but if an oracle gives us the certificate $\{6, 700, 417, 641\}$, it is easy to carry out in polynomial time the multiplication of these two numbers and check that it is equal to $4, 294, 967, 297$. Finding a certificate is usually (very) hard, but checking that it works is easy. This is the point of certificates.

We conclude this section with a brief discussion of the complexity classes $\text{coNP}$ and $\mathcal{EXP}$.

By definition,

$$\text{coNP} = \{ \overline{L} \mid L \in \mathcal{NP} \},$$

where $\overline{L}$ is the complement of $L$. In other words, $\text{coNP}$ contains all languages whose complements are in $\mathcal{NP}$. The class $\text{coNP}$ is of great interest because it contains many natural problems that are believed to be hard, but for which no efficient algorithms are known. For example, the problem of determining whether a given graph is not a clique is in $\text{coNP}$.

Additionally, the class $\text{EXP}$ is defined as the class of languages decidable in exponential time. It is known that $\mathcal{NP} \subseteq \text{coNP} \subseteq \text{EXP}$, and it is an open question whether these inclusions are proper. If $\mathcal{NP} = \text{coNP}$, then $\text{coNP} = \text{EXP}$, which would imply the existence of problems that are hard to solve but easy to verify. However, it is widely believed that $\mathcal{NP} \neq \text{coNP}$, and thus $\text{coNP} \neq \text{EXP}$.

The study of $\text{coNP}$ and $\text{EXP}$ is important because it helps us understand the limits of efficient computation and the nature of complexity classes.
that is, \( \text{coNP} \) consists of all complements of languages in \( \text{NP} \). Since \( \mathcal{P} \subseteq \text{NP} \) and \( \mathcal{P} \) is closed under complementation,

\[
\mathcal{P} \subseteq \text{coNP},
\]

but nobody knows whether \( \text{NP} \) is closed under complementation, that is, nobody knows whether \( \text{NP} = \text{coNP} \).

A language \( L \) is \( \text{coNP} \)-hard if every language in \( \text{coNP} \) is polynomial-time reducible to \( L \), and \( \text{coNP} \)-complete if \( L \in \text{coNP} \) and \( L \) is \( \text{coNP} \)-hard.

What can be shown is that if \( \text{NP} \neq \text{coNP} \), then \( \mathcal{P} \neq \text{NP} \). However it is possible that \( \mathcal{P} \neq \text{NP} \) and yet \( \text{NP} = \text{coNP} \), although this is considered unlikely.

Of course, \( \mathcal{P} \subseteq \text{NP} \cap \text{coNP} \). There are problems in \( \text{NP} \cap \text{coNP} \) not known to be in \( \mathcal{P} \). One of the most famous in the following problem:

**Integer factorization problem:**

Given an integer \( N \geq 3 \), and another integer \( M \) (a budget) such that \( 1 < M < N \), does \( N \) have a factor \( d \) with \( 1 < d \leq M \)?

**Proposition 9.2.** *The problem Integer factorization is in \( \text{NP} \cap \text{coNP} \).*

**Proof.** That Integer factorization is in \( \text{NP} \) is clear. To show that Integer factorization is in \( \text{coNP} \), we can guess a factorization of \( N \) into distinct factors all greater than \( M \), check that they are prime using the results of Chapter 10 showing that testing primality is in \( \text{NP} \) (even in \( \mathcal{P} \), but that’s much harder to prove), and then check that the product of these factors is \( N \). \( \square \)

It is widely believed that Integer factorization does not belong to \( \mathcal{P} \), which is the technical justification for saying that this problem is hard. Most cryptographic algorithms rely on this unproven fact. If Integer factorization was either \( \text{NP} \)-complete or \( \text{coNP} \)-complete, then we would have \( \text{NP} = \text{coNP} \), which is considered very unlikely.

**Remark:** If \( \sqrt{N} \leq M < N \), the above problem is equivalent to asking whether \( N \) is prime.

A natural instance of a problem in \( \text{coNP} \) is the unsatisfiability problem for propositions \( \text{UNSAT} = \neg \text{SAT} \), namely deciding that a proposition \( P \) has no satisfying assignment.

**Definition 9.4.** A proposition \( P \) (in CNF) is **falsifiable** if there is some truth assignment \( v \) such that \( \hat{v}(P) = F \).

It is obvious that the set of falsifiable propositions is in \( \text{NP} \). Since a proposition \( P \) is valid iff \( P \) is not falsifiable, the validity (or tautology) problem TAUT for propositions is in \( \text{coNP} \). In fact, the following result holds.

**Proposition 9.3.** *The problem TAUT is \( \text{coNP} \)-complete.*
Proof. See Papadimitriou [27]. Since SAT is $\mathcal{NP}$-complete, for every language $L \in \mathcal{NP}$, there is a polynomial-time computable function $f: \Sigma^* \to \Sigma^*$ such that $x \in L$ iff $f(x) \in \text{SAT}$. Then $x \notin L$ iff $f(x) \notin \text{SAT}$, that is, $x \notin \overline{L}$ iff $f(x) \in \neg \text{SAT}$, which means that every language $\overline{L} \in \text{co}\mathcal{NP}$ is polynomial-time reducible to $\neg \text{SAT} = \text{UNSAT}$. But TAUT = \{\neg P \mid P \in \text{UNSAT}\}, so we have the polynomial-time computable function $g$ given by $g(x) = \neg f(x)$ which gives us the reduction $x \in \overline{L}$ iff $g(x) \in \text{TAUT}$, which shows that TAUT is $\text{co}\mathcal{NP}$-complete.

Despite the fact that this problem has been extensively studied, not much is known about its exact complexity.

The reasoning used to show that TAUT is $\text{co}\mathcal{NP}$-complete can also be used to show the following interesting result.

**Proposition 9.4.** If a language $L$ is $\mathcal{NP}$-complete, then its complement $\overline{L}$ is $\text{co}\mathcal{NP}$-complete.

**Proof.** By definition, since $L \in \mathcal{NP}$, we have $\overline{L} \in \text{co}\mathcal{NP}$. Since $L$ is $\mathcal{NP}$-complete, for every language $L_2 \in \mathcal{NP}$, there is a polynomial-time computable function $f: \Sigma^* \to \Sigma^*$ such that $x \in L_2$ iff $f(x) \in L$. Then $x \notin L_2$ iff $f(x) \notin L$, that is, $x \in \overline{L_2}$ iff $f(x) \in \overline{L}$, which means that $\overline{L}$ is $\text{co}\mathcal{NP}$-hard as well, thus $\text{co}\mathcal{NP}$-complete.

The class $\mathcal{EXP}$ is defined as follows.

**Definition 9.5.** A deterministic Turing machine $M$ is said to be exponentially bounded if there is a polynomial $p(X)$ such that for every input $x \in \Sigma^*$, there is no ID ID$_n$ such that ID$_0 \vdash \text{ID}_1 \vdash \cdots \text{ID}_{n-1} \vdash \text{ID}_n$, with $n > 2^{p(|x|)}$.

The class $\mathcal{EXP}$ is the class of all languages that are accepted by some exponentially bounded deterministic Turing machine.

**Remark:** We can also define the class $\mathcal{NEXP}$ as in Definition 9.5, except that we allow nondeterministic Turing machines.

One of the interesting features of $\mathcal{EXP}$ is that it contains $\mathcal{NP}$.

**Theorem 9.5.** We have the inclusion $\mathcal{NP} \subseteq \mathcal{EXP}$.

**Sketch of proof.** Let $M$ be some nondeterministic Turing machine accepting $L$ in polynomial time bounded by $p(X)$. We can construct a deterministic Turing machine $M'$ that operates as follows: for every input $x$, $M'$ simulates $M$ on all computations of length 1, then on all possible computations of length 2, and so on, up to all possible computations of length $p(|x|) + 1$. At this point, either an accepting computation has been discovered or all computations have halted rejecting. We claim that $M'$ operates in time bounded by $2^{q(|x|)}$ for some polynomial $q(X)$. First, let $r$ be the degree of nondeterminism of $M$, that is, the maximum number of triples $(b,m,q)$ such that a quintuple $(p,q,b,m,q)$ is an instructions of $M$. Then
to simulate a computation of $M$ of length $\ell$, $M'$ needs $O(\ell)$ steps—to copy the input, to produce a string $c$ in \{1, \ldots, r\}^\ell$, and so simulate $M$ according to the choices specified by $c$. It follows that $M'$ can carry out the simulation of $M$ on an input $x$ in

$$p(|x|)+1 \sum_{\ell=1}^{p(|x|)+1} r^\ell \leq (r + 1)^{p(|x|)+1}$$

steps. Including the $O(\ell)$ extra steps for each $\ell$, we obtain the bound $(r + 2)^{p(|x|)+1}$. Then we can pick a constant $k$ such that $2^k > r + 2$, and with $q(X) = k(p(X) + 1)$, we see that $M'$ operates in time bounded by $2^{q(|x|)}$.

It is also immediate to see that $\mathcal{E}\mathcal{X}\mathcal{P}$ is closed under complementation. Furthermore the strict inclusion $\mathcal{P} \subset \mathcal{E}\mathcal{X}\mathcal{P}$ holds.

**Theorem 9.6.** We have the strict inclusion $\mathcal{P} \subset \mathcal{E}\mathcal{X}\mathcal{P}$.

*Sketch of proof.* We use a diagonalization argument to produce a language $E$ such that $E \notin \mathcal{P}$, yet $E \in \mathcal{E}\mathcal{X}\mathcal{P}$. We need to code a Turing machine as a string, but this can certainly be done using the techniques of Chapter 2. Let $(M)$ be the code of Turing machine $M$ and let $(x)$ be the code of $x$. Define $E$ as

$$E = \{(M)(x) \mid M \text{ accepts input } x \text{ after at most } 2^{|x|} \text{ steps}\}.$$ 

We claim that $E \notin \mathcal{P}$. We proceed by contradiction. If $E \in \mathcal{P}$, then so is the language $E_1$ given by

$$E_1 = \{(M) \mid M \text{ accepts } (M) \text{ after at most } 2^{|(M)|} \text{ steps}\}.$$ 

Since $\mathcal{P}$ is closed under complementation, we also have $\overline{E_1} \in \mathcal{P}$. Let $M^*$ be a deterministic Turing machine accepting $\overline{E_1}$ in time $p(X)$, for some polynomial $p(X)$. Since $p(X)$ is a polynomial, there is some $n_0$ such that $p(n) \leq 2^n$ for all $n \geq n_0$. We may also assume that $|\#(M^*)| \geq n_0$, since if not we can add $n_0$ “dead states” to $M^*$.

Now what happens if we run $M^*$ on its own code $(M^*)$?

It is easy to see that we get a contradiction, namely $M^*$ accepts $(M^*)$ iff $M^*$ rejects $(M^*)$. We leave this verification as an exercise.

In conclusion, $E_1 \notin \mathcal{P}$, which in turn implies that $E \notin \mathcal{P}$.

It remains to prove that $E \in \mathcal{E}\mathcal{X}\mathcal{P}$. This is because we can construct a Turing machine that can in exponential time simulate any Turing machine $M$ on input $x$ for $2^{|x|}$ steps.}

In summary, we have the chain of inclusions

$$\mathcal{P} \subseteq \mathcal{N}\mathcal{P} \subseteq \mathcal{E}\mathcal{X}\mathcal{P},$$
where the inclusions $\mathcal{P} \subset \mathcal{X} \mathcal{P}$ is strict but the left inclusion and the right inclusion are both open problems, and we know that at least one of these two inclusions is strict.

We also have the inclusions

$$\mathcal{P} \subseteq \mathcal{N} \mathcal{P} \subseteq \mathcal{X} \mathcal{P} \subseteq \mathcal{N} \mathcal{X} \mathcal{P},$$

where the inclusions $\mathcal{P} \subset \mathcal{X} \mathcal{P}$ and $\mathcal{N} \mathcal{P} \subset \mathcal{N} \mathcal{X} \mathcal{P}$ are strict. The left inclusion and the right inclusion in $\mathcal{N} \mathcal{P} \subseteq \mathcal{X} \mathcal{P} \subseteq \mathcal{N} \mathcal{X} \mathcal{P}$ are both open problems, but we know that at least one of these two inclusions is strict. It can be shown that if $\mathcal{X} \mathcal{P} \neq \mathcal{N} \mathcal{X} \mathcal{P}$, then $\mathcal{P} \neq \mathcal{N} \mathcal{P}$; see Papadimitriou [27].
Chapter 10

Primality Testing is in $\mathcal{NP}$

10.1 Prime Numbers and Composite Numbers

Prime numbers have fascinated mathematicians and more generally curious minds for thousands of years. What is a prime number? Well, 2, 3, 5, 7, 11, 13, $\ldots$, 9973 are prime numbers.

Definition 10.1. A positive integer $p$ is prime if $p \geq 2$ and if $p$ is only divisible by 1 and $p$. Equivalently, $p$ is prime if and only if $p$ is a positive integer $p \geq 2$ that is not divisible by any integer $m$ such that $2 \leq m < p$. A positive integer $n \geq 2$ which is not prime is called composite.

Observe that the number 1 is considered neither a prime nor a composite. For example, $6 = 2 \cdot 3$ is composite. Is 3215031751 composite? Yes, because $3215031751 = 151 \cdot 751 \cdot 28351$.

Even though the definition of primality is very simple, the structure of the set of prime numbers is highly nontrivial. The prime numbers are the basic building blocks of the natural numbers because of the following theorem bearing the impressive name of fundamental theorem of arithmetic.

Theorem 10.1. Every natural number $n \geq 2$ has a unique factorization

$$n = p_1^{i_1} p_2^{i_2} \cdots p_k^{i_k},$$

where the exponents $i_1, \ldots, i_k$ are positive integers and $p_1 < p_2 < \cdots < p_k$ are primes.

Every book on number theory has a proof of Theorem 10.1. The proof is not difficult and uses induction. It has two parts. The first part shows the existence of a factorization. The second part shows its uniqueness. For example, see Apostol [1] (Chapter 1, Theorem 1.10).

How many prime numbers are there? Many! In fact, infinitely many.
Theorem 10.2. The set of prime numbers is infinite.

Proof. The following proof attributed to Hermite only use the fact that every integer greater than 1 has some prime divisor. We prove that for every natural number \( n \geq 2 \), there is some prime \( p > n \). Consider \( N = n! + 1 \). The number \( N \) must be divisible by some prime \( p \) (\( p = N \) is possible). Any prime \( p \) dividing \( N \) is distinct from 2, 3, \ldots, \( n \), since otherwise \( p \) would divide \( N - n! = 1 \), a contradiction. \( \square \)

The problem of determining whether a given integer is prime is one of the better known and most easily understood problems of pure mathematics. This problem has caught the interest of mathematicians again and again for centuries. However, it was not until the 20th century that questions about primality testing and factoring were recognized as problems of practical importance and a central part of applied mathematics. The advent of cryptographic systems that use large primes, such as RSA, was the main driving force for the development of fast and reliable methods for primality testing. Indeed, in order to create RSA keys, one needs to produce large prime numbers.

10.2 Methods for Primality Testing

The general strategy to test whether an integer \( n > 2 \) is prime or composite is to choose some property, say \( A \), implied by primality, and to search for a counterexample \( a \) to this property for the number \( n \), namely some \( a \) for which property \( A \) fails. We look for properties for which checking that a candidate \( a \) is indeed a countexample can be done quickly.

A simple property that is the basis of several primality testing algorithms is the Fermat test, namely

\[
a^{n-1} \equiv 1 \pmod{n},
\]

which means that \( a^{n-1} - 1 \) is divisible by \( n \) (see Definition 10.2 for the meaning of the notation \( a \equiv b \pmod{n} \)). If \( n \) is prime, and if \( \gcd(a, n) = 1 \), then the above test is indeed satisfied; this is Fermat’s little theorem, Theorem 10.7.

Typically, together with the number \( n \) being tested for primality, some candidate countexample \( a \) is supplied to an algorithm which runs a test to determine whether \( a \) is really a countexample to property \( A \) for \( n \). If the test says that \( a \) is a countexample, also called a witness, then we know for sure that \( n \) is composite.

For example, using the Fermat test, if \( n = 10 \) and \( a = 3 \), we check that

\[
3^9 = 19683 = 10 \cdot 1968 + 3,
\]

so \( 3^9 - 1 \) is not divisible by 10, which means that

\[
a^{n-1} = 3^9 \not\equiv 1 \pmod{10},
\]
and the Fermat test fails. This shows that 10 is not prime and that \( a = 3 \) is a witness of this fact.

If the algorithm reports that \( a \) is not a witness to the fact that \( n \) is composite, does this imply that \( n \) is prime? Unfortunately, no. This is because, there may be some composite number \( n \) and some candidate counterexample \( a \) for which the test says that \( a \) is not a countexample. Such a number \( a \) is called a liar.

For example, using the Fermat test for \( n = 91 = 7 \cdot 13 \) and \( a = 3 \), we can check that

\[
a^{n-1} = 3^{90} \equiv 1 \pmod{91},
\]

so the Fermat test succeeds even though 91 is not prime. The number \( a = 3 \) is a liar.

The other reason is that we haven’t tested all the candidate counterexamples \( a \) for \( n \). In the case where \( n = 91 \), it can be shown that \( 2^{90} - 64 \) is divisible by 91, so the Fermat test fails for \( a = 2 \), which confirms that 91 is not prime, and \( a = 2 \) is a witness of this fact.

Unfortunately, the Fermat test has the property that it may succeed for all candidate counterexamples, even though \( n \) is composite. The number \( n = 561 = 3 \cdot 11 \cdot 17 \) is such a devious number. It can be shown that for all \( a \in \{2, \ldots, 560\} \) such that \( \gcd(a, 561) = 1 \), we have

\[
a^{560} \equiv 1 \pmod{561},
\]

so all these \( a \) are liars.

Such composite numbers for which the Fermat test succeeds for all candidate counterexamples are called Carmichael numbers, and unfortunately there are infinitely many of them. Thus the Fermat test is doomed. There are various ways of strengthening the Fermat test, but we will not discuss this here. We refer the interested reader to Crandall and Pomerance [5] and Gallier and Quaintance [12].

The remedy is to make sure that we pick a property \( A \) such that if \( n \) is composite, then at least some candidate \( a \) is not a liar, and to test all potential countexamples \( a \). The difficulty is that trying all candidate countexamples can be too expensive to be practical.

There are two classes of primality testing algorithms:

1. Algorithms that try all possible countexamples and for which the test does not lie. These algorithms give a definite answer: \( n \) is prime or \( n \) is composite. Until 2002, no algorithms running in polynomial time were known. The situation changed in 2002 when a paper with the title “PRIMES is in \( \mathbf{P} \),” by Agrawal, Kayal and Saxena, appeared on the website of the Indian Institute of Technology at Kanpur, India. In this paper, it was shown that testing for primality has a deterministic (nonrandomized) algorithm that runs in polynomial time.

We will not discuss algorithms of this type here, and instead refer the reader to Crandall and Pomerance [5] and Ribenboim [31].
Randomized algorithms. To avoid having problems with infinite events, we assume that we are testing numbers in some large finite interval $I$. Given any positive integer $m \in I$, some candidate witness $a$ is chosen at random. We have a test which, given $m$ and a potential witness $a$, determines whether or not $a$ is indeed a witness to the fact that $m$ is composite. Such an algorithm is a Monte Carlo algorithm, which means the following:

1. **If the test is positive, then $m \in I$ is composite.** In terms of probabilities, this is expressed by saying that the conditional probability that $m \in I$ is composite given that the test is positive is equal to 1. If we denote the event that some positive integer $m \in I$ is composite by $C$, then we can express the above as

$$\Pr(C \mid \text{test is positive}) = 1.$$ 

2. **If $m \in I$ is composite, then the test is positive for at least 50\% of the choices for $a$.** We can express the above as

$$\Pr(\text{test is positive} \mid C) \geq \frac{1}{2}.$$ 

This gives us a degree of confidence in the test.

The contrapositive of (1) says that if $m \in I$ is prime, then the test is negative. If we denote by $P$ the event that some positive integer $m \in I$ is prime, then this is expressed as

$$\Pr(\text{test is negative} \mid P) = 1.$$ 

If we repeat the test $\ell$ times by picking independent potential witnesses, then the conditional probability that the test is negative $\ell$ times given that $n$ is composite, written $\Pr(\text{test is negative } \ell \text{ times} \mid C)$, is given by

$$\Pr(\text{test is negative } \ell \text{ times} \mid C) = \Pr(\text{test is negative} \mid C)^\ell = (1 - \Pr(\text{test is positive} \mid C))^\ell \leq \left(1 - \frac{1}{2}\right)^\ell = \left(\frac{1}{2}\right)^\ell,$$

where we used Property (2) of a Monte Carlo algorithm that

$$\Pr(\text{test is positive} \mid C) \geq \frac{1}{2}$$

and the independence of the trials. This confirms that if we run the algorithm $\ell$ times, then $\Pr(\text{test is negative } \ell \text{ times} \mid C)$ is very small. In other words, it is very unlikely that the test will lie $\ell$ times (is negative) given that the number $m \in I$ is composite.
10.3. MODULAR ARITHMETIC, THE GROUPS $\mathbb{Z}/n\mathbb{Z}$, $(\mathbb{Z}/n\mathbb{Z})^*$

If the probability $\Pr(P)$ of the event $P$ is known, which requires knowledge of the distribution of the primes in the interval $\mathcal{I}$, then the conditional probability

$$\Pr(P \mid \text{test is negative } \ell \text{ times})$$

can be determined using Bayes’s rule.

A Monte Carlo algorithm does not give a definite answer. However, if $\ell$ is large enough (say $\ell = 100$), then the conditional probability that the number $n$ being tested is prime given that the test is negative $\ell$ times, is very close to 1.

Two of the best known randomized algorithms for primality testing are the *Miller–Rabin test* and the *Solovay–Strassen test*. We will not discuss these methods here, and we refer the reader to Gallier and Quaintance [12].

However, what we will discuss is a nondeterministic algorithm that checks that a number $n$ is prime by guessing a certain kind of tree that we call a Lucas tree (because this algorithm is based on a method due to E. Lucas), and then verifies in polynomial time (in the length $\log_2 n$ of the input given in binary) that this tree constitutes a “proof” that $n$ is indeed prime. This shows that primality testing is in $\mathcal{NP}$, a fact that is not obvious at all. Of course, this is a much weaker result than the AKS algorithm, but the proof that the AKS works in polynomial time (in $\log_2 n$) is much harder.

The Lucas test, and basically all of the primality-testing algorithms, use modular arithmetic and some elementary facts of number theory such as the Euler-Fermat theorem, so we proceed with a review of these concepts.

### 10.3 Modular Arithmetic, the Groups $\mathbb{Z}/n\mathbb{Z}$, $(\mathbb{Z}/n\mathbb{Z})^*$

Recall the fundamental notion of congruence modulo $n$ and its notation due to Gauss (circa 1802).

**Definition 10.2.** For any $a, b \in \mathbb{Z}$, we write $a \equiv b \pmod{m}$ iff $a - b = km$, for some $k \in \mathbb{Z}$ (in other words, $a - b$ is divisible by $m$), and we say that $a$ and $b$ are congruent modulo $m$.

For example, $37 \equiv 1 \pmod{9}$, since $37 - 1 = 36 = 4 \cdot 9$. It can also be shown that $200^{250} \equiv 1 \pmod{251}$, but this is impossible to do by brute force, so we will develop some tools to either avoid such computations, or to make them tractable.

It is easy to check that congruence is an equivalence relation but it also satisfies the following properties.

**Proposition 10.3.** For any positive integer $m$, for all $a_1, a_2, b_1, b_2 \in \mathbb{Z}$, the following properties hold. If $a_1 \equiv b_1 \pmod{m}$ and $a_2 \equiv b_2 \pmod{m}$, then

1. $a_1 + a_2 \equiv b_1 + b_2 \pmod{m}$. 

(2) \(a_1 - a_2 \equiv b_1 - b_2 \pmod{m}\).

(3) \(a_1 a_2 \equiv b_1 b_2 \pmod{m}\).

Proof. We only check (3), leaving (1) and (2) as easy exercises. Because \(a_1 \equiv b_1 \pmod{m}\) and \(a_2 \equiv b_2 \pmod{m}\), we have \(a_1 = b_1 + k_1 m\) and \(a_2 = b_2 + k_2 m\), for some \(k_1, k_2 \in \mathbb{Z}\), so we obtain

\[
a_1 a_2 - b_1 b_2 = a_1(a_2 - b_2) + (a_1 - b_1)b_2 = (a_1 k_2 + k_1 b_2)m.
\]

Proposition 10.3 allows us to define addition, subtraction, and multiplication on equivalence classes modulo \(m\).

**Definition 10.3.** Given any positive integer \(m\), we denote by \(\mathbb{Z}/m\mathbb{Z}\) the set of equivalence classes modulo \(m\). If we write \(\overline{a}\) for the equivalence class of \(a \in \mathbb{Z}\), then we define addition, subtraction, and multiplication on residue classes as follows:

\[
\overline{a} + \overline{b} = \overline{a + b} \\
\overline{a} - \overline{b} = \overline{a - b} \\
\overline{a} \cdot \overline{b} = \overline{ab}.
\]

The above operations make sense because \(\overline{a + b}\) does not depend on the representatives chosen in the equivalence classes \(\overline{a}\) and \(\overline{b}\), and similarly for \(\overline{a - b}\) and \(\overline{ab}\). Each equivalence class \(\overline{a}\) contains a unique representative from the set of remainders \(\{0, 1, \ldots, m-1\}\), modulo \(m\), so the above operations are completely determined by \(m \times m\) tables. Using the arithmetic operations of \(\mathbb{Z}/m\mathbb{Z}\) is called modular arithmetic.

The addition tables of \(\mathbb{Z}/n\mathbb{Z}\) for \(n = 2, 3, 4, 5, 6, 7\) are shown below.

\[
\begin{array}{c|cc}
+ & 0 & 1 \\
0 & 0 & 1 \\
1 & 1 & 0
\end{array}
\quad
\begin{array}{c|cccc}
+ & 0 & 1 & 2 \\
0 & 0 & 1 & 2 \\
1 & 1 & 2 & 0 \\
2 & 2 & 0 & 1
\end{array}
\quad
\begin{array}{c|cccc}
+ & 0 & 1 & 2 & 3 \\
0 & 0 & 1 & 2 & 3 \\
1 & 1 & 2 & 3 & 0 \\
2 & 2 & 3 & 0 & 1 \\
3 & 3 & 0 & 1 & 2
\end{array}
\quad
\begin{array}{c|cccc}
+ & 0 & 1 & 2 & 3 & 4 \\
0 & 0 & 1 & 2 & 3 & 4 \\
1 & 1 & 2 & 3 & 4 & 0 \\
2 & 2 & 3 & 4 & 0 & 1 \\
3 & 3 & 4 & 0 & 1 & 2 \\
4 & 4 & 0 & 1 & 2 & 3
\end{array}
\]

\[
\begin{array}{c|cccccc}
+ & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 1 & 2 & 3 & 4 & 5 \\
1 & 1 & 2 & 3 & 4 & 5 & 0 \\
2 & 2 & 3 & 4 & 5 & 0 & 1 \\
3 & 3 & 4 & 5 & 0 & 1 & 2 \\
4 & 4 & 5 & 0 & 1 & 2 & 3 \\
5 & 5 & 0 & 1 & 2 & 3 & 4
\end{array}
\quad
\begin{array}{c|cccccc}
+ & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
0 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
1 & 1 & 2 & 3 & 4 & 5 & 6 & 0 \\
2 & 2 & 3 & 4 & 5 & 6 & 0 & 1 \\
3 & 3 & 4 & 5 & 6 & 0 & 1 & 2 \\
4 & 4 & 5 & 6 & 0 & 1 & 2 & 3 \\
5 & 5 & 6 & 0 & 1 & 2 & 3 & 4 \\
6 & 6 & 0 & 1 & 2 & 3 & 4 & 5
\end{array}
\]
10.3. MODULAR ARITHMETIC, THE GROUPS $\mathbb{Z}/n\mathbb{Z}$, $(\mathbb{Z}/n\mathbb{Z})^*$

It is easy to check that the addition operation $+$ is commutative (abelian), associative, that 0 is an identity element for $+$, and that every element $a$ has $-a$ as additive inverse, which means that

$$a + (-a) = (-a) + a = 0.$$ 

The set $\mathbb{Z}/n\mathbb{Z}$ of residue classes modulo $n$ is a group under addition, a notion defined formally in Definition 10.4.

It is easy to check that the multiplication operation $\cdot$ is commutative (abelian), associative, that 1 is an identity element for $\cdot$, and that $\cdot$ is distributive on the left and on the right with respect to addition. We usually suppress the dot and write $ab$ instead of $a \cdot b$. The multiplication tables of $\mathbb{Z}/n\mathbb{Z}$ for $n = 2, 3, \ldots, 9$ are shown below. Since $0 \cdot m = m \cdot 0 = 0$ for all $m$, these tables are only given for nonzero arguments.

$$
\begin{array}{c|cccc}
 n = 2 & \cdot & 1 & 2 \\
 \hline 
 1 & 1 & 2 \\
 2 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 3 & \cdot & 1 & 2 & 3 \\
 \hline 
 1 & 1 & 2 & 3 \\
 2 & 2 & 0 & 2 \\
 3 & 3 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 4 & \cdot & 1 & 2 & 3 & 4 \\
 \hline 
 1 & 1 & 2 & 3 & 4 \\
 2 & 2 & 4 & 1 & 3 \\
 3 & 3 & 1 & 4 & 2 \\
 4 & 4 & 3 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 5 & \cdot & 1 & 2 & 3 & 4 & 5 \\
 \hline 
 1 & 1 & 2 & 3 & 4 & 5 \\
 2 & 2 & 4 & 0 & 2 & 4 \\
 3 & 3 & 0 & 3 & 0 & 3 \\
 4 & 4 & 2 & 0 & 4 & 2 \\
 5 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 6 & \cdot & 1 & 2 & 3 & 4 & 5 & 6 \\
 \hline 
 1 & 1 & 2 & 3 & 4 & 5 & 6 \\
 2 & 2 & 4 & 6 & 0 & 2 & 4 \\
 3 & 3 & 6 & 2 & 5 & 1 & 4 \\
 4 & 4 & 1 & 5 & 2 & 6 & 3 \\
 5 & 5 & 3 & 1 & 6 & 4 & 2 \\
 6 & 6 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 7 & \cdot & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
 \hline 
 1 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
 2 & 2 & 4 & 6 & 0 & 2 & 4 & 6 \\
 3 & 3 & 6 & 1 & 4 & 7 & 2 & 5 \\
 4 & 4 & 0 & 4 & 0 & 4 & 0 & 4 \\
 5 & 5 & 2 & 7 & 4 & 1 & 6 & 3 \\
 6 & 6 & 4 & 2 & 0 & 6 & 4 & 2 \\
 7 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
$$

$$
\begin{array}{c|cccc}
 n = 9 & \cdot & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
 \hline 
 1 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
 2 & 2 & 4 & 6 & 8 & 1 & 3 & 5 & 7 \\
 3 & 3 & 6 & 0 & 3 & 6 & 0 & 3 & 6 \\
 4 & 4 & 8 & 3 & 7 & 2 & 6 & 1 & 5 \\
 5 & 5 & 1 & 6 & 2 & 7 & 3 & 8 & 4 \\
 6 & 6 & 3 & 0 & 6 & 3 & 0 & 6 & 3 \\
 7 & 7 & 5 & 3 & 1 & 8 & 6 & 4 & 2 \\
 8 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
$$
Examining the above tables, we observe that for \( n = 2, 3, 5, 7 \), which are primes, every element has an inverse, which means that for every nonzero element \( a \), there is some (actually, unique) element \( b \) such that
\[
a \cdot b = b \cdot a = 1.
\]
For \( n = 2, 3, 5, 7 \), the set \( \mathbb{Z}/n\mathbb{Z} - \{0\} \) is an abelian group under multiplication (see Definition 10.4). When \( n \) is composite, there exist nonzero elements whose product is zero. For example, when \( n = 6 \), we have \( 3 \cdot 2 = 0 \), when \( n = 8 \), we have \( 4 \cdot 4 = 0 \), when \( n = 9 \), we have \( 6 \cdot 6 = 0 \).

For \( n = 4, 6, 8, 9 \), the elements \( a \) that have an inverse are precisely those that are relatively prime to the modulus \( n \) (that is, \( \gcd(a, n) = 1 \)).

These observations hold in general. Recall the Bezout theorem: two nonzero integers \( m, n \in \mathbb{Z} \) are relatively prime (\( \gcd(m, n) = 1 \)) iff there are integers \( a, b \in \mathbb{Z} \) such that
\[
am + bn = 1.
\]

**Proposition 10.4.** Given any integer \( n \geq 1 \), for any \( a \in \mathbb{Z} \), the residue class \( \overline{a} \in \mathbb{Z}/n\mathbb{Z} \) is invertible with respect to multiplication iff \( \gcd(a, n) = 1 \).

**Proof.** If \( \overline{a} \) has inverse \( \overline{b} \) in \( \mathbb{Z}/n\mathbb{Z} \), then \( \overline{a} \overline{b} = 1 \), which means that
\[
ab \equiv 1 \pmod{n},
\]
that is \( ab = 1 + nk \) for some \( k \in \mathbb{Z} \), which is the Bezout identity
\[
ab - nk = 1
\]
and implies that \( \gcd(a, n) = 1 \). Conversely, if \( \gcd(a, n) = 1 \), then by Bezout’s identity there exist \( u, v \in \mathbb{Z} \) such that
\[
au + nv = 1,
\]
so \( au = 1 - nv \), that is,
\[
au \equiv 1 \pmod{n},
\]
which means that \( \overline{a} \overline{u} = 1 \), so \( \overline{a} \) is invertible in \( \mathbb{Z}/n\mathbb{Z} \).

We have alluded to the notion of a group. Here is the formal definition.

**Definition 10.4.** A group is a set \( G \) equipped with a binary operation \( \cdot : G \times G \to G \) that associates an element \( a \cdot b \in G \) to every pair of elements \( a, b \in G \), and having the following properties: \( \cdot \) is associative, has an identity element \( e \in G \), and every element in \( G \) is invertible (w.r.t. \( \cdot \)). More explicitly, this means that the following equations hold for all \( a, b, c \in G \):

- \( a \cdot (b \cdot c) = (a \cdot b) \cdot c \) \hspace{2cm} \text{(associativity)};
- \( a \cdot e = e \cdot a = a \) \hspace{1cm} \text{(identity)};
For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. (inverse).

A group $G$ is abelian (or commutative) if

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$ 

It is easy to show that the element $e$ satisfying property (G2) is unique, and for any $a \in G$, the element $a^{-1} \in G$ satisfying $a \cdot a^{-1} = a^{-1} \cdot a = e$ required to exist by (G3) is actually unique. This element is called the inverse of $a$.

The set of integers $\mathbb{Z}$ with the addition operation is an abelian group with identity element 0. The set $\mathbb{Z}/n\mathbb{Z}$ of residues modulo $m$ is an abelian group under addition with identity element 0. In general, $\mathbb{Z}/n\mathbb{Z} - \{0\}$ is not a group under multiplication, because some nonzero elements may not have an inverse. However, by Proposition 10.4, if $p$ is prime, then $\mathbb{Z}/n\mathbb{Z} - \{0\}$ is an abelian group under multiplication.

When $p$ is not prime, the subset of elements, shown in boldface in the multiplication tables, forms an abelian group under multiplication.

**Definition 10.5.** The group (under multiplication) of invertible elements of the ring $\mathbb{Z}/n\mathbb{Z}$ is denoted by $(\mathbb{Z}/n\mathbb{Z})^*$. Note that this group is abelian and only defined if $n \geq 2$.

**Definition 10.6.** If $G$ is a finite group, the number of elements in $G$ is called the order of $G$.

Given a group $G$ with identity element $e$, and any element $g \in G$, we often need to consider the powers of $g$ defined as follows.

**Definition 10.7.** Given a group $G$ with identity element $e$, for any nonnegative integer $n$, it is natural to define the power $g^n$ of $g$ as follows:

$$g^0 = e,$$
$$g^{n+1} = g \cdot g^n.$$

Using induction, it is easy to show that

$$g^m g^n = g^{n+m}$$

for all $m, n \in \mathbb{N}$.

Since $g$ has an inverse $g^{-1}$, we can extend the definition of $g^n$ to negative powers. For $n \in \mathbb{Z}$, with $n < 0$, let

$$g^n = (g^{-1})^{-n}.$$
Then it is easy to prove that
\[
g^i \cdot g^j = g^{i+j} \\
(g^i)^{-1} = g^{-i} \\
g^i \cdot g^j = g^j \cdot g^i
\]
for all \( i, j \in \mathbb{Z} \).

Given a finite group \( G \) of order \( n \), for any element \( a \in G \), it is natural to consider the set of powers \( \{ e, a^1, a^2, \ldots, a^k, \ldots \} \). A crucial fact is that there is a smallest positive \( s \in \mathbb{N} \) such that \( a^s = e \), and that \( s \) divides \( n \).

**Proposition 10.5.** Let \( G \) be a finite group of order \( n \). For every element \( a \in G \), the following facts hold:

1. There is a smallest positive integer \( s \leq n \) such that \( a^s = e \).
2. The set \( \{ e, a, \ldots, a^{s-1} \} \) is an abelian group denoted \( \langle a \rangle \).
3. We have \( a^n = e \), and the positive integer \( s \) divides \( n \). More generally, for any positive integer \( m \), if \( a^m = e \), then \( s \) divides \( m \).

**Proof.**

1. Consider the sequence of \( n + 1 \) elements
\[
(e, a^1, a^2, \ldots, a^n).
\]
Since \( G \) only has \( n \) distinct elements, by the pigeonhole principle, there exist \( i, j \) such that \( 0 \leq i < j \leq n \) such that
\[
a^i = a^j.
\]
By multiplying both sides by \((a^i)^{-1} = a^{-i}\), we get
\[
e = a^i(a^i)^{-1} = a^j(a^i)^{-1} = a^j a^{-i} = a^{j-i}.
\]
Since \( 0 \leq i < j \leq n \), we have \( 0 \leq j - i \leq n \) with \( a^{j-i} = e \). Thus there is some \( s \) with \( 0 < s \leq n \) such that \( a^s = e \), and thus a smallest such \( s \).

2. Since \( a^s = e \), for any \( i, j \in \{ 0, \ldots, s-1 \} \) if we write \( i + j = sq + r \) with \( 0 \leq r \leq s - 1 \), we have
\[
a^i a^j = a^{i+j} = a^{sq+r} = a^{sq} a^r = (a^s)^q a^r = e^q a^r = a^r,
\]
so \( \langle a \rangle \) is closed under multiplication. We have \( e \in \langle a \rangle \) and the inverse of \( a^i \) is \( a^{s-i} \), so \( \langle a \rangle \) is a group. This group is obviously abelian.

3. For any element \( g \in G \), let \( g \langle a \rangle = \{ ga^k \mid 0 \leq k \leq s - 1 \} \). Observe that for any \( i \in \mathbb{N} \), we have
\[
a^i \langle a \rangle = \langle a \rangle.
\]
We claim that for any two elements \( g_1, g_2 \in G \), if \( g_1 \langle a \rangle \cap g_2 \langle a \rangle \neq \emptyset \), then \( g_1 \langle a \rangle = g_2 \langle a \rangle \).
10.3. MODULAR ARITHMETIC, THE GROUPS $\mathbb{Z}/n\mathbb{Z}$, $\left(\mathbb{Z}/n\mathbb{Z}\right)^*$

Proof of the claim. If $g \in g_1\langle a \rangle \cap g_2\langle a \rangle$, then there exist $i, j \in \{0, \ldots, s - 1\}$ such that

$$g_1a^i = g_2a^j.$$  

Without loss of generality, we may assume that $i \geq j$. By multiplying both sides by $(a^j)^{-1}$, we get

$$g_2 = g_1a^{i-j}.$$  

Consequently

$$g_2\langle a \rangle = g_1a^{i-j}\langle a \rangle = g_1\langle a \rangle,$$

as claimed. \hfill \square

It follows that the pairwise disjoint nonempty subsets of the form $g\langle a \rangle$, for $g \in G$, form a partition of $G$. However, the map $\varphi_g$ from $\langle a \rangle$ to $g\langle a \rangle$ given by $\varphi_g(a^i) = ga^i$ has for inverse the map $\varphi_g^{-1}$, so $\varphi_g$ is a bijection, and thus the subsets $g\langle a \rangle$ all have the same number of elements $s$. Since these subsets form a partition of $G$, we must have $n = sq$ for some $q \in \mathbb{N}$, which implies that $a^n = e$.

If $g^m = 1$, then writing $m = sq + r$, with $0 \leq r < s$, we get

$$1 = g^m = g^{sq+r} = (g^s)^q \cdot g^r = g^r,$$

so $g^r = 1$ with $0 \leq r < s$, contradicting the minimality of $s$, so $r = 0$ and $s$ divides $m$. \hfill \square

**Definition 10.8.** Given a finite group $G$ of order $n$, for any $a \in G$, the smallest positive integer $s \leq n$ such that $a^s = e$ in (1) of Proposition 10.5 is called the **order** of $a$.

The **Euler $\varphi$-function** plays an important role in the theory of the groups $(\mathbb{Z}/n\mathbb{Z})^*$.

**Definition 10.9.** Given any positive integer $n \geq 1$, the **Euler $\varphi$-function** (or Euler totient function) is defined such that $\varphi(n)$ is the number of integers $a$, with $1 \leq a \leq n$, which are relatively prime to $n$; that is, with $\gcd(a, n) = 1$.\(^1\)

If $p$ is prime, then by definition

$$\varphi(p) = p - 1.$$  

We leave it as an exercise to show that if $p$ is prime and if $k \geq 1$, then

$$\varphi(p^k) = p^{k-1}(p - 1).$$  

It can also be shown that if $\gcd(m, n) = 1$, then

$$\varphi(mn) = \varphi(m)\varphi(n).$$

\(^1\)We allow $a = n$ to accommodate the special case $n = 1$.  


The above properties yield a method for computing \( \varphi(n) \), based on its prime factorization. If \( n = p_1^{i_1} \cdots p_k^{i_k} \), then

\[
\varphi(n) = p_1^{i_1-1} \cdots p_k^{i_k-1}(p_1 - 1) \cdots (p_k - 1).
\]

For example, \( \varphi(17) = 16 \), \( \varphi(49) = 7 \cdot 6 = 42 \), \( \varphi(900) = \varphi(2^2 \cdot 3^2 \cdot 5^2) = 2 \cdot 3 \cdot 5 \cdot 1 \cdot 2 \cdot 4 = 240 \).

Proposition 10.4 shows that \( (\mathbb{Z}/n\mathbb{Z})^* \) has \( \varphi(n) \) elements. It also shows that \( \mathbb{Z}/n\mathbb{Z} \setminus \{0\} \) is a group (under multiplication) iff \( n \) is prime.

For any integer \( n \geq 2 \), let \( (\mathbb{Z}/n\mathbb{Z})^* \) be the group of invertible elements of the ring \( \mathbb{Z}/n\mathbb{Z} \). This is a group of order \( \varphi(n) \). Then Proposition 10.5 yields the following result.

**Theorem 10.6.** (Euler) For any integer \( n \geq 2 \) and any \( a \in \{1, \ldots, n-1\} \) such that \( \gcd(a, n) = 1 \), we have

\[
a^{\varphi(n)} \equiv 1 \pmod{n}.
\]

In particular, if \( n \) is a prime, then \( \varphi(n) = n - 1 \), and we get Fermat’s little theorem.

**Theorem 10.7.** (Fermat’s little theorem) For any prime \( p \) and any \( a \in \{1, \ldots, p-1\} \), we have

\[
a^{p-1} \equiv 1 \pmod{p}.
\]

Since 251 is prime, and since \( \gcd(200, 252) = 1 \), Fermat’s little theorem implies our earlier claim that \( 200^{250} \equiv 1 \pmod{251} \), without making any computations.

Proposition 10.5 suggests considering groups of the form \( \langle g \rangle \).

**Definition 10.10.** A finite group \( G \) is cyclic iff there is some element \( g \in G \) such that \( G = \langle g \rangle \). An element \( g \in G \) with this property is called a generator of \( G \).

Even though, in principle, a finite cyclic group has a very simple structure, finding a generator for a finite cyclic group is generally hard. For example, it turns out that the multiplicative group \( (\mathbb{Z}/p\mathbb{Z})^* \) is a cyclic group when \( p \) is prime, but no efficient method for finding a generator for \( (\mathbb{Z}/p\mathbb{Z})^* \) is known (besides a brute-force search).

Examining the multiplication tables for \( (\mathbb{Z}/n\mathbb{Z})^* \) for \( n = 3, 4, \ldots, 9 \), we can check the following facts:

1. 2 is a generator for \( (\mathbb{Z}/3\mathbb{Z})^* \).
2. 3 is a generator for \( (\mathbb{Z}/4\mathbb{Z})^* \).
3. 2 is a generator for \( (\mathbb{Z}/5\mathbb{Z})^* \).
4. 5 is a generator for $(\mathbb{Z}/6\mathbb{Z})^*$. 

5. 3 is a generator for $(\mathbb{Z}/7\mathbb{Z})^*$. 

6. Every element of $(\mathbb{Z}/8\mathbb{Z})^*$ satisfies the equation $a^2 = 1 \pmod{8}$, thus $(\mathbb{Z}/8\mathbb{Z})^*$ has no generators. 

7. 2 is a generator for $(\mathbb{Z}/9\mathbb{Z})^*$. 

More generally, it can be shown that the multiplicative groups $(\mathbb{Z}/p^k\mathbb{Z})^*$ and $(\mathbb{Z}/2p^k\mathbb{Z})^*$ are cyclic groups when $p$ is an odd prime and $k \geq 1$. 

**Definition 10.11.** A generator of the group $(\mathbb{Z}/n\mathbb{Z})^*$ (when there is one), is called a **primitive root modulo** $n$. 

As an exercise, the reader should check that the next value of $n$ for which $(\mathbb{Z}/n\mathbb{Z})^*$ has no generator is $n = 12$. 

The following theorem due to Gauss can be shown. For a proof, see Apostol [1] or Gallier and Quaintance [12].

**Theorem 10.8.** (Gauss) For every odd prime $p$, the group $(\mathbb{Z}/p\mathbb{Z})^*$ is cyclic of order $p - 1$. It has $\varphi(p - 1)$ generators. 

According to Definition 10.11, the generators of $(\mathbb{Z}/p\mathbb{Z})^*$ are the **primitive roots modulo** $p$.

### 10.4 The Lucas Theorem

In this section we discuss an application of the existence of primitive roots in $(\mathbb{Z}/p\mathbb{Z})^*$ where $p$ is an odd prime, known as the **$n - 1$ test**. This test due to E. Lucas determines whether a positive odd integer $n$ is prime or not by examining the prime factors of $n - 1$ and checking some congruences.

The $n - 1$ test can be described as the construction of a certain kind of tree rooted with $n$, and it turns out that the number of nodes in this tree is bounded by $2\log_2 n$, and that the number of modular multiplications involved in checking the congruences is bounded by $2\log_2^2 n$.

When we talk about the complexity of algorithms dealing with numbers, we assume that all inputs (to a Turing machine) are strings representing these numbers, typically in base 2. Since the length of the binary representation of a natural number $n \geq 1$ is $\lfloor \log_2 n \rfloor + 1$ (or $\lceil \log_2(n + 1) \rceil$, which allows $n = 0$), the complexity of algorithms dealing with (nonzero) numbers $m, n, \text{ etc.}$ is expressed in terms of $\log_2 m, \log_2 n, \text{ etc.}$ Recall that for any real
number \( x \in \mathbb{R} \), the \textit{floor} of \( x \) is the greatest integer \( \lfloor x \rfloor \) that is less than or equal to \( x \), and the \textit{ceiling} of \( x \) is the least integer \( \lceil x \rceil \) that is greater than or equal to \( x \).

If we choose to represent numbers in base 10, since for any base \( b \) we have \( \log_b x = \frac{\ln x}{\ln b} \), we have
\[
\log_2 x = \frac{\ln 10}{\ln 2} \log_{10} x.
\]
Since \( (\ln 10) / (\ln 2) \approx 3.322 \approx 10 / 3 \), we see that the number of decimal digits needed to represent the integer \( n \) in base 10 is approximately 30% of the number of bits needed to represent \( n \) in base 2.

Since the Lucas test yields a tree such that the number of modular multiplications involved in checking the congruences is bounded by \( 2 \log_2 2 n \), it is not hard to show that testing whether or not a positive integer \( n \) is prime, a problem denoted \( \text{PRIMES} \), belongs to the complexity class \( \mathcal{NP} \). This result was shown by V. Pratt [29] (1975), but Peter Freyd told me that it was “folklore.” Since 2002, thanks to the AKS algorithm, we know that \( \text{PRIMES} \) actually belongs to the class \( \mathcal{P} \), but this is a much harder result.

Here is Lehmer’s version of the Lucas result, from 1876.

\textbf{Theorem 10.9.} (Lucas theorem) Let \( n \) be a positive integer with \( n \geq 2 \). Then \( n \) is prime iff there is some integer \( a \in \{1, 2, \ldots, n - 1\} \) such that the following two conditions hold:

1. \( a^{n-1} \equiv 1 \pmod{n} \).
2. If \( n > 2 \), then \( a^{(n-1)/q} \not\equiv 1 \pmod{n} \) for all prime divisors \( q \) of \( n - 1 \).

\textit{Proof.} First assume that Conditions (1) and (2) hold. If \( n = 2 \), since 2 is prime, we are done. Thus assume that \( n \geq 3 \), and let \( r \) be the order of \( a \) (we are working in the abelian group \( (\mathbb{Z}/n\mathbb{Z})^* \)). We claim that \( r = n - 1 \). The condition \( a^{n-1} \equiv 1 \pmod{n} \) implies that \( r \) divides \( n - 1 \). Suppose that \( r < n - 1 \), and let \( q \) be a prime divisor of \((n - 1)/r\) (so \( q \) divides \( n - 1 \)). Since \( r \) is the order of \( a \) we have \( a^r \equiv 1 \pmod{n} \), so we get
\[
a^{(n-1)/q} \equiv a^{r(n-1)/(rq)} \equiv (a^r)^{(n-1)/(rq)} \equiv 1^{(n-1)/(rq)} \equiv 1 \pmod{n},
\]
contradicting Condition (2). Therefore, \( r = n - 1 \), as claimed.

We now show that \( n \) must be prime. Now \( a^{n-1} \equiv 1 \pmod{n} \) implies that \( a \) and \( n \) are relatively prime so by Euler’s theorem (Theorem 10.6),
\[
a^{\varphi(n)} \equiv 1 \pmod{n}.
\]
Since the order of \( a \) is \( n - 1 \), we have \( n - 1 \leq \varphi(n) \). If \( n \geq 3 \) is not prime, then \( n \) has some prime divisor \( p \), but \( n \) and \( p \) are integers in \( \{1, 2, \ldots, n\} \) that are not relatively prime to \( n \), so by definition of \( \varphi(n) \), we have \( \varphi(n) \leq n - 2 \), contradicting the fact that \( n - 1 \leq \varphi(n) \). Therefore, \( n \) must be prime.

Conversely, assume that \( n \) is prime. If \( n = 2 \), then we set \( a = 1 \). Otherwise, pick \( a \) to be any primitive root modulo \( p \). \( \square \)
Clearly, if \( n > 2 \) then we may assume that \( a \geq 2 \). The main difficulty with the \( n - 1 \) test is not so much guessing the primitive root \( a \), but finding a complete prime factorization of \( n - 1 \). However, as a nondeterministic algorithm, the \( n - 1 \) test yields a “proof” that a number \( n \) is indeed prime which can be represented as a tree, and the number of operations needed to check the required conditions (the congruences) is bounded by \( c \log_2 n \) for some positive constant \( c \), and this implies that testing primality is in \( NP \).

Before explaining the details of this method, we sharpen slightly Lucas theorem to deal only with odd prime divisors.

**Theorem 10.10.** Let \( n \) be a positive odd integer with \( n \geq 3 \). Then \( n \) is prime iff there is some integer \( a \in \{2, \ldots, n-1\} \) (a guess for a primitive root modulo \( n \)) such that the following two conditions hold:

1. \( a^{(n-1)/2} \equiv -1 \pmod{n} \).
2. If \( n - 1 \) is not a power of 2, then \( a^{(n-1)/2q} \not\equiv -1 \pmod{n} \) for all odd prime divisors \( q \) of \( n - 1 \).

**Proof.** Assume that Conditions (1b) and (2b) of Theorem 10.10 hold. Then we claim that Conditions (1) and (2) of Theorem 10.9 hold. By squaring the congruence \( a^{(n-1)/2} \equiv -1 \pmod{n} \), we get \( a^{n-1} \equiv 1 \pmod{n} \), which is Condition (1) of Theorem 10.9. Since \( a^{(n-1)/2} \equiv -1 \pmod{n} \), Condition (2) of Theorem 10.9 holds for \( q = 2 \). Next, if \( q \) is an odd prime divisor of \( n - 1 \), let \( m = a^{(n-1)/2q} \). Condition (1b) means that

\[
m^q \equiv a^{(n-1)/2} \equiv -1 \pmod{n}.
\]

Now if \( m^2 \equiv a^{(n-1)/q} \equiv 1 \pmod{n} \), since \( q \) is an odd prime, we can write \( q = 2k + 1 \) for some \( k \geq 1 \), and then

\[
m^q \equiv m^{2k+1} \equiv (m^2)^k m \equiv 1^k m \equiv m \pmod{n},
\]

and since \( m^q \equiv -1 \pmod{n} \), we get

\[
m \equiv -1 \pmod{n}
\]

(regardless of whether \( n \) is prime or not). Thus we proved that if \( m^q \equiv -1 \pmod{n} \) and \( m^2 \equiv 1 \pmod{n} \), then \( m \equiv -1 \pmod{n} \). By contrapositive, we see that if \( m \not\equiv -1 \pmod{n} \), then \( m^2 \not\equiv 1 \pmod{n} \) or \( m^q \not\equiv -1 \pmod{n} \), but since \( m^q \equiv a^{(n-1)/2} \equiv -1 \pmod{n} \) by Condition (1a), we conclude that \( m^2 \equiv a^{(n-1)/q} \not\equiv 1 \pmod{n} \), which is Condition (2) of Theorem 10.9. But then Theorem 10.9 implies that \( n \) is prime.

Conversely, assume that \( n \) is an odd prime, and let \( a \) be any primitive root modulo \( n \). Then by little Fermat we know that

\[
a^{n-1} \equiv 1 \pmod{n},
\]
so

\[(a^{(n-1)/2} - 1)(a^{(n-1)/2} + 1) \equiv 0 \pmod{n}.
\]

Since \(n\) is prime, either \(a^{(n-1)/2} \equiv 1 \pmod{n}\) or \(a^{(n-1)/2} \equiv -1 \pmod{n}\), but since \(a\) generates \((\mathbb{Z}/n\mathbb{Z})^*\), it has order \(n-1\), so the congruence \(a^{(n-1)/2} \equiv 1 \pmod{n}\) is impossible, and Condition (1b) must hold. Similarly, if we had \(a^{(n-1)/2q} \equiv -1 \pmod{n}\) for some odd prime divisor \(q\) of \(n-1\), then by squaring we would have

\[a^{(n-1)/q} \equiv 1 \pmod{n},\]

and \(a\) would have order at most \((n-1)/q < n-1\), which is absurd. \(\square\)

### 10.5 Lucas Trees

If \(n\) is an odd prime, we can use Theorem 10.10 to build recursively a tree which is a proof, or certificate, of the fact that \(n\) is indeed prime. We first illustrate this process with the prime \(n = 1279\).

**Example 10.1.** If \(n = 1279\), then we easily check that \(n-1 = 1278 = 2 \cdot 3^2 \cdot 71\). We build a tree whose root node contains the triple \((1279, ((2, 1), (3, 2), (71, 1)), 3)\), where \(a = 3\) is the guess for a primitive root modulo 1279. In this simple example, it is clear that 3 and 71 are prime, but we must supply proofs that these number are prime, so we recursively apply the process to the odd divisors 3 and 71.

Since \(3 - 1 = 2^1\) is a power of 2, we create a one-node tree \((3, ((2, 1)), 2)\), where \(a = 2\) is a guess for a primitive root modulo 3. This is a leaf node.

Since \(71 - 1 = 70 = 2 \cdot 5 \cdot 7\), we create a tree whose root node is \((71, ((2, 1), (5, 1), (7, 1)), 7)\), where \(a = 7\) is the guess for a primitive root modulo 71. Since \(5 - 1 = 4 = 2^2\), and \(7 - 1 = 6 = 2 \cdot 3\), this node has two successors \((5, ((2, 2)), 2)\) and \((7, ((2, 1), (3, 1)), 3)\), where 2 is the guess for a primitive root modulo 5, and 3 is the guess for a primitive root modulo 7.

Since \(4 = 2^2\) is a power of 2, the node \((5, ((2, 2)), 2)\) is a leaf node.

Since \(3 - 1 = 2^1\), the node \((7, ((2, 1), (3, 1)), 3)\) has a single successor, \((3, ((2, 1)), 2)\), where \(a = 2\) is a guess for a primitive root modulo 3. Since \(2 = 2^1\) is a power of 2, the node \((3, ((2, 1)), 2)\) is a leaf node.
To recap, we obtain the following tree:

```
(1279, ((2, 1), (3, 2), (71, 1)), 3)
  ↓
(p3, (2, 1), 2)  (71, ((2, 1), (5, 1), (7, 1)), 7)
  ↓
(5, ((2, 2), 2)  (7, ((2, 1), (3, 1)), 3)
   ↓
(3, ((2, 1)), 2)
```

We still have to check that the relevant congruences hold at every node. For the root node \((1279, ((2, 1), (3, 2), (71, 1)), 3)\), we check that

\[
3^{1278/2} \equiv 3^{864} \equiv -1 \pmod{1279} \quad (1b)
\]
\[
3^{1278/(2\cdot3)} \equiv 3^{213} \equiv 775 \pmod{1279} \quad (2b)
\]
\[
3^{1278/(2\cdot71)} \equiv 3^9 \equiv 498 \pmod{1279}. \quad (2b)
\]

Assuming that 3 and 71 are prime, the above congruences check that Conditions (1a) and (2b) are satisfied, and by Theorem 10.10 this proves that 1279 is prime. We still have to certify that 3 and 71 are prime, and we do this recursively.

For the leaf node \((3, ((2, 1)), 2)\), we check that

\[
2^{2/2} \equiv -1 \pmod{3}. \quad (1b)
\]

For the node \((71, ((2, 1), (5, 1), (7, 1)), 7)\), we check that

\[
7^{70/2} \equiv 7^{35} \equiv -1 \pmod{71} \quad (1b)
\]
\[
7^{70/(2\cdot5)} \equiv 7^7 \equiv 14 \pmod{71} \quad (2b)
\]
\[
7^{70/(2\cdot7)} \equiv 7^5 \equiv 51 \pmod{71}. \quad (2b)
\]

Now we certified that 3 and 71 are prime, assuming that 5 and 7 are prime, which we now establish.

For the leaf node \((5, ((2, 2)), 2)\), we check that

\[
2^{4/2} \equiv 2^2 \equiv -1 \pmod{5}. \quad (1b)
\]
For the node \((7, ((2, 1), (3, 1)), 3)\), we check that
\[
3^{6/2} \equiv 3^3 \equiv -1 \pmod{7} \quad (1b)
\]
\[
3^{6/(2 \cdot 3)} \equiv 3^1 \equiv 3 \pmod{7}. \quad (2b)
\]
We have certified that 5 and 7 are prime, given that 3 is prime, which we finally verify.

At last, for the leaf node \((3, ((2, 1)), 2)\), we check that
\[
2^{2/2} \equiv -1 \pmod{3}. \quad (1b)
\]

The above example suggests the following definition.

**Definition 10.12.** Given any odd integer \(n \geq 3\), a pre-Lucas tree for \(n\) is defined inductively as follows:

1. It is a one-node tree labeled with \((n, ((2, i_0)), a)\), such that \(n - 1 = 2^{i_0}\), for some \(i_0 \geq 1\) and some \(a \in \{2, \ldots, n - 1\}\).

2. If \(L_1, \ldots, L_k\) are \(k\) pre-Lucas (with \(k \geq 1\)), where the tree \(L_j\) is a pre-Lucas tree for some odd integer \(q_j \geq 3\), then the tree \(L\) whose root is labeled with \((n, ((2, i_0), ((q_1, i_1), \ldots, (q_k, i_k)), a)\) and whose \(j\)th subtree is \(L_j\) is a pre-Lucas tree for \(n\) if
\[
n - 1 = 2^{i_0}q_1^{i_1} \cdots q_k^{i_k},
\]
for some \(i_0, i_1, \ldots, i_k \geq 1\), and some \(a \in \{2, \ldots, n - 1\}\).

Both in (1) and (2), the number \(a\) is a guess for a primitive root modulo \(n\).

A pre-Lucas tree for \(n\) is a Lucas tree for \(n\) if the following conditions are satisfied:

3. If \(L\) is a one-node tree labeled with \((n, ((2, i_0)), a)\), then
\[
a^{(n-1)/2} \equiv -1 \pmod{n}.
\]

4. If \(L\) is a pre-Lucas tree whose root is labeled with \((n, ((2, i_0), ((q_1, i_1), \ldots, (q_k, i_k)), a)\), and whose \(j\)th subtree \(L_j\) is a pre-Lucas tree for \(q_j\), then \(L_j\) is a Lucas tree for \(q_j\) for \(j = 1, \ldots, k\), and
\[
\begin{align*}
(a) & \quad a^{(n-1)/2} \equiv -1 \pmod{n}. \\
(b) & \quad a^{(n-1)/2q_j} \not\equiv -1 \pmod{n} \text{ for } j = 1, \ldots, k.
\end{align*}
\]

Since Conditions (3) and (4) of Definition 10.12 are Conditions (1b) and (2b) of Theorem 10.10, we see that Definition 10.12 has been designed in such a way that Theorem 10.10 yields the following result.

**Theorem 10.11.** An odd integer \(n \geq 3\) is prime iff it has some Lucas tree.

The issue is now to see how long it takes to check that a pre-Lucas tree is a Lucas tree. For this, we need a method for computing \(x^n \mod n\) in polynomial time in \(\log_2 n\). This is the object of the next section.
10.6 Algorithms for Computing Powers Modulo $m$

Let us first consider computing the $n$th power $x^n$ of some positive integer. The idea is to look at the parity of $n$ and to proceed recursively. If $n$ is even, say $n = 2k$, then

$$x^n = x^{2k} = (x^k)^2,$$

so, compute $x^k$ recursively and then square the result. If $n$ is odd, say $n = 2k + 1$, then

$$x^n = x^{2k+1} = (x^k)^2 \cdot x,$$

so, compute $x^k$ recursively, square it, and multiply the result by $x$.

What this suggests is to write $n \geq 1$ in binary, say

$$n = b_\ell \cdot 2^\ell + b_{\ell-1} \cdot 2^{\ell-1} + \cdots + b_1 \cdot 2^1 + b_0,$$

where $b_\ell \in \{0, 1\}$ with $b_\ell = 1$ or, if we let $J = \{j \mid b_j = 1\}$, as

$$n = \sum_{j \in J} 2^j.$$

Then we have

$$x^n \equiv x^{\sum_{j \in J} 2^j} = \prod_{j \in J} x^{2^j} \mod m.$$

This suggests computing the residues $r_j$ such that

$$x^{2^j} \equiv r_j \pmod m,$$

because then,

$$x^n \equiv \prod_{j \in J} r_j \pmod m,$$

where we can compute this latter product modulo $m$ two terms at a time.

For example, say we want to compute $999^{179} \mod 1763$. First, we observe that

$$179 = 2^7 + 2^5 + 2^4 + 2^1 + 1,$$

and we compute the powers modulo 1763:

- $999^{2^1} \equiv 143 \pmod{1763}$
- $999^{2^2} \equiv 143^2 \equiv 1056 \pmod{1763}$
- $999^{2^3} \equiv 1056^2 \equiv 920 \pmod{1763}$
- $999^{2^4} \equiv 920^2 \equiv 160 \pmod{1763}$
- $999^{2^5} \equiv 160^2 \equiv 918 \pmod{1763}$
- $999^{2^6} \equiv 918^2 \equiv 10 \pmod{1763}$
- $999^{2^7} \equiv 10^2 \equiv 100 \pmod{1763}$.
Consequently,

\[ 999^{179} \equiv 999 \cdot 143 \cdot 160 \cdot 918 \cdot 100 \pmod{1763} \]
\[ \equiv 54 \cdot 160 \cdot 918 \cdot 100 \pmod{1763} \]
\[ \equiv 1588 \cdot 918 \cdot 100 \pmod{1763} \]
\[ \equiv 1546 \cdot 100 \pmod{1763} \]
\[ \equiv 1219 \pmod{1763}, \]

and we find that

\[ 999^{179} \equiv 1219 \pmod{1763}. \]

Of course, it would be impossible to exponentiate \( 999^{179} \) first and then reduce modulo 1763. As we can see, the number of multiplications needed is bounded by \( 2 \log_2 n \), which is quite good.

The above method can be implemented without actually converting \( n \) to base 2. If \( n \) is even, say \( n = 2k \), then \( n/2 = k \), and if \( n \) is odd, say \( n = 2k + 1 \), then \( (n - 1)/2 = k \), so we have a way of dropping the unit digit in the binary expansion of \( n \) and shifting the remaining digits one place to the right without explicitly computing this binary expansion. Here is an algorithm for computing \( x^n \pmod{m} \), with \( n \geq 1 \), using the repeated squaring method.

**An Algorithm to Compute** \( x^n \pmod{m} **Using Repeated Squaring**

begin
\[ u := 1; \ a := x; \]
while \( n > 1 \) do
\[ \text{if even}(n) \text{ then } e := 0 \text{ else } e := 1; \]
\[ \text{if } e = 1 \text{ then } u := a \cdot u \pmod{m}; \]
\[ a := a^2 \pmod{m}; \ n := (n - e)/2 \]
endwhile;
\[ u := a \cdot u \pmod{m} \]
end

The final value of \( u \) is the result. The reason why the algorithm is correct is that after \( j \) rounds through the while loop, \( a = x^{2^j} \pmod{m} \) and

\[ u = \prod_{i \in J \mid i < j} x^{2^i} \pmod{m}, \]

with this product interpreted as 1 when \( j = 0 \).
Observe that the while loop is only executed \( n - 1 \) times to avoid squaring once more unnecessarily and the last multiplication \( a \cdot u \) is performed outside of the while loop. Also, if we delete the reductions modulo \( m \), the above algorithm is a fast method for computing the \( n \)th power of an integer \( x \) and the time speed-up of not performing the last squaring step is more significant. We leave the details of the proof that the above algorithm is correct as an exercise.

### 10.7 PRIMES is in \( \mathcal{NP} \)

Exponentiation modulo \( n \) can be performed by repeated squaring, as explained in Section 10.6. In that section, we observed that computing \( x^m \mod n \) requires at most \( 2 \log_2 m \) modular multiplications. Using this fact, we obtain the following result adapted from Crandall and Pomerance [5].

**Proposition 10.12.** If \( p \) is any odd prime, then any pre-Lucas tree \( L \) for \( p \) has at most \( \log_2 p \) nodes, and the number \( M(p) \) of modular multiplications required to check that the pre-Lucas tree \( L \) is a Lucas tree is less than \( 2 \log_2^2 p \).

**Proof.** Let \( N(p) \) be the number of nodes in a pre-Lucas tree for \( p \). We proceed by complete induction. If \( p = 3 \), then \( p - 1 = 2^1 \), any pre-Lucas tree has a single node, and \( 1 < \log_2 3 \).

Suppose the results holds for any odd prime less than \( p \). If \( p - 1 = 2^i \), then any Lucas tree has a single node, and \( 1 < \log_2 3 < \log_2 p \). If \( p - 1 \) has the prime factorization

\[
p - 1 = 2^i q_1^i_1 \cdots q_k^i_k,
\]

then by the induction hypothesis, each pre-Lucas tree \( L_j \) for \( q_j \) has less than \( \log_2 q_j \) nodes, so

\[
N(p) = 1 + \sum_{j=1}^k N(q_j) < 1 + \sum_{j=1}^k \log_2 q_j = 1 + \log_2 (q_1 \cdots q_k) \leq 1 + \log_2 \left( \frac{p-1}{2} \right) < \log_2 p,
\]

establishing the induction hypothesis.

If \( r \) is one of the odd primes in the pre-Lucas tree for \( p \), and \( r < p \), then there is some other odd prime \( q \) in this pre-Lucas tree such that \( r \) divides \( q - 1 \) and \( q \leq p \). We also have to show that at some point, \( a^{(q-1)/2r} \not\equiv -1 \pmod q \) for some \( a \), and at another point, that \( b^{(r-1)/2} \equiv -1 \pmod r \) for some \( b \). Using the fact that the number of modular multiplications required to exponentiate to the power \( m \) is at most \( 2 \log_2 m \), we see that the number of multiplications required by the above two exponentiations does not exceed

\[
2 \log_2 \left( \frac{q-1}{2r} \right) + 2 \log_2 \left( \frac{r-1}{2} \right) = 2 \log_2 \left( \frac{(q-1)(r-1)}{4r} \right) < 2 \log_2 q - 4 < 2 \log_2 p.
\]
As a consequence, we have

\[ M(p) < 2 \log_2 \left( \frac{p-1}{2} \right) + (N(p) - 1)2 \log_2 p \]

\[ < 2 \log_2 p + (\log_2 p - 1)2 \log_2 p = 2 \log^2_2 p, \]

as claimed.

The following impressive example is from Pratt [29].

**Example 10.2.** Let \( n = 474\,397\,531 \). It is easy to check that \( n - 1 = 474\,397\,531 - 1 = 474\,397\,530 = 2 \cdot 3 \cdot 5 \cdot 251^3 \). We claim that the following is a Lucas tree for \( n = 474\,397\,531 \):

\[ (474\,397\,531, ((2, 1), (3, 1), (5, 1), (251, 3)), 2) \]

\[ (3, ((2, 1)), 2) \quad (5, ((2, 2)), 2) \quad (251, ((2, 1), (5, 3)), 6) \]

\[ (5, ((2, 2)), 2) \]

To verify that the above pre-Lucas tree is a Lucas tree, we check that 2 is indeed a primitive root modulo 474\,397\,531 by computing (using **Mathematica**) that

\[ 2^{474\,397\,530/2} \equiv 2^{237\,198\,765} \equiv -1 \pmod{474\,397\,531} \] (1)

\[ 2^{474\,397\,530/(2 \cdot 3)} \equiv 2^{79\,066\,255} \equiv 9\,583\,569 \pmod{474\,397\,531} \] (2)

\[ 2^{474\,397\,530/(2 \cdot 5)} \equiv 2^{47\,439\,753} \equiv 91\,151\,207 \pmod{474\,397\,531} \] (3)

\[ 2^{474\,397\,530/(2 \cdot 251)} \equiv 2^{945\,015} \equiv 282\,211\,150 \pmod{474\,397\,531}. \] (4)

The number of modular multiplications is: 27 in (1), 26 in (2), 25 in (3) and 19 in (4).

We have \( 251 - 1 = 250 = 2 \cdot 5^3 \), and we verify that 6 is a primitive root modulo 251 by computing:

\[ 6^{250/2} \equiv 6^{125} \equiv -1 \pmod{251} \] (5)

\[ 6^{250/(2 \cdot 5)} \equiv 6^{10} \equiv 175 \pmod{251}. \] (6)

The number of modular multiplications is: 6 in (5), and 3 in (6).

We have \( 5 - 1 = 4 = 2^2 \), and 2 is a primitive root modulo 5, since

\[ 2^{4/2} \equiv 2^2 \equiv -1 \pmod{5}. \] (7)

This takes one multiplication.

We have \( 3 - 1 = 2 = 2^1 \), and 2 is a primitive root modulo 3, since

\[ 2^{2/2} \equiv 2^1 \equiv -1 \pmod{3}. \] (8)

This takes 0 multiplications.

Therefore, 474\,397\,531 is prime.
As nice as it is, Proposition 10.12 is deceiving, because finding a Lucas tree is hard.

**Remark:** Pratt [29] presents his method for finding a certificate of primality in terms of a proof system. Although quite elegant, we feel that this method is not as transparent as the method using Lucas trees, which we adapted from Crandall and Pomerance [5]. Pratt’s proofs can be represented as trees, as Pratt sketches in Section 3 of his paper. However, Pratt uses the basic version of Lucas’ theorem, Theorem 10.9, instead of the improved version, Theorem 10.10, so his proof trees have at least twice as many nodes as ours.

As nice as it is, Proposition 10.12 is deceiving, because finding a Lucas tree is hard.

The following nice result was first shown by V. Pratt in 1975 [29].

**Theorem 10.13.** The problem PRIMES (testing whether an integer is prime) is in $NP$.

**Proof.** Since all even integers besides 2 are composite, we can restrict our attention to odd integers $n \geq 3$. By Theorem 10.11, an odd integer $n \geq 3$ is prime iff it has a Lucas tree. Given any odd integer $n \geq 3$, since all the numbers involved in the definition of a pre-Lucas tree are less than $n$, there is a finite (very large) number of pre-Lucas trees for $n$. Given a guess of a Lucas tree for $n$, checking that this tree is a pre-Lucas tree can be performed in $O(\log_2 n)$, and by Proposition 10.12, checking that it is a Lucas tree can be done in $O(\log_2^2 n)$. Therefore PRIMES is in $NP$. 

Of course, checking whether a number $n$ is composite is in $NP$, since it suffices to guess to factors $n_1, n_2$ and to check that $n = n_1 n_2$, which can be done in polynomial time in $\log_2 n$. Therefore, PRIMES $\in NP \cap coNP$. As we said earlier, this was the situation until the discovery of the AKS algorithm, which places PRIMES in $P$.

**Remark:** Although finding a primitive root modulo $p$ is hard, we know that the number of primitive roots modulo $p$ is $\varphi(\varphi(p))$. If $p$ is large enough, this number is actually quite large. According to Crandal and Pomerance [5] (Chapter 4, Section 4.1.1), if $p$ is a prime and if $p > 200560490131$, then $p$ has more than $p/(2\ln\ln p)$ primitive roots.
CHAPTER 10. PRIMALITY TESTING IS IN $\mathcal{NP}$
Chapter 11

Polynomial- Space Complexity; \( \mathcal{PS} \) and \( \mathcal{NPS} \)

11.1 The Classes \( \mathcal{PS} \) (or PSPACE) and \( \mathcal{NPS} \) (NPSPACE)

In this chapter we consider complexity classes based on restricting the amount of space used by the Turing machine rather than the amount of time.

**Definition 11.1.** A deterministic or nondeterministic Turing machine \( M \) is polynomial-space bounded if there is a polynomial \( p(X) \) such that for every input \( x \in \Sigma^* \), no matter how much time it uses, the machine \( M \) never visits more than \( p(|x|) \) tape cells (symbols). Equivalently, for every ID \( uav \) arising during the computation, we have \( |uav| \leq p(|x|) \).

The class of languages \( L \subseteq \Sigma^* \) accepted by some deterministic polynomial-space bounded Turing machine is denoted by \( \mathcal{PS} \) or PSPACE. Similarly, the class of languages \( L \subseteq \Sigma^* \) accepted by some nondeterministic polynomial-space bounded Turing machine is denoted by \( \mathcal{NPS} \) or NPSPACE.

Obviously \( \mathcal{PS} \subseteq \mathcal{NPS} \). Since a (time) polynomially bounded Turing machine can’t visit more tape cells (symbols) than one plus the number of moves it makes, we have

\[
\mathcal{P} \subseteq \mathcal{PS} \quad \text{and} \quad \mathcal{NP} \subseteq \mathcal{NPS}.
\]

Nobody knows whether these inclusions are strict, but this is the most likely assumption. Unlike the situation for time-bounded Turing machines where the big open problem is whether \( \mathcal{P} \neq \mathcal{NP} \), for time-bounded Turing machines, we have

\[
\mathcal{PS} = \mathcal{NPS}.
\]

Walter Savitch proved this result in 1970 (and it is known as Savitch’s theorem).
Now Definition 11.1 does not say anything about the time-complexity of the Turing machine, so such a machine could even run forever. However, the number of ID’s that a polynomial-space bounded Turing machine can visit started on input $x$ is a function of $|x|$ of the form $sp(|x|)t^{p(|x|)}$ for some constants $s > 0$ and $t > 0$, so by the pigeonhole principle, it the number of moves is larger than a certain constant ($c^{1+p(|x|)}$ with $c = s + t$), then some ID must repeat. This fact can be used to show that there is a shorter computation accepting $x$ of length at most $c^{1+p(|x|)}$.

**Proposition 11.1.** For any deterministic or nondeterministic polynomial-space bounded Turing machine $M$ with polynomial space bound $p(X)$, there is a constant $c > 1$ such that for every input $x \in \Sigma^*$, if $M$ accepts $x$, then $M$ accepts $x$ in at most $c^{1+p(|x|)}$ steps.

**Proof.** Suppose there are $t$ symbols in the tape alphabet and $s$ states. Then the number of distinct ID’s when only $p(|x|)$ tape cells are used is at most $sp(|x|)^{t^{p(|x|)}}$, because we can choose one of $s$ states, place the reading head in any of $p(|x|)$ distinct positions, and there are $t^{p(|x|)}$ strings of tape symbols of length $p(|x|)$. If we let $c = s + t$, by the binomial formula we have

$$c^{1+p(|x|)} = (s + t)^{1+p(|x|)} = \sum_{k=0}^{1+p(|x|)} \binom{1+p(|x|)}{k} s^{k} t^{1+p(|x|)-k}$$

$$= t^{1+p(|x|)} + (1 + p(|x|))st^{p(|x|)} + \cdots$$

Obviously $(1 + p(|x|))st^{p(|x|)} > sp(|x|)^{t^{p(|x|)}}$, so if the number of ID’s in the computation is greater than $c^{1+p(|x|)}$, by the pigeonhole principle, two ID’s must be identical. By considering a shortest accepting sequence of ID’s with $n$ steps, we deduce that $n \leq c^{1+p(|x|)}$, since otherwise the preceding argument shows that the computation would be of the form

$$ID_0 \vdash \cdots \vdash ID_h \vdash \vdash ID_k \vdash \vdash ID_n$$

with $ID_h = ID_k$, so we would have an even shorter computation

$$ID_0 \vdash \cdash \cdash \vdash ID_h \vdash \vdash ID_n,$$

contradicting the minimality of the original computation. \qed

Proposition 11.1 implies that languages in $\mathcal{NPS}$ are computable (in fact, primitive recursive, and even in $\mathcal{EAXP}$). This still does not show that languages in $\mathcal{NPS}$ are accepted by polynomial-space Turing machines that *always halt* within some time $c^{q(|x|)}$ for some polynomial $q(X)$. Such a result can be shown using a simulation involving a Turing machine with two tapes.

**Proposition 11.2.** For any language $L \in \mathcal{PS}$ (resp. $L \in \mathcal{NPS}$), there is deterministic (resp. nondeterministic) polynomial-space bounded Turing machine $M$, a polynomial $q(X)$ and a constant $c > 1$, such that for every input $x \in \Sigma^*$, $M$ accepts $x$ in at most $c^{q(|x|)}$ steps.

A proof of Proposition 11.2 can be found in Hopcroft, Motwani and Ullman [20] (Section 11.2.2, Theorem 11.4).

We now turn to Savitch’s theorem.
11.2 Savitch’s Theorem: \( \mathcal{PS} = \mathcal{NP} \)

The key to the fact that \( \mathcal{PS} = \mathcal{NP} \) is that given a polynomial-space bounded nondeterministic Turing machine \( M \), there is a recursive method to check whether \( I \vdash_k J \) with \( 0 \leq k \leq m \) using at most \( \log_2 m \) recursive calls, for any two ID’s \( I \) and \( J \) and any natural number \( m \geq 1 \), (that is, whether there is some computation of \( k \) \( \leq m \) steps from \( I \) to \( J \)).

The idea is reminiscent of binary search, namely, to recursively find some intermediate ID \( K \) such that \( I \vdash_1 K \) and \( K \vdash_{m/2} J \) with \( m_1 \leq m/2 \) and \( m_2 \leq m/2 \) (here \( m/2 \) is the integer quotient obtained by dividing \( m \) by 2). Because the Turing machine \( M \) is polynomial-space bounded, for a given input \( x \), we know from Proposition 11.1 that there are at most \( c^{1+p(|x|)} \) distinct ID’s, so the search is finite. We will initially set \( m = c^{1+p(|x|)} \), so at most \( \log_2 c^{1+p(|x|)} = O(p(|x|)) \) recursive calls will be made. We will show that each stack frame takes \( O(p(|x|)) \) space, so altogether the search uses \( O(p(|x|)^2) \) amount of space. This is the crux of Savitch’s argument.

The recursive procedure that deals with stack frames of the form \([I, J, m]\) is shown below.

**function** reach\((I, J, m)\): boolean

**begin**

if \( m = 1 \) then

if \( I = J = K \) or \( I \vdash 1 J \) then

reach = true

else

reach = false

endif

else

for each possible ID \( K \) do

if reach\((I, K, m/2)\) and reach\((K, J, m/2)\) then

reach = true

else

reach = false

endif

endfor

end

end

Even though the above procedure makes two recursive calls, they are performed sequentially, so the maximum number of stack frames that may arise corresponds to the sequence

\([I_1, J_1, m], [I_2, J_2, m/2], [I_3, J_3, m/4], [I_4, J_4, m/8], \ldots, [I_k, J_k, m/2^{k-1}], \ldots\)

which has length at most \( \log_2 m \). Using the procedure \( \text{search} \), we obtain Savitch’s theorem.
Theorem 11.3. (Savitch, 1970) The complexity classes \( \mathcal{P} \mathcal{S} \) and \( \mathcal{N} \mathcal{P} \mathcal{S} \) are identical. In fact, if \( L \) is accepted by the polynomial-space bounded nondeterministic Turing machine \( M \) with space bound \( p(X) \), then there is a polynomial-space bounded deterministic Turing machine \( D \) accepting \( L \) with space bound \( O(p(X)^2) \).

Sketch of proof. Assume that \( L \) is accepted by the polynomial-space bounded nondeterministic Turing machine \( M \) with space bound \( p(X) \). By Proposition 11.1 we may assume that \( M \) accepts any input \( x \in L \) in at most \( c + p(|x|) \) steps (for some \( c > 1 \)). Set \( m = c + p(|x|) \).

We can design a deterministic Turing machine \( D \) which determines (using the function \( \text{search} \)) whether \( I_0 \vdash^k J \) with \( k \leq m \) where \( I_0 = q_0x \) is the starting ID, for all accepting ID’s \( J \), by enumerate all accepting ID’s \( J \) using at most \( p(|x|) \) tape cells, using a scratch tape.

As we explained above, the function \( \text{search} \) makes no more than \( \log_2 (1 + p(|x|)) = O(p(|x|)) \) recursive calls, Each stack frame takes \( O(p(|x|)) \) space. The reason is that every ID has at most \( 1 + p(|x|) \) tape cells and that if we write \( m = c + p(|x|) \) in binary, this takes \( \log_2 m = O(p(|x|)) \) tape cells. Since at most \( O(p(|x|)) \) stack frames may arise and since each stack frame has size at most \( O(p(|x|)) \), the deterministic TM \( D \) uses at most \( O(p(|x|)^2) \) space. For more details, see Hopcroft, Motwani and Ullman [20] (Section 11.2.3, Theorem 11.5).

Savitch’s theorem and Proposition 11.1 show that \( \mathcal{P} \mathcal{S} = \mathcal{N} \mathcal{P} \mathcal{S} \subseteq \mathcal{E} \mathcal{X} \mathcal{P} \). Whether this inclusion is strict is an open problem. The present status of the relative containments of the complexity classes that we have discussed so far is illustrated in Figure 11.1

Savitch’s theorem shows that nondeterminism does not help as far as polynomial space is concerned, but we still don’t have a good example of a language in \( \mathcal{P} \mathcal{S} = \mathcal{N} \mathcal{P} \mathcal{S} \) which is not known to be in \( \mathcal{N} \mathcal{P} \). The next section is devoted to such a problem. This problem also turns out to be \( \mathcal{P} \mathcal{S} \)-complete, so we discuss this notion as well.

11.3 A Complete Problem for \( \mathcal{P} \mathcal{S} \): QBF

Logic is a natural source of problems complete with respect to a number of complexity classes: SAT is \( \mathcal{N} \mathcal{P} \)-complete (see Theorem 8.8), TAUT is \( \text{co} \mathcal{N} \mathcal{P} \)-complete (see Proposition 9.3). It turns out that the validity problem for quantified boolean formulae is \( \mathcal{P} \mathcal{S} \)-complete.

We will describe this problem shortly, but first we define \( \mathcal{P} \mathcal{S} \)-completeness.

Definition 11.2. A language \( L \subseteq \Sigma^* \) is \( \mathcal{P} \mathcal{S} \)-complete if:

1. \( L \in \mathcal{P} \mathcal{S} \).

2. For every language \( L_2 \in \mathcal{P} \mathcal{S} \), there is a polynomial-time computable function \( f : \Sigma^* \rightarrow \Sigma^* \) such that \( x \in L_2 \) iff \( f(x) \in L \), for all \( x \in \Sigma^* \).

Observe that we require the reduction function \( f \) to be polynomial-time computable rather than polynomial-space computable. The reason for this is that with this stronger form of reduction we can prove the following proposition whose simple proof is left as an exercise.
11.3. A COMPLETE PROBLEM FOR PS: QBF

Proposition 11.4. Suppose \( L \) is a \( \mathcal{PS} \)-complete language. Then the following facts hold:

1. If \( L \in \mathcal{P} \), then \( \mathcal{P} = \mathcal{PS} \).
2. If \( L \in \mathcal{NP} \), then \( \mathcal{NP} = \mathcal{PS} \).

The premises in Proposition 11.4 are very unlikely, but we never know!

We now define the class of quantified boolean formulae. These are actually second-order formulae because we are allowed to quantify over propositional variables, which are 0-ary (constant) predicate symbols. As we will see, validity is still decidable, but the fact that we allow alternation of the quantifiers \( \forall \) and \( \exists \) makes the problem harder, in the sense that testing validity or nonvalidity no longer appears to be doable in \( \mathcal{NP} \) (so far, nobody knows how to do this!).

Recall from Section 8.5 that we have a countable set \( \mathbf{PV} \) of propositional (or boolean)
Definition 11.3. A quantified boolean formula (for short QBF) is an expression $A$ defined inductively as follows:

1. The constants $\top$ and $\bot$ and every propositional variable $x_i$ are QBF’s called *atomic* QBF’s.

2. If $B$ is a QBF, then $\neg B$ is a QBF.

3. If $B$ and $C$ are QBF’s, then $(B \lor C)$ is a QBF.

4. If $B$ and $C$ are QBF’s, then $(B \land C)$ is a QBF.

5. If $B$ is a QBF and if $x$ is a propositional variable, then $\forall x B$ is a QBF. The variable $x$ is said to be *universally bound by* $\forall$.

6. If $B$ is a QBF and if $x$ is a propositional variable, then $\exists x B$ is a QBF. The variable $x$ is said to be *existentially bound by* $\exists$.

7. If allow the connective $\Rightarrow$, and if $B$ and $C$ are QBF’s, then $(B \Rightarrow C)$ is a QBF.

Example 11.1. The following formula is a QBF:

$$A = \forall x (\exists y (x \land y) \lor \forall z (\neg x \lor z)).$$

As usual, we can define inductively the notion of free and bound variable as follows.

Definition 11.4. Given any QBF $A$, we define the set $FV(A)$ of *variables free in* $A$ and the set $BV(A)$ of *variables bound in* $A$ as follows:

$$FV(\bot) = FV(\top) = \emptyset,$$

$$FV(x_i) = \{x_i\},$$

$$FV(\neg B) = FV(B),$$

$$FV((B \ast C)) = FV(B) \cup FV(C), \quad \ast \in \{\lor, \land, \Rightarrow\},$$

$$FV(\forall x B) = FV(B) - \{x\},$$

$$FV(\exists x B) = FV(B) - \{x\},$$

and

$$BV(\bot) = BV(\top) = \emptyset,$$

$$BV(x_i) = \emptyset,$$

$$BV(\neg B) = BV(B),$$

$$BV((B \ast C)) = BV(B) \cup BV(C), \quad \ast \in \{\lor, \land, \Rightarrow\},$$

$$BV(\forall x B) = BV(B) \cup \{x\},$$

$$BV(\exists x B) = BV(B) \cup \{x\}.$$
11.3. A COMPLETE PROBLEM FOR \( \mathcal{P}S \): QBF

It should be noted that \( FV(A) \) and \( BV(A) \) may not be disjoint! For example, if

\[
A = x_1 \lor \forall x_1 (\neg x_1 \lor x_2),
\]

then \( FV(A) = \{x_1, x_2\} \) and \( BV(A) = \{x_1\} \). This situation is somewhat undesirable. Intuitively, \( A \) is "equivalent" to the QBF

\[
A' = x_1 \lor \forall x_3 (\neg x_3 \lor x_2),
\]

with \( FV(A') = \{x_1, x_2\} \) and \( BV(A') = \{x_3\} \). Here equivalent means that \( A \) and \( A' \) have the same truth value for all truth assignments. To make all this precise we proceed as follows.

**Definition 11.5.** A substitution is a set of pairs \( \varphi = \{(y_1, A_1), \ldots, (y_m, A_m)\} \) where the variables \( y_1, \ldots, y_m \) are distinct and \( A_1, \ldots, A_m \) are arbitrary QBF's. We write \( \varphi = [y_1 := A_1, \ldots, y_m := A_m] \). For any QBF \( B \), we also denote by \( \varphi[y_i := B] \) the substitution such that \( y_i := A_i \) is replaced by \( y_i := B \). In particular, \( \varphi[y_i := x_i] \) leaves \( x_i \) unchanged.

Given a QBF \( A \), the result of applying the substitution \( \varphi = [y_1 := A_1, \ldots, y_m := A_m] \) to \( A \), denoted \( A[\varphi] \), is defined inductively as follows:

\[
\begin{align*}
\bot[\varphi] &= \bot \\
\top[\varphi] &= \top \\
x[\varphi] &= A_i \quad \text{if } x = y_i, 1 \leq i \leq m \\
x[\varphi] &= x \quad \text{if } x \notin \{y_1, \ldots, y_m\} \\
(\neg B)[\varphi] &= (\neg B)[\varphi] \\
(B \ast C)[\varphi] &= (B[\varphi] \ast C[\varphi]), \quad \ast \in \{\lor, \land, \Rightarrow\} \\
(\forall x B)[\varphi] &= \forall x B[\varphi[y_i := y_i]] \quad \text{if } x = y_i, 1 \leq i \leq m \\
(\forall x B)[\varphi] &= \forall x B[\varphi] \quad \text{if } x \notin \{y_1, \ldots, y_m\} \\
(\exists x B)[\varphi] &= \exists x B[\varphi[y_i := y_i]] \quad \text{if } x = y_i, 1 \leq i \leq m \\
(\exists x B)[\varphi] &= \exists x B[\varphi] \quad \text{if } x \notin \{y_1, \ldots, y_m\}.
\end{align*}
\]

**Definition 11.6.** A QBF \( A \) is rectified if distinct quantifiers bind distinct variables and if \( BV(A) \cap FV(A) = \emptyset \).

Given a QBF \( A \) and any finite set \( V \) of variables, we can define recursively a new rectified QBF \( A' \) such that \( BV(A') \cap V = \emptyset \).

1. If \( A = \top \), or \( A = \bot \), or \( A = x_i \), then \( A' = A \).
2. If \( A = \neg B \), then \( A' = A \).
3. If \( A = (B \lor C) \), then first we find recursively some rectified QBF \( B_1 \) such that \( BV(B_1) \cap V = \emptyset \), then we find recursively some rectified QBF \( C_1 \) such that \( BV(C_1) \cap (FV(B_1) \cup BV(B_1) \cup V) = \emptyset \), and we set \( A' = (B_1 \lor C_1) \). We proceed similarly if \( A = (B \land C) \) or \( A = (B \Rightarrow C) \), with \( \lor \) replaced by \( \land \) or \( \Rightarrow \).
(4) If \( A = \forall x B \), first we find recursively some rectified QBF \( B_1 \) such that \( BV(B_1) \cap V = \emptyset \), and then we let \( A' = \forall z B_1[x := z] \) for some new variable \( z \) such that \( z \notin FV(B_1) \cup BV(B_1) \cup V \). Note that in this step it is possible that \( x \notin FV(B) \).

(5) If \( A = \exists x B \), first we find recursively some rectified QBF \( B_1 \) such that \( BV(B_1) \cap V = \emptyset \), and then we let \( A' = \exists z B_1[x := z] \) for some new variable \( z \) such that \( z \notin FV(B_1) \cup BV(B_1) \cup V \). Note that in this step it is possible that \( x \notin FV(B) \).

Given any QBF \( A \), we find a rectified QBF \( A' \) by applying the above procedure recursively starting with \( A \) and \( V = \emptyset \).

Recall that a truth assignment or valuation is a function \( v : PV \to \{T, F\} \). We also let \( \overline{F} = F \) and \( \overline{T} = T \).

**Definition 11.7.** Given a valuation \( v : PV \to \{T, F\} \), we define truth value \( A[v] \) of a QBF \( A \) inductively as follows.

\[
\begin{align*}
\bot[v] &= F \\
\top[v] &= T \\
x[v] &= v(x) \\
(\neg B)[v] &= \overline{B[v]} = F \text{ if } B[v] = T \text{ else } T \text{ if } B[v] = F \\
(B \lor C)[v] &= B[v] \text{ or } C[v] \\
(B \land C)[v] &= B[v] \text{ and } C[v] \\
(B \Rightarrow C)[v] &= \overline{B[v]} \text{ or } C[v] \\
(\forall x B)[v] &= B[v[x := T]] \text{ and } B[v[x := F]] \\
(\exists x B)[v] &= B[v[x := T]] \text{ or } B[v[x := F]].
\end{align*}
\]

If \( A[v] = T \), we write say that \( v \) satisfies \( A \) and we write \( v \models A \). If \( A[v] = T \) for all valuations \( v \), we say that \( A \) is valid and we write \( \models A \).

As usual, we write \( A \equiv B \) iff \( (A \Rightarrow B) \land (B \Rightarrow A) \) is valid.

In Clause (5) when evaluating \( (B \lor C)[v] \), if \( B[v] = T \), then we don’t need to evaluate \( C[v] \), since \( T \lor b = T \) independently of \( b \in \{T, F\} \), and so \( (B \lor C)[v] = T \). If \( B[v] = F \), then we need to evaluate \( C[v] \), and \( (B \lor C)[v] = T \iff C[v] = T \). Even though the above method is more economical, we usually evaluate both \( B[v] \) and \( C[v] \) and then compute \( B[v] \text{ or } C[v] \).

A similar discussion applies to evaluating \( (\exists x B)[v] \) in Clause (9). If \( B[v[x := T]] = T \), then we don’t need to evaluate \( B[v[x := F]] \) and \( (\exists x B)[v] = T \). If \( B[v[x := T]] = F \), then we need to evaluate \( B[v[x := F]] \), and \( (\exists x B)[v] = T \iff B[v[x := F]] = T \). Even though the above method is more economical, we usually evaluate both \( B[v[x := T]] \) and \( B[v[x := F]] \) and then compute \( B[v[x := T]] \text{ or } B[v[x := F]] \).
11.3. A COMPLETE PROBLEM FOR $\mathcal{PS}$: QBF

Example 11.2. Let us show that the QBF

$$A = \forall x (\exists y (x \land y) \lor \forall z (\neg x \lor z))$$

from Example 11.1 is valid. This is a closed formula so $v$ is irrelevant. By Clause (8) of Definition 11.7, we need to evaluate $A[x := T]$ and $A[x := F]$.

To evaluate $A[x := T]$, by Clause (5) of Definition 11.7, we need to evaluate $(\exists y (x \land y))[x := T]$ and $(\forall z (\neg x \lor z))[x := T]$.

To evaluate $(\exists y (x \land y))[x := T]$, by Clause (9) of Definition 11.7, we need to evaluate $(x \land y)[x := T, y := T]$ and $(x \land y)[x := T, y := F]$.

We have (by Clause (6)) $(x \land y)[x := T, y := T] = T$ and $(x \land y)[x := T, y := F] = F = T$, so

$$(\exists y (x \land y))[x := T] = (x \land y)[x := T, y := T] \lor (x \land y)[x := T, y := F] = T \lor F = T. \quad (1)$$

To evaluate $(\forall z (\neg x \lor z))[x := T]$, by Clause (8) of Definition 11.7, we need to evaluate $(\neg x \lor z)[x := T, z := T]$ and $(\neg x \lor z)[x := T, z := F]$.

Using Clauses (4) and (5) of Definition 11.7, we have $(\neg x \lor z)[x := T, z := T] = T$ and $(\neg x \lor z)[x := T, z := F] = T$. So

$$(\forall z (\neg x \lor z))[x := T] = (\neg x \lor z)[x := T, z := T] \land (\neg x \lor z)[x := T, z := F] = F. \quad (2)$$

By (1) and (2) we have

$$A[x := T] = (\exists y (x \land y))[x := T] \lor (\forall z (\neg x \lor z))[x := T] = T \lor F = T. \quad (3)$$

Now we need to evaluate $A[x := F]$. By Clause (5) of Definition 11.7, we need to evaluate $(\exists y (x \land y))[x := F]$ and $(\forall z (\neg x \lor z))[x := F]$.

To evaluate $(\exists y (x \land y))[x := F]$, by Clause (9) of Definition 11.7, we need to evaluate $(x \land y)[x := F, y := T]$ and $(x \land y)[x := F, y := F]$.

We have (by Clause (6)) $(x \land y)[x := F, y := T] = F$ and $(x \land y)[x := F, y := F] = F = F$, so

$$(\exists y (x \land y))[x := F] = (x \land y)[x := F, y := T] \lor (x \land y)[x := F, y := F] = F \lor F = F. \quad (4)$$

To evaluate $(\forall z (\neg x \lor z))[x := F]$, by Clause (8) of Definition 11.7, we need to evaluate $(\neg x \lor z)[x := F, z := T]$ and $(\neg x \lor z)[x := F, z := F]$.

Using Clauses (4) and (5) of Definition 11.7, we have $(\neg x \lor z)[x := F, z := T] = F$ and $(\neg x \lor z)[x := F, z := F] = F \lor F = T$, so

$$(\forall z (\neg x \lor z))[x := F] = (\neg x \lor z)[x := F, z := T] \land (\neg x \lor z)[x := F, z := F] = T \quad (5)$$
By (4) and (5) we have
\[ A[x := F] = (\exists y (x \land y))[x := F] \text{ or } (\forall z (\neg x \lor z))[x := F] = F \text{ or } T = T. \] (6)

Finally, by (3) and (6) we get
\[ A[x := T] \text{ and } A[x := F] = T \text{ and } T = T, \] (7)
so \( A \) is valid.

The reader should observe that in evaluating
\[ (\exists x B)[v] = B[v[x := T]] \text{ or } B[v[x := F]], \]
if \((\exists x B)[v] = T\), it is only necessary to guess which of \(B[v[x := T]]\) or \(B[v[x := F]]\) evaluates to \(T\), so we can view the computation of \(A[v]\) as an AND/OR tree, where an AND node corresponds to the evaluation of a formula \((\forall x B)[v]\), and an OR node corresponds to the evaluation of a formula \((\exists x B)[v]\).

Evaluating the truth value \(A[v]\) of a QBF \(A\) can take exponential time in the size \(n\) of \(A\), but we will see that it only requires \(O(n^2)\) space. Also, the validity of QBF's of the form
\[ \exists x_1 \exists x_2 \cdots \exists x_m B \]
where \(B\) is quantifier-free and \(FV(B) = \{x_1, \ldots, x_m\}\) is equivalent to SAT (the satisfiability problem), and the validity of QBF's of the form
\[ \forall x_1 \forall x_2 \cdots \forall x_m B \]
where \(B\) is quantifier-free and \(FV(B) = \{x_1, \ldots, x_m\}\) is equivalent to TAUT (the validity problem). This is why the validity problem for QBF's is as hard as both SAT and TAUT.

We mention the following technical results. Part (1) and Part (2) are used all the time.

**Proposition 11.5.** Let \(A\) be any QBF.

1. For any two valuations \(v_1\) and \(v_2\), if \(v_1(x) = v_2(x)\) for all \(x \in FV(A)\), then \(A[v_1] = A[v_2]\). In particular, if \(A\) is a sentence, then \(A[v]\) is independent of \(v\).

2. If \(A'\) is any rectified QBF obtained from \(A\), then \(A[v] = A'[v]\) for all valuations \(v\); that is, \(A \equiv A'\).

3. For any QBF \(A\) of the form \(A = \forall x B\) and any QBF \(C\) such that \(BV(B) \cap FV(C) = \emptyset\), if \(A\) is valid, then \(B[x := C]\) is also valid.

4. For any QBF \(B\) and any QBF \(C\) such that \(BV(B) \cap FV(C) = \emptyset\), if \(B[x := C]\) is valid, then \(\exists x B\) is also valid.
We also repeat Proposition 4.13 which states that the connectives $\land, \lor, \neg$ and $\exists$ are definable in terms of $\Rightarrow$ and $\forall$. This shows the power of the second-order quantifier $\forall$.

**Proposition 11.6.** The connectives $\land, \lor, \neg, \bot$ and $\exists$ are definable in terms of $\Rightarrow$ and $\forall$, which means that the following equivalences are valid, where $x$ is not free in $B$ or $C$:

\[
\begin{align*}
B \land C &\equiv \forall x((B \Rightarrow (C \Rightarrow x)) \Rightarrow x) \\
B \lor C &\equiv \forall x((B \Rightarrow x) \Rightarrow ((C \Rightarrow x) \Rightarrow x)) \\
\bot &\equiv \forall xx \\
\neg B &\equiv B \Rightarrow \forall xx \\
\exists y B &\equiv \forall x((\forall y(B \Rightarrow x)) \Rightarrow x).
\end{align*}
\]

We now prove the first step in establishing that the validity problem for QBF’s is $\mathcal{PS}$-complete.

**Proposition 11.7.** Let $A$ be any QBF of length $n$. Then for any valuation $v$, the truth value $A[v]$ can be evaluated in $O(n^2)$ space. Thus the validity problem for closed QBF’s is in $\mathcal{PS}$.

**Proof.** The clauses of Definition 11.7 show that $A[v]$ is evaluated recursively. In clauses (5)-(9), even though two recursive calls are performed, it is only necessary to save one of the two stack frames at a time. It follows that the stack will never contain more than $n$ stack frames, and each stack frame has size at most $n$. Thus only $O(n^2)$ space is needed. For more details, see Hopcroft, Motwani and Ullman [20] (Section 11.3.4, Theorem 11.10).

Finally we state the main theorem proven by Meyer and Stockmeyer (1973).

**Theorem 11.8.** The validity problem for closed QBF’s is $\mathcal{PS}$-complete.

We will not prove Theorem 11.8, mostly because it requires simulating the computation of a polynomial-space bounded deterministic Turing machine, and this is very technical and tedious. Most details of such a proof can be found in Hopcroft, Motwani and Ullman [20] (Section 11.3.4, Theorem 11.11).

Let us simply make the following comment which gives a clue as to why QBF’s are helpful in describing the simulation (for details, see Hopcroft, Motwani and Ullman [20] (Theorem 11.11)). It turns out that the idea behind the function $reach$ presented in Section 11.2 plays a key role. It is necessary to express for any two ID’s $I$ and $J$ and any $i \geq 1$, that $I \vdash^k J$ with $k \leq i$. This is achieved by defining $N_{2i}(I, J)$ as the following QBF:

\[
N_{2i}(I, J) = \exists K \forall R \forall S \left( ((R = I \land S = K) \lor (R = K \land S = J)) \Rightarrow N_i(R, S) \right).
\]

Another interesting $\mathcal{PS}$-complete problem due to Karp (1972) is the following. Given any alphabet $\Sigma$, decide whether a regular expression $R$ denotes $\Sigma^*$; that is, $L[R] = \Sigma^*$.

We conclude with some comments regarding some remarkable results of Statman regarding the connection between validity of closed QBF’s and provability in intuitionistic propositional logic.
11.4 Complexity of Provability in Intuitionistic Propositional Logic

Recall that intuitionistic logic is obtained from classical logic by taking away the proof-by-contradiction rule. The reader is strongly advised to review Chapter ??, especially Sections ??, ??, ??, ?? and ??, before proceeding.

Statman [34] shows how to reduce the validity problem for QBFs to provability in intuitionistic propositional logic. To simplify the construction we may assume that we consider QBF’s in prenex form, which means that they are of the form

\[ A = Q_n x_n Q_{n-1} x_{n-1} \cdots Q_1 x_1 B_0 \]

where \( B_0 \) is quantifier-free and \( Q_i \in \{\forall, \exists\} \) for \( i = 1, \ldots, n \). We also assume that \( A \) is rectified. It is easy to show that any QBF \( A \) is equivalent to some QBF \( A' \) in prenex form by adapting the method for converting a first-order formula to prenex form; see Gallier [15] or Shoenfield [33].

Statman’s clever trick is to exploit some properties of intuitionistic provability that do not hold for classical logic. One of these properties is that if a proposition \( B \lor C \) is provable intuitionistically, we write \( \vdash_I B \lor C \), then either \( \vdash_I B \) or \( \vdash_I C \), that is, either \( B \) is provable or \( C \) is provable (of course, intuitionistically). This fact is used in the “easy direction” of the proof of Theorem 11.9.

To illustrate the power of the above fact, in his construction, Statman associates the proposition

\[(x \Rightarrow B) \lor (\neg x \Rightarrow B) \quad (*)\]

to the QBF \( \exists x B \). Classically this is useless, because \( (*) \) is classically valid, but if \( (*) \) is intuitionistically provable, then either \( x \Rightarrow B \) is provable or \( \neg x \Rightarrow B \) is intuitionistically provable, but this implies that either \( x \Rightarrow B \) is classically provable or \( \neg x \Rightarrow B \) is classically provable, and so either \( B[x := T] \) is valid or \( B[x := F] \) is valid, which means that \( \exists x B \) is valid.

As a first step, Statman defines the proposition \( B_k^+ \) inductively as follows: for all \( k \) such that \( 0 \leq k \leq n - 1 \),

\[ B_0^+ = \neg \neg B_0 \]
\[ B_{k+1}^+ = (x_{k+1} \lor \neg x_{k+1}) \Rightarrow B_k^+ \quad \text{if } Q_{k+1} = \forall \]
\[ B_{k+1}^+ = (x_{k+1} \Rightarrow B_k^+) \lor (\neg x_{k+1} \Rightarrow B_k^+) \quad \text{if } Q_{k+1} = \exists \]

and set \( A^+ = B_n^+ \). Obviously \( A^+ \) is quantifier-free. We also let \( B_{k+1} = Q_{k+1} x_{k+1} B_k \) for \( k = 0, \ldots, n - 1 \), so that \( A = B_n \).

The following example illustrates the above definition.
Example 11.3. Consider the QBF is prenex form

\[ A = \exists x_3 \forall x_2 \exists x_1 ((x_1 \lor x_2) \land (\neg x_1 \lor \neg x_2) \land (\neg x_3 \lor x_1)) \]

It is indeed valid, as we see by setting \( x_3 = \text{T} \), and if \( x_2 = \text{T} \) then \( x_1 = \text{T} \), else if \( x_2 = \text{F} \) then \( x_1 = \text{F} \). We have

\[
\begin{align*}
B_0^+ &= \neg\neg((x_1 \lor x_2) \land (\neg x_1 \lor \neg x_2) \land (\neg x_3 \lor x_1)) \\
B_1^+ &= (x_1 \Rightarrow B_0^+) \lor (\neg x_1 \Rightarrow B_0^+) \\
B_2^+ &= (x_2 \lor \neg x_2) \Rightarrow B_1^+ \\
B_3^+ &= (x_3 \Rightarrow B_2^+) \lor (\neg x_3 \Rightarrow B_2^+) ,
\end{align*}
\]

and \( A^+ = B_3^+ \).

Statman proves the following remarkable result (Statman [34], Proposition 1).

**Theorem 11.9.** For any closed QBF \( A \) in prenex form, \( A \) is valid iff \( \vdash I A^+ \); that is, \( A^+ \) is intuitionistically provable.

**Proof sketch.** Here is a sketch of Statman’s proof using the QBF of Example 11.3. First assume the QBF \( A \) is valid. The first step is to eliminate existential quantifiers using a variant of what is known as Skolem functions; see Gallier [15] or Shoenfield [33].

The process is to assign to the \( j \)th existential quantifier \( \exists x_k \) from the left in the formula \( Q_n x_n \cdots Q_1 x_1 B_0 \) a boolean function \( C_j \) depending on the universal quantifiers \( \forall x_{i_1}, \ldots, \forall x_{i_p} \) to the left of \( \exists x_k \) and defined such that \( Q_n x_n \cdots Q_{k+1} x_{k+1} \exists x_k B_{k-1} \) is valid iff \( \forall x_{i_1} \cdots \forall x_{i_q} B_{k-1} \) is valid, where \( B_{k-1} \) is the result of substituting the functions \( C_1, \ldots, C_j \) associated with the \( j \) existential quantifiers from the left for these existentially quantified variables.

We associate with \( \exists x_3 \) the constant \( C_1 \) such that \( C_1 = \text{F} \), and with \( \exists x_1 \) the boolean function \( C_2(x_2) \) given by

\[ C_2(\text{T}) = \text{F}, \quad C_2(\text{F}) = \text{T}. \]

The constant \( C_1 \) and the function \( C_2 \) are chosen so that

\[ A = \exists x_3 \forall x_2 \exists x_1 ((x_1 \lor x_2) \land (\neg x_1 \lor \neg x_2) \land (\neg x_3 \lor x_1)) \]

is valid iff

\[ A^+ = \forall x_2((C_2(x_2) \lor x_2) \land (\neg C_2(x_2) \lor x_2) \land (\neg C_1 \lor C_2(x_2))) \tag{S} \]

is valid. Indeed, since \( C_1 = \text{F} \), the clause \((\neg C_1 \lor C_2(x_2))\) evaluates to \( \text{T} \) regardless of the value of \( x_2 \), and by definition of \( C_2 \), the expression

\[ \forall x_2(C_2(x_2) \lor x_2) \land (\neg C_2(x_2) \lor x_2) \]

also evaluates to \( \text{T} \). We now build a tree of Gentzen sequents (from the root up) from the expression in \( (S) \) which guides us in deciding which disjunct to pick when dealing with a proposition \( B_k^+ \) associated with an existential quantifier. Here is the tree.
\[ \neg x_3, x_2, \neg x_1 \rightarrow B_0^+ \]
\[ \neg x_3, x_2 \rightarrow \neg x_1 \Rightarrow B_0^+ \]
\[ \neg x_3, x_2 \rightarrow (x_1 \Rightarrow B_0^+) \lor (\neg x_1 \Rightarrow B_0^+) \]
\[ \neg x_3, \neg x_2 \rightarrow x_1 \Rightarrow B_0^+ \]
\[ \neg x_3, \neg x_2 \rightarrow (x_1 \Rightarrow B_0^+) \lor (\neg x_1 \Rightarrow B_0^+) \]
\[ \neg x_3, x_2 \lor \neg x_2 \rightarrow B_1^+ \]
\[ \neg x_3 \rightarrow (x_2 \lor \neg x_2) \Rightarrow B_1^+ \]
\[ \rightarrow \neg x_3 \Rightarrow B_2^+ \]
\[ \rightarrow (x_3 \Rightarrow B_2^+) \lor (\neg x_3 \Rightarrow B_2^+) \]

We will see that by adding subtrees proving the sequents in the leaf nodes, this tree becomes an intuitionistic proof of \( A^+ \). Note that this a proof in a Gentzen sequent style formulation of intuitionistic logic (see Kleene [21], Gallier [11], Takeuti [36]), not a proof in a natural deduction style proof system as in Section ??.

The tree is constructed from the bottom-up starting with \( \rightarrow A^+ \). For every leaf node in the tree where a sequent is of the form
\[ \ell_n, \ldots, \ell_{k+1} \rightarrow (x_k \Rightarrow B_{k-1}^+) \lor (\neg x_k \Rightarrow B_{k-1}^+) \]
where \( \ell_n, \ldots, \ell_{k+1} \) are literals, we know that \( Q_k = \exists \) is the \( j \)th existential quantifier from the left, so we use the boolean function \( C_j \) to determine which of the two disjuncts \( x_k \Rightarrow B_{k-1}^+ \) or \( \neg x_k \Rightarrow B_{k-1}^+ \) to keep. The function \( C_j \) depends on the value of the literals \( \ell_n, \ldots, \ell_{k+1} \) associated with universal quantifiers (where \( \ell_i \) has the value \( T \) if \( \ell_i = x_i \) and \( \ell_i \) has the value \( F \) if \( \ell_i = \neg x_i \)). Even though \( C_j \) is independent of the value of the literals \( \ell_i \) associated with existential quantifiers, to simplify notation we write \( C_j(\ell_n, \ldots, \ell_{k+1}) \) for the value of the function \( C_j \). If \( C_j(\ell_n, \ldots, \ell_{k+1}) = T \), then we pick the disjunct \( x_k \Rightarrow B_{k-1}^+ \), else if \( C_j(\ell_n, \ldots, \ell_{k+1}) = F \), then we pick the disjunct \( \neg x_k \Rightarrow B_{k-1}^+ \). Denote the literal corresponding to the chosen disjunct by \( \ell_k \) (\( \ell_k = x_k \) in the first case, \( \ell_k = \neg x_k \) in the second case). Then we grow two new nodes
\[ \ell_n, \ldots, \ell_{k+1} \rightarrow \ell_k \Rightarrow B_{k-1}^+ \]
and
\[ \ell_n, \ldots, \ell_{k+1}, \ell_k \rightarrow B_{k-1}^+ \]
above the (leaf) node
\[ \ell_n, \ldots, \ell_{k+1} \rightarrow (x_k \Rightarrow B_{k-1}^+) \lor (\neg x_k \Rightarrow B_{k-1}^+) \]

For every leaf node of the form
\[ \ell_n, \ldots, \ell_{k+1} \rightarrow (x_k \lor \neg x_k) \Rightarrow B_{k-1}^+ \]
we grow the new node
\[ \ell_n, \ldots, \ell_{k+1}, x_k \lor \neg x_k \rightarrow B_{k-1}^+ \].
and then the two new nodes (both descendants of the above node, so there is branching in
the tree),
\[ \ell_n, \ldots, \ell_{k+1}, x_k \rightarrow B^+_{k-1} \quad \text{and} \quad \ell_n, \ldots, \ell_{k+1}, \neg x_k \rightarrow B^+_{k-1}. \]

By induction from the bottom-up, since \( A \) is valid and since the tree was constructed in
terms of the constant \( C_1 \) and the function \( C_2 \) which ensure the validity of \( A \), it is easy to see
that for every node \( \ell_n, \ldots, \ell_{k+1} \rightarrow B^+_k \), the sequent \( \ell_n, \ldots, \ell_{k+1} \rightarrow B_k \) (note, the right-hand
side is the original formula \( B_k \) is classically valid, and thus classically provable (by the
completeness theorem for propositional logic). Consequently every leaf \( \ell_n, \ldots, \ell_1 \rightarrow B_0 \) is
classically provable, so by Glivenko’s theorem (see Kleene [21] (Theorem 59), or Gallier [11]
(Section 13)), the sequent \( \ell_n, \ldots, \ell_1 \rightarrow \neg\neg B_0 \) is intuitionistically provable. But this is the
sequent \( \ell_n, \ldots, \ell_1 \rightarrow B^+_0 \) so all the leaves of the tree are intuitionistically provable, and since
the tree is a deduction tree in a Gentzen sequent style formulation of intuitionistic logic (see
Kleene [21], Gallier [11], Takeuti [36]), the root \( A^+ = B^+_n \) is intuitionistically provable.

In the other direction, assume that \( A^+ \) is intuitionistically provable. We use the fact that if
\[ \ell_n, \ldots, \ell_k \rightarrow A \lor B \]
is intuitionistically provable and the \( \ell_i \) are literals, then either \( \ell_n, \ldots, \ell_k \rightarrow A \) is intuition-
istically provable or \( \ell_n, \ldots, \ell_k \rightarrow B \) is intuitionistically provable, and other proof rules of
intuitionistic logic (see Kleene [21], Gallier [11], Takeuti [36]), to build a proof tree just
as we did before. Then every sequent \( \ell_n, \ldots, \ell_{k+1} \rightarrow B_k^+ \) is intuitionistically provable, thus
classically provable, and consequently classically valid. But this immediately implies (by
induction starting from the leaves) that \( \ell_n, \ldots, \ell_{k+1} \rightarrow B_k \) is also classically valid for all \( k \),
and thus \( A = B^n \) is valid.

Statman does not specifically state which proof system of intuitionistic logic is used in
Theorem 11.9. Careful inspection of the proof shows that we can construct proof trees in a
Gentzen sequent calculus as described in Gallier [11] (system \( G_i \), Section 4) or Kleene [21]
(system \( G_3a \), Section 80, pages 481-482). This brings up the following issue: could we use
instead proofs in natural deduction style, as in Prawitz [30] or Gallier [11]? The answer is
yes, because there is a polynomial-time translation of intuitionistic proofs in Gentzen sequent
style to intuitionistic proofs in natural deduction style, as shown in Gallier [11], Section 5.
So Theorem 11.9 applies to a Gentzen sequent style proof system or to a natural deduction
style proof system.

The problem with the translation \( A \mapsto A^+ \) is that \( A^+ \) may not have size polynomial in
the size (the length of \( A \) as a string) of \( A \) because in the case of an existential quantifier
the length of the formula \( B^+_{k+1} \) is more than twice the length of the formula \( B^+_k \), so Statman
introduces a second translation.

The proposition \( B^+_k \) is defined inductively as follows. Let \( y_0, y_1, \ldots, y_n \) be \( n + 1 \) new
propositional variables. For all \( k \) such that \( 0 \leq k \leq n - 1 \),

\[
\begin{align*}
B_0^\dagger &= \neg \neg B_0 \equiv y_0 \\
B_{k+1}^\dagger &= ((x_{k+1} \lor \neg x_{k+1}) \Rightarrow y_k) \equiv y_{k+1} & \text{if } Q_{k+1} = \forall \\
B_{k+1}^\dagger &= ((x_{k+1} \Rightarrow y_k) \lor (\neg x_{k+1} \Rightarrow y_k)) \equiv y_{k+1}, & \text{if } Q_{k+1} = \exists
\end{align*}
\]

and set

\[ A^* = B_0^\dagger \Rightarrow (B_1^\dagger \Rightarrow (\cdots (B_n^\dagger \Rightarrow y_n) \cdots)). \]

It is easy to see that the translation \( A \mapsto A^* \) can be done in polynomial space. Statman proves the following result (Statman [34], Proposition 2).

**Theorem 11.10.** For any closed QBF \( A \) in prenex form, \( \vdash I A^+ \) iff \( \vdash I A^* \); that is, \( A^+ \) is intuitionistically provable iff \( A^* \) is intuitionistically provable.

**Proof.** First suppose the sequent \( \vdash A^+ \) is provable (in Kleene G3a). We claim that the sequent

\[ B_0^1, \ldots, B_k^1 \rightarrow B_k^+ \equiv y_k \]

is provable for \( k = 0, \ldots, n \). We proceed by induction on \( k \). For the base case \( k = 0 \), we have \( B_0^1 = (\neg \neg B_0 \equiv y_0) \) and \( B_0^+ = \neg \neg B_0 \), so \( B_0^1 \rightarrow (B_0^+ \equiv y_0) = (B_0^+ \equiv y_0) \rightarrow (B_0^+ \equiv y_0) \), which is an axiom.

For the induction step, if \( Q_{k+1} = \forall \), then

\[ B_{k+1}^+ = (x_{k+1} \lor \neg x_{k+1}) \Rightarrow B_k^+, \quad B_{k+1}^1 = ((x_{k+1} \lor \neg x_{k+1}) \Rightarrow y_k) \equiv y_{k+1}, \]

by the induction hypothesis

\[ B_0^1, \ldots, B_k^1 \rightarrow B_k^+ \equiv y_k \]

is provable, and since the sequent

\[ B_0^1, \ldots, B_k^1, B_{k+1}^1 \rightarrow B_{k+1}^1 \]

is an axiom, by substituting \( B_{k+1}^+ \) for \( y_k \) in \( B_{k+1}^1 = ((x_{k+1} \lor \neg x_{k+1}) \Rightarrow y_k) \equiv y_{k+1} \) in the conclusion of the above sequent, we deduce that

\[ B_0^1, \ldots, B_k^1, B_{k+1}^1 \rightarrow ((x_{k+1} \lor \neg x_{k+1}) \Rightarrow B_k^+) \equiv y_{k+1} \]

is provable. Since \( B_{k+1}^+ = (x_{k+1} \lor \neg x_{k+1}) \Rightarrow B_k^+ \), we conclude that

\[ B_0^1, \ldots, B_k^1, B_{k+1}^1 \rightarrow B_{k+1}^+ \equiv y_{k+1} \]

is provable.

If \( Q_{k+1} = \exists \), then

\[ B_{k+1}^+ = (x_{k+1} \Rightarrow B_k^+) \lor (\neg x_{k+1} \Rightarrow B_k^+), \quad B_{k+1}^1 = ((x_{k+1} \Rightarrow y_k) \lor (x_{k+1} \Rightarrow y_k)) \equiv y_{k+1}, \]
by the induction hypothesis
\[ B_0^+, \ldots, B_k^+ \rightarrow B_k^+ \equiv y_k \]
is provable, and since the sequent
\[ B_0^+, \ldots, B_k^+, B_{k+1}^+ \rightarrow B_{k+1}^+ \]
is an axiom, by substituting \( B_k^+ \) for \( y_k \) in \( B_{k+1}^+ = ((x_{k+1} \Rightarrow y_k) \lor (x_{k+1} \Rightarrow y_k)) \equiv y_{k+1} \) in the conclusion of the above sequent, we deduce that
\[ B_0^+, \ldots, B_k^+, B_{k+1}^+ \rightarrow ((x_{k+1} \Rightarrow B_{k+1}^+) \lor (\neg x_{k+1} \Rightarrow B_{k+1}^+)) \equiv y_{k+1} \]
is provable. Since \( B_{k+1}^+ = (x_{k+1} \Rightarrow B_k^+) \lor (\neg x_{k+1} \Rightarrow B_k^+) \), we conclude that
\[ B_0^+, \ldots, B_k^+, B_{k+1}^+ \rightarrow B_{k+1}^+ \equiv y_{k+1} \]
is provable. Therefore the induction step holds. For \( k = n \), we see that the sequent
\[ B_0^+, \ldots, B_n^+ \rightarrow (B_n^+ \equiv y_n) = B_0^+, \ldots, B_n^+ \rightarrow (A^+ \equiv y_n) \]
is provable, and since by hypothesis \( \rightarrow A^+ \) is provable, we deduce that
\[ B_0^+, \ldots, B_n^+ \rightarrow y_n \]
is provable. Finally we deduce that
\[ A^* = B_0^+ \Rightarrow (B_1^+ \Rightarrow (\cdots (B_n^+ \Rightarrow y_n) \cdots)) \]
is provable intuitionistically.

Conversely assume that \( A^* = B_0^+ \Rightarrow (B_1^+ \Rightarrow (\cdots (B_n^+ \Rightarrow y_n) \cdots)) \) is provable intuitionistically. Then using basic properties of intuitionistic provability, the sequent
\[ B_0^+, \ldots, B_n^+ \rightarrow y_n \]
is provable intuitionistically. Now if we substitute \( B_{k+1}^+ \) for \( y_{k+1} \) in \( B_{k+1}^+ \) for \( k = 0, \ldots, n - 1 \), we see immediately that
\[ B_{k+1}^+[y_{k+1} := B_{k+1}^+] = B_{k+1}^+ \equiv B_{k+1}^+ \]
so the proof of
\[ B_0^+, \ldots, B_n^+ \rightarrow y_n \]
yields a proof of
\[ B_0^+ \equiv B_0^+, \ldots, B_n^+ \equiv B_n^+ \rightarrow B_n^+ \]
that is, a proof (intuitionistic) of \( B_n^+ = A^+ \).

**Remark:** Note that we made implicit use of the cut rule several times, but by Gentzen’s cut-elimination theorem this does not matter (see Gallier [11]).
CHAPTER 11. POLYNOMIAL-SPACE COMPLEXITY; \( \mathcal{PS} \) AND \( \mathcal{NPS} \)

Using Theorems 11.9 and 11.10 we deduce from the fact that validity of QBF’s is \( \mathcal{PS} \)-complete that provability in propositional intuitionistic logic is \( \mathcal{PS} \)-hard (every problem in \( \mathcal{PS} \) reduces in polynomial time to provability in propositional intuitionistic logic). Using results of Tarski and Ladner, it is can be shown that validity in Kripke models for propositional intuitionistic logic belongs to \( \mathcal{PS} \), so Statman proves the following result (Statman [34], Section 2, Theorem).

**Theorem 11.11.** The problem of deciding whether a proposition is valid in all Kripke models is \( \mathcal{PS} \)-complete.

Theorem 11.11 also applies to any proof system for intuitionistic logic which is sound and complete for Kripke semantics.

**Theorem 11.12.** The problem of deciding whether a proposition is intuitionistically provable in any sound and complete proof system (for Kripke semantics) is \( \mathcal{PS} \)-complete.

Theorem 11.12 applies to Gallier’s system \( \mathcal{G}_i \), to Kleene’s system \( \mathcal{G}3a \), and to natural deduction systems. To prove that \( \mathcal{G}_i \) is complete for Kripke semantics it is better to convert proofs in \( \mathcal{G}_i \) to proofs in a system due to Takeuti, the system denoted \( \mathcal{GKT} \) in Gallier [11]; see Section 9, Definition 9.3. Since there is a polynomial-time translation of proofs in \( \mathcal{G}_i \) to proofs in natural deduction, the latter system is also complete. This is also proven in van Dalen [38].

Statman proves an even stronger remarkable result, namely that \( \mathcal{PS} \)-completeness holds even for propositions using only the connective \( \Rightarrow \) (Statman [34], Section 2, Proposition 3).

**Theorem 11.13.** There is an algorithm which given any proposition \( A \) constructs another proposition \( A^\# \) only involving \( \bot, \Rightarrow \), such that that \( \vdash I A \iff \vdash I A^\# \).

Theorem 11.13 is somewhat surprising in view of the fact that \( \lor, \land, \Rightarrow \) are independent connectives in propositional intuitionistic logic. Finally Statman obtains the following beautiful result (Statman [34], Section 2, Theorem).

**Theorem 11.14.** The problem of deciding whether a proposition only involving \( \bot, \Rightarrow \) is valid in all Kripke models, and intuitionistically provable in any sound and complete proof system, is \( \mathcal{PS} \)-complete.

We highly recommend reading Statman [34], but we warn the reader that this requires perseverance.
Bibliography


277


Symbol Index

$(M N)$, 89
$M[\varphi]$, 93
$[x_1 := N_1, \ldots, x_n := N_n]$, 92
$\leftarrow_\beta$, 94
$\rightarrow_\beta$, 94
$\lambda x. M$, 89
$\ast \rightarrow_\beta$, 94
$+ \rightarrow_\beta$, 94
Index

α-conversion, 93
   immediate, 93
α-reduction, 93
   immediate, 93
β-conversion, 94
β-redex, 94
β-reduction, 94
   immediate, 94
λ-abstraction, 89
λ-calculus, 89
   pure, 89
   untyped, 89
λ-term
   closed, 91
   raw simply-typed, 89
application, 89
combinator, 91
congruent, 239
modular arithmetic, 239, 240
redex, 94
repeated squaring, 254
substitution, 92
type, 115
   second order, 115
   polymorphic, 115
variable, 89
   bound, 91
   capture, 93
   free, 91