An Introduction to Artificial Intelligence

— Lecture Notes in Progress —

Prof. Dr. Karl Stroetmann

January 29, 2019

These lecture notes, their LaTeX sources, and the programs discussed in these lecture notes are all available at

https://github.com/karlstroetmann/Artificial-Intelligence.

In particular, the lecture notes are found in the directory Lecture-Notes-Python in the file artificial-intelligence.pdf. The lecture notes are subject to continuous change. Provided the program `git` is installed on your computer, the repository containing the lecture notes can be cloned using the command

```
git clone https://github.com/karlstroetmann/Artificial-Intelligence.git
```

Once you have cloned the repository, the command

```
git pull
```

can be used to load the current version of these lecture notes from `github`. As this is the third time that I give this lecture, these lecture notes are still incomplete and will be changed frequently during the semester. If you find any typos, errors, or inconsistencies, please contact me via discord or, if that is not possible, email me at:

`karl.stroetmann@dhbw-mannheim.de`.

I am currently rewriting these lecture notes as I am moving from the programming language SetlX to *Python*.

Contents

Chapter 1

Introduction

1.1 What is Artificial Intelligence?

Before we start to dive into the subject of Artificial Intelligence we have to answer the following question:

What is Artificial Intelligence?

Historically, there have been a number of different answers to this question [RN09]. We will look at these different answers and discuss them.

1. Artificial Intelligence is the study of creating machines that think like humans.

   As we have a working prototype of intelligence, namely humans, it is quite natural to try to build machines that work in a way similar to humans, thereby creating artificial intelligence. As a first step in this endeavor we would have to study how humans actually think and thus we would have to study the brain. Unfortunately, as of today, no one really knows how the brain works. Although there are branches of science devoted to studying the human thought processes and the human brain, namely cognitive science and computational neuroscience, this approach has not proven to be fruitful for creating thinking machines, the reason being that the current understanding of the human thought processes is just not sufficient.

2. Artificial Intelligence is the science of machines that act like humans.

   Since we do not know how humans think, we cannot build machines that think like people. Therefore, the next best thing might be to build machines that act and behave like humans. Actually, the Turing Test is based on this idea: Turing suggested that if we want to know whether we have succeeded in building an intelligent machine, we should place it at the other end of a chat line. If we cannot distinguish the computer from a human, then we have succeeded at creating intelligence.

   However, with respect to the kind of Artificial Intelligence that is needed in industry, this approach isn't very useful. To illustrate the point, consider an analogy with aerodynamics: In aerodynamics we try to build planes that fly fast and efficiently, not planes that flap their wings like birds do, as the later approach has failed historically, e.g. remember what happened to Daedalus and Icarus.

3. Artificial Intelligence is the science of creating machines that think logically.

   The idea with this approach is to create machines that are based on mathematical logic. If a goal is given to these machines, then these machines use logical reasoning in order to deduce those actions that need to be performed in order to best achieve the given goals. Unfortunately, this approach had only limited success: In playing games the approach was quite successful for dealing with games like checkers or chess. However, the approach was mostly unsuccessful for dealing with many real world problems. There were two main reasons for its failure:

   (a) In order for the logical approach to be successful, the environment has to be completely described by mathematical axioms. It has turned out that our knowledge of the real world is often not sufficient to completely describe the environment via axioms.

(b) Even if we have complete knowledge, it often turns out that describing every possible case via logic formulae is just unwieldy. Consider the following formula:

$$\forall x : \big(\texttt{bird}(x) \to \texttt{flies}(x)\big)$$

The problem with this formula is that although it appears to be common sense, there are a number of counter examples:

i. Penguins, emus, and ostriches don't fly.
However, if we put a penguin into a plane, it turns out the penguin will fly.

ii. Birds that have just hatched do not fly.

iii. Birds with clipped wings do not fly.

It is easy to extend this list with ever more contrived examples and counter examples. This shows that trying to model all eventualities with logic is just too unwieldy to be practical.

(c) In real life situations we often deal with uncertainty. Classical logic does not perform well when it has to deal with uncertainties. Instead, we need statistics.

4. Artificial Intelligence is the science of creating machines that act rationally.

All we really want is to build machines that, given the knowledge we have, try to optimize the expected results: In our world, there is lots of uncertainty. We cannot hope to create machines that always make the decisions that turn out to be optimal. What we can hope is to create machines that will make decisions that turn out to be good on average. For example, suppose we try to create a program for asset management: We cannot hope to build a machine that always buys the best company share in the stock market. Rather, our goal should be to build a program that maximizes our expected profits in the long term.

It has turned out that the main tool needed for this approach is not mathematical logic but rather numerical analysis and mathematical statistics. The shift from logic to numerical analysis and statistics has been the most important reason for the spectacular success of Artificial Intelligence in the recent years. Another important factor is the enhanced performance of modern hardware.

Now that we have clarified the notion of artificial intelligence, we should set its goals. As we can never achieve more than what we aim for, we have every reason to be ambitious here. For example, my personal vision of Artificial Intelligence goes like this: Imagine 70 years from now you (not feeling too well) have a conversation with Siri. Instead of asking Siri for the best graveyard in the vicinity, you think about all the sins you have committed. As Siri has accompanied you for your whole life, she knows about these sins better than you. Hence, the conversation with Siri works out as follows:

You (with trembling voice):	Hey Siri, does God exist?
Siri (with the voice of Darth Vader): After a small pause which almost drains the battery of your phone completely, Siri gets back with a soothing announcement:	Your voice seems troubled, let me think $\cdots$
	You don't have to worry any more, I have fixed the problem. He is dead now.

My apologies to all those infidels that do not get the point.

1.2 Overview

This lecture consists of two parts.

1. The common theme of the first part is declarative programming. The main idea of declarative programming is that we start with a problem specification. This is usually a short description of the problem that is to be solved. This description is then fed into an automatic problem solver that returns a solution of the problem. Originally, declarative programming was a very general approach to problem solving. The idea was that in order to solve a problem, the problem would first be formulated as a logic formula and an automated theorem prover would then be able to solve the problem. Unfortunately, in this generality the idea of declarative programming has turned out to be unsuitable for real world problems for two reasons:

   (a) First, it is very difficult to specify practical problems completely in a logical framework.

   (b) Second, even in those cases where a complete logic based specification of a problem is feasible, automatic theorem proving is generally not powerful enough to find a solution automatically.

   However, there are a number of domains where the approach of declarative programming has turned out to be useful. In particular, we show how declarative programming can be used to solve problems in the following domains.

   (a) Search problems are problems where the task is to find a path in a graph. A typical example of a search problem is the fifteen puzzle. We discuss various state-of-the-art algorithms that can solve search problems.

   (b) Games like chess or poker can be specified quite easily and there are various techniques for computers to find optimal strategies for playing adversarial games.

   (c) Constraint satisfaction problems have great practical importance. Today, very efficient constraint solvers have been developed to solve various constraint satisfaction problems that often occur in practice.

2. In the second part of this lecture we discuss machine learning. In the last ten years, a number of advances in machine learning have made it into the headlines of the news. It is fair to say that currently machine learning is the hottest topic in computer science. Among others, we discuss the following algorithms:

   (a) Linear regression is one of the most fundamental machine learning algorithms. In machine learning, we are given a number of data pairs of the form $\langle \mathbf{x}_i, y_i \rangle$ where $i \in \{1, \cdots, N\}$ and for all $i \in \{1, \cdots, N\}$ we have $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$. We assume that there is an unknown function $f : \mathbb{R}^m \to \mathbb{R}$ such that $y_i \approx f(\mathbf{x}_i)$. Our task is to find an approximation for the function f. When using linear regression, we assume that the function f is linear in its arguments. Although this sounds like a strong assumption, we will see that linear regression is surprisingly powerful in practice.

   (b) In a classification problem we again have N pairs of the form $\langle \mathbf{x}_i, y_i \rangle$. As before, we have $\mathbf{x}_i \in \mathbb{R}^m$, but now $y_i \in \mathbb{B}$, where $\mathbb{B}$ is the set of Boolean values, i.e. we have $\mathbb{B} = \{\texttt{true}, \texttt{false}\}$. The task is then to find a function $f : \mathbb{R}^m \to \mathbb{B}$ such that the equation $y_i = f(\mathbf{x}_i)$ is true for most $i \in \{1, \cdots, N\}$. A typical classification problem is spam detection. The first algorithm we introduce to solve classification problems is logistic regression.

   (c) Finally, we discuss neural networks. As an example, we will build a neural network that is able to recognize digits.

1.3 Literature

The main sources of these lecture notes are the following:

1. A course on artificial intelligence that was offered on the EDX platform. The course materials are available at

   http://ai.berkeley.edu/home.html.

2. The book

   Introduction to Artificial Intelligence

   written by Stuart Russell and Peter Norvig [RN09].

3. A course on artificial intelligence that is offered on Udacity. The title of the course is

   Intro to Artificial Intelligence

   and the course is given by Peter Norvig, who is director of research at Google and Sebastian Thrun, who is the chairman of Udacity.

The programs presented in these lecture notes are expected to run with version 2.7 of SETLX.

Chapter 2

Search

In this chapter we discuss various search algorithms. First, we define the notion of a search problem. We will discuss the sliding puzzle as our running example. Then we introduce various algorithms for solving search problems. In particular, we present

1. breadth first search,
2. depth first search,
3. iterative deepening,
4. bidirectional breadth first search,
5. A* search,
6. bidirectional A* search,
7. iterative deepening A* search, and
8. A*-IDA* search.

Definition 1 (Search Problem) A search problem is a tuple of the form

$$\mathcal{P} = \langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal} \rangle$$

where

1. Q is the set of states, also known as the state space.
2. `nextStates` is a function taking a state as input and returning the set of those states that can be reached from the given state in one step, i.e. we have

   $$\texttt{nextStates} : Q \to 2^Q.$$

   The function `nextStates` gives rise to the transition relation R, which is a relation on Q, i.e. $R \subseteq Q \times Q$. This relation is defined as follows:

   $$R := \big\{ \langle s_1, s_2 \rangle \in Q \times Q \mid s_2 \in \texttt{nextStates}(s_1) \big\}.$$

   If either $\langle s_1, s_2 \rangle \in R$ or $\langle s_2, s_1 \rangle \in R$, then s_1 and s_2 are called neighboring states.
3. `start` is the start state, hence $\texttt{start} \in Q$.
4. `goal` is the goal state, hence $\texttt{goal} \in Q$.

   Sometimes, instead of a single `goal` there is a set of goal states `Goals`.

A path is a list $[s_1, \cdots, s_n]$ such that $s_{i+1} \in \texttt{nextStates}(s_i)$ for all $i \in \{1, \cdots, n-1\}$. The length of this path is defined as the length of the list. A path $[s_1, \cdots, s_n]$ is a solution to the search problem P iff the following conditions are satisfied:

1. $s_1 = \texttt{start}$, i.e. the first element of the path is the start state.

2. $s_n = \texttt{goal}$, i.e. the last element of the path is the goal state.

   If instead of a single `goal` we have a set of `Goals`, then the last condition is changed into

   $$s_n \in \texttt{Goals}.$$

A path $p = [s_1, \cdots, s_n]$ is a minimal solution to the search problem $\mathcal{P}$ iff it is a solution and, furthermore, the length of p is minimal among all other solutions. ◇

Remark: In the literature, a state is often called a node. In these lecture notes, I will also refer to states as nodes. ◇

Example: We illustrate the notion of a search problem with the following example, which is also known as the missionaries and cannibals problem: Three missionaries and three infidels have to cross a river that runs from the north to the south. Initially, both the missionaries and the infidels are on the western shore. There is just one small boat and that boat has only room for at most two passengers. Both the missionaries and the infidels can steer the boat. However, if at any time the missionaries are confronted with a majority of infidels on either shore of the river, then the missionaries have a problem.

```
1  problem := [m, i] |-> m > 0 && m < i;
2
3  noProblemAtAll := [m, i] |-> !problem(m, i) && !problem(3 - m, 3 - i);
4
5  nextStates := procedure(s) {
6      [m, i, b] := s;
7      if (b == 1) {  // The boat is on the western shore.
8          return { [m - mb, i - ib, 0]
9                 : mb in {0 .. m}, ib in {0 .. i}
10                | mb + ib in {1, 2} && noProblemAtAll(m - mb, i - ib)
11                };
12     } else {
13         return { [m + mb, i + ib, 1]
14                : mb in {0 .. 3 - m}, ib in {0 .. 3 - i}
15                | mb + ib in {1, 2} && noProblemAtAll(m + mb, i + ib)
16                };
17     }
18 };
19 start := [3, 3, 1];
20 goal  := [0, 0, 0];
```

Figure 2.1: The missionary and cannibals problem coded as a search problem.

Figure 2.1 shows a formalization of the missionaries and cannibals problem as a search problem. We discuss this formalization line by line.

1. Line 1 defines the auxiliary function `problem`.

   If m is the number of missionaries on a given shore, while i is the number of infidels on that same shore, then $problem(m, i)$ is `true` iff the missionaries have a problem on that shore.

2. Line 3 defines the auxiliary function `noProblemAtAll`.

   If m is the number of missionaries on the western shore and i is the number of infidels on that shore, then the expression `noProblemAtAll`(m, i) is true, if there is no problem for the missionaries on either shore.

   The implementation of this function uses the fact that if m is the number of missionaries on the western shore, then $3 - m$ is the number of missionaries on the eastern shore. Similarly, if i is the number of infidels on the western shore, then the number of infidels on the eastern shore is $3 - i$.

3. Lines 5 to 18 define the function `nextStates`. A state s is represented as a triple of the form

   $$s = [m, i, b] \quad \text{where } m \in \{0, 1, 2, 3\},\ i \in \{0, 1, 2, 3\},\ b \in \{0, 1\}.$$

   Here m is the number of missionaries on the western shore, i is the number of infidels on the western shore, and b is the number of boats on the western shore.

   (a) Line 6 extracts the components m, i, and b from the state s.

   (b) Line 7 checks whether the boat is on the western shore.

   (c) If this is the case, then the states reachable from the given state s are those states where `mb` missionaries and `ib` infidels cross the river. After `mb` missionaries and `ib` infidels have crossed the river and reached the eastern shore, `m` $-$ `mb` missionaries and `i` $-$ `ib` infidels remain on the western shore. Of course, after the crossing the boat is no longer on the western shore. Therefore, the new state has the form

   ```
   [m - mb, i - ib, 0].
   ```

   This explains line 8.

   (d) Since the number `mb` of missionaries leaving the western shore can not be greater than the number m of all missionaries on the western shore, we have the condition

   $$\texttt{mb} \in \{0, \cdots, \texttt{m}\}.$$

   There is a similar condition for the number of infidels crossing:

   $$\texttt{ib} \in \{0, \cdots, \texttt{i}\}.$$

   This explains line 9.

   (e) Furthermore, we have to check that the number of persons crossing the river is at least 1 and at most 2. This explains the condition

   $$\texttt{mb} + \texttt{ib} \in \{1, 2\}.$$

   Finally, there should be no problem in the new state on either shore. This is checked using the expression

   ```
   noProblemAtAll(m - mb, i - ib).
   ```

   These two checks are performed in line 10.

4. If the boat is on the eastern shore instead, then the missionaries and the infidels will be crossing the river from the eastern shore to the western shore. Therefore, the number of missionaries and infidels on the western shore is now increased. Hence, in this case the new state has the form

   ```
   [m + mb, i + ib, 1].
   ```

   As the number of missionaries on the eastern shore is $3 - \texttt{m}$ and the number of infidels on the eastern shore is $3 - \texttt{i}$, `mb` is now a member of the set $\{0, \cdots, 3 - \texttt{m}\}$, while `ib` is a member of the set $\{0, \cdots, 3 - \texttt{i}\}$.

5. Finally the start state and the goal state are defined in line 19 and line 20.

The code in Figure 2.1 does not define the set of states Q of the search problem. The reason is that, in order to solve the problem, we do not need to define this set. If we wanted to, we could define the set of states as follows:

```
States := { [m,i,b] : m in {0..3}, i in {0..3}, b in {0,1} | noProblemAtAll(m, i) };
```

Figure 2.2 shows a graphical representation of the transition relation of the missionaries and cannibals puzzle. In that figure, for every state both the western and the eastern shore are shown. The start state is covered with a blue ellipse, while the goal state is covered with a green ellipse. The figure clearly shows that the problem is solvable and that there is a solution involving just 11 crossings of the river. ◇

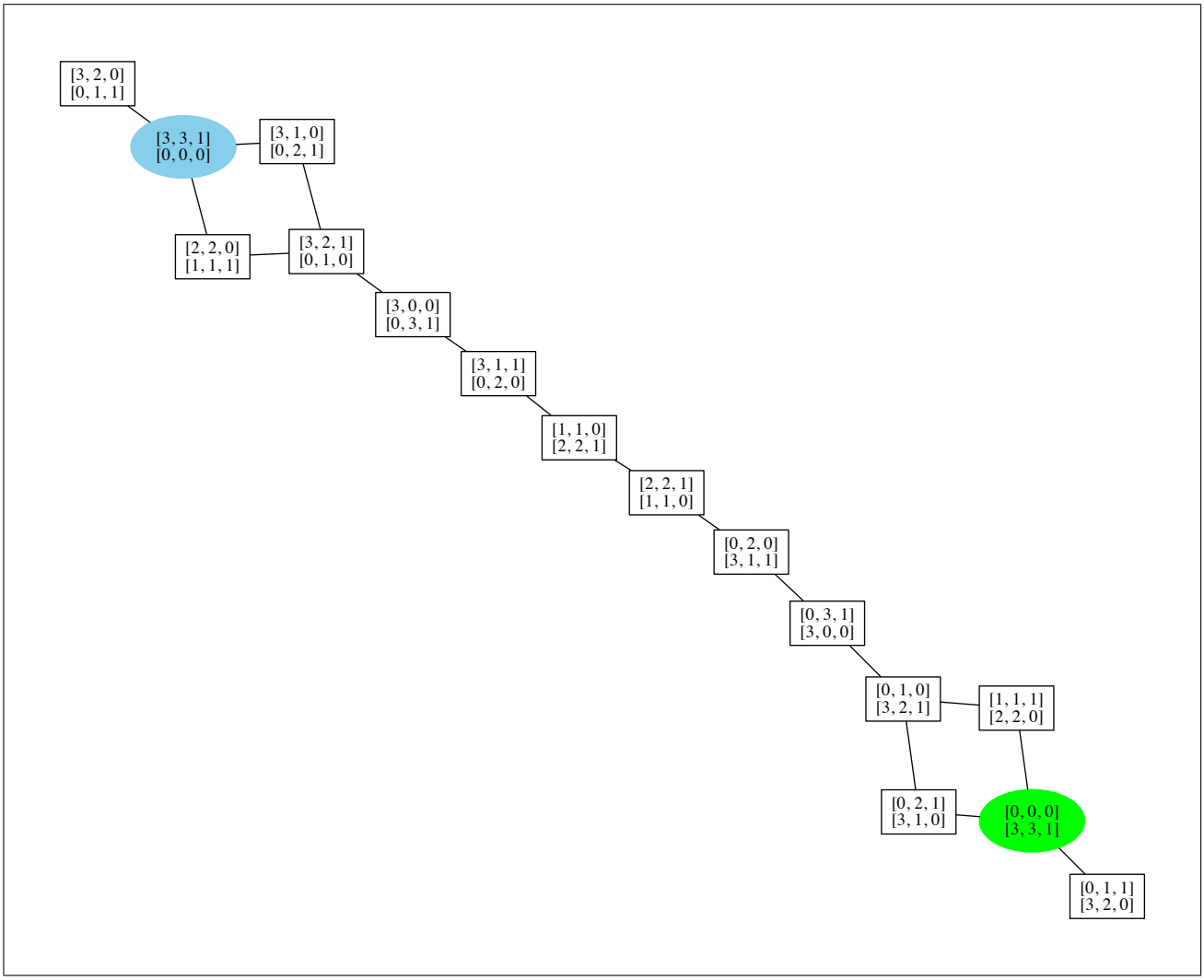

Figure 2.2: A graphical representation of the missionaries and cannibals problem.

2.1 The Sliding Puzzle

The 3×3 sliding puzzle uses a square board, where each side has a length of 3. This board is subdivided into $3 \times 3 = 9$ squares of length 1. Of these 9 squares, 8 are occupied with square tiles that are numbered from 1 to 8. One square remains empty. Figure 2.3 on page 11 shows two possible states of this sliding puzzle. The 4×4 sliding puzzle is similar to the 3×3 sliding puzzle but it is played on a square board of length 4 instead. The 4×4 sliding puzzle is also known as the 15 puzzle, while the 3×3 puzzle is called the 8 puzzle.

In order to solve the 3×3 sliding puzzle shown in Figure 2.3 we have to transform the state shown on the left of Figure 2.3 into the state shown on the right of this figure. The following operations are permitted when transforming a state of the sliding puzzle:

1. If a tile is to the left of the free square, this tile can be moved to the right.

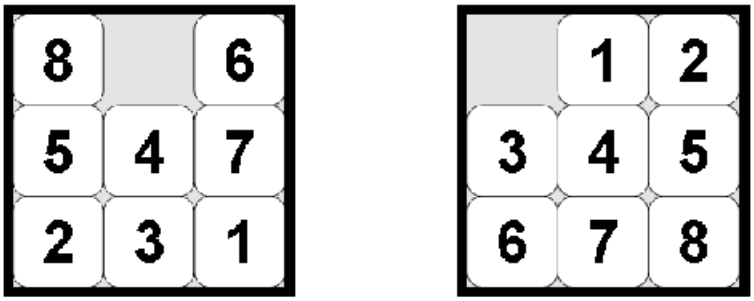

Figure 2.3: The 3×3 sliding puzzle.

2. If a tile is to the right of the free square, this tile can be moved to the left.
3. If a tile is above the free square, this tile can be moved down.
4. If a tile is below the free square, this tile can be moved up.

In order to get a feeling for the complexity of the sliding puzzle, you can check the page

http://mypuzzle.org/sliding.

The sliding puzzle is much more complex than the missionaries and cannibals problem because the state space is much larger. For the case of the 3×3 sliding puzzle, there are 9 squares that can be positioned in 9! different ways. It turns out that only half of these positions are reachable from a given start state. Therefore, the effective number of states for the 3×3 sliding puzzle is

$$9!/2 = 181,440.$$

This is already a big number, but 181,440 states can still be stored on a modern computer. However, the 4×4 sliding puzzle has

$$16!/2 = 10,461,394,944,000$$

different states reachable from a given start state. If a state is represented as matrix containing 16 numbers and we store every number using just 4 bits, we still need $16 \cdot 4 = 64$ bits or 8 bytes for every state. Hence we would need a total of

$$(16!/2) \cdot 8 = 83,691,159,552,000$$

bytes to store every state. We would thus need about 84 Terabytes to store the set of all states. As few computers are equipped with this kind of memory, it is obvious that we won't be able to store the entire state space in memory.

Figure 2.4 shows how the 3×3 sliding puzzle can be formulated as a search problem. We discuss this program line by line.

1. `findTile` is an auxiliary procedure that takes a `number` and a `State` and returns the row and column where the tile labelled with `number` can be found.

   Here, a state is represented as a list of lists. For example, the states shown in Figure 2.3 are represented as shown in line 26 and line 30. The empty tile is coded as 0.

2. `moveDir` takes a `State`, the `row` and the `column` where to find the empty square and a direction in which the empty square should be moved. This direction is specified via the two variables `dx` and `dy`. The tile at the position $\langle$`row` + `dx`, `col` + `dy`$\rangle$ is moved into the position $\langle$`row`, `col`$\rangle$, while the tile at position $\langle$`row` + `dx`, `col` + `dy`$\rangle$ becomes empty.

3. Given a `State`, the procedure `nextStates` computes the set of all states that can be reached in one step from `State`. The basic idea is to find the position of the empty tile and then try to move the empty

```
findTile := procedure(number, State) {
    n := #State;
    L := [1 .. n];
    for (row in L, col in L | State[row][col] == number) {
        return [row, col];
    }
};
moveDir := procedure(State, row, col, dx, dy) {
    State[row     ][col     ] := State[row + dx][col + dy];
    State[row + dx][col + dy] := 0;
    return State;
};
nextStates := procedure(State) {
    n          := #State;
    [row, col] := findTile(0, State);
    newStates  := [];
    directions := [ [1, 0], [-1, 0], [0, 1], [0, -1] ];
    L          := [1 .. n];
    for ([dx, dy] in directions) {
        if (row + dx in L && col + dy in L) {
            newStates += [ moveDir(State, row, col, dx, dy) ];
        }
    }
    return newStates;
};
start := [ [8, 0, 6],
           [5, 4, 7],
           [2, 3, 1]
         ];
goal := [ [0, 1, 2],
          [3, 4, 5],
          [6, 7, 8]
        ];
```

Figure 2.4: The 3×3 sliding puzzle.

tile in all possible directions. If the empty tile is found at position $[\texttt{row}, \texttt{col}]$ and the direction of the movement is given as $[\texttt{dx}, \texttt{dy}]$, then in order to ensure that the empty tile can be moved to the position $[\texttt{row} + \texttt{dx}, \texttt{col} + \texttt{dy}]$, we have to ensure that both

$$\texttt{row} + \texttt{dx} \in \{1, \cdots, n\} \quad \text{and} \quad \texttt{col} + \texttt{dy} \in \{1, \cdots, n\}$$

hold, where n is the size of the board.

Next, we want to develop an algorithm that can solve puzzles of the kind described so far. The most basic algorithm to solve search problems is breadth first search. We discuss this algorithm next.

2.2 Breadth First Search

Informally, breadth first search, abbreviated as Bfs, works as follows:

1. Given a search problems $\langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal} \rangle$, we initialize a set `Frontier` to contain the state `start`.

In general, `Frontier` contains those states that have just been discovered and whose successors have not yet been seen.

2. As long as the set `Frontier` does not contain the state `goal`, we recompute this set by adding all states to it that can be reached in one step from a state in `Frontier`. Then, the states that had been previously present in `Frontier` are removed. These old states are then saved into a set `Visited`.

In order to avoid loops, an implementation of breadth first search keeps track of those states that have been visited. These states are collected in a set `Visited`. Once a state has been added to the set `Visited`, it will never be revisited again. Furthermore, in order to keep track of the path leading to the goal, we utilize a dictionary called `Parent`. For every state s that is in `Frontier`, `Parent`$[s]$ is the state that caused s to be added to the set `Frontier`, i.e. for all states $s \in$ `Visited` $\cup$ `Frontier` we have

$$s \in \texttt{nextStates}(\texttt{Parent}[s]).$$

```
1  search := procedure(start, goal, nextStates) {
2      Frontier := { start };
3      Visited  := {}; // set of nodes that have been expanded
4      Parent   := {};
5      while (Frontier != {}) {
6          NewFrontier := {};
7          for ( s in Frontier, ns in nextStates(s)
8              | !(ns in Visited) && !(ns in Frontier)
9              )
10         {
11             NewFrontier += { ns };
12             Parent[ns]  := s;
13             if (ns == goal) {
14                 return pathTo(goal, Parent);
15             }
16         }
17         Visited  += Frontier;
18         Frontier := NewFrontier;
19     }
20 };
```

Figure 2.5: Breadth first search.

Figure 2.5 on page 13 shows an implementation of breadth first search in SetlX. We discuss this implementation line by line:

1. `Frontier` is the set of all those states that have been encountered but whose neighbours have not yet been explored. Initially, it contains the state `start`.

2. `Visited` is the set of all those states, all whose neighbours have already been added to the set `Frontier`. In order to avoid infinite loops, these states must not be visited again.

3. `Parent` is a dictionary keeping track of the state leading to a given state.

4. As long as the set `Frontier` is not empty, we add all neighbours of states in `Frontier` that have not yet been visited to the set `NewFrontier`. When doing this, we keep track of the path leading to a new state `ns` by storing its parent in the dictionary `Parent`.

5. If the new state happens to be the state `goal`, we return a path leading from `start` to `goal`. The procedure `pathTo`() is shown in Figure 2.6 on page 14.

6. After we have collected all successors of states in `Frontier`, the states in the set `Frontier` have been visited and are therefore added to the set `Visited`, while the `Frontier` is updated to `NewFrontier`.

```
pathTo := procedure(state, Parent) {
    Path := [];
    while (state != om) {
        Path  += [state];
        state := Parent[state];
    }
    return reverse(Path);
};
```

Figure 2.6: The procedure `pathTo`().

The procedure call `pathTo`(`state`, `Parent`) constructs a path reaching from `start` to `state` in reverse by looking up the parent states.

If we try breadth first search to solve the missionaries and cannibals problem, we immediately get the solution shown in Figure 2.7. 15 nodes had to be expanded to find this solution. To keep this in perspective, we note that Figure 2.2 shows that the entire state space contains 16 states. Therefore, with the exception of one state, we have inspected all the states. This is a typical behaviour for breadth first search.

```
MMM   KKK   B      |~~~~~|
                   >  KK >
MMM   K            |~~~~~|              KK  B
                   <  K  <
MMM   KK    B      |~~~~~|               K
                   >  KK >
MMM                |~~~~~|             KKK  B
                   <  K  <
MMM   K     B      |~~~~~|              KK
                   > MM  >
M     K            |~~~~~|        MM    KK  B
                   < M K <
MM    KK    B      |~~~~~|         M     K
                   > MM  >
      KK           |~~~~~|       MMM     K  B
                   <  K  <
      KKK   B      |~~~~~|       MMM
                   >  KK >
      K            |~~~~~|       MMM    KK  B
                   <  K  <
      KK    B      |~~~~~|       MMM     K
                   >  KK >
                   |~~~~~|       MMM   KKK  B
```

Figure 2.7: A solution of the missionaries and cannibals problem.

Next, let us try to solve the 3×3 sliding puzzle. It takes about 6 seconds to solve this problem on my computer[1], while 181439 states are touched. Again, we see that breadth first search touches nearly all the

[1] I happen to own an iMac from 2017. This iMac is equipped with 32 Gigabytes of main memory and a quad core 3.4 GHz "Intel Core i5" processor. I suspect this to be the I5-7500 (Kaby Lake) processor.

states reachable from the start state.

2.2.1 A Queue Based Implementation of Breadth First Search

In the literature, for example in Figure 3.11 of Russell & Norvig [RN09], breadth first search is often implemented using a queue data structure. Figure 2.8 on page 15 shows an implementation of breadth first search that uses a queue to store the set `Frontier`. However, when we run this version, it turns out that the solution of the 3×3 sliding puzzle needs about 58 seconds, which is a lot slower than our set based implementation that has been presented in Figure 2.5.

```
1  search := procedure(start, goal, nextStates) {
2      Queue   := [ start ];
3      Visited := {};
4      Parent  := {};
5      while (Queue != []) {
6          state := Queue[1];
7          Queue := Queue[2..];
8          if (state == goal) {
9              return pathTo(state, Parent);
10         }
11         Visited   += { state };
12         newStates := nextStates(state);
13         for (ns in newStates | !(ns in Visited) && Parent[ns] == om) {
14             Parent[ns] := state;
15             Queue      += [ ns ];
16         }
17     }
18 };
```

Figure 2.8: A queue based implementation of breadth first search.

The solution of the 3×3 sliding puzzle that is found by breadth first search is shown in Figure 2.9 and Figure 2.10.

We conclude our discussion of breadth first search by noting the two most important properties of breadth first search.

1. Breadth first search is complete: If there is a solution to the given search problem, then breadth first search is going to find it.

2. The solution found by breadth first search is optimal, i.e. it is one of the shortest possible solutions.

Proof: Both of these claims can be shown simultaneously. Consider the implementation of breadth first search shown in Figure 2.5. An easy induction on the number of iterations of the `while` loop shows that after n iterations of the `while` loop, the set `Frontier` contains exactly those states that have a distance of n to the state `start`. This claim is obviously true before the first iteration of the while loop as in this case, `Frontier` only contains the state `start`. In the induction step we assume the claim is true after n iterations. Then, in the next iteration all states that can be reached in one step from a state in `Frontier` are added to the new `Frontier`, provided there is no shorter path to these states. There is a shorter path to these states if these states are already a member of the set `Visited`. Hence, the claim is true after $n + 1$ iterations also.

Now, if there is a path from `start` to `goal`, there must also be a shortest path. Assume this path has a length of k. Then, `goal` is reached in the iteration number k and the shortest path is returned. □

The fact that breadth first search is both complete and the path returned is optimal is rather satisfying. However, breadth first search still has a big downside that makes it unusable for many problems: If the `goal` is

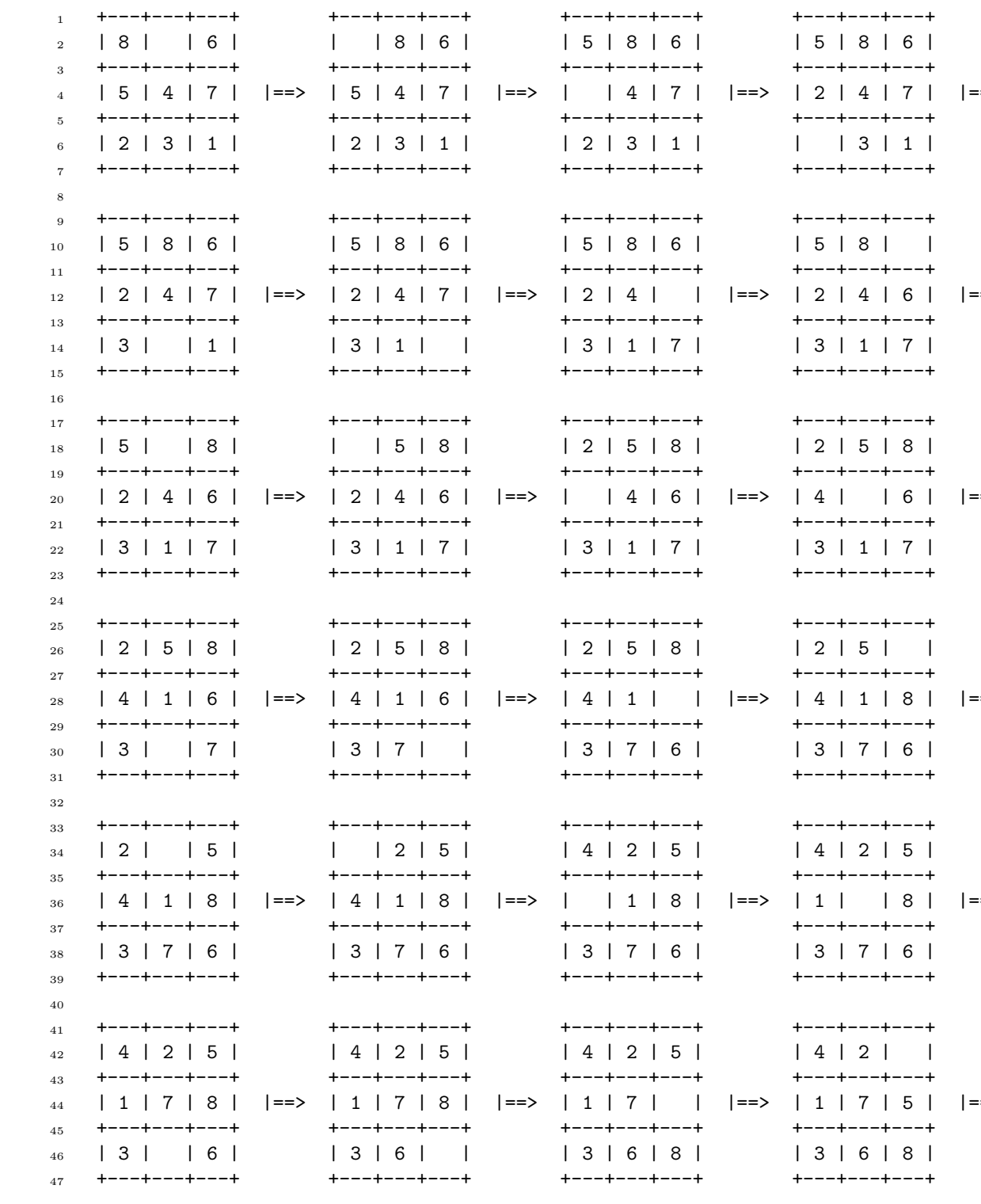

Figure 2.9: The first 24 steps in the solution of the 3×3 sliding puzzle.

far from the `start`, breadth first search will use a lot of memory because it will store a large part of the state space in the set `Visited`. In many cases, the state space is so big that this is not possible. For example, it is

```
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 4 |   | 2 |       |   | 4 | 2 |       | 1 | 4 | 2 |       | 1 | 4 | 2 |
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 1 | 7 | 5 | |==>  | 1 | 7 | 5 | |==>  |   | 7 | 5 | |==>  | 3 | 7 | 5 | |==>
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 3 | 6 | 8 |       | 3 | 6 | 8 |       | 3 | 6 | 8 |       |   | 6 | 8 |
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+

+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 1 | 4 | 2 |       | 1 | 4 | 2 |       | 1 |   | 2 |       |   | 1 | 2 |
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 3 | 7 | 5 | |==>  | 3 |   | 5 | |==>  | 3 | 4 | 5 | |==>  | 3 | 4 | 5 |
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
| 6 |   | 8 |       | 6 | 7 | 8 |       | 6 | 7 | 8 |       | 6 | 7 | 8 |
+---+---+---+       +---+---+---+       +---+---+---+       +---+---+---+
```

Figure 2.10: The last 7 steps in the solution of the 3×3 sliding puzzle.

impossible to solve the more interesting cases of the 4×4 sliding puzzle using breadth first search.

2.3 Depth First Search

To overcome the memory limitations of breadth first search, the depth first search algorithm has been developed. The basic idea is to replace the queue of Figure 2.8 by a stack. The resulting algorithm is shown in Figure 2.11 on page 17. Basically, in this implementation, a path is searched to its end before trying an alternative. This way, we might be able to find a `goal` that is far away from `start` without exploring the whole state space.

```
search := procedure(start, goal, nextStates) {
    Stack  := [ start ];
    Parent := {};
    while (Stack != []) {
        state := Stack[-1];
        Stack := Stack[..-2];
        if (state == goal) {
            return pathTo(state, Parent);
        }
        newStates := nextStates(state);
        for (ns in newStates | ns != start && Parent[ns] == om) {
            Parent[ns] := state;
            Stack      += [ns];
        }
    }
};
```

Figure 2.11: The depth first search algorithm.

Actually, it is not necessary to understand the details of the implementation shown in Figure 2.11 on page 17. The reason is that the recursive implementation of depth first search that is presented in the following subsection is superior to the implementation shown in Figure 2.11 on page 17. When we test the implementation shown above with the 3×3 sliding puzzle, it takes about 68 seconds to find a solution. The solution that is found has

a length of 41,553 steps. As the shortest path from `start` to `goal` has 31 steps, the solution found by depth first search is very far from optimal. All this is rather disappointing news. The only good news is that there is no longer a need to keep the set `Visited` around. However, we still have to maintain the set `Parent`. If we were more ambitious, we could eliminate the use of this dictionary also, but the resulting implementation would be rather unwieldy. Fortunately, we will be able to get rid of the set `Parent` with next to no effort when we develop a recursive implementation of depth first search in the following subsection.

2.3.1 Getting Rid of the `Parent` Dictionary

It can be argued that the implementation of depth first search discussed previously is not really depth first search because it uses the dictionary `Parent`. As states are only added to `Parent` and never removed, at the end of the search this dictionary will contain all states that have been visited. This defeats the most important advantage of depth first search which is the fact that it should only store the current path that is investigated. Therefore, it has been suggested (for example compare Russel and Norvig [RN09]) that instead of storing single states, the stack should store the full paths leading to these states. This leads to the implementation shown in Figure 2.12 on page 18.

```
1  search := procedure(start, goal, nextStates) {
2      Stack  := [ [start] ];
3      while (Stack != []) {
4          Path  := Stack[-1];
5          Stack := Stack[..-2];
6          state := Path[-1];
7          if (state == goal) {
8              return Path;
9          }
10         newStates := nextStates(state);
11         for (ns in newStates | !(ns in Path)) {
12             Stack += [ Path + [ns] ];
13         }
14     }
15 };
```

Figure 2.12: An path-based implementation of depth first search.

Unfortunately, it turns out that the paths get very long and hence need a lot of memory to be stored and this fact defeats the main idea of this implementation. As a result, the procedure `search` that is given in Figure 2.12 on page 18 is not able to solve the instance of the 3×3 sliding puzzle that was shown in Figure 2.3 on page 11.

Exercise 1: Assume that n is a positive natural number and that the set of states Q_n is defined as

$$Q_n := \big\{\langle a, b\rangle \mid a \in \{0, \cdots, n\} \wedge b \in \{0, \cdots, n\}\big\}.$$

Furthermore, the states `start` and `goal` are defined as

$$\texttt{start} := \langle 0, 0\rangle \quad \text{and} \quad \texttt{goal} := \langle n, n\rangle.$$

Next, the function `nextStates` is defined as

$$\texttt{nextStates}\big(\langle a, b\rangle\big) := \begin{cases} \big\{\langle a+1, b\rangle, \langle a, b+1\rangle\big\} & \text{if } a < n \text{ and } b < n \\ \big\{\langle a+1, b\rangle, \langle a+1, 0\rangle\big\} & \text{if } a < n \text{ and } b = n \\ \big\{\langle 0, b+1\rangle, \langle a, b+1\rangle\big\} & \text{if } a = n \text{ and } b < n \\ \{\} & \text{if } a = n \text{ and } b = n \end{cases}$$

Finally, the search problem $\mathcal{P}$ is defined as

$\mathcal{P} := \langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal} \rangle$.

Assume that states can be stored using 8 bytes. Furthermore, assume that we use the algorithm given in Figure 2.12 on page 18 to solve the search problem $\mathcal{P}$. Finally, assume that the path found by depth first search is the longest path leading from `start` to `goal`.

How many bytes do we need to store all the states on the stack in the moment that the `goal` is reached? How big is this number if $n = 100$?

Note that the question only asks for the memory needed to store the states. The memory needed to store the stack itself and the various lists on the stack is on top of the memory needed to store the states. However, to answer this exercise correctly, you should ignore this kind of memory. ◇

2.3.2 A Recursive Implementation of Depth First Search

Sometimes, the depth first search algorithm is presented as a recursive algorithm, since this leads to an implementation that is slightly shorter and is easier to understand. What is more, we no longer need the dictionary `Parent` to record the parent of each node. The resulting implementation is shown in Figure 2.13 on page 19.

```
1  search := procedure(start, goal, nextStates) {
2      return dfs(start, goal, nextStates, [start]);
3  };
4  dfs := procedure(state, goal, nextStates, Path) {
5      if (state == goal) {
6          return Path;
7      }
8      newStates := nextStates(state);
9      for (ns in newStates | !(ns in Path)) {
10         result := dfs(ns, goal, nextStates, Path + [ns]);
11         if (result != om) {
12             return result;
13         }
14     }
15 };
```

Figure 2.13: A recursive implementation of depth first search.

The only purpose of the procedure `search` is to call the procedure `dfs`, which needs one additional argument. This argument is called `Path`. The idea is that `Path` is a path leading from the state `start` to the current `state` that is the first argument of the procedure `dfs`. On the first invocation of `dfs`, the parameter `state` is equal to `start` and therefore `Path` is initialized as the list containing only `start`.

The implementation of `dfs` works as follows:

1. If `state` is equal to `goal`, our search is successful. Since by assumption the list `Path` is a path connecting `start` and `state` and we have checked that `state` is equal to `goal`, we can return `Path` as our solution.

2. Otherwise, `newStates` is the set of states that are reachable from `state` in one step. Any of the states `ns` in this set could be the next state on a path that leads to `goal`. Therefore, we try recursively to reach `goal` from every state `ns`. Note that we have to change `Path` to the list

   ```
   Path + [ns]
   ```

   when we call the procedure `dfs` recursively. This way, we retain the invariant of `dfs` that the list `Path` is a path connecting `start` with `state`.

3. We still have to avoid running in circles. In the recursive version of depth first search, this is achieved by checking that the state `ns` is not already a member of the list `Path`. In the non-recursive version of depth

first search, we had used the set `Parent` instead. The current implementation no longer has a need for the dictionary `Parent`. This is very fortunate since it reduces the memory requirements of depth first search considerably.

4. If one of the recursive calls of `dfs` returns a list, this list is a solution to our search problem and hence it is returned. However, if instead the undefined value `om` is returned, the `for` loop needs to carry on and test the other successors of `state`.

5. Note that the recursive invocation of `dfs` returns `om` if the end of the `for` loop is reached and no solution has been returned so far. The reason is that there is no `return` statement at the end of the procedure `dfs`. Hence, if the last line of the procedure `dfs` is reached, `om` is returned by default.

For the 3×3 puzzle, it takes about one second to compute the solution. In this case, the length of the solution is still 3653 steps, which is unsatisfying. The good news is that this program does not need much memory. The only variable that uses considerable memory is the variable `Path`. If we can somehow keep the list `Path` short, then the recursive version of depth first search uses only a tiny fraction of the memory needed by breadth first search.

2.4 Iterative Deepening

The fact that the recursive version of depth first search took just one second to find a solution is very impressive. The questions is whether it might be possible to force depth first search to find the shortest solution. The answer to this question leads to an algorithm that is known as iterative deepening. The main idea behind iterative deepening is to run depth first with a depth limit d. This limit enforces that a solution has at most a length of d. If no solution is found at a depth of d, the new depth $d + 1$ can be tried next and the process can be continued until a solution is found. The program shown in Figure 2.14 on page 20 implements this strategy. We proceed to discuss the details of this program.

```
1  search := procedure(start, goal, nextStates) {
2      limit := 1;
3      while (true) {
4          Path := depthLimitedSearch(start, goal, nextStates, limit);
5          if (Path != om) { return Path; }
6          limit += 1;
7      }
8  };
9  depthLimitedSearch := procedure(start, goal, nextStates, limit) {
10     Stack := [ [start] ];
11     while (Stack != []) {
12         Path  := Stack[-1];
13         Stack := Stack[..-2];
14         state := Path[-1];
15         if (state == goal)  { return Path; }
16         if (#Path >= limit) { continue;    }
17         for (ns in nextStates(state) | !(ns in Path)) {
18             Stack += [ Path + [ns] ];
19         }
20     }
21 };
```

Figure 2.14: Iterative deepening implemented in SetlX.

1. The procedure `search` initializes the variable `limit` to 1 and tries to find a solution to the search problem that has a length that is less than or equal to `limit`. If a solution is found, it is returned. Otherwise, the variable `limit` is incremented by one and a new instance of depth first search is started. This process continues until either
   - a solution is found or
   - the sun rises in the west.
2. The procedure `depthLimitedSearch` implements depth first search but takes care to compute only those paths that have a length of at most `limit`. The implementation shown in Figure 2.14 is stack based. In this implementation, the stack contains paths leading from `start` to the state at the end of a given path. Hence it is similar to the implementation of depth first search shown in Figure 2.11 on page 17.
3. The stack is initialized to contain the path [`start`].
4. In the `while`-loop, the first thing that happens is that the `Path` on top of the stack is removed from the stack. The state at the end of this `Path` is called `state`. If this `state` happens to be the `goal`, a solution to the search problem has been found and this solution is returned.
5. Otherwise, we check the length of `Path`. If this length is greater than or equal to the `limit`, the `Path` can be discarded as we have already checked that it does not end in the `goal`.
6. Otherwise, the neighbours of `state` are computed. For every neighbour `ns` of `state` that has not yet been encountered in `Path`, we extend `Path` to a new list that ends in `ns`.
7. This process is iterated until the `Stack` is exhausted.

The nice thing about the program presented in this section is the fact that it does not use much memory. The reason is that the stack can never have a size that is longer than `limit` and therefore the overall memory that is needed can be bounded by $\mathcal{O}(\texttt{limit}^2)$. However, when we run this program to solve the 3×3 sliding puzzle, the algorithm takes about 25 minutes. There are two reasons for this:

1. First, it is quite wasteful to run the search for a depth limit of 1, 2, 3, $\cdots$ all the way up to 31. Essentially, all the computations done with a limit less than 31 are wasted.
2. Given a state s that is reachable from the `start`, there often are a huge number of different paths that lead from `start` to s. The version of iterative deepening presented in this section tries all of these paths and hence needs a large amount of time.

Exercise 2: Assume the set of states Q is defined as

$$Q := \big\{\langle a, b\rangle \mid a \in \mathbb{N} \wedge b \in \mathbb{N}\big\}.$$

Furthermore, the states `start` and `goal` are defined as

$$\texttt{start} := \langle 0, 0\rangle \quad \text{and} \quad \texttt{goal} := \langle n, n\rangle \text{ where } n \in \mathbb{N}.$$

Next, the function `nextStates` is defined as

$$\texttt{nextStates}\big(\langle a, b\rangle\big) := \big\{\langle a+1, b\rangle, \langle a, b+1\rangle\big\}.$$

Finally, the search problem $\mathcal{P}$ is defined as

$$\mathcal{P} := \langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal}\rangle.$$

Given a natural number n, compute the number of different solutions of this search problem and prove your claim.

Hint: The expression giving the number of different solutions contains factorials. In order to get a better feeling for the asymptotic growth of this expression we can use Stirling's approximation of the factorial. Stirling's approximation of $n!$ is given as follows:

$$n! \sim \sqrt{2 \cdot \pi \cdot n} \cdot \left(\frac{n}{e}\right)^n. \qquad \diamond$$

Exercise 3: If there is no solution, the implementation of iterative deepening that is shown in Figure 2.14 does not terminate. The reason is that the function `depthLimitedSearch` does not distinguish between the following two reasons for failure:

1. It can fail to find a solution because the depth limit is reached.
2. It can also fail because it has tried all paths without hitting the depth limit but the `Stack` is exhausted.

Improve the implementation of iterative deepening so that it will always terminate eventually, provided the state space is finite. $\diamond$

2.4.1 A Recursive Implementation of Iterative Deepening

If we implement iterative deepening recursively, then we know that the call stack is bounded by the length of the shortest solution. Figure 2.15 on page 22 shows a recursive implementation of iterative deepening. Unfortunately, the running time of the recursive implementation of iterative deepening is still quite big: On my computer, the recursive implementation takes about 23 minutes.

```
1  search := procedure(start, goal, nextStates) {
2      limit := 1;
3      while (true) {
4          result := dfsLimited(start, goal, nextStates, [start], limit);
5          if (result != om) {
6              return result;
7          }
8          limit += 1;
9      }
10 };
11 dfsLimited := procedure(state, goal, nextStates, Path, limit) {
12     if (state == goal) {
13         return Path;
14     }
15     if (limit == 0) {
16         return;  // limit execceded
17     }
18     for (ns in nextStates(state) | !(ns in Path)) {
19         result := dfsLimited(ns, goal, nextStates, Path + [ns], limit - 1);
20         if (result != om) {
21             return result;
22         }
23     }
24 };
```

Figure 2.15: A recursive implementation of iterative deepening.

2.5 Bidirectional Breadth First Search

Breadth first search first visits all states that have a distance of 1 from start, then all states that have a distance of 2, then of 3 and so on until finally the goal is found. If the length of the shortest path from `start` to `goal` is d, then all states that have a distance of at most d will be visited. In many search problems, the number of

states grows exponentially with the distance, i.e. there is a branching factor b such that the set of all states that have a distance of at most d from `start` is roughly

$$1 + b + b^2 + b^3 + \cdots + b^d = \frac{b^{d+1} - 1}{b - 1} = \mathcal{O}(b^d).$$

At least this is true in the beginning of the search. As the size of the memory that is needed is the most constraining factor when searching, it is important to cut down this size. One simple idea is to start searching both from the node `start` and the node `goal` simultaneously. The justification is that we can hope that the path starting from `start` and the path starting from `goal` will meet in the middle and hence they will both have a size of approximately $d/2$. If this is the case, only

$$2 \cdot \frac{b^{d/2+1} - 1}{b - 1}$$

nodes need to be explored and even for modest values of b this number is much smaller than

$$\frac{b^{d+1} - 1}{b - 1}$$

which is the number of nodes expanded in breadth first search. For example, assume that the branching factor $b = 2$ and that the length of the shortest path leading from `start` to `goal` is 40. Then we need to explore

$$2^{41} - 1 = 2,199,023,255,551$$

in breadth first search, while we only have to explore

$$2 \cdot \left(2^{40/2+1} - 1\right) = 4,194,302$$

with bidirectional breadth first search. While it is certainly feasible to keep four million states in memory, keeping a two trillion states in memory is impossible on most devices.

Figure 2.16 on page 24 shows the implementation of bidirectional breadth first search. Essentially, we have to copy the breadth first program shown in Figure 2.5. Let us discuss the details of the implementation.

1. The variable `FrontierA` is the frontier that starts from the state `start`, while `FrontierB` is the frontier that starts from the state `goal`.

2. `VisitedA` is the set of states that have been visited starting from `start`, while `VisitedB` is the set of states that have been visited starting from `goal`.

3. For every state s that is in `FrontierA`, `ParentA`$[s]$ is the state that caused s to be added to the set `FrontierA`. Similarly, for every state s that is in `FrontierB`, `ParentB`$[s]$ is the state that caused s to be added to the set `FrontierB`.

4. The bidirectional search keeps running for as long as both sets `FrontierA` and `FrontierB` are non-empty and a path has not yet been found.

5. Initially, the `while` loop adds the frontier sets to the visited sets as all the neighbours of the frontier sets will now be explored.

6. Then the `while` loop computes those states that can be reached from `FrontierA` and have not been visited from `start`. If a state `ns` is a neighbour of a state `s` from the set `FrontierA` and the state `ns` has already been encountered during the search that started from `goal`, then a path leading from `start` to `goal` has been found and this path is returned. The function `combinePaths` that computes this path by combining the path that leads from `start` to `ns` and then from `ns` to `goal` to is shown in Figure 2.17 on page 24.

7. Next, the same computation is done with the role of the states `start` and `goal` exchanged.

On my computer, bidirectional breadth first search solves the 3×3 sliding puzzle in less than a second! However, bidirectional breadth first search is still not able to solve the 4×4 sliding puzzle since the portion of the search space that needs to be computed is just too big to fit into memory.

```
search := procedure(start, goal, nextStates) {
    FrontierA := { start };
    VisitedA  := {}; // set of nodes expanded starting from start
    ParentA   := {};
    FrontierB := { goal };
    VisitedB  := {}; // set of nodes expanded starting from goal
    ParentB   := {};
    while (FrontierA != {} && FrontierB != {}) {
        VisitedA += FrontierA;
        VisitedB += FrontierB;
        NewFrontier := {};
        for (s in FrontierA, ns in nextStates(s) | !(ns in VisitedA)) {
            NewFrontier += { ns };
            ParentA[ns] := s;
            if (ns in VisitedB) {
                return combinePaths(ns, ParentA, ParentB);
            }
        }
        FrontierA   := NewFrontier;
        NewFrontier := {};
        for (s in FrontierB, ns in nextStates(s) | !(ns in VisitedB)) {
            NewFrontier += { ns };
            ParentB[ns] := s;
            if (ns in VisitedA) {
                return combinePaths(ns, ParentA, ParentB);
            }
        }
        FrontierB := NewFrontier;
    }
};
```

Figure 2.16: Bidirectional breadth first search.

```
combinePaths := procedure(node, ParentA, ParentB) {
    Path1 := pathTo(node, ParentA);
    Path2 := pathTo(node, ParentB);
    return Path1[..-2] + reverse(Path2);
};
```

Figure 2.17: Combining two paths.

2.6 Best First Search

Up to now, all the search algorithms we have discussed were essentially blind. Given a state s and all of its neighbours, they had no idea which of the neighbours they should pick because they had no conception which of these neighbours might be more promising than the other neighbours. If a human tries to solve a problem, she usually will develop a feeling that certain states are more favourable than other states because they seem to be closer to the solution. In order to formalise this procedure, we next define the notion of a heuristic.

Definition 2 (Heuristic) Given a search problem

$$\mathcal{P} = \langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal} \rangle,$$

a heuristic is a function

$$h : Q \to \mathbb{R}$$

that computes an approximation of the distance of a given state s to the goal state goal. The heuristic is admissible if it never overestimates the true distance, i.e. if the function

$$d : Q \to \mathbb{R}$$

computes the true distance from a state s to the goal, then we must have

$$h(s) \leq d(s) \quad \text{for all } s \in Q.$$

Hence, the heuristic is admissible iff it is optimistic: It must never overestimate the distance to the goal, but it is free to underestimate this distance.

Finally, the heuristic h is called consistent iff we have

$$h(\texttt{goal}) = 0 \quad \text{and} \quad h(s_1) \leq 1 + h(s_2) \quad \text{for all } s_2 \in \texttt{nextStates}(s_1). \qquad \diamond$$

Let us explain the idea behind the notion of consistency. First, if we are already at the goal, the heuristic should notice this and therefore we need to have $h(\texttt{goal}) = 0$. Secondly, assume we are at the state s_1 and s_2 is a neighbour of s_1, i.e. we have that

$$s_2 \in \texttt{nextStates}(s_1).$$

Now if our heuristic h assumes that the distance of s_2 from the goal is $h(s_2)$, then the distance of s_1 from the goal can be at most $1 + h(s_2)$ because starting from s_1 we can first go to s_2 in one step and then from s_2 to goal in $h(s_2)$ steps for a total of $1 + h(s_2)$ steps. Of course, it is possible that there exists a shorter path from s_1 leading to the goal than the one that visits s_2 first. Hence, we have the inequality

$$h(s_1) \leq 1 + h(s_2).$$

Theorem 3 Every consistent heuristic is an admissible heuristic.

Proof: Assume that the heuristic h is consistent. Assume further that $s \in Q$ is some state such that there is a path p from s to the goal. Assume this path has the form

$$p = [s_n, s_{n-1}, \cdots, s_1, s_0], \quad \text{where } s_n = s \text{ and } s_0 = \texttt{goal}.$$

Then the length of the path p is n and we have to show that $h(s) \leq n$. In order to prove this claim, we show that we have

$$h(s_k) \leq k \quad \text{for all } k \in \{0, 1, \cdots, n\}.$$

This claim is shown by induction on k.

B.C.: $k = 0$.

We have $h(s_0) = h(\texttt{goal}) = 0 \leq 0$ because the fact that h is consistent implies $h(\texttt{goal}) = 0$.

I.S.: $k \mapsto k + 1$.

We have to show that $h(s_{k+1}) \leq k + 1$ holds. This is shown as follows:

$$\begin{array}{rcll} h(s_{k+1}) & \leq & 1 + h(s_k) & \text{because } s_k \in \texttt{nextStates}(s_{k+1}) \text{ and } h \text{ is consistent,} \\ & \leq & 1 + k & \text{because } h(s_k) \leq k \text{ by induction hypotheses.} \end{array}$$

□

It is natural to ask whether the last theorem can be reversed, i.e. whether every admissible heuristic is also consistent. The answer to this question is negative since there are some *contorted* heuristics that are admissible but that fail to be consistent. However, in practice it turns out that most admissible heuristics are

also consistent. Therefore, when we construct consistent heuristics later, we will start with admissible heuristics, since these are easy to find. We will then have to check that these heuristics are also consistent.

Examples: In the following, we will discuss several heuristics for the sliding puzzle.

1. The simplest heuristic that is admissible is the function $h(s) := 0$. Since we have

   $$0 \leq 1 + 0,$$

   this heuristic is obviously consistent, but this heuristic is too trivial to be of any use.

2. The next heuristic is the number of misplaced tiles heuristic. For a state s, this heuristic counts the number of tiles in s that are not in their final position, i.e. that are not in the same position as the corresponding tile in `goal`. For example, in Figure 2.3 on page 11 in the state depicted to the left, only the tile with the label 4 is in the same position as in the state depicted to the right. Hence, there are 7 misplaced tiles.

   As every misplaced tile must be moved at least once and every step in the sliding puzzle moves at most one tile, it is obvious that this heuristic is admissible. It is also consistent. First, the `goal` has no misplaced tiles, hence its heuristic is 0. Second, in every step of the sliding puzzle only one tile is moved. Therefore the number of misplaced tiles in two neighbouring state can differ by at most one.

   Unfortunately, the number of misplaced tiles heuristic is very crude and therefore not particularly useful.

3. The Manhattan heuristic improves on the previous heuristic. For two points $\langle x_1, y_1\rangle, \langle x_2, y_2\rangle \in \mathbb{R}^2$ the Manhattan distance of these points is defined as

   $$d_1\big(\langle x_1, y_1\rangle, \langle x_2, y_2\rangle\big) := |x_1 - x_2| + |y_1 - y_2|.$$

   The Manhattan distance is also called the L_1 norm of the difference vector $\langle x_2 - x_1, y_2 - y_1\rangle$. If we associate Cartesian coordinates with the tiles of the sliding puzzle such that the tile in the upper left corner has coordinates $\langle 1, 1\rangle$ and the coordinates of the tile in the lower right corner is $\langle 3, 3\rangle$, then the Manhattan distance of two positions measures how many steps it takes to move a tile from the first position to the second position if we are allowed to move the tile horizontally or vertically regardless of the fact that the intermediate positions might be blocked by other tiles. To compute the Manhattan heuristic for a state s with respect to the `goal`, we first define the function $\texttt{pos}(t, s)$ for all tiles $t \in \{1, \cdots, 8\}$ in a given state s as follows:

   $$\texttt{pos}(t, s) = \langle \texttt{row}, \texttt{col}\rangle \stackrel{\text{def}}{\iff} s[\texttt{row}][\texttt{col}] = t,$$

   i.e. given a state s, the expression $\texttt{pos}(t, s)$ computes the Cartesian coordinates of the tile t with respect to the state s. Then we can define the Manhattan heuristic h for the 3×3 puzzle as follows:

   $$h(s) := \sum_{t=1}^{8} d_1\big(\texttt{pos}(t, s), \texttt{pos}(t, \texttt{goal})\big).$$

   The Manhattan heuristic measure the number of moves that would be needed if we wanted to put every tile of s into its final positions and if we were allowed to slide tiles over each other. Figure 2.18 on page 27 shows how the Manhattan distance can be computed. The code given in that figure works for a general $n \times n$ sliding puzzle. It takes two states `stateA` and `stateB` and computes the Manhattan distance between these states.

   (a) First, the size `n` of the puzzle is computed by checking the number of rows of `stateA`.

   (b) Next, the `for` loop iterates over all rows and columns of `stateA` that do not contain a blank tile. Remember that the blank tile is coded using the number 0. The tile at position $\langle$`rowA`, `colA`$\rangle$ in `stateA` is computed using the expression `stateA[rowA][colA]` and the corresponding position $\langle$`rowB`, `colB`$\rangle$ of this tile in state `stateB` is computed using the function `findTile`.

   (c) Finally, the Manhattan distance between the two positions $\langle$`rowA`, `colA`$\rangle$ and $\langle$`rowB`, `colB`$\rangle$ is added to the `result`.

```
manhattan := procedure(stateA, stateB) {
    n := #stateA;
    L := [1 .. n];
    result := 0;
    for (rowA in L, colA in L | stateA[rowA][colA] != 0) {
        [rowB, colB] := findTile(stateA[rowA][colA], stateB);
        result += abs(rowA - rowB) + abs(colA - colB);
    }
    return result;
};
```

Figure 2.18: The Manhattan distance between two states.

The Manhattan distance is admissible. The reason is that if $s_2 \in \texttt{nextStates}(s_1)$, then there can be only one tile t such that the position of t in s_1 is different from the position of t in s_2. Furthermore, this position differs by either one row or one column. Therefore,

$$|h(s_1) - h(s_2)| = 1$$

and hence $h(s_1) \leq 1 + h(s_2)$. □

Now we are ready to present best first search. This algorithm is derived from the stack based version of depth first search. However, instead of using a stack, the algorithm uses a priority queue. In this priority queue, the paths are ordered with respect to the estimated distance of the state at the end of the path from the `goal`. We always expand the path next that seems to be closest to the goal.

```
bestFirstSearch := procedure(start, goal, nextStates, heuristic) {
    PrioQueue := { [heuristic(start, goal), [start]] };
    while (PrioQueue != {}) {
        [_, Path] := fromB(PrioQueue);
        state     := Path[-1];
        if (state == goal) { return Path; }
        for (ns in nextStates(state) | !(ns in Path)) {
            PrioQueue += { [heuristic(ns, goal), Path + [ns]] };
        }
    }
};
```

Figure 2.19: The best first search algorithm.

The procedure `bestFirstSearch` shown in Figure 2.19 on page 27 takes four parameters. The first three of these parameters are the same as in the previous search algorithms. The last parameter `heuristic` is a function that takes to states and then estimates the distance between these states. Later, we will use the Manhattan distance to serve as this `heuristic`. The details of the implementation are as follows:

1. The variable `PrioQueue` serves as a priority queue. We take advantage of the fact that SetlX stores sets as ordered binary trees that store their elements in increasing order. Hence, the smallest element of a set is the first element.

   Furthermore, SetlX orders pairs lexicographically. Hence, we store the paths in `PrioQueue` as pairs of the form

   $$\langle \texttt{estimate}, \texttt{Path} \rangle.$$

Here `Path` is a list of states starting from the node `start`. If the last node on this list is called `state`, then we have

$$\texttt{estimate} = \texttt{heuristic}(\texttt{state}, \texttt{goal}),$$

i.e. `estimate` is the estimated distance between `state` and `goal` and hence an estimate of the number of steps needed to complete `Path` into a solution. This ensures, that the best path, i.e. the path whose last state is nearest to the `goal` is at the beginning of the set `PrioQueue`.

2. As long as `PrioQueue` is not empty, we take the `Path` from the beginning of this priority queue and remove it from the queue. The state at the end of `Path` is named `state`.

   In SetlX, a function call of the form `fromB`(S) returns the smallest element from the set S. Furthermore, this element is removed from the set S.

3. If `state` is the `goal`, a solution has been found and is returned.

4. Otherwise, the states reachable from `state` are inserted into the priority queue. When these states are inserted, we have to compute their estimated distance to `goal` since this distance is used as the priority in `PrioQueue`.

Best first search solves the instance of the 3×3 puzzle shown in Figure 2.3 on page 11 in less than a quarter of a second. However, the solution that is found takes 75 steps. While this is not as ridiculous as the solution found by depth first search, it is still far from an optimal solution. Furthermore, best first search is still not strong enough to solve the 4×4 puzzle shown in Figure 2.23 on page 36.

It should be noted that the fact that the Manhattan distance is a consistent heuristic is of no consequence for best first search. Only the A* algorithm, which is presented next, makes use of this fact.

2.7 The A* Search Algorithm

We have seen that best first search can be very fast. However, the solution returned by best first search is not optimal. In contrast, the A* algorithm described next guarantees that a shortest path is found, provided the heuristic used is consistent. The A* search algorithm works similar to best first search, but instead of using the heuristic as the priority, the priority $f(s)$ of every state s is given as

$$f(s) := g(s) + h(s),$$

where $g(s)$ computes the length of the path leading from `start` to s and $h(s)$ is the heuristical estimate of the distance from s to `goal`. Hence, $f(s)$ is the estimate of the total distance that a path connecting `start` and `goal` while passing through the state s would use. The details of the A* algorithm are given in Figure 2.20 on page 29 and discussed below. The function `aStarSearch` takes 4 parameters:

1. `start` is a state. This state represents the start state of the search problem.

2. `goal` is the goal state.

3. `nextStates` is a function that takes a state as a parameter. For a state s,

   $$\texttt{nextStates}(s)$$

   computes the set of all those states that can be reached from s in a single step.

4. `heuristic` is a function that takes two parameters. For two states s_1 and s_2, the expression

   $$\texttt{heuristic}(s_1, s_2)$$

   computes an estimate of the distance between s_1 and s_2.

The function `aStarSearch` maintains 5 variables that are crucial for the understanding of the algorithm.

```
aStarSearch := procedure(start, goal, nextStates, heuristic) {
    Parent   := {};                    // back pointers, represented as dictionary
    Distance := { [start, 0] };
    estGoal  := heuristic(start, goal);
    Estimate := { [start, estGoal] };  // estimated distances
    Frontier := { [estGoal, start] };  // priority queue
    while (Frontier != {}) {
        [estimate, state] := fromB(Frontier);
        if (state == goal) {
            return pathTo(state, Parent);
        }
        stateDist := Distance[state];
        for (neighbour in nextStates(state)) {
            oldEstimate := Estimate[neighbour];
            newEstimate := stateDist + 1 + heuristic(neighbour, goal);
            if (oldEstimate == om || newEstimate < oldEstimate) {
                Parent[neighbour]   := state;
                Distance[neighbour] := stateDist + 1;
                Estimate[neighbour] := newEstimate;
                Frontier            += { [newEstimate, neighbour] };
                if (oldEstimate != om) {
                    Frontier -= { [oldEstimate, neighbour] };
                }
            }
        }
    }
};
```

Figure 2.20: The A* search algorithm.

1. `Parent` is a dictionary associating a parent state with those states that have already been encountered during the search, i.e. we have

   $$\texttt{Parent}[s_2] = s_1 \;\Rightarrow\; s_2 \in \texttt{nextStates}(s_1).$$

   Once the goal has been found, this dictionary is used to compute the path from `start` to `goal`.

2. `Distance` is a dictionary. For every state s that is encountered during the search, this dictionary records the length of the shortest path from `start` to s.

3. `Estimate` is a dictionary. For every state s encountered in the search, $\texttt{Estimate}[s]$ is an estimate of the length that a path from `start` to `goal` would have if it would pass through the state s. This estimate is calculated using the equation

   $$\texttt{Estimate}[s] = \texttt{Distance}[s] + \texttt{heuristic}(s, \texttt{goal}).$$

   Instead of recalculating this sum every time we need it, we store it in the dictionary `Estimate`. This increases the efficiency of the algorithm.

4. `Frontier` is a priority queue. The elements of `Frontier` are pairs of the form

   $$[d, s] \quad \text{such that} \quad d = \texttt{Estimate}[s],$$

   i.e. if $[d, s] \in \texttt{Frontier}$, then the state s has been encountered in the search and it is estimated that a path leading from `start` to `goal` and passing through s would have a length of d.

After the variables described above have been initialized, the A* algorithm runs in a `while` loop that does only terminate if either a solution is found or the priority queue `Frontier` is exhausted and hence the state `goal` is not reachable from the state `start`.

1. First, the `state` with the smallest estimated distance for a path connecting `start` with `goal` and passing through `state` is chosen from the priority queue `Frontier`. Note that the call of the function `fromB` does not only return the pair

   [`estimate`, `state`]

   from `Frontier` that has the lowest value of `estimate`, but also removes this pair from the priority queue.

2. Now if this `state` is the `goal`, then a solution has been found. The path corresponding to this solution is computed by the function `pathTo`, which utilizes the dictionary `Parent`. The resulting path is then returned and the function terminates.

3. Otherwise, we retrieve the length of the path leading from `start` to `state` from the dictionary `Distance` and store this length in the variable `stateDist`. We will later prove that `stateDist` is not just the length of a some path connecting `start` and `state` but rather the length of the shortest path connecting these two states.

4. Next, we have a `for` loop that iterates over all states neighbouring `state`.

   (a) For every `neighbour` we check the estimated length of a solution passing through `neighbour` and store this length in `oldEstimate`. Note that `oldEstimate` is undefined, i.e. it has the value `om`, if we haven't yet encountered the node `neighbour` in our search.

   (b) If a solution connects `start` with `goal` while passing first through `state` and then through `neighbour`, the estimated length of this solution would be

   $$\texttt{stateDist} + 1 + \texttt{heuristic}(\texttt{neighbour}, \texttt{goal}).$$

   Therefore this value is stored in `newEstimate`.

   (c) Next, we need to check whether this new solution that first passes through `state` and then proceeds to `neighbour` is better than the previous solution that passes through `neighbour`. This check is done by comparing `newEstimate` and `oldEstimate`. Note that we have to take care of the fact that there might be no valid `oldEstimate`, i.e. the variable `oldEstimate` might have the value `om`.

   In case the new solution seems to be better than the old solution, we have to update the `Parent` dictionary, the `Distance` dictionary, and the `Estimate` dictionary. Furthermore, we have to update the priority queue `Frontier`. Here, we have to take care to remove the previous entry for the state `neighbour` if it exists, which is the case if `oldEstimate` is different from `om`.

The A* algorithm has been discovered by Hart, Nilsson, and Raphael and was first published in 1968 [HNR68]. However, there was a subtle bug in the first publication which was corrected in 1972 [HNR72].

When we run A* on the 3×3 sliding puzzle, it takes about 17 seconds to solve the instance shown in Figure 2.3 on page 11. If we just look at the time, this seems to be disappointing. However, the good news is that:

1. The path which is found is optimal.

2. Only $10,061$ states are touched in the search for a solution. This is more than a tenfold reduction when compared with breadth first search.

The fact that the running time is, nevertheless, quite high results from the complexity of computing the Manhattan distance.

2.7.1 Completeness and Optimality of A* Search

In order to prove the completeness and the optimality of the A* search algorithm it is convenient to reformulate the algorithm: In particular, we replace the dictionary `Parent` by a dictionary `PathDict`. For every state s

that is reached during A* search, `PathDict`$[s]$ returns a path that connects the node `start` with the node s. Furthermore, we have added a set `Explored` to the implementation. This set is only needed for the proof of the optimality of the path found by A* search. Figure 2.21 on page 31 shows this version of the algorithm.

```
1  aStarSearch := procedure(start, goal, nextStates, heuristic) {
2      estGoal  := heuristic(start, goal);
3      Estimate := { [start, estGoal] };    // estimated total distance
4      Frontier := { [estGoal, start] };    // priority queue
5      PathDict := { [start, [start]] };
6      Distance := { [start,       0] };
7      Explored := {};
8      while (Frontier != {}) {
9          [_, state] := fromB(Frontier);
10         Explored   += { state };
11         if (state == goal) {
12             return PathDict[state];
13         }
14         stateDist := Distance[state];
15         for (neighbour in nextStates(state)) {
16             oldEstimate := Estimate[neighbour];
17             newEstimate := stateDist + 1 + heuristic(neighbour, goal);
18             if (oldEstimate == om || newEstimate < oldEstimate) {
19                 Distance[neighbour] := stateDist + 1;
20                 Estimate[neighbour] := newEstimate;
21                 Frontier            += { [newEstimate, neighbour] };
22                 PathDict[neighbour] := PathDict[state] + [neighbour];
23                 if (oldEstimate != om) {
24                     Frontier -= { [oldEstimate, neighbour] };
25                 }
26             }
27         }
28     }
29 };
```

Figure 2.21: A path based implementation of A* search.

Theorem 4 (Completeness and Optimality of A* Search) Assume

$$\mathcal{P} = \langle Q, \texttt{nextStates}, \texttt{start}, \texttt{goal} \rangle$$

is a search problem and

$$\texttt{heuristic} : Q \to \mathbb{N}$$

is a consistent heuristic for the search problem $\mathcal{P}$. Then the A* search algorithm is both complete and optimal, i.e. if there is a path from `start` to `goal`, then the search is successful and, furthermore, the solution that is computed is a shortest path leading from `start` to `goal`.

Proof: In order to prove this theorem, we first need to establish a number of notions that are needed in order to improve our understanding of the A* algorithm. The first of these notions is the notion of explored states. A state is said to have been explored iff it has been removed from the priority queue `Frontier`. To emphasize this notion I have added the set `Explored` in the implementation of A* search that is shown in Figure 2.21 on page 31. A state is explored once it has been added to the set `Explored`. Note that the variable `Explored` serves no purpose in the implementation of A* search: This variable is only written to but the algorithm never reads this

variable. I have added the variable `Explored` to aid in our proof. Furthermore, we define a state as visited iff it has been entered into the dictionary `Distance`, i.e. a state s is visited iff `Distance`$[s]$ is defined. Before we prove the theorem, let us establish the following auxiliary claim:

Claim One: If a state s is visited and $P :=$ `PathDict`$[s]$, then P is a path from `start` to s and the length of P is equal to `Distance`$[s]$.

We establish this claim by a straightforward computational induction.

BC: The first path that is entered into the dictionary `PathDict` is the path [`start`]. This path connects the node `start` with the node `start`. Obviously, this path has the length 0 and that is exactly the value that is entered in the dictionary `Distance` in line 6.

IS: If we add in line 22 the path

$$\texttt{PathDict}[\texttt{state}] + [\texttt{neighbour}]$$

as a path leading to the state `neighbour`, then by our induction hypotheses we know that

$$P := \texttt{PathDict}[\texttt{state}]$$

is a path leading from `start` to `state` and that, furthermore, the length of the path P is equal to `Distance`[`state`]. When we append the state `neighbour` to this path P, the length of the resulting path is bigger by one than the length of P and it is obvious that this new path

$$P + [\texttt{neighbour}]$$

connects `start` with `neighbour`. The induction hypotheses tell us that `Distance`[`state`] is equal to the length of P. Therefore, the length of the new path is `Distance`[`state`] $+ 1$ and since `stateDist` is defined as `Distance`[`state`], the correct length is stored in line 19.

This concludes the proof of Claim One.

Another notion we need to establish is a function g that is defined for all states s and returns the distance of s from the state `start`, i.e. we have

$$g(s) := \texttt{dist}(\texttt{start}, s) \quad \text{for all } s \in Q.$$

Here, the expression `dist`(`start`, s) returns the length of the shortest path from `start` to s. From the definition of the dictionary `Distance` and Claim One it is obvious that

$$g(s) \leq \texttt{Distance}[s].$$

The reason is that every time an entry for a state s is added to the dictionary `Distance`, we have found a path from `start` to s that has the length `Distance`$[s]$. This might not be the shortest path, hence $g(s)$ might be less than the length of this path.

Next, we introduce the notion of the estimated total distance of a state s that has been visited during the search. The estimated total distance of a state s is written as $f(s)$ and denotes the estimated length of a path starting in `start` and ending in `goal` that visits the state s in between. Formally, $f(s)$ is defined as follows:

$$f(s) := \texttt{Distance}[s] + \texttt{heuristic}(s, \texttt{goal}).$$

Here, `Distance`$[s]$ is the number of steps that it takes to reach the state s from `start` when following the path stored in `PathDict`$[s]$, while `heuristic`$(s, \texttt{goal})$ is the estimated distance of a path from s to `goal`. Note that the priority queue `Frontier` is ordered according to the estimated total distance. The shorter the estimated total distance of a node s is, the higher the priority of s. In order to prove the theorem we need to establish the following claim.

Claim Two: If a state $s \in Q$ has been explored, then we have

$$\texttt{Distance}[s] = g(s),$$

i.e. the path leading from `start` to s is guaranteed to be a shortest path.

Proof of Claim Two: The proof of Claim Two is a proof by contradiction. We assume that s is a state that

has been explored such that

$$g(s) < \texttt{Distance}[s],$$

i.e. we assume that $P_1 := \texttt{PathDict}[s]$ is not a shortest path from `start` to s. Then, there must be another path P_2 from `start` to s that is shorter than P_1. Both P_1 and P_2 start at the same state `start`. Hence, these two paths must have the form

$$P_1 = [\texttt{start}, x_1, \cdots, x_k, x_{k+1}, \cdots, x_{n-1}, s], \quad P_2 = [\texttt{start}, x_1, \cdots, x_k, y_{k+1}, \cdots, y_{m-1}, s]$$

where $x_{k+1} \neq y_{k+1}$. Here, x_k is the state at which the paths P_1 and P_2 diverge. The index k could be 0. In that case the two paths would already diverge at the state `start`, but in general k will be some non-negative integer. Since we assume that P_1 is longer than P_2 we will have $m < n$. Furthermore, according to Claim One we have

$$\texttt{Distance}[s] = n.$$

Now it is time to make use of the fact that the heuristic h is consistent. First, since $s \in \texttt{nextStates}(y_{m-1})$ we have that

$$\texttt{heuristic}(y_{m-1}) \leq 1 + \texttt{heuristic}(s).$$

Next, as we have $y_{m-1} \in \texttt{nextStates}(y_{m-2})$ the consistency of `heuristic` implies

$$\texttt{heuristic}(y_{m-2}) \leq 1 + \texttt{heuristic}(y_{m-1}) \leq 2 + \texttt{heuristic}(s).$$

Continuing this way we conclude that

$$\texttt{heuristic}(y_{m-i}) \leq i + \texttt{heuristic}(s) \quad \text{for all } i \in \{1, \cdots, m-k-1\}.$$

Define $i := m - k - 1$. Since $m - i = m - (m - k - 1) = k + 1$ we have shown that

$$\texttt{heuristic}(y_{k+1}) \leq m - k - 1 + \texttt{heuristic}(s).$$

This implies

$$\begin{aligned} f(y_{k+1}) &= \texttt{Distance}[y_{k+1}] + \texttt{heuristic}(y_{k+1}) \\ &\leq \texttt{Distance}[y_{k+1}] + m - k - 1 + \texttt{heuristic}(s) \\ &= (k+1) + m - k - 1 + \texttt{heuristic}(s) && \text{since } \texttt{Distance}[y_{k+1}] = k+1 \\ &= m + \texttt{heuristic}(s). \end{aligned}$$

Next, we compute $f(s)$. We have

$$f(s) = \texttt{Distance}[s] + \texttt{heuristic}(s) = n + \texttt{heuristic}(s).$$

However, since we have $m < n$ this implies

$$f(y_{k+1}) = m + \texttt{heuristic}(s) < n + \texttt{heuristic}(s) = f(s).$$

Now $f(s)$ is the priority of the state s in the priority queue, while $f(y_{k+1})$ is the priority attached to the state y_{k+1}. At the latest, the state y_{k+1} is put onto the priority queue `Frontier` immediately after the state x_k is explored. But then the priority of y_{k+1} is higher than the priority of the state s and hence it is explored prior to s. This means that the state y_{k+2} is explored before the state s is explored. Similarly to the proof that $f(y_{k+2}) \leq m + \texttt{heuristic}(s)$, we can see that

$$f(y_{k+2}) \leq k + 2 + (m - k - 2) + \texttt{heuristic}(s) = m + \texttt{heuristic}(s).$$

Hence, the state y_{k+2} is explored before the state s is visited. In general, we have

$$f(y_j) \leq m + \texttt{heuristic}(s) \quad \text{for all } j \in \{k+1, \cdots, m-1\}.$$

Since all states of the form

$$y_j \quad \text{for } j \in \{k+1, \cdots, m-1\}$$

have a priority that is higher than the state s, these states are all explored prior to the state s. In particular,

after the state y_{m-1} has been explored, we will have

$$\texttt{Distance}[s] = m \quad \text{instead of} \quad \texttt{Distance}[s] = n.$$

This contradiction proves that our assumption

$$\texttt{Distance}[s] > g(s)$$

must be wrong and therefore we know that the equation $\texttt{Distance}[s] = g(s)$ must hold for all states s that have been explored. Hence the validity of Claim Two has been established.

Claim Two implies that if the A* algorithm finds a path from `start` to `goal`, then this path must be optimal. This follows from the fact that before we check whether the search has reached the `goal`, the `state` that is compared to `goal` has been explored and, according to Claim Two, must therefore have $\texttt{Distance}[\texttt{goal}] = g(\texttt{goal})$. As $g(\texttt{goal})$ is the true distance of the state `goal` form the state `start`, this implies that the path that has been found is optimal.

Finally, we have to show that the algorithm is complete, i.e. we have to show that if there is a path connecting `start` and `goal`, then we will find a path. Therefore, let us assume that there is a path P connecting `start` and `goal` and that, furthermore, this path is shorter than any other such path and has length n. By reasoning analogously to the proof of Claim Two it can be shown that the states making up this path P all have a priority less or equal than n. Therefore, these states will all be explored before a state with a priority greater than n is explored. However, the set of states that have a priority of at most n is finite, because it is a subset of the set of states that have a distance from `start` that is at most n and this latter set is obviously finite. Therefore, unless the `goal` is reached before, all the states on the path P will eventually be explored and this implies that the `goal` is eventually found. □

2.8 Bidirectional A* Search

So far, the best search algorithm we have encountered is bidirectional breadth first search. However, in terms of memory consumption, the A* algorithm also looks very promising. Hence, it might be a good idea to combine these two algorithms. Figure 2.22 on page 35 shows the resulting program. This program relates to the A* algorithm shown in Figure 2.20 on page 29 as the algorithm for bidirectional search shown in Figure 2.16 on page 24 relates to breadth first search shown in Figure 2.5 on page 13. The only new idea is that we alternate between the A* search starting from `start` and the A* search starting from `goal` depending on the estimated total distance:

(a) As long as the search starting from `start` is more promising, we remove states from `FrontierA`.

(b) Once the total estimated distance of a path starting from `goal` is less than the best total estimated distance of a path starting from `start`, we switch and remove states from `FrontierB`.

When we run bidirectional A* search for the 3×3 sliding puzzle shown in Figure 2.3 on page 11, the program takes one second but only uses $2,963$ states. Therefore, I have tried to solve the 4×4 sliding puzzle shown in Figure 2.23 on page 36 using bidirectional A* search. A solution of 44 steps was found in 50 seconds. Only $20,624$ states had to be processed to compute this solution! None of the other algorithms presented so far was able to compute the solution.

2.9 Iterative Deepening A* Search

So far, we have combined A* search with bidirectional search and achieved good results. When memory space is too limited for bidirectional A* search to be possible, we can instead combine A* search with iterative deepening. The resulting search technique is known as iterative deepening A* search and is commonly abbreviated as IDA*. It has been invented by Richard Korf [Kor85]. Figure 2.24 on page 36 shows an implementation of IDA* in SetlX. We proceed to discuss this program.

1. As in the A* search algorithm, the function `idaStarSearch` takes four parameters.

```
aStarSearch := procedure(start, goal, nextStates, heuristic) {
    estimate  := heuristic(start, goal);
    ParentA   := {};                    ParentB   := {};
    DistanceA := { [start, 0] };        DistanceB := { [goal,  0] };
    EstimateA := { [start, estimate] }; EstimateB := { [goal,  estimate] };
    FrontierA := { [estimate, start] }; FrontierB := { [estimate, goal ] };
    while (FrontierA != {} && FrontierB != {}) {
        [guessA, stateA] := first(FrontierA); stateADist := DistanceA[stateA];
        [guessB, stateB] := first(FrontierB); stateBDist := DistanceB[stateB];
        if (guessA <= guessB) {
            FrontierA -= { [guessA, stateA] };
            for (neighbour in nextStates(stateA)) {
                oldEstimate := EstimateA[neighbour];
                newEstimate := stateADist + 1 + heuristic(neighbour, goal);
                if (oldEstimate == om || newEstimate < oldEstimate) {
                    ParentA[neighbour]   := stateA;
                    DistanceA[neighbour] := stateADist + 1;
                    EstimateA[neighbour] := newEstimate;
                    FrontierA            += { [newEstimate, neighbour] };
                    if (oldEstimate != om) {
                        FrontierA -= { [oldEstimate, neighbour] };
                    }
                }
                if (DistanceB[neighbour] != om) {
                    return combinePaths(neighbour, ParentA, ParentB);
                }
            }
        } else {
            FrontierB -= { [guessB, stateB] };
            for (neighbour in nextStates(stateB)) {
                oldEstimate := EstimateB[neighbour];
                newEstimate := stateBDist + 1 + heuristic(start, neighbour);
                if (oldEstimate == om || newEstimate < oldEstimate) {
                    ParentB[neighbour]   := stateB;
                    DistanceB[neighbour] := stateBDist + 1;
                    EstimateB[neighbour] := newEstimate;
                    FrontierB            += { [newEstimate, neighbour] };
                    if (oldEstimate != om) {
                        FrontierB -= { [oldEstimate, neighbour] };
                    }
                }
                if (DistanceA[neighbour] != om) {
                    return combinePaths(neighbour, ParentA, ParentB);
                }
            }
        }
    }
};
```

Figure 2.22: Bidirectional A* search.

```
start := [ [  1, 2,  0,  4 ],
           [ 14, 7, 12, 10 ],
           [  3, 5,  6, 13 ],
           [ 15, 9,  8, 11 ]
         ];
goal  := [ [  1,  2,  3,  4 ],
           [  5,  6,  7,  8 ],
           [  9, 10, 11, 12 ],
           [ 13, 14, 15,  0 ]
         ];
```

Figure 2.23: A start state and a goal state for the 4×4 sliding puzzle.

```
idaStarSearch := procedure(start, goal, nextStates, heuristic) {
    limit := heuristic(start, goal);
    while (true) {
        Path := search([start], goal, nextStates, limit, heuristic);
        if (isList(Path)) {
            return Path;
        }
        limit := Path;
    }
};
search := procedure(Path, goal, nextStates, limit, heuristic) {
    state    := Path[-1];
    distance := #Path - 1;
    total    := distance + heuristic(state, goal);
    if (total > limit) {
        return total;
    }
    if (state == goal) {
        return Path;
    }
    smallest := mathConst("Infinity");
    for (ns in nextStates(state) | !(ns in Path) ) {
        result := search(Path + [ns], goal, nextStates, limit, heuristic);
        if (isList(result)) {
            return result;
        }
        smallest := min([result, smallest]);
    }
    return smallest;
};
```

Figure 2.24: Iterative deepening A* search.

(a) `start` is a state. This state represents the start state of the search problem.

(b) `goal` is the goal state.

(c) `nextStates` is a function that takes a state s as a parameter and computes the set of all those states that can be reached from s in a single step.

(d) `heuristic` is a function that takes two parameters s_1 and s_2, where s_1 and s_2 are states. The expression

$$\texttt{heuristic}(s_1, s_2)$$

computes an estimate of the distance between s_1 and s_2. It is assumed that this estimate is optimistic, i.e. `heuristic` has to be admissible.

2. The function `idaStarSearch` initializes `limit` to be an estimate of the distance between `start` and `goal`. As we assume that the function `heuristic` is optimistic, we know that there is no path from `start` to `goal` that is shorter than `limit`. Hence, we start our search by assuming that we might find a path that has a length of `limit`.

3. Next, we start a `while` loop. In this loop, we call the function `search` to compute a path from `start` to `goal` that has a length of at most `limit`. The function `search` uses A* search and is described in detail below. When the function `search` returns, there are two cases:

   (a) `search` does find a path. In this case, this path is returned in the variable `Path` and the value of this variable is a list. This list is returned as the solution to the search problem.

   (b) `search` is not able to find a path within the given `limit`. In this case, `search` will not return a list representing a path but instead it will return a number. This number will specify the minimal length that any path leading from `start` to `goal` needs to have. This number is then used to update the `limit` which is used for the next invocation of `search`.

   Note that the fact that `search` is able to compute this new `limit` is a significant enhancement over iterative deepening. While we had to test every single possible length in iterative deepening, now the fact that we can intelligently update the `limit` results in a considerable saving of computation time.

We proceed to discuss the function `search`. This function takes 5 parameters, which we describe next.

1. `Path` is a list of states. This list starts with the state `start`. If `state` is the last state on this list, then `Path` represents a path leading from `start` to `state`.

2. `goal` is another state. The purpose of the recursive invocations of `search` is to find a path from `state` to `goal`, where `state` is the last element of the list `Path`.

3. `nextStates` is a function that takes a state s as input and computes the set of states that are reachable from s in one step.

4. `limit` is the upper bound for the path from `start` to `goal`. If the procedure `search` is not able to find a path from `start` to `goal` that has a length of at most `limit`, then the search is unsuccessful. In that case, instead of a path the function `search` returns a new estimate for the distance between `start` and `goal`. Of course, this new estimate will be bigger than `limit`.

5. `heuristic` is a function taking to states as arguments. The invocation $\texttt{heuristic}(s_1, s_2)$ computes an *estimate* of the distance between the states s_1 and s_2. It is assumed that this estimate is optimistic, i.e. the value returned by $\texttt{heuristic}(s_1, s_2)$ is less or equal than the true distance between s_1 and s_2.

We proceed to describe the implementation of the function `search`.

1. The variable `state` is assigned the last element of `Path`. Hence, `Path` connects `start` and `state`.

2. The length of the path connecting `start` and `state` is stored in `distance`.

3. Since `heuristic` is assumed to be optimistic, if we want to extend `Path`, then the best we can hope for is to find a path from `start` to `goal` that has a length of

$$\texttt{distance} + \texttt{heuristic}(\texttt{state}, \texttt{goal}).$$

   This length is computed and saved in the variable `total`.

4. If `total` is bigger than `limit`, it is not possible to find a path from `start` to `goal` passing through `state` that has a length of at most `limit`. Hence, in this case we return `total` to communicate that the limit needs to be increased to have at least a value of `total`.

5. If we are lucky and `state` is equal to `goal`, the search is successful and `Path` is returned.

6. Otherwise, we iterate over all nodes `ns` reachable from `state` that have not already been visited by `Path`. If `ns` is a node of this kind, we extend the `Path` so that this node is visited next. The resulting path has the form

   ```
   Path + [ns].
   ```

   Next, we recursively start a new search starting from the node `ns`. If this search is successful, the resulting path is returned. Otherwise, the search returns the minimum distance that is needed to reach the state `goal` from the state `start` on a path via the state `ns`. If this distance is smaller than the distance `smallest` that is returned from visiting previous neighbouring nodes, the variable `smallest` variable is updated accordingly. This way, if the `for` loop is not able to return a path leading to `goal`, the variable `smallest` contains a lower bound for the distance that is needed to reach `goal` by a path that extends the given `Path`.

   Note: At this point, a natural question is to ask whether the `for` loop should collect all paths leading to `goal` and then only return the path that is shortest. However, this is not necessary: Every time the function `search` is invoked it is already guaranteed that there is no path that is shorter than the parameter `limit`. Therefore, if `search` is able to find a path that has a length of at most `limit`, this path is known to be optimal.

Iterative deepening A* is a complete search algorithm that does find an optimal path, provided that the employed heuristic is optimistic. On the instance of the 3×3 sliding puzzle shown on Figure 2.3 on page 11, this algorithm takes about 1.7 seconds to solve the puzzle. For the 4×4 sliding puzzle, the algorithm takes about 154 seconds. Although this is more than the time needed by bidirectional A* search, the good news is that the IDA* algorithm does not need much memory since basically only the path discovered so far is stored in memory. Hence, IDA* is a viable alternative if the available memory is not sufficient to support the bidirectional A* algorithm.

2.10 The A*-IDA* Search Algorithm

So far, from all of the search algorithms we have tried, the bidirectional A* search has performed best. However, bidirectional A* search is only feasible if sufficient memory is available. While IDA* requires more time, its memory consumption is much lower than the memory consumption of bidirectional A*. Hence, it is natural to try to combine the A* algorithm and the IDA* algorithm. Concretely, the idea is to run an A* search from the `start` node until memory is more or less exhausted. Then, we start IDA* from the `goal` node and search until we find any of the nodes discovered by the A* search that had been started from the `start` node.

An implementation of the A*-IDA* algorithm is shown in Figure 2.25 on page 39 and Figure 2.26 on page 40. We begin with a discussion of the procedure `aStarIdaStarSearch`.

1. The procedure takes 5 arguments.

   (a) `start` and `goal` are nodes. The procedure tries to find a path connecting `start` and `goal`.

   (b) `nextStates` is a function that takes a state s as input and computes the set of states that are reachable from s in one step.

   (c) `heuristic` computes an *estimate* of the distance between s_1 and s_2. It is assumed that this estimate is optimistic, i.e. the value returned by `heuristic`(s_1, s_2) is less or equal than the true distance between s_1 and s_2.

   (d) `size` is the maximal number of states that the A* search is allowed to explore before the algorithm switches over to IDA* search.

```
aStarIdaStarSearch := procedure(start, goal, nextStates, heuristic, size) {
    Parent   := {};               // dictionary: state |-> predecessor(state)
    Distance := { [start, 0] };   // dictionary: state |-> distance(start,state)
    est      := heuristic(start, goal);
    Estimate := { [start, est] }; // dictionary: state |-> distance(start,state,goal)
    Frontier := { [est, start] }; // priority queue
    while (#Distance < size && Frontier != {}) {
        [_, state] := fromB(Frontier);
        if (state == goal) {
             return pathTo(state, Parent);
        }
        stateDist := Distance[state];
        for (neighbour in nextStates(state)) {
            oldEstimate := Estimate[neighbour];
            newEstimate := stateDist + 1 + heuristic(neighbour, goal);
            if (oldEstimate == om || newEstimate < oldEstimate) {
                Parent[neighbour]   := state;
                Distance[neighbour] := stateDist + 1;
                Estimate[neighbour] := newEstimate;
                Frontier            += { [newEstimate, neighbour] };
                if (oldEstimate != om) {
                    Frontier -= { [oldEstimate, neighbour] };
                }
            }
        }
    }
    Path := deepeningSearch(goal, start, nextStates, heuristic, Distance);
    return pathTo(Path[-1], Parent) + reverse(Path)[2..];
};
```

Figure 2.25: The A*-IDA* search algorithm, part I.

2. The basic idea behind the A*-IDA* algorithm is to first use A* search to find a path from `start` to `goal`. If this is successfully done without visiting more than `size` nodes, the algorithm terminates and returns the path that has been found. Otherwise, the algorithm switches over to an IDA* search that starts from `goal` and tries to connect goal to any of the nodes that have been encountered during the A* search. To this end, the procedure `aStarIdaStarSearch` maintains the following variables.

   (a) `Parent` is a dictionary associating a parent state with those states that have already been encountered during the search, i.e. we have

   $$\texttt{Parent}[s_2] = s_1 \;\Rightarrow\; s_2 \in \texttt{nextStates}(s_1).$$

   Once the goal has been found, this dictionary is used to compute the path from `start` to `goal`.

   (b) `Distance` is a dictionary that remembers for every state s that is encountered during the A* search the length of the shortest path from `start` to s.

   (c) `Estimate` is a dictionary. For every state s encountered in the A* search, `Estimate`$[s]$ is an estimate of the length that a path from `start` to `goal` would have if it would pass through the state s. This estimate is calculated using the equation

   $$\texttt{Estimate}[s] = \texttt{Distance}[s] + \texttt{heuristic}(s, \texttt{goal}).$$

   Instead of recalculating this sum every time we need it, we store it in the dictionary `Estimate`.

   (d) `Frontier` is a priority queue. The elements of `Frontier` are pairs of the form

$$[d, s] \quad \text{such that} \quad d = \texttt{Estimate}[s],$$

i.e. if $[d, s] \in \texttt{Frontier}$, then the state s has been encountered in the A* search and it is estimated that a path leading from `start` to `goal` and passing through s would have a length of d.

3. The A* search runs exactly as discussed previously. The only difference is that the `while` loop is terminated once the dictionary `Distance` has more than `size` entries. If we are lucky, the A* search is already able to find the `goal` and the algorithm terminates.

4. Otherwise, the procedure `deepeningSearch` is called. This procedure starts an iterative deepening A* search from the node `goal`. This search terminates as soon as a state is found that has already been encountered during the A* search. The set of these nodes is given to the procedure `deepeningSearch` via the parameter `Distance`. The procedure `deepeningSearch` returns a pair. The first component of this pair is the state `s`. This is the state in `Distance` that has been reached by the IDA* search. The second component is the path `P` that leads from the node `s` to the node `goal` but that does not include the node `s`. In order to compute a path from `start` to `goal`, we still have to compute a path from `start` to s. This path is then combined with the path `P` and the resulting path is returned.

```
1  deepeningSearch := procedure(goal, start, nextStates, heuristic, Distance) {
2      limit := 0;
3      while (true) {
4          print("limit = $limit$");
5          Path := search([goal], start, nextStates, limit, heuristic, Distance);
6          if (isList(Path)) {
7              return Path;
8          }
9          limit := Path;
10     }
11 };
12 search := procedure(Path, start, nextStates, limit, heuristic, Distance) {
13     state := Path[-1];
14     total := #Path - 1 + heuristic(state, start);
15     if (total > limit) {
16         return total;
17     }
18     if (Distance[state] != om) {
19         return Path;
20     }
21     smallest := mathConst("Infinity");
22     for (ns in nextStates(state) | !(ns in Path) ) {
23         result := search(Path + [ns], start, nextStates, limit, heuristic, Distance);
24         if (isList(result)) {
25             return result;
26         }
27         smallest := min([smallest, result]);
28     }
29     return smallest;
30 };
```

Figure 2.26: The A*-IDA* search algorithm, part II.

5

Iterative deepening A*-IDA* is a complete search algorithm. On the instance of the 3×3 sliding puzzle shown on Figure 2.3 on page 11, this algorithm takes about 0.8 seconds to solve the puzzle. For the 4×4 sliding puzzle, if the algorithm is allowed to visit at most 3 000 states, the algorithm takes less than 5 seconds.

Exercise 4: The eight queens puzzle is the problem of placing eight chess queens on a chessboard so that no two queens can attack each other. In chess, a queen can attack by moving horizontally, vertically, or diagonally. Reformulate the eight queens puzzle as a search problem and solve this puzzle via breadth first search. For extra credit, compute all 92 solutions of the eight queens puzzle.

Hint: It is easiest to encode states as lists. For example, the solution of the eight queens puzzle that is shown in Figure 2.27 would be represented as the list

$$[6, 4, 7, 1, 8, 2, 5, 3]$$

because the queen in the first row is positioned in column 6, the queen in the second row is positioned in column 4, and so on. The start state would then be the empty list and given a state L, all states from the set `nextState`(L) would be lists of the form $L + [c]$. If $\#L = k$, then the state $l + [c]$ has $k + 1$ queens, where the queen in row $k + 1$ has been placed in column c. A frame containing a function that can be used to print a board can be found at:

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/queens-frame.stlx

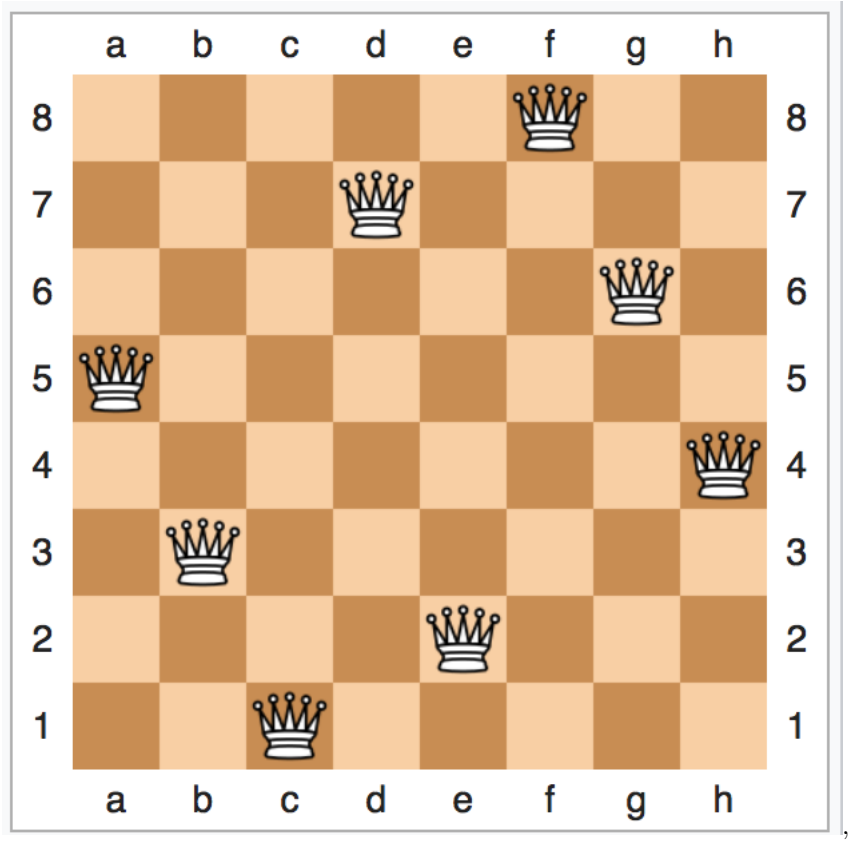

Figure 2.27: A solution of the eight queens puzzle.

Exercise 5: The founder of Taoism, the Chinese philosopher Laozi once said:

> *"A journey of a thousand miles begins but with a single step".*

This proverb is the foundation of taoistic search. The idea is, instead of trying to reach the goal directly, we rather define some intermediate states which are easier to reach than the goal state and that are nearer to the goal than the start state. To make this idea more precise, consider the following instance of the 15-puzzle, where the states `Start` and `Goal` are given as follows:

```
Start :=  +---+---+---+---+    Goal :=  +---+---+---+---+
          |15 |14 | 8 |12 |              | 1 | 2 | 3 | 4 |
          +---+---+---+---+              +---+---+---+---+
          |10 |11 |13 | 9 |              | 5 | 6 | 7 | 8 |
          +---+---+---+---+              +---+---+---+---+
          | 6 | 2 | 5 | 1 |              | 9 |10 |11 |12 |
          +---+---+---+---+              +---+---+---+---+
          | 3 | 4 | 7 |   |              |13 |14 |15 |   |
          +---+---+---+---+              +---+---+---+---+
```

In order to solve the puzzle, we could try to first move the tiles numbered with 1 and 2 into the upper left corner. The resulting state would have the following form:

```
+---+---+---+---+
| 1 | 2 | * | * |
+---+---+---+---+
| * | * | * | * |
+---+---+---+---+
| * | * | * | * |
+---+---+---+---+
| * | * | * | * |
+---+---+---+---+
```

Here, the character "`*`" is used as a wildcard character, i.e. we do not care about the actual character in the state, for we only want to ensure that the first two tiles are positioned correctly. Once we have reached a state specified by the pattern given above, we could the proceed to reach a state that is described by the following pattern:

```
+---+---+---+---+
| 1 | 2 | * | * |
+---+---+---+---+
| 5 | 6 | * | * |
+---+---+---+---+
| * | * | * | * |
+---+---+---+---+
| * | * | * | * |
+---+---+---+---+
```

This way, slowly but surely we will reach the goal. I have prepared a framework for taoistic search. The file

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/sliding-puzzle-frame.stlx

contains a framework for the sliding puzzle where some functions are left unimplemented. The file

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/a-star-lao-tzu.stlx

is also required. It contains an adapted form of A* search. More details will be given in the lecture. Your task is to implement the missing functions in the file `sliding-puzzle-frame.stlx`. ◇

Chapter 3

Constraint Satisfaction

In this chapter we discuss algorithms for solving constraint satisfaction problems. In a constraint satisfaction problem we have been given a number of formulæ and search for values that can be assigned to the variables occurring in these formulæ so that the formulæ evaluate as true. Constraint satisfaction problems can be seen as a refinement of the search problems discussed in the previous chapter: In a search problem, the states are abstract and therefore have no structure that can be exploited to guide the search, while in a constraint satisfaction problem, the states have a structure, as these states are variable assignments. This structure can be exploited to guide the search. This chapter is structured as follows:

1. The first section defines the notion of a constraint satisfaction problem. In order to illustrate this notion, two examples of constraint satisfaction problems are presented. After that, we discuss applications of constraint satisfaction problems.

2. The simplest algorithm to solve a constraint satisfaction problem is via brute force search. The idea behind brute force search is to test all possible variable assignments.

3. In most cases, the search space is so large that it is not possible to enumerate all potential solutions. Backtracking search improves on brute force search by mixing the generation of variable assignments with the testing of the constraints. In many cases, this approach can drastically improve the performance of the search algorithm.

4. Backtracking search can be refined by using constraint propagation and by using the most restricted variable heuristic.

5. Furthermore, checking the consistency of the values assigned to different variables can reduce the size of the search space considerably.

6. Finally, local search is a completely different approach to solve constraint satisfaction problems. This approach is especially useful if the constraint satisfaction problem is big, but the problem is not too complicated.

3.1 Formal Definition of Constraint Satisfaction Problems

Formally, we define a constraint satisfaction problem as a triple

$$\mathcal{P} := \langle \texttt{Vars}, \texttt{Values}, \texttt{Constraints} \rangle$$

where

1. `Vars` is a set of strings which serve as variables,

2. `Values` is a set of values that can be assigned to the variables in `Vars`.

3. `Constraints` is a set of formulæ from first order logic. Each of these formulæ is called a constraint of $\mathcal{P}$.

   In order to be able to interpret these formulæ, we need a first order structure $\mathcal{S} = \langle \mathcal{U}, \mathcal{J} \rangle$. Here, $\mathcal{U}$ is the universe of $\mathcal{S}$ and we will assume that this universe is identical to the set `Values`, that is we have

   $$\mathcal{U} = \texttt{Values}.$$

   The second component $\mathcal{J}$ defines the interpretations of the function symbols and predicate symbols that are used in the formulæ defining the constraints. In what follows we assume that these interpretations are understood from the context of the constraint satisfaction problem $\mathcal{P}$.

In the following, the abbreviation Csp is short for constraint satisfaction problem. Given a Csp

$$\mathcal{P} = \langle \texttt{Vars}, \texttt{Values}, \texttt{Constraints} \rangle,$$

a variable assignment for $\mathcal{P}$ is a function

$$A : \texttt{Vars} \to \texttt{Values}.$$

A variable assignment A is a solution of the Csp $\mathcal{P}$ if, given the assignment A, all constraints of $\mathcal{P}$ are satisfied, i.e. we have

$$\texttt{eval}(f, A) = \texttt{true} \quad \text{for all } f \in \texttt{Constraints}.$$

Finally, a partial variable assignment B for $\mathcal{P}$ is a function

$$B : \texttt{Vars} \to \texttt{Values} \cup \{\Omega\} \quad \text{where } \Omega \text{ denotes the undefined value.}$$

Hence, a partial variable assignment does not assign values to all variables. Instead, it assigns values only to a subset of the set `Vars`. The domain $\texttt{dom}(B)$ of a partial variable assignment B is the set of those variables that are assigned a value different from Ω, i.e. we define

$$\texttt{dom}(B) := \big\{x \in \texttt{Vars} \mid B(x) \neq \Omega\big\}.$$

We proceed to illustrate the definitions given so far by presenting two examples.

3.1.1 Example: Map Colouring

In map colouring a map showing different state borders is given and the task is to colour the different states such that no two states that have a common border share the same colour. Figure 3.1 on page 45 shows a map of Australia. There are seven different states in Australia:

1. Western Australia, abbreviated as WA,
2. Northern Territory, abbreviated as NT,
3. South Australia, abbreviated as SA,
4. Queensland, abbreviated as Q,
5. New South Wales, abbreviated as NSW,
6. Victoria, abbreviated as V, and
7. Tasmania, abbreviated as T.

Figure 3.1 would certainly look better if different states had been coloured with different colours. For the purpose of this example let us assume that we have only three colours available. The question then is whether it is possible to colour the different states in a way that no two neighbouring states share the same colour. This problem can be formalized as a constraint satisfaction problem. To this end we define:

1. $\texttt{Vars} := \{\text{WA}, \text{NT}, \text{SA}, \text{Q}, \text{NSW}, \text{V}, \text{T}\}$,
2. $\texttt{Values} := \{\texttt{red}, \texttt{green}, \texttt{blue}\}$,

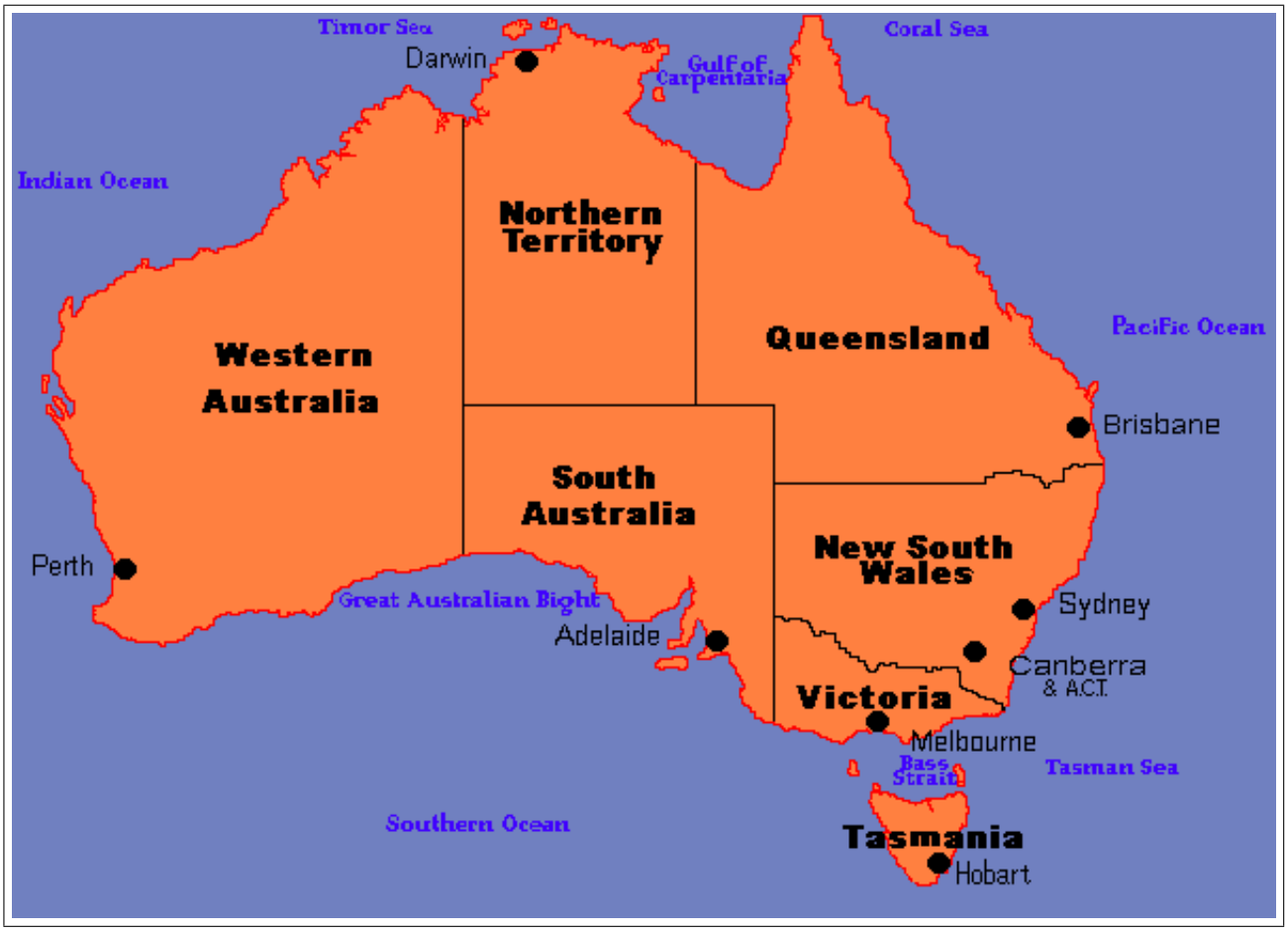

Figure 3.1: A map of Australia.

3. $\texttt{Constraints} :=$
$$\{\text{WA} \neq \text{NT}, \text{WA} \neq \text{SA}, \text{NT} \neq \text{SA}, \text{NT} \neq \text{Q}, \text{SA} \neq \text{Q}, \text{SA} \neq \text{NSW}, \text{SA} \neq \text{V}, \text{Q} \neq \text{NSW}, \texttt{NSW} \neq \texttt{V}, \text{V} \neq \text{T}\}$$

Then $\mathcal{P} := \langle \texttt{Vars}, \texttt{Values}, \texttt{Constraints} \rangle$ is a constraint satisfaction problem. If we define the assignment A such that

1. $A(\text{WA}) = \texttt{blue}$,
2. $A(\text{NT}) = \texttt{red}$,
3. $A(\text{SA}) = \texttt{green}$,
4. $A(\text{Q}) = \texttt{blue}$,
5. $A(\text{NSW}) = \texttt{red}$,
6. $A(\text{V}) = \texttt{blue}$,
7. $A(\text{T}) = \texttt{red}$,

then you can check that the assignment A is indeed a solution to the constraint satisfaction problem $\mathcal{P}$.

3.1.2 Example: The Eight Queens Puzzle

The eight queens problem asks to put 8 queens onto a chessboard such that no queen can attack another queen. In chess, a queen can attack all pieces that are either in the same row, the same column, or the same diagonal. If we want to put 8 queens on a chessboard such that no two queens can attack each other, we have to put exactly one queen in every row: If we would put more than one queen in a row, the queens in that row can attack each other. If we would leave a row empty, then, given that the other rows contain at most one queen, there would be less than 8 queens on the board. Therefore, in order to model the eight queens problem as a constraint satisfaction problem, we will use the following set of variables:

$$\texttt{Vars} := \{\texttt{V}_1, \texttt{V}_2, \texttt{V}_3, \texttt{V}_4, \texttt{V}_5, \texttt{V}_6, \texttt{V}_7, \texttt{V}_8\},$$

where for $i \in \{1, \cdots, 8\}$ the variable $\texttt{V}_i$ specifies the column of the queen that is placed in row i. As the columns run from one to eight, we define the set `Values` as

$$\texttt{Values} := \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

Next, let us define the constraints. There are two different types of constraints.

1. We have constraints that express that no two queens positioned in different rows share the same column. To capture these constraints, we define

   $$\texttt{SameRow} := \big\{\texttt{V}_i \neq \texttt{V}_j \;\big|\; i \in \{1, \cdots, 8\} \wedge j \in \{1, \cdots, 8\} \wedge j < i\big\}.$$

   Here the condition $i < j$ ensures that, for example, we have the constraint $\texttt{V}_2 \neq \texttt{V}_1$ but not the constraint $\texttt{V}_1 \neq \texttt{V}_2$, as the latter constraint would be redundant if the former constraint has already been established.

2. We have constraints that express that no two queens positioned in different rows share the same diagonal. The queens in row i and row j share the same diagonal iff the equation

   $$|i - j| = |V_i - V_j|$$

   holds. The expression $|i - j|$ is the absolute value of the difference of the rows of the queens in row i and row j, while the expression $|V_i - V_j|$ is the absolute value of the difference of the columns of these queens. To capture these constraints, we define

   $$\texttt{SameDiagonal} := \big\{|i - j| \neq |\texttt{V}_i - \texttt{V}_j| \;\big|\; i \in \{1, \cdots, 8\} \wedge j \in \{1, \cdots, 8\} \wedge j < i\big\}.$$

   A more detailed explanation of this constraint is given in my lecture notes on logic.

Then, the set of constraints is defined as

$$\texttt{Constraints} := \texttt{SameRow} \cup \texttt{SameRising} \cup \texttt{SameFalling}$$

and the eight queens problem can be stated as the constraint satisfaction problem

$$\mathcal{P} := \langle \texttt{Vars}, \texttt{Values}, \texttt{Constraints} \rangle.$$

If we define the assignment A such that

$$A(V_1) := 4,\ A(V_2) := 8,\ A(V_3) := 1,\ A(V_4) := 2,\ A(V_5) := 6,\ A(V_6) := 2,\ A(V_7) := 7,\ A(V_8) := 5,$$

then it is easy to see that this assignment is a solution of the eight queens problem. This solution is shown in Figure 3.2 on page 47.

Later, when we implement procedures to solve CSPs, we will represent variable assignments and partial variable assignments as binary relations. For example, A would then be represented as the relation

$$A = \big\{\langle \texttt{V}_1, 4\rangle, \langle \texttt{V}_2, 8\rangle, \langle \texttt{V}_3, 1\rangle, \langle \texttt{V}_4, 2\rangle, \langle \texttt{V}_5, 6\rangle, \langle \texttt{V}_6, 2\rangle, \langle \texttt{V}_7, 7\rangle, \langle \texttt{V}_8, 5\rangle\big\}.$$

If we define

$$B := \big\{\langle \texttt{V}_1, 4\rangle, \langle \texttt{V}_2, 8\rangle, \langle \texttt{V}_3, 1\rangle\big\},$$

then B is a partial assignment and $\texttt{dom}(B) = \{\texttt{V}_1, \texttt{V}_2, \texttt{V}_3\}$. This partial assignment is shown in Figure 3.3 on page 47.

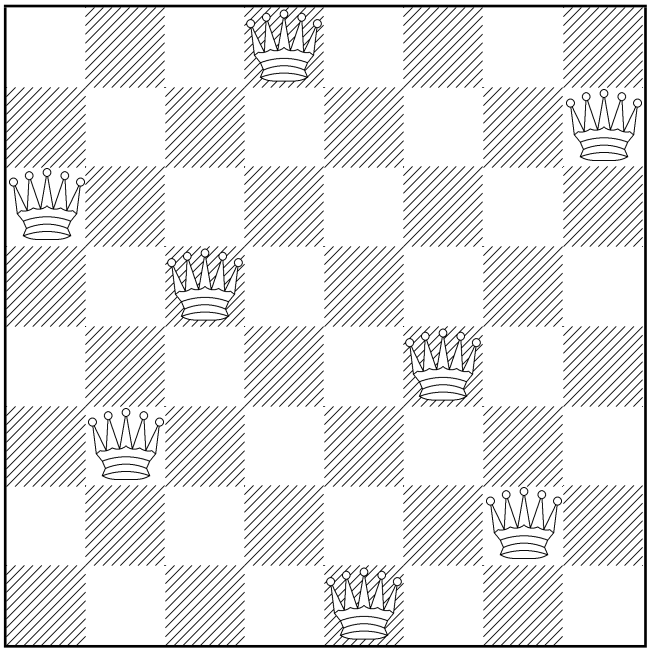

Figure 3.2: A solution of the eight queens problem.

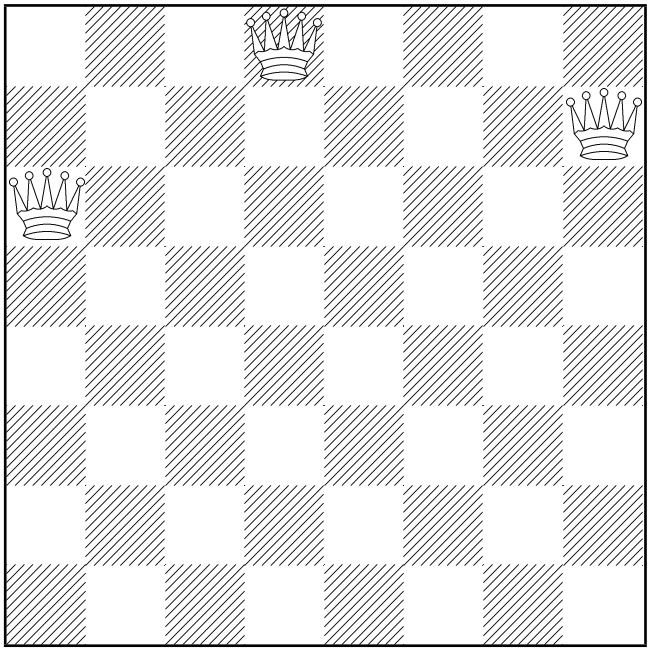

Figure 3.3: The partial assignment $\{\langle \mathrm{V}_1, 4\rangle, \langle \mathrm{V}_2, 8\rangle, \langle \mathrm{V}_3, 1\rangle\}$.

Figure 3.4 on page 48 shows a SetlX program that can be used to create the eight queens puzzle as a Csp. The code shown in this figure is more general than the eight queens puzzle: Given a natural number n, the function call `queensCSP`(n) creates a constraint satisfaction problem $\mathcal{P}$ that generalizes the eight queens problem to the problem of putting n queens on a board of size n times n.

The beauty of constraint programming is the fact that we will be able to develop a so called constraint solver that takes as input a Csp like the one produced by the program shown in Figure 3.4 and that is then

```
createCSP := procedure(n) {
    S            := { 1 .. n };           // used as indices
    Variables    := { "V$i$" : i in S };
    Values       := { 1 .. n };
    SameRow      := { "V$i$ != V$j$" : i in S, j in S | i < j };
    SameDiagonal := { "abs($i$-$j$)!=abs(V$i$-V$j$)" : i in S, j in S | i < j };
    return [Variables, Values, SameRow + SameDiagonal];
};
```

Figure 3.4: SetlX code to create the CSP representing the eight-queens puzzle.

capable of computing a solution. In effect, this enables us to use declarative programming: Instead of specifying an algorithm that solves a problem we confine ourselves to specifying the problem precisely and then let a general purpose problem solver do the job of finding the solution. This approach of declarative programming was one of the main ideas incorporated in the programming language Prolog. While Prolog could not live up to the promisses made in declarative programming as a viable general purpose programming method, constraint programming has proven to be very useful in a number of domains.

3.1.3 Applications

Besides the toy problems discussed so far, there are a number of industrial applications of constraint satisfaction problems. The most important application seem to be variants of scheduling problems. A simple example of a scheduling problem is the problem of generating a time table for a school. A school has various teachers, each of which can teach some subjects but not others. Furthermore, there are a number of classes that must be taught in different subjects. The problem is then to assign teachers to classes and to create a time table. In practice, crew scheduling is an important problem. For example, airlines have to solve a crew scheduling problem in order to efficiently assign crews to aircrafts.

3.2 Brute Force Search

The most straightforward algorithm to solve a Csp is to test all possible combinations of assigning values to variables. If there are n different values that can be assigned to k variables, this amounts to checking n^k different assignments. For example, for the eight queens problem there are 8 variables and 8 possible values leading to

$$8^8 = 2^{24} = 16,777,216$$

different assignments that need to be tested. Given the clock speed of modern computers, checking a million assignments per second is plausible. Hence, this approach is be able to solve the eight queens problem in less than 5 minutes. The approach of testing all possible combinations is known as brute force search. An implementation of brute force search is shown in Figure 3.5 on page 49.

The procedure `solve` gets a constraint satisfaction problem `CSP` as its input. This `CSP` is given as a triple of the form

$$\texttt{CSP} = [\texttt{Variables}, \texttt{Values}, \texttt{Constraints}].$$

The sole purpose of `search` is to call the function `brute_force_search`. The implementation of this function is recursive and the function takes two arguments.

1. `Assignment` is a partial assignment of values to variables. Initially, this assignment will be empty. Every recursive call of `brute_force_search` adds the assignment of one variable to the given assignment.

2. `CSP` is a triple of the form

   $$\texttt{CSP} = [\texttt{Variables}, \texttt{Values}, \texttt{Constraints}].$$

```
solve := procedure(CSP) {
    return brute_force_search({}, CSP);
};
brute_force_search := procedure(Assignment, CSP) {
    [Variables, Values, Constraints] := CSP;
    if (#Assignment == #Variables) {  // all variables have been assigned
        if (check_all_constraints(Assignment, Constraints)) {
            return Assignment;
        }
        return;
    }
    var := rnd(Variables - domain(Assignment));
    for (value in Values) {
        NewAss := Assignment + { [var, value] };
        result := brute_force_search(NewAss, CSP);
        if (result != om) {
            return result;
        }
    }
};
```

Figure 3.5: Solving a CSP via brute force search.

Here, `Constraints` is a set of first order logic formulæ that are given as strings. These strings have to follow the syntax of SETLX formulæ so that they can be evaluated using the SETLX function `eval`.

The implementation of `brute_force_search` works as follows:

1. If all variables have been assigned a value, the dictionary `Assignment` will have the same number of entries as the set `Variables` has elements. Hence, in that case `Assignment` is a complete assignment of all variables and we now have to test whether all constraints are satisfied. This is done using the auxiliary procedure `check_all_constraints` that is shown in Figure 3.6 on page 50. If the current `Assignment` does indeed satisfy all constraints, it is a solution to the given `CSP` and is therefore returned.

   If, instead, some constraint is violated, then `brute_force_search` returns the undefined value Ω.

2. If the assignment is not yet complete, we randomly[1] pick a variable `var` from the set of those `Variables` that still have no value assigned. Then, for every possible `value` in the set `Values`, we augment the current partial `Assignment` to the new assignment

   `Assignment + { [var, value] }`.

   Next, the algorithm recursively tries to find a solution for this new partial assignment. If this recursive call succeeds, the solution it has computed is returned. Otherwise, another `value` is tried.

The function `check_all_constraints` takes a complete variable `Assignment` as its first input. The second input is the set of `Constraints`. For all formulæ `f` from the set `Constraints`, it checks whether `f` is satisfied provided all variables are assigned as specified in `Assignment`. This check is done using the auxiliary function `eval_constraint`.

The procedure `eval_constraint` takes an `Assignment` and a `Formula` that is supposed to be a constraint. Its purpose is to evaluate `Formula` using the `Assignment`. In order for this to be possible we have to assume that

[1] Experimental evidence suggests that picking a random variable performs much better than sorting the variables and always picking the first unassigned variable.

```
21  check_all_constraints := procedure(Assignment, Constraints) {
22      return forall(f in Constraints | eval_constraint(Assignment, f));
23  };
24  eval_constraint := procedure(Assignment, Formula) {
25      for ([var, value] in Assignment) {
26          execute("$var$ := $value$;");
27      }
28      return eval(Formula);
29  };
```

Figure 3.6: Auxiliary procedures for brute force search.

$$\texttt{var(Formula)} \subseteq \texttt{dom(Assignment)},$$

i.e. all variables occurring in `Formula` have a value assigned in `Assignment`.

The `Formula` is evaluated by first assigning the variables assigned in `Assignment` to their corresponding values. Once these assignments have been executed in the `for`-loop, the `Formula` can be evaluated using the procedure `eval`, which is one of the predefined procedures in SetlX.

When I tested this brute force search with the eight queens problem, it took about 300 seconds to compute a solution. In contrast, the seven queens problem took roughly 14 seconds. As we have

$$\frac{8^8}{7^7} \approx 20 \qquad \frac{300}{14} \approx 21$$

this shows that the computation time does indeed grow with the number of possible assignments that have to be checked. However, the correspondence is not exact. The reason is that we stop our search as soon as a solution is found. If we are lucky and the given Csp is easy to solve, this might happen when we have checked only a small portion of the set of all possible assignments.

3.3 Backtracking Search

Due to the combinatorial explosion of the search space, brute force search is only viable when we deal with small problems containing just a handful of variables. One approach to solve a Csp that is both conceptually simple and at least more efficient than brute force search is backtracking. The idea is to try to evaluate constraints as soon as possible: If C is a constraint and B is a partial assignment such that all the variables occurring in C have already been assigned a value in B and the evaluation of C fails, then there is no point in trying to complete the variable assignment B. Hence, in backtracking we evaluate a constraint C as soon as all of its variables have been assigned a value. If C is not valid, we backtrack and discard the current partial variable assignment. This approach can result in huge time savings.

Figure 3.7 on page 51 shows a simple Csp solver that employs backtracking. We discuss this program next. The procedure `solve` takes a constraint satisfaction problem `CSP` as input and tries to find a solution.

1. First, the `CSP` is split into its components.

2. Next, for every constraint `f` of the given `CSP`, we compute the set of variables that are used in `f`. This is done using the procedure `collectVars` that is shown in Figure 3.8 on page 52. These variables are then stored together with the constraint `f` and the correspondingly modified data structure is stored in `CSP` and is called an augmented Csp.

   The reason to compute and store these variables is efficiency: When we later check whether a constraint `f` is satisfied for a partial variable assignment `Assignment` where `Assignment` is stored as a dictionary, we only need to check the constraint `f` iff all of the variables occurring in `f` are elements of the domain of `Assignment`. It would be wasteful to compute these variables every time.

```
solve := procedure(CSP) {
    [Vars, Vals, Constrs] := CSP;
    CSP := [Vars, Vals, { [f, collectVars(f)] : f in Constrs }];
    check {
        return bt_search({}, CSP);
    }
};
bt_search := procedure(Assignment, CSP) {
    [Variables, Values, Constraints] := CSP;
    if (#Assignment == #Variables) {
        return Assignment;
    }
    var := rnd(Variables - domain(Assignment));
    for (value in Values) {
        check {
            if (is_consistent(var, value, Assignment, Constraints)) {
                return bt_search(Assignment + { [var, value] }, CSP);
            }
        }
    }
    backtrack;
};
```

Figure 3.7: A backtracking Csp solver.

3. Next, we call the procedure `bt_search` to compute a solution of `CSP`. This procedure is enclosed in a `check`-block. Conceptually, this `check`-block is the same as if we had enclosed the call to `bt_search` in a `try-catch`-block as shown below:

   ```
   try {
       return bt_search({}, csp);
   } catch(e) {}
   ```

   The point is that the procedure `bt_search` either returns a solution or, if it is not able to find a solution, it throws a special kind of exception, a so called backtrack exception. The `check`-block ensures that this exception is silently discarded. It is just a syntactical convenience that is more concise than using `try` and `catch`.

Next, we discuss the implementation of the procedure `bt_search`. This procedure receives a partial assignment `Assignment` as input together with an augmented `CSP`. This partial assignment is consistent with `CSP`: If `f` is a constraint of `CSP` such that all the variables occurring in `f` are members of `dom(Assignment)`, then evaluating `f` using `Assignment` yields `true`. Initially, this partial assignment is empty and hence trivially consistent. The idea is to extend this partial assignment until it is a complete assignment that satisfies all constraints of the given `CSP`.

1. First, the augmented `CSP` is split into its components.
2. Next, if `Assignment` is already a complete variable assignment, i.e. if the dictionary `Assignment` has as many elements as there are variables, then it must be a solution of the `CSP` and, therefore, it is returned.
3. Otherwise, we have to extend the partial `Assignment`. In order to do so, we first have to select a variable `var` that has not yet been assigned a value in `Assignment` so far.

```
collectVars := procedure(f) {
    return varsTerm(parseTerm(f));
};
varsTerm := procedure(f) {
    match (f) {
        case v | isVariable(v): return { varName(v) };
        case n | isNumber(n):   return {};
        case lhs + rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs - rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs * rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs / rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs \ rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs % rhs:         return varsTerm(lhs) + varsTerm(rhs);
        case lhs == rhs:        return varsTerm(lhs) + varsTerm(rhs);
        case lhs != rhs:        return varsTerm(lhs) + varsTerm(rhs);
        case !formula:          return varsTerm(formula);
        case lhs && rhs:        return varsTerm(lhs) + varsTerm(rhs);
        case lhs || rhs:        return varsTerm(lhs) + varsTerm(rhs);
        case lhs => rhs:        return varsTerm(lhs) + varsTerm(rhs);
        case lhs <==> rhs:      return varsTerm(lhs) + varsTerm(rhs);
        case lhs <!=> rhs:      return varsTerm(lhs) + varsTerm(rhs);
        default:                return +/ [ varsTerm(t) : t in args(f) ];
    }
};
```

Figure 3.8: The procedure `collectVars`.

4. Next, it is tried to assign a `value` to the selected variable `var`. After assigning a `value` to `var`, we immediately check whether this assignment would be consistent with the constraints using the procedure `is_consistent`. If the partial `Assignment` turns out to be consistent, the partial `Assignment` is extended to the new partial assignment

   `Assignment + { [var, value] }`.

   Then, the procedure `bt_search` is called recursively to complete this new partial assignment. If this is successful, the resulting assignment is a solution that is returned. Otherwise, the recursive call of `bt_search` will instead raise an exception. This exception is muted by the `check`-block that surrounds the call to `bt_search`. In that case, the `for`-loop generates a new possible `value` that can be assigned to the variable `var`. If all possible values have been tried and none was successful, the `for`-loop ends and the `backtrack`-statement is executed. Effectively, this statement raises an exception that is caught by one of the `check`-blocks that are above the `backtrack`-statement in the call stack.

```
is_consistent := procedure(var, value, Assignment, Constraints) {
    NA := Assignment + { [var, value] };
    return forall([F, Vars] in Constraints |
                   var in Vars && Vars <= domain(NA) => eval_constraint(NA, F)
                  );
};
```

Figure 3.9: Auxiliary procedures for the Csp solver shown in Figure 3.7

We still need to discuss the implementation of the auxiliary procedure `is_consistent` shown in 3.9. This procedure takes a variable `var`, a `value`, a partial `Assignment` and a set of `Constraints`. It is assumed that `Assignment` is partially consistent with respect to the set `Constraints`, i.e. for every formula `f` occurring in `Constraints` such that

$$\texttt{vars(f)} \subseteq \texttt{dom(Assignment)}$$

holds, the formula `f` evaluates to `true` given the `Assignment`. The purpose of `is_consistent` is to check, whether the extended assignment

$$\texttt{NA} := \texttt{Assignment} \cup \{\langle \texttt{var}, \texttt{value} \rangle\}$$

that assigns `value` to the variable `var` is still partially consistent with `Constraints`. To this end, the `for`-loop iterates over all `Formula` in `Constraints`. However, we only have to check those `Formula` that contain the variable `var` and, furthermore, have the property that

$$\texttt{Vars(Formula)} \subseteq \texttt{dom(NA)},$$

i.e. all variables occurring in `Formula` need to have a value assigned in `NA`. The reasoning is as follows:

1. If `var` does not occur in `Formula`, then adding `var` to `Assignment` cannot change the result of evaluating `Formula` and as `Assignment` is assumed to be partially consistent with respect to `Formula`, `NA` is also partially consistent with respect to `Formula`.
2. If $\texttt{dom(NA)} \not\subseteq \texttt{Vars(Formula)}$, then `Formula` can not be evaluated anyway.

If we use backtracking, we can solve the 8 queens problem in less than a second.

3.4 Constraint Propagation

Once we choose a value for a variable, this choice influences the values that are still available for other variables. For example, suppose that in order to solve the n queens problem we place the queen in row 1 in the second column, then no other queen can be placed in that column. Additionally, the queen in row 2 can then not be placed in any of the first three columns. Abstractly, constraint propagation works as follows.

1. Before the search is started, we create a dictionary `ValuesPerVar`. For every variable `x`, the set

   `ValuesPerVar[x]`

   contains those values that can be assigned to `x` without violating a constraint.
2. As long as the problem is not solved, we choose a variable `x` that has not been assigned a value yet. This variable is chosen using the most constrained variable heuristic: We choose a variable `x` such that the number of values in the set

   `ValuesPerVar[x]`

   is minimal. This is done because we have to find values for all variables. If the current partial variable assignment can not be completed into a solution, then we want to find out this fact as soon as possible. Therefore, we try to find the values for the most difficult variables first. A variable is more difficult to get right if it has only a few values left that can be used to instantiate it.
3. Once we have picked a variable `x`, we next iterate over all values `v` in `ValuesPerVar[x]`. If assigning the value `v` to the variable `x` does not violate a constraint, we propagate the consequences of this assignment:
   (a) For every constraint `F` that mentions only the variable `x` and one other variable `y` that has not yet been instantiated, we compute the set `Legal` of those values from `ValuesPerVar[y]` that can be assigned to `y` without violating the constraint `F`.
   (b) Then, the set `ValuesPerVar[y]` is updated to the set `Legal` and we go back to step 2.

As described above, this method performs binary constraint propagation, because we only use binary constraints to restrict the values of variables. More advanced methods of constraint propagation are discussed in section 3.5.

It turns out that elaborating the idea outline above can enhance the performance of backtracking search considerably. Figure 3.10 on page 54 shows an implementation of constraint propagation. This implementation is also able to handle unary constraints, i.e. constraints that contain only a single variable.

```
1  solve := procedure(CSP) {
2      [Variables, Values, Constrs] := CSP;
3      ValuesPerVar := { [v, Values] : v in Variables };
4      Annotated    := { [f, collectVars(f)] : f in Constrs };
5      UnaryConstrs := { [f, V] : [f, V] in Annotated | #V == 1 };
6      OtherConstrs := { [f, V] : [f, V] in Annotated | #V >= 2 };
7      check {
8          for ([f, V] in UnaryConstrs) {
9              var               := arb(V);
10             ValuesPerVar[var] := solve_unary(f, var, ValuesPerVar[var]);
11         }
12         return bt_search({}, ValuesPerVar, OtherConstrs);
13     }
14 };
```

Figure 3.10: Constraint Propagation.

In order to implement constraint propagation, it is necessary to administer the values that can be used to instantiate the different variables separately, i.e. for every variable `x` we need to know which values are admissible for `x`. To this end, we need a dictionary `ValuesPerVar` that contains the set of possible values for every variable `x`. Initially, this dictionary assigns the set `Values` to every variable. Next, we take care of the unary constraints and shrink these sets so that the unary constraints are satisfied. Then, whenever we assign a value to a variable `x`, we inspect the binary constraints mentioning the variable `x` and shrink the set of values `ValuesPerVar`[`y`] that can be assigned to those variables `y` that are constrained by the variable `x`.

1. The procedure `solve` receives a `CSP`. This `CSP` is first split into its three components.
2. The first task of `solve` is to create the dictionary `ValuesPerVar`. Given a variable `v`, this dictionary assigns the set of values that can be used to instantiate this variable. Initially, this set is the same for all variables and is equal to `Values`.
3. Next, for every constraint `f`, the dictionary `Annotated` associates the set of variables occurring in this constraint. This dictionary is only needed for efficiency: We do not want to recompute the variables of a constraint every time the set of variables is needed.
4. In order to solve the unary constraints we first have to find them. The set `UnaryConstrs` contains all those pairs [`f`, `V`] from the set of annotated constraints `Annotated` such that the set of variables `V` contains just a single variable.
5. Similarly, the set `OtherConstrs` contains those constraints that involve two or more variables.
6. In order to solve the unary constraints, we iterate over all unary constraints and shrink the set of values associated with the variable occurring in the constraint as dictated by the constraint. This is done using the function `solve_unary`.
7. Then, we start backtracking search using the function `bt_search`.

```
solve_unary := procedure(f, var, Values) {
    Legal := {value : value in Values | eval_constraint({[var,value]}, f)};
    if (Legal == {}) { backtrack; } // game over
    return Legal;
};
```

Figure 3.11: Implementation of `solve_unary`.

The function `solve_unary` shown in Figure 3.11 on page 55 takes a unary constraint `f`, a variable `var` and the set of values `Values` that can be assigned to this variable. It returns the subset of values that can be substituted for the variable `var` without violating the given constraint `f`.

1. To achieve its goal, `solve_unary` iterates over all possible `Values`.
2. Next, for every `value` in the set `Values`, an assignment is created that assign the `value` to the variable `var`. This assignment has the form

   `{[var, value]}`.
3. The set `Legal` collects all those values that satisfy the constraint `f`.
4. If the set `Legal` is empty, then the unary constraint `f` is unsatisfiable and hence the Csp is unsolvable. In that case, the Csp is unsolvable. Hence, an exception is raised which is caught in the procedure `solve`. After catching the exception, the procedure `solve` returns the undefined value Ω to signal the given Csp has no solution.
5. Otherwise, the set `Legal` is returned.

```
bt_search := procedure(Assignment, ValuesPerVar, Constraints) {
    if (#Assignment == #ValuesPerVar) {
        return Assignment;
    }
    x := most_constrained_variable(Assignment, ValuesPerVar);
    for (val in ValuesPerVar[x]) {
        check {
            if (is_consistent(x, val, Assignment, Constraints)) {
                NewVals := propagate(x, val, Assignment, Constraints, ValuesPerVar);
                return bt_search(Assignment + {[x, val]}, NewVals, Constraints);
            }
        }
    }
    backtrack;
};
```

Figure 3.12: Implementation of `bt_search`.

The procedure `bt_search` shown in Figure 3.12 on page 55 is called with a partial `Assignment` that is guaranteed to be consistent, a set of `Variables`, a dictionary `ValuesPerVar` associating every variable with the set of values that are admissible for this variable and a set of annotated `Constraints`. It tries to complete `Assignment` and thereby compute a solution of the `CSP`.

1. If the partial `Assignment` is already complete, i.e. if it assigns a value to every variable, then a solution to the given `CSP` has been found and this solution is returned. As the dictionary `ValuesPerVar` has an entry

for every variable, its size is the same as the number of variables. Therefore, `Assignment` is complete iff it has the same size as `ValuesPerVar`.

2. Otherwise, we choose a variable `x` such that the number of values that can still be used to instantiate `x` is minimal. This variable is computed using the procedure `most_constrained_variable` that is shown in Figure 3.13 on page 56.

3. Next, all values that are still available for `x` are tried. Note that since `ValuesPerVar[x]` is, in general, smaller than the set of all values of the `CSP`, the `for`-loop in this version of backtracking search is more efficient than the version of backtracking search discussed in the previous section.

4. If it is consistent to assign `val` to the variable `x`, we propagate the consequences of this assignment using the procedure `propagate` shown in Figure 3.14 on page 57. This procedure updates the values that are still allowed for the variables of the `CSP` once the value `val` has been assigned to the variable `x`.

5. Finally, the partial variable `Assignment` is updated to include the assignment of `val` to `x` and the recursive call to `bt_search` tries to complete this new assignment.

```
1  most_constrained_variable := procedure(Assignment, ValuesPerVar) {
2      Unassigned := { [x, U] : [x, U] in ValuesPerVar | Assignment[x] == om };
3      minSize    := min({ #U : [x, U] in Unassigned });
4      return rnd({ x : [x, U] in Unassigned | #U == minSize });
5  };
```

Figure 3.13: Finding a most constrained variable.

Figure 3.13 on page 56 show the implementation of the procedure `most_constrained_variable`. The procedure `most_constrained_variable` takes a partial `Assignment` and a dictionary `ValuesPerVar` returning for all variables the set of values that are still admissible for this variable.

1. First, this procedure computes the set of `Unassigned` variables. For every variable `x` that has not yet been assigned a value in `Assignment` this set contains the pair [`x`, `U`], where `U` is the set of admissible values for the variable `x`.

2. Next, it computes from all unassigned variables `x` the number of values `#U` that can be assigned to `x`. From these numbers, the minimum is computed and stored in `minSize`.

3. Finally, the set of variables that are maximally constrained is computed. These are all those variables `x` such that `ValuesPerVar[x]` has size `minSize`. From these variables, a random variable is returned.

   The logic behind choosing a maximally constrained variables is that these variables are the most difficult to get right. If we have a partial assignment that is inconsistent, then we will discover this fact earlier if we try the most difficult variables first. This might save us a lot of unnecessary backtracking.

The function `propagate` shown in Figure 3.14 on page 57 takes the following inputs:

(a) `x` is a variable and `v` is a value that is assigned to the variable `x`. The purpose of the function `propagate` is to restrict the values of variables different from `x` that result from setting `x` to the value `v`.

(b) `Assignment` is a partial assignment that contains some assignments for variables that are different from `x`.

(c) `Constraints` is a set of annotated constraints, i.e. this set contains pairs of the form `[F, Vars]`, where `F` is a constraint and `Vars` is the set of variables occurring in `F`.

(d) `ValuesPerVar` is a dictionary assigning sets of values to all variables.

The function `propagate` updates the dictionary `ValuesPerVar` by taking into account the consequences of assigning the value `v` to the variable `x`. The implementation of `propagate` proceeds as follows.

```
1  propagate := procedure(x, v, Assignment, Constraints, ValuesPerVar) {
2      ValuesPerVar[x] := { v };
3      BoundVars := domain(Assignment);
4      for ([F, Vars] in Constraints | x in Vars) {
5          UnboundVars := Vars - BoundVars - { x };
6          if (#UnboundVars == 1) {
7              y := arb(UnboundVars);  // y is the remaining unbound variable
8              Legal := { w : w in ValuesPerVar[y]
9                           | eval_constraint({[x,v], [y,w]} + Assignment, F)
10                     };
11             if (Legal == {}) { backtrack; }
12             ValuesPerVar[y] := Legal;
13         }
14     }
15     return ValuesPerVar;
16 };
```

Figure 3.14: Constraint Propagation.

1. As `x` is assigned the value `v`, the corresponding entry in the dictionary `ValuesPerVar` is changed accordingly.
2. `BoundVars` is the set of those variable that already have a value assigned.
3. Next, `propagate` iterates over all constraints `F` such that the variable `x` occurs in `F`.
4. `UnboundVars` is the set of those variables occurring in `F` that are different from `x` and that do not yet have a value assigned. These variables are called *unbound variables* since we still need to assign values for these variables.
5. If there is exactly one unbound variable `y` in the constraint `F`, then the possible values for this variable will be restricted by `F`. Hence, in this case we need to recompute the set `ValuesPerVar[x]`.
6. As the set `UnboundVars` contains just a single variable in line 7, the function `arb` returns this variable.
7. In order to recompute the set `ValuesPerVar[x]`, all values `w` in `ValuesPerVar[y]` are tested. The set `Legal` contains all values `w` that can be assigned to the variable `y` without violating the constraint `F`.
8. If it turns out that `Legal` is the empty set, then this means that the constraint `F` is inconsistent with assigning the value `v` to the variable `x`. Hence, in this case the search has to `backtrack`.
9. Otherwise, the set of admissible values for `y` is updated to be the set `Legal`.
10. Finally, the updated dictionary `ValuesPerVar` is returned.

I have tested the program described in this section using the n queens puzzle. The procedure is able to solve the 32 queens problem, taking about 4 seconds on average, while the version of backtracking search that does not use constraint propagation took more than 3 minutes on average for solving the 32 queens problem.

Exercise 6: There are many different versions of the *zebra puzzle*. The version below is taken from *Wikipedia*. The puzzle reads as follows:

1. There are five houses.
2. The Englishman lives in the red house.
3. The Spaniard owns the dog.

4. Coffee is drunk in the green house.
5. The Ukrainian drinks tea.
6. The green house is immediately to the right of the ivory house.
7. The Old Gold smoker owns snails.
8. Kools are smoked in the yellow house.
9. Milk is drunk in the middle house.
10. The Norwegian lives in the first house.
11. The man who smokes Chesterfields lives in the house next to the man with the fox.
12. Kools are smoked in the house next to the house where the horse is kept.
13. The Lucky Strike smoker drinks orange juice.
14. The Japanese smokes Parliaments.
15. The Norwegian lives next to the blue house.
16. Who drinks water?
17. Who owns the zebra?

In order to solve the puzzle, we also have to know the following facts:

1. Each of the five houses is painted in a different colour.
2. The inhabitants of the five houses are of different nationalities,
3. they own different pets,
4. they drink different beverages, and
5. they smoke different brands of cigarettes.

Formulate the zebra puzzle as a constraint satisfaction problem and solve the puzzle using the program discussed in this section. You should also try to solve the puzzle using the constraint solver discussed in the previous section that uses only backtracking. Compare the results. A framework for solving the zebra puzzle can be found at

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/zebra-csp.stlx

When coding the zebra puzzle as a Csp you should be careful to name the variables correctly. A single spelling error in one of the constraints can easily result in an unsolvable problem. ◇

Exercise 7: Table 3.1 on page 59 shows a sudoku that I have taken from the Zeit Online magazine. Write a program that transforms this sudoku into a constraint satisfaction problem and solve the resulting Csp using the constraint satisfaction solver developed in this section. A framework for solving sudoku puzzles can be found at:

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/sudoku-csp.stlx

The data describing the sudoku shown in Table 3.1 is given in the file:

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/sudoku-data.stlx ◇

	3	9						7
			7			4	9	2
				6	5		8	3
			6		3	2	7	
				4		8		
5	6							
		5	2		9			1
	2	1					4	
7						5		

Table 3.1: A super hard sudoku from the magazine "Zeit Online".

3.5 Consistency Checking

So far, the constraints in the constraints satisfaction problems discussed are either unary constraints or binary constraints: A unary constraint is a constraint `f` such that the formula `f` contains only one variable, while a binary constraint contains two variables. If we have a constraint satisfaction problem that involves also constraints that mention more than two variables, then the constraint propagation shown in the previous section is not as effective as it is only used for a constraint `f` if all but one variable of `f` have been assigned. For example, consider the cryptarithmetic puzzle shown in Figure 3.15 on page 59. The idea is that the letters "`S`", "`E`", "`N`", "`D`", "`M`", "`O`", "`R`", "`Y`" are interpreted as variables ranging over the set of decimal digits, i.e. these variables can take values in the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Then, the string "`SEND`" is interpreted as a decimal number, i.e. it is interpreted as the number

$$\texttt{S} \cdot 10^3 + \texttt{E} \cdot 10^2 + \texttt{N} \cdot 10^1 + \texttt{D} \cdot 10^0.$$

The strings "`MORE` and "`MONEY`" are interpreted similarly. To make the problem interesting, the assumption is that different variables have different values. Furthermore, the digits at the beginning of a number should be different from 0.

Figure 3.15: A cryptarithmetic puzzle

A naïve approach to solve this problem would be to code it as a constraint satisfaction problem that has, among others, the following constraint:

$$(\texttt{S} \cdot 10^3 + \texttt{E} \cdot 10^2 + \texttt{N} \cdot 10 + \texttt{D}) + (\texttt{M} \cdot 10^3 + \texttt{O} \cdot 10^2 + \texttt{R} \cdot 10 + \texttt{E}) = \texttt{M} \cdot 10^4 + \texttt{O} \cdot 10^3 + \texttt{N} \cdot 10^2 + \texttt{E} \cdot 10 + \texttt{Y}.$$

The problem with this constraint is that it involves far too many variables. As this constraint can only be checked when all the variables have values assigned to them, the backtracking search would essentially boil down to a mere brute force search. We would have 8 variables and hence we would have to test 8^{10} possible assignments. In order to do better, we have to perform the addition in Figure 3.15 column by column, just as it is taught in elementary school. Figure 3.16 on page 60 shows how this can be implemented in SetlX.

Notice that we have introduced three additional variables "`C1`", "`C2`", "`C3`". These variables serve as the carry digits. For example, "`C1`" is the carry digit that we get when we do the addition of the last places of the two numbers, i.e. we have

$$\texttt{D} + \texttt{E} = \texttt{C1} \cdot 10 + \texttt{Y}.$$

```
createCSP := procedure() {
    Variables   := {"S", "E", "N", "D", "M", "O", "R", "Y", "C1", "C2", "C3"};
    Values      := { 0 .. 9 };
    Constraints := allDifferent({ "S", "E", "N", "D", "M", "O", "R", "Y" });
    Constraints += { '(D + E) % 10 == Y', '(D + E) \ 10 == C1',
                     '(N + R + C1) % 10 == E', '(N + R + C1) \ 10 == C2',
                     '(E + O + C2) % 10 == N', '(E + O + C2) \ 10 == C3',
                     '(S + M + C3) % 10 == O', '(S + M + C3) \ 10 == M'
                   };
    Constraints += { "S != 0", "M != 0" };
    return [Variables, Values, Constraints];
};
```

Figure 3.16: Formulating "SEND + MORE = MONEY" as a CSP.

This equation still contains four variables. We can reduce it to two equations that each involve only three variables as follows:

$$(\texttt{D} + \texttt{E})\ \%\ 10 = \texttt{Y} \quad \text{and} \quad (\texttt{D} + \texttt{E})\ \backslash\ 10 = \texttt{C1}.$$

Here, the symbol "\" denotes integer division, e.g. we have $7\backslash 3 = 2$. If we try to solve the cryptarithmetic puzzle as coded in Figure 3.16 on page 60 using the constraint solver developed in the previous section, we will be disappointed: Solving the puzzle takes about 17 seconds on my computer. The reason is that most constraints involve either three or four variables and therefore the effects of constraint propagation kick only in when many variables have already been initialized. However, we can solve the problem in a few seconds if we add the following constraints for the variables "C1", "C2", "C3":

$$\texttt{"C1 < 2", "C2 < 2", "C3 < 2"}.$$

Although these constraints are certainly true, the problem with this approach is that we would prefer if our constraint solver could figure out these constraints by itself. After all, since D and E are both less than 10, there sum is obviously less than 20 and hence the carry C1 has to be less than 2. This line of reasoning is known as consistency maintenance: Assume that the formula f is a constraint and the set of variables occurring in f has the form

$$\texttt{Var}(f) = \{x\} \cup R \quad \text{where } x \notin R,$$

i.e. the variable x occurs in the constraint f and, furthermore, R is the set of all variables occurring in f that are different from x. Furthermore, assume that we have a dictionary ValuesPerVar such that for every variable y, the dictionary entry $\texttt{ValuesPerVar}[y]$ is the set of values that can be substituted for the variable y. Now a value v is consistent for x with respect to the constraint f iff the partial assignment $\{\langle x, v\rangle\}$ can be extended to an assignment A satisfying the constraint f, i.e. for every variable $y \in R$ we have to find a value $w \in \texttt{ValuesPerVar}[y]$ such that the resulting assignment A satisfies the equations

$$\texttt{evaluate}(f, A) = \texttt{true}.$$

Here, the function evaluate takes a formula f and an assignment A and evaluates f using the assignment A. Now, consistency maintenance works as follows.

1. The dictionary ValuesPerVar is initialized as follows:

   $$\texttt{ValuesPerVar}[x] := \texttt{Values} \quad \text{for all } x \in \texttt{Variables},$$

   i.e. initially every variable x can take any value from the set of Values.

2. Next, the set UncheckedVariables is initialized to the set of all Variables:

   $$\texttt{UncheckedVariables} := \texttt{Variables}.$$

3. As long as the set `UncheckedVariables` is not empty, we remove one variable x from this set:

   $$x := \texttt{from}(\texttt{UncheckedVariables})$$

4. We iterate over all constraints f such that x occurs in f.

   (a) For every value $v \in \texttt{ValuesPerVar}[x]$ we check whether v is consistent with f.

   (b) If v is not consistent with f, then v is removed from $\texttt{ValuesPerVar}[x]$. Furthermore, if R is the set of all variables occurring in f that are different from x, then this set R is added to the set of `UncheckedVariables`.

5. Once `UncheckedVariables` is empty, the algorithm terminates. Otherwise, we jump back to step 3 and remove the next variable from the set `UncheckedVariables`.

The algorithm terminates as every iteration removes either a variable from the set `UncheckedVariables` or it removes a value from one of the sets $\texttt{ValuesPerVar}[y]$ for some variable y. Although the set `UncheckedVariables` can grow during the algorithm, the union

$$\bigcup_{x \in \texttt{Vars}} \texttt{ValuesPerVar}[x]$$

can never grow: Every time the set `UncheckedVariables` grows, for some variable x the set $\texttt{ValuesPerVar}[x]$ shrinks. As the sets $\texttt{ValuesPerVar}[x]$ are finite for all variables x, the set `UncheckedVariables` can only grow a finite number of times. Once the set `UncheckedVariables` does not grow any more, every iteration of the algorithm removes one variable from this set and hence the algorithm terminates eventually.

```
1  enforceConsistency := procedure(rw ValuesPerVar, Var2Formulas, Annotated, Connected) {
2      UncheckedVars := domain(Var2Formulas);
3      while (UncheckedVars != {}) {
4          variable    := from(UncheckedVars);
5          Constraints := Var2Formulas[variable];
6          Values      := ValuesPerVar[variable];
7          RemovedVals := {};
8          for (f in Constraints) {
9              OtherVars := Annotated[f] - { variable };
10             for (value in Values) {
11                 if(!existValues(variable, value, f, OtherVars, ValuesPerVar)) {
12                     RemovedVals   += { value };
13                     UncheckedVars += Connected[variable];
14                 }
15             }
16         }
17         Remaining := Values - RemovedVals;
18         if (Remaining == {}) { backtrack; }
19         ValuesPerVar[variable] := Remaining;
20     }
21 };
```

Figure 3.17: Consistency maintenance in SetlX.

Figure 3.17 on page 61 shows how consistency maintenance can be implemented in SetlX. The procedure `enforceConsistency` takes four arguments.

(a) `ValuesPerVar` is a dictionary associating the set of possible values with each variable.

(b) `Var2Formulas` is a dictionary. For every variable v, $\texttt{Var2Formulas}[v]$ is the set of those constraints f such that v occurs in f.

(c) `Annotated` is the set of annnotated constraints, i.e. this set contains pairs of the form $\langle f, V \rangle$ where f is a constraint and V is the set of all variables occurring in f. This argument is needed only for efficiency: In order to avoid computing the set V of variables occurring in a constraint f every time the constraint f is encountered, we compute these sets at the beginning of our computation and store them in the dictionary `Annnotated`.

(d) `Connected` is a dictionary that takes a variable x and returns the set V of all variables that are related to x via a common constraint f, i.e. we have $y \in$ `Connected`$[x]$ if there exists a constraint f such that both x and y occur in f and, furthermore, $x \neq y$.

The procedure `enforceConsistency` modifies the dictionary `ValuesPerVar` so that once the procedure has terminated, for every variable x the set `ValuesPerVar`$[x]$ is consistent with the constraints for x. The implementation works as follows:

1. Initially, all variables need to be checked for consistency. Therefore, `UncheckedVars` is defined to be the set of all variables that occur in any of the constraints.
2. The `while`-loop iterates as long as there are still variables x left in `UncheckedVars` such that the consistency of `ValuesPerVar`$[x]$ has not been established.
3. Next, a `variable` is selected and removed from `UncheckedVars`.
4. `Constraints` is the set of all constraints f such that this `variable` occurs in f.
5. `Values` is the set of those values that can be assigned to `variable`.
6. `RemovedVals` is the subset of those values that are found to be inconsistent with some constraint.
7. We iterate over all constraints `f` $\in$ `Constraints`.
8. `OtherVars` is the set of variables occurring in `f` that are different from the chosen `variable`.
9. We iterate over all `value` $\in$ `Values` that can be substituted for `variable` and check whether `value` is consistent with `f`. To this end, we need to find values that can be assigned to the variables in the set `OtherVars` such that `f` evaluates as `true`. This is checked using the function `existValues`.
10. If we do not find such values, then `value` is inconsistent for `variable` w.r.t. `f` and needs to be removed from the set `ValuesPerVar`[`variable`]. Furthermore, all variables that are connected to `variables` have to be added to the set `UncheckedVars`. The reason is that once a value is removed for `variable`, the value assigned to another variable y occurring in a constraint that mentions both `variable` and y might now become inconsistent.
11. If there are no consistent values for `variable` left, we have to backtrack.

Figure 3.18 on page 63 shows the implementation of the function `existValues` that is used in the implementation of `enforceConsistency`. This procedure is called with five arguments.

(a) `var` is variable.

(b) `val` is a value that is to be assigned to `var`.

(c) `f` is a constraint such that `var` occurs in `f`

(d) `Vars` is the set of all those other variables occurring in `f`, i.e. the set of those variables that occur in `f` but that are different from `var`.

(e) `ValuesPerVar` is a dictionary associating the set of possible values with each variable.

The procedure checks whether the partial assignment $\{\langle$`var`$,$`val`$\rangle\}$ can be extended so that the constraint `f` is satisfied. To this end it needs to create the set of all possible assignments. This set is generated using the function `createAllAssignments`. This function gets a set of variables `Vars` and a dictionary that assigns to every variable `var` in `Vars` the set of values that might be assigned to `var`.

```
existValues := procedure(var, val, f, Vars, ValuesPerVar) {
    return exists( A in createAllAssignments(Vars, ValuesPerVar)
                   | eval_constraint(A + { [var, val] }, f)        );
};
createAllAssignments := procedure(Vars, ValuesPerVar) {
    if (Vars == {}) {
        return { {} };  // set containing empty assignment
    }
    var         := from(Vars);
    Values      := ValuesPerVar[var];
    Assignments := createAllAssignments(Vars, ValuesPerVar);
    return { { [var, val] } + A : val in Values, A in Assignments };
};
```

Figure 3.18: The implementation of `existValues`.

1. If the set of variables `Vars` is empty, the empty set can serve as a dictionary that assigns a value to every variable in `Vars`.
2. Otherwise, we remove a variable `var` from `Vars` and get the set of `Values` that can be assigned to `var`.
3. Recursively, we create the set of all `Assignments` that associate values with the remaining variables.
4. Finally, the set of all possible assignments is the set of all combinations of assigning a value `val` $\in$ `Values` to `var` and assigning the remaining variables according to an assignment `A` $\in$ `Assignments`.

On one hand, consistency checking creates a lot of overhead.[2] Therefore, it might actually slow down the solution of some constraint satisfaction problems that are easy to solve using just backtracking and propagation. On the other hand, many difficult constraint satisfaction problems can not be solved without consistency checking.

3.6 Local Search

There is another approach to solve constraint satisfaction problems. This approach is known as local search. The basic idea is simple: Given as constraint satisfaction problem $\mathcal{C}$ of the form

$$\mathcal{P} := \langle \texttt{Variables}, \texttt{Values}, \texttt{Constraints} \rangle,$$

local search works as follows:

1. Initialize the values of the variables in `Variables` randomly.
2. If all `Constraints` are satisfied, return the solution.
3. For every $x \in$ `Variables`, count the number of unsatisfied constraints that involve the variable x.
4. Set `maxNum` to be the maximum of these numbers, i.e. `maxNum` is the maximal number of unsatisfied constraints for any variable.
5. Compute the set `maxVars` of those variables that have `maxNum` unsatisfied constraints.
6. Randomly choose a variable x from the set `maxVars`.

[2] To be fair, the implementation shown in this section is far from optimal. In particular, by remembering which combinations of variables and values work for a given formula, the overhead can be reduced significantly. I have refrained from implementing this optimization because I did not want the code to get too complex.

7. Find a value $d \in$ `Values` such that by assigning d to the variable x, the number of unsatisfied constraints for the variable x is minimized.

   If there is more than one value d with this property, choose the value d randomly from those values that minimize the number of unsatisfied constraints.

8. Goto step 2 and repeat until either a solution is found or the sun rises in the west.

```
1  solve := procedure(n) {
2      Queens := [];
3      for (row in [1 .. n]) {    // randomly place n queens
4          Queens[row] := rnd({1 .. n});
5      }
6      iteration := 0;
7      lastRow   := 0;
8      while (n > 3) {  // there is no solution for n <= 3
9          // printBoard(Queens);
10         Conflicts := { [numConflicts(Queens, row), row] : row in {1..n} - {lastRow} };
11         [maxNum, _] := last(Conflicts);  // select queen with most conflicts
12         if (maxNum == 0) { // no conflicts, solution found
13             return Queens;
14         }
15         if (iteration % 30 != 0) { // avoid infinite loops
16             row := rnd({ row : [num,row] in Conflicts | num == maxNum });
17         } else {
18             row := rnd({ 1 .. n });
19         }
20         lastRow   := row;
21         Conflicts := {};
22         for (col in [1 .. n]) {
23             Board      := Queens;
24             Board[row] := col;
25             Conflicts  += { [numConflicts(Board, row), col] };
26         }
27         [minNum, _] := first(Conflicts);
28         newColumn   := rnd({ col : [num, col] in Conflicts | num == minNum });
29         Queens[row] := newColumn;
30         iteration += 1;
31     }
32 };
```

Figure 3.19: Solving the n queens problem using local search.

Figure 3.19 on page 64 shows an implementation of these ideas in SetlX. Instead of solving an arbitrary constraint satisfaction problem, the program solves the n queens problem. We proceed to discuss this program line by line.

1. The procedure `solve` takes one parameter `n`, which is the size of the chess board. If the computation is successful, `solve(n)` returns a list of length `n`. Lets call this list `Queens`. For every row `r` $\in \{1, \cdots, n\}$, the value `Queens[r]` specifies that the queen that resides in row `r` is positioned in column `Queens[r]`.

2. The `for` loop initializes the positions of the queens to random values from the set $\{1, \cdots, n\}$. Effectively, for every row on the chess board, this puts a queen in a random column.

3. The variable `iteration` counts the number of times that we need to reassign a queen in a given row.

4. The variable `lastRow` remembers the row that was changed in the last iteration. We have to remember this value as it would not make sense to reposition a queen in the same row twice without changing any other queen in between.

5. All the remaining statements are surrounded by a `while` loop that is only terminated once a solution has been found. The `while` loop checks that `n` is bigger than 3 since otherwise the problem is unsolvable and local search can not figure out that a problem is unsolvable.

6. The variable `Conflicts` is a set of pairs of the form $[c, r]$, where c is the number of times the queen in row r is attacked by other queens. Hence, c is the same as the number of unsatisfied constraints for the variable specifying the column of the queen in row r.

7. `maxNum` is the maximum of the number of conflicts for any row. The computation of `maxNum` rests on the fact that sets are ordered in SetlX and therefore the last element of the set `Conflicts` contains the pair $[c, r]$ such that the number of conflicts is maximal.

8. If this number is 0, then all constraints are satisfied and the list `Queens` is a solution to the `n` queens problem.

9. Otherwise, we compute those rows that exhibit the maximal number of conflicts. From these rows we select one `row` randomly.

10. The reason for enclosing the assignment to `row` in an `if` statement is explained later. On a first reading of this program, this `if` statement should be ignored.

11. Now that we have identified the `row` where the number of conflicts is biggest, we need to reassign `Queens[row]`. Of course, when reassigning this variable, we would like to have fewer conflicts after the reassignment. Hence, we test all columns to find the best column that can be assigned for the queen in the given `row`. This is done in a `for` loop that runs over all possible columns. The set `Conflicts` that is maintained in this loop is a set of pairs of the form $[k, c]$ where k is the number of times the queen in `row` would be attacked if it would be placed in column c.

12. We compute the minimum number of conflicts that is possible for the queen in `row` and assign it to `minNum`.

13. From those columns that minimize the number of violated constraints, we choose a column randomly and assign it for the specified `row`.

There is a technical issue, that must be addressed: It is possible there is just one row that exhibits the maximum number of conflicts. It is further possible that, given the placements of the other queens, there is just one optimal column for this row. In this case, the procedure `solve` would loop forever. To avoid this case, every 30 iterations we pick a random row to change.

The procedure `numConficts` shown in Figure 3.20 on page 66 implements the function `numConficts`. Given a board `Queens` that specifies the positions of the queens on the board and a `row`, this function computes the number of ways that the queen in `row` is attacked by other queens. If all queens are positioned in different rows, then there are only three ways left that a queen can be attacked by another queen.

1. The queen in row `r` could be positioned in the same column as the queen in `row`.

2. The queen in row `r` could be positioned in the same falling or rising diagonal as the queen in `row`. These diagonals are specified in line 6 of Figure 3.20.

Using the program discussed in this section, the n queens problem can be solved for a $\texttt{n} = 1000$ in 30 minutes. As the memory requirements for local search are small, even much higher problem sizes can be tackled if sufficient time is available. Hence, it seems that local search is much better than the algorithms discussed previously. However, we have to note that local search is incomplete: If a constraint satisfaction problem $\mathcal{P}$ has no solution, then local search loops forever. Therefore, in practise a dual approach is used to solve a constraint satisfaction problem. The constraint solver starts two threads: The first search does local search, the second thread tries

```
numConflicts := procedure(Queens, row) {
    n      := #Queens;
    result := 0;
    for (r in {1 .. n} | r != row) {
        if ( Queens[r] == Queens[row]                      ||
             abs(row - r) == abs(Queens[row] - Queens[r])
           )
        { result += 1; }
    }
    return result;
};
```

Figure 3.20: The procedure `numConflicts`.

to solve the problem via some refinement of backtracking. The first thread that terminates wins. The resulting algorithm is complete and, for a solvable problem, will have a performance that is similar to the performance of local search. If the problem is unsolvable, this will eventually be discovered by backtracking. Note, however, that the constraint satisfaction problem is NP-complete. Hence, it is unlikely that there is an efficient algorithm that works always. However, today many practically relevant constraint satisfaction problems can be solved in a reasonably short time.

Chapter 4

Playing Games

One major breakthrough for the field of artificial intelligence happened in 1997 when the chess-playing computer Deep Blue was able to beat the World Chess Champion Garry Kasparov by $3\frac{1}{2} - 2\frac{1}{2}$. While Deep Blue was based on special hardware, according to the computer chess rating list of the 24th of March 2018, the chess program Stockfish runs on ordinary desktop computers and has an Elo rating of 3445. To compare, according to the Fide list of March 2018, the current World Chess Champion Magnus Carlsen has an Elo rating of just 2843. Hence, he wouldn't stand a chance to win a game against Stockfish. In 2017, at the Future of Go Summit, the computer program AlphaGo was able to beat Ke Jie, who is currently[1] considered to be the best human Go player in the world. Besides Go and chess, there are many other games where today the performance of a computer exceeds the performance of human players. To name just one more example, at the beginning of 2017 the program Libratus was able to beat four professional poker players resoundingly.

In this chapter we want to investigate how a computer can play a game. To this end we define a game $\mathcal{G}$ as a six-tuple

$$\mathcal{G} = \langle \texttt{States}, s_0, \texttt{Players}, \texttt{nextStates}, \texttt{finished}, \texttt{utility} \rangle$$

where the components are interpreted as follows:

1. `States` is the set of all possible states of the game.

2. $s_0 \in \texttt{States}$ is the start state.

3. `Players` is the list of players of the game. The first element in `Players` is the player to start the game and after that the players take turns. As we only consider two person games, we assume that `Players` is a list of length two.

4. `nextStates` is a function that takes a state $s \in \texttt{States}$ and a player $p \in \texttt{Players}$ and returns the set of states that can be reached if the player p has to make a move in the state s. Hence, the signature of `nextStates` is given as follows:

   $$\texttt{nextStates} : \texttt{States} \times \texttt{Players} \to 2^{\texttt{States}}.$$

5. `finished` is a function that takes a state s and decides whether the games is finished. Therefore, the signature of `finished` is

   $$\texttt{finished} : \texttt{States} \to \mathbb{B}.$$

   Here, $\mathbb{B}$ is the set of Boolean values, i.e. we have $\mathbb{B} := \{\texttt{true}, \texttt{false}\}$.

   Using the function `finished`, we define the set `TerminalStates` as the set of those states such that the game has finished, i.e. we define

   $$\texttt{TerminalStates} := \{s \in \texttt{States} \mid \texttt{finished}(s)\}.$$

[1] This assessment is of March 2018.

6. `utility` is a function that takes a state $s \in$ `TerminalStates` and a player $p \in$ `Players`. It returns the value that the game has for player p. In general, a value is a real number, but in all of our examples, this value will be an element from the set $\{-1, 0, +1\}$. The value -1 indicates that player p has lost the game, if the value is $+1$ the player p has won the game, and if this value is 0, then the game is a draw. Hence the signature of `utility` is

   $$\texttt{utility} : \texttt{TerminalStates} \times \texttt{Players} \rightarrow \{-1, 0, +1\}.$$

In this chapter we will only consider so called two person, zero sum games. This means that the list `Players` has exactly two elements. If we call these players A and B, i.e. if we have

$$\texttt{Players} = [\texttt{A}, \texttt{B}],$$

then the game is called a zero sum game iff we have

$$\forall s \in \texttt{TerminalStates} : \texttt{utility}(s, \texttt{A}) + \texttt{utility}(s, \texttt{B}) = 0,$$

i.e. the losses of player A are compensated by the wins of player `B` and vice versa. Games like Go and chess are two person, zero sum games. We proceed to discuss an example.

4.1 Tic-Tac-Toe

The game tic-tac-toe is played on a square board of size 3×3. On every turn, one player puts an "`X`" on one of the free squares of the board, while the other player puts an "`O`" onto a free square when it is his turn. If the first player manages to place three `X`s in a row, column, or diagonal, she has won the game. Similarly, if the second player manages to put three `O`s in a row, column, or diagonal, this player is the winner. Otherwise, the game is drawn. Figure 4.1 on page 69 shows a SetlX implementation of tic-tac-toe.

1. The function `players` returns the list `Players`. Traditionally, the players in tic-tac-toe are called "`X`" and "`O`". As this function does not take any arguments, you might well ask why we use a function to define the list of players. The reason is that SetlX does not provide global variables. Therefore, the definition

   ```
   players := [ "X", "O" ];
   ```

   would not work, as the variable `players` would be undefined inside of the function `play_game` where it is needed later. This function is defined in Figure 4.2 and will be discussed in Section 4.2.

2. The function `startState` returns the start state, which is an empty board. States are represented as list of lists. The entries in these lists are the characters "`X`", "`O`", and the blank character " ". As the `StartState` is the empty board, it is represented as a list of three lists containing three blanks each:

   ```
   [ [" ", " ", " "],
     [" ", " ", " "],
     [" ", " ", " "]
   ].
   ```

3. The function `nextStates` takes a `State` and a `player` and computes the set of states that can be reached from `State` if `player` is to move next. To this end, it first computes the set of empty positions, i.e. those positions that have not yet been marked by either player.. Every position is represented as pair of the form [`row`, `col`] where `row` specifies the row and `col` specifies the column of the position. The position [`row`, `col`] is empty in `State` iff

   $$\texttt{State}[\texttt{row}][\texttt{col}] = \texttt{" "}.$$

   The computation of the empty position has been sourced out to the function `empty`. The function `nextStates` then iterates over the set of empty positions. For every empty position [`row`, `col`] it creates a new state `NextState` that results from the current `State` by putting the mark of `player` in this position. The resulting states are collected in the set `Result` and returned.

4. The function `empty` takes a state `S` and returns the set of positions that are empty in this state.

```
players    := [] |-> [ "X", "O" ];
startState := [] |-> [ [ " " : col in [1..3]] : row in [1..3] ];
nextStates := procedure(State, player) {
    Empty  := empty(State);
    Result := {};
    for ([row, col] in Empty) {
        NextState           := State;
        NextState[row][col] := player;
        Result              += { NextState };
    }
    return Result;
};
empty := S |-> {[row,col] : row in [1..3], col in [1..3] | S[row][col] == " "};
utility := procedure(State, player) {
    for (Pairs in all_lines()) {
        Marks := { State[row][col] : [row, col] in Pairs };
        if (#Marks == 1 && Marks != { " " }) {
            if (Marks == { player }) { return  1; } else { return -1; }
        }
    }
    if (forall(row in [1..3], col in [1..3] | State[row][col] != " ")) {
        return 0;
    }
};
all_lines := closure() {
    Lines := { { [row, col] : col in [1..3] } : row in [1..3] };
    Lines += { { [row, col] : row in [1..3] } : col in [1..3] };
    Lines += { { [idx, idx] : idx in [1..3] } };
    Lines += { { [1, 3], [2, 2], [3, 1] } };
    return Lines;
};
finished := procedure(State) {
    return utility(State, "X") != om;
};
```

Figure 4.1: A SetlX description of tic-tac-toe.

5. The function `utility` takes a `State` and a `player`. If the game is finished in the given `State`, it returns the value that this `State` has for the current `player`. If the outcome of the game is not yet decided, the undefined value Ω is returned instead.

   In order to achieve its goal, the procedure first computes the set of all sets of coordinate pairs that either specify a horizontal, vertical, or diagonal line on a 3×3 tic-tac-toe board. Concretely, the function `all_lines` returns the following set:

$$\left\{ \begin{array}{l} \{[1,1],[1,2],[1,3]\},\ \{[2,1],[2,2],[2,3]\},\ \{[3,1],[3,2],[3,3]\}, \\ \{[1,1],[2,1],[3,1]\},\ \{[1,2],[2,2],[3,2]\},\ \{[1,3],[2,3],[3,3]\}, \\ \{[1,1],[2,2],[3,3]\},\ \{[3,1],[2,2],[1,3]\} \end{array} \right\}$$

   The first line in this expression gives the set of pairs defining the rows, the second line defines the columns, and the last line yields the diagonals. Given a state `State` and a set `Pairs`, the set

```
Marks := { State[row][col] :  [row, col] in Pairs }
```

is the set of all marks in the line specified by `Pairs`. For example, if

```
Pairs := { [1, 1], [2, 2], [3, 3] },
```

then `Marks` is the set of marks on the falling diagonal. The game is decided if all entries in a set of the form

```
Marks := { State[row][col] :  [row, col] in Pairs }
```

where `Pairs` is a set from `all_lines` either have the value "`X`" or the value "`O`". In this case, the set `Marks` has exactly one element which is different from the blank. If this element is the same as `player`, then the game is won by `player`, otherwise it must be the mark of his opponent and hence the game is lost for him.

Finally, if there are no more empty squares left, then the game is a draw.

6. The auxiliary procedure `all_lines` computes the sets of coordinate pairs that either specify a horizontal, vertical, or diagonal line.
   (a) For any row $\in \{1, 2, 3\}$ the set
   ```
   { [row, col] :  col in [1..3]}
   ```
   forms a horizontal line.
   (b) Likewise, for any col $\in \{1, 2, 3\}$ the set
   ```
   { [row, col] :  row in [1..3]}
   ```
   forms a vertical line.
   (c) Given that the top row is indexed with 1, and the bottom row is row number 3, the set
   ```
   { [idx, idx] :  idx in [1..3] };
   ```
   is the falling diagonal.
   (d) Finally, the set
   ```
   { [3, 1], [2, 2], [1, 3] }
   ```
   is the rising diagonal.
7. The procedure `finished` takes a `State` and checks whether the game is finished. To this end it computes the `utility` of the state for the player "`X`". If this `utility` is different from Ω, the game is finished. Note that it does make no difference whether we take the utility of the state for the player "`X`" or for the player "`O`": If the game is finished for "`X`", then it is also finished for "`O`" and vice versa.

4.2 The Minimax Algorithm

Having defined the notion of a game, our next task is to come up with an algorithm that can play a game. The algorithm that is easiest to explain is the minimax algorithm. This algorithm is based on the notion of the value of a state. To this end, we define a function

$$\texttt{value} : \texttt{States} \times \texttt{Players} \to \{-1, 0, +1\}$$

that takes a state $s \in \texttt{States}$ and a player $p \in \texttt{Players}$ and returns the value of s provided both the player p and his opponent play optimally. The easiest way to define this function is via recursion. The base case is simple:

$$\texttt{finished}(s) \to \texttt{value}(s, p) = \texttt{utility}(s, p).$$

If the game is not yet finished, assume that player o is the opponent of player p. Then we define

$$\neg\texttt{finished}(s) \to \texttt{value}(s, p) = \max\bigl(\bigl\{-\texttt{value}(n, o) \mid n \in \texttt{nextStates}(s, p)\bigr\}\bigr).$$

The reason is that, if the game is not finished yet, the player p has to evaluate all possible moves. From these, the player p will choose the move that maximizes the value of the game for herself. In order to do so, the player p computes the set `nextStates`(s, p) of all states that can be reached from the state s in any one move of the player p. Now if n is a state that results from player p making some move, then it is the turn of the other player o to make a move. Hence, in order to evaluate the state n, we have to call the function `value` recursively as `value`(n, o). Since the gains of the other player o are the losses of the player p, we have to take the negative of `value`(n, o). Figure 4.2 on page 71 shows an implementation of this strategy.

```
1  other := p |-> ([o : o in players() | o != p])[1];
2  value := cachedProcedure(State, player) {  // dynamic programming
3      if (finished(State)) {
4          return utility(State, player);
5      }
6      return max({ -value(ns, other(player)): ns in nextStates(State,player) });
7  };
8  best_move := procedure(State, player) {
9      NS      := nextStates(State, player);
10     bestVal := value(State, player);
11     return [bestVal, rnd({ ns : ns in NS | -value(ns, other(player)) == bestVal })];
12 };
13 play_game := procedure() {
14     State := startState();
15     print(stateToString(State));
16     while (true) {
17         firstPlayer  := players()[1];
18         [val, State] := best_move(State, firstPlayer);
19         print("For me, the game has the value $val$. My move:");
20         print(stateToString(State));
21         if (finished(State)) {
22             final_msg(State);
23             break;
24         }
25         State := getMove(State);
26         print(stateToString(State));
27         if (finished(State)) {
28             final_msg(State);
29             break;
30         }
31     }
32 };
```

Figure 4.2: The Minimax algorithm.

1. Given a player `p`, the function `other` computes the other player, which is the first element of the list of all players that are different from p. This works because we assume that there are just two players. Therefore the list of all players different from the player `p` contains exactly one element.

2. The implementation of the function `value` follows the reasoning outlined above. However, note that we have implemented the function `value` as a `cachedProcedure`, i.e. as a procedure that memorizes its results. Hence, when the function `value` is called a second time with the same pair of arguments, it does not recompute the value but rather the value is looked up in a so called cache that stores all previous results computed by the function `value`. To understand why this is important, let us consider how many

states would be explored in the case of tic-tac-toe if we would not use the idea of memorizing previous results, a technique which is know as memoization. In this case, we have 9 moves for player `X` from the start state, then 8 moves for player `O`, then again 7 moves for player `O`. If we disregard the fact that some games are decided after fewer than 9 moves, the function `value` needs to consider

$$9 \cdot 8 \cdot 7 \cdot \ldots \cdot 2 \cdot 1 = 9! = 362\,880$$

moves. However, if we count the number of possibilities of putting 5 "`O`"s and 4 "`X`"s on a 3×3 board, we see that there are only

$$\binom{9}{5} = \frac{9!}{5! \cdot 4!} = 126$$

possibilities, because we only have to count the number of ways to put 5 "`O`"s on 9 positions and that number is the same as the number of subsets of five elements from a set of nine elements. Therefore, if we disregard the fact that some games are decided after fewer than nine moves, there are a factor of $5! \cdot 4! = 2880$ less terminal states to evaluate if we use memoization!

As we have to evaluate not just terminal states but all states, the saving is actually a bit smaller that 2880. The next exercise explores this in more detail.

3. The function `best_move` takes a `State` and a `player` and returns a pair $\langle v, s \rangle$ where s is a state that is optimal for the `player` and such that s can be reached in one step from `State`. Furthermore, v is the value of this state.

   (a) To this end, it first computes the set `NS` of all states that can be reached from the given `State` in one step if `player` is to move next.

   (b) `bestValue` is the best value that `player` can achieve in the given `State`.

   (c) The function returns randomly one of those states `ns` $\in$ `NS` such that the value of `ns` is optimal, i.e. is equal to `bestValue`. We use randomization here since we want to have more interesting games. If we would always choose the first state that achieves the best value, then our program would always make the same move in a given state. Hence, playing the program would get boring much sooner.

4. The function `play_game` is used to play a game.

   (a) Initially, `State` is the `startState`.

   (b) As long as the game is not finished, the procedure keeps running.

   (c) We assume that the computer goes first and therefore define `firstPlayer` as the first element of the list `players`(). Next, the function `best_move` is used to compute the state that results from the best move of `firstPlayer`.

   (d) After that, it is checked whether the game is finished.

   (e) If the game is not yet finished, the user is asked to make its move via the function `getMove` that takes a `State`, displays it, and asks the user to enter a move. The state resulting from this move is then returned and displayed.

   Note that we do not check the legality of the move entered by the user. This feature can be used for exploration and cheating.

   (f) Next, we have to check whether the game is finished after the move of the user has been executed.

   (g) The `while`-loop keeps iterating until the game is finished. We do not have to put a test into the condition of this `while`-loop as we call the function `finished`(`State`) every time that a new `State` has been reached. If the game is finished, a message giving the result of the game is printed.

In order to better understand the reason for using memoization in the implementation of the function `value` we introduce the following notions.

Definition 5 (Game Tree) Assume that

$$\mathcal{G} = \langle \texttt{States}, s_0, \texttt{Players}, \texttt{nextStates}, \texttt{finished}, \texttt{utility} \rangle$$

is a game. Then a play of length n is a list of states of the form

$$[s_0, s_1, \cdots, s_n] \quad \text{such that} \quad \forall i \in \{0, \cdots, n-1\} : s_{i+1} \in \texttt{nextStates}(s_i, p_i),$$

where the players p_i are defined such that for all even $i \in \{0, \cdots, n-1\}$ we have that p_i is the first element in the list `Players`, while p_i is the second element otherwise. The game tree of the game $\mathcal{G}$ is the set of all possible plays. ◇

The following exercise shows why memoization is important.

Exercise 8: In simplified tic-tac-toe the game only ends when there are no more empty squares left. The player `X` wins if she has more rows, columns, or diagonals of three `X`s than the player `O` has rows, columns, or diagonals of three `O`s. Similarly, the player `O` wins if he has more rows, columns, or diagonals of three `O`s than the player `X` has rows, columns, or diagonals of three `X`s. Otherwise, the game is a draw.

(a) Derive a formula to compute the size of the game tree of simplified tic-tac-toe.

(b) Write a short program to evaluate the formula derived in part (c) of this exercise.

(c) Derive a formula that gives the number of all states of simplified tic-tac-toe.

Notice that this question does not ask for the number of all terminal states but rather asks for all states.

(d) Write a short program to evaluate the formula derived in part (a) of this exercise.

4.3 α-β-Pruning

The efficiency of the minimax algorithm can be improved if we provide two additional arguments to the function `value`. Traditionally, these arguments are called α and β. In order to be able to distinguish between the old function `value` and its improved version, we call the improved version `alphaBeta`. The idea is that the function `alphaBeta` and the function `value` are related by the following requirements:

1. As long as $\texttt{value}(s, p)$ is between α and β, the function `alphaBeta` computes the same result as the function `value`, i.e. we have

   $$\alpha \leq \texttt{value}(s, p) \leq \beta \;\rightarrow\; \texttt{alphaBeta}(s, p, \alpha, \beta) = \texttt{value}(s, p).$$

2. If $\texttt{value}(s, p) < \alpha$, we require that the value returned by `alphaBeta` is less than or equal to α, i.e. we have

   $$\texttt{value}(s, p) < \alpha \;\rightarrow\; \texttt{alphaBeta}(s, p, \alpha, \beta) \leq \alpha.$$

3. Similarly, if $\texttt{value}(s, p) > \beta$, we require that the value returned by `valueAlphaBeta` is bigger than or equal to β, i.e. we have

   $$\beta < \texttt{value}(s, p) \;\rightarrow\; \beta \leq \texttt{alphaBeta}(s, p, \alpha, \beta).$$

Therefore, $\texttt{alphaBeta}(\texttt{State}, \texttt{player})$ is only an approximation of $\texttt{value}(\texttt{State}, \texttt{player})$. However, it turns out that this approximation is all that is needed. Figure 4.3 on page 74 shows an implementation of the function `alphaBeta` that satisfies the specification given above. Once the function `alphaBeta` is implemented, the function `value` can then be computed as

$$\texttt{value}(s, p) := \texttt{alphaBeta}(s, p, -1, +1).$$

The reason is that we already know that $-1 \leq \texttt{value}(s, p) \leq +1$ and hence the first case of the specification of `alphaBeta` guarantees that the equation

$$\texttt{value}(s, p) = \texttt{alphaBeta}(s, p, -1, +1)$$

holds. Since `alphaBeta` is implemented as a recursive procedure, the fact that the implementation of `alphaBeta` shown in Figure 4.3 on page 74 satisfies the specification given above can be established by computational induction. A proof by computational induction can be found in an article by Donald E. Knuth and Ronald W. Moore [KM75].

```
1  alphaBeta := cachedProcedure(State, player, alpha := -1, beta := 1) {
2      if (finished(State)) {
3          return utility(State, player);
4      }
5      val := alpha;
6      for (ns in nextStates(State, player)) {
7          val := max({ val, -alphaBeta(ns, other(player), -beta, -alpha) });
8          if (val >= beta) {
9              return val;
10         }
11         alpha := max({ val, alpha });
12     }
13     return val;
14 };
```

Figure 4.3: α-β-Pruning.

We proceed to discuss the implementation of the function `alphaBeta`.

1. If `State` is a terminal state, the function returns the `utility` of the given `State` with respect to `player`.
2. The variable `val` is supposed to store the maximum of the values of all states that can be reached from the given `State` if `player` makes one move.

   According to the specification of `alphaBeta`, we are not interested in values that are less than `alpha`. Hence, it suffices to initialize `val` with `alpha`. This way, in the case that we have

   $$\texttt{value}(\texttt{State}, \texttt{player}) < \alpha,$$

   instead of returning the true value of the given `State`, the function $\texttt{alphaBeta}(\texttt{State}, \texttt{player}, \alpha, \beta)$ will instead return the value α, which is permitted by its specification.
3. Next, we iterate over all successor states $\texttt{ns} \in \texttt{nextStates}(\texttt{State}, \texttt{player})$.
4. We have to recursively evaluate the states `ns` with respect to the opponent $\texttt{other}(\texttt{player})$. Since the value of a state for the opponent is the negative of the value for `player`, we have to exchange the roles of α and β and prefix them with a negative sign.
5. As the specification of `alphaBeta` ask us to compute the value of `State` only in those cases where it is less than or equal to β, once we find a successor state s that has a value `val` that is at least as big as β we can stop any further evaluation of the successor states and return the value `val`.

   In practice, this shortcut results in significant savings of computation time!
6. Once we have found a successor state that has a value `val` greater than `alpha`, we can increase `alpha` to the value `val`. The reason is, that once we know we can achieve a value of `val` we are no longer interested in any values that are less than `val`. This is the reason for assigning to `alpha` the maximum of `val` and `alpha`.

4.4 Depth Limited Search

In practice, most games are far too complex to be evaluated completely, i.e. the size of the set `States` is so big that even the fastest computer does not stand a chance to explore this set completely. For example, it is

believed[2] that in chess there are about 10^{40} different states that could occur in a reasonable game. Hence, it is impossible to explore all possible states in chess. Instead, we have to limit the exploration in a way that is similar to the way professional players evaluate their game: Usually, a player considers all variations of the game for, say, the next three moves. After a given number of moves, the value of a position is estimated using an evaluation function. This function approximates the true value of a given state via a heuristic.

In order to implement this idea, we add a parameter `limit` to the procedure `value`. On every recursive invocation of the function `value`, the function `limit` is decreased. Once the limit reaches 0, instead of invoking the function `value` again recursively, we try to estimate the value of the given `State` using our evaluation function. This leads to the code shown in Figure 4.4 on page 75.

```
1  value := cachedProcedure(State, player, limit, alpha := -1, beta := 1) {
2      if (finished(State)) {
3          return utility(State, player);
4      }
5      if (limit == 0) { return heuristic(State, player); }
6      val := alpha;
7      for (ns in nextStates(State, player)) {
8          val := max({val, -value(ns, other(player), limit - 1, -beta, -alpha)});
9          if (val >= beta) {
10             return val;
11         }
12         alpha := max({val, alpha});
13     }
14     return val;
15 };
```

Figure 4.4: Depth-limited α-β-pruning.

For a game like tic-tac-toe it is difficult to come up with a decent heuristic. A very crude approach would be to define:

```
heuristic := [State, player] |-> 0;
```

This heuristic would simply estimate the value of all states to be 0. As this heuristic is only called after it has been tested that the game has not yet been decided, this approach is not utterly unreasonable. For a more complex game like chess, the heuristic could instead be a weighted count of all pieces. Concretely, the algorithm for estimating the value of a state would work as follows:

1. Initially, the variable `sum` is set to 0:

   ```
   sum := 0;
   ```

2. We would count the number of white rooks $\texttt{Rook}_{\text{white}}$ and black rooks $\texttt{Rook}_{\text{black}}$, subtract these numbers from each other and multiply the difference by 5. The resulting number would be added to `sum`:

   $$\texttt{sum} \mathrel{+}= (\texttt{Rook}_{\text{white}} - \texttt{Rook}_{\text{black}}) \cdot 5;$$

3. We would count the number of white bishops $\texttt{Bishop}_{\text{white}}$ and black bishops $\texttt{Bishop}_{\text{black}}$, subtract these numbers from each other and multiply the difference by 3. The resulting number would be added to `sum`:

   $$\texttt{sum} \mathrel{+}= (\texttt{Bishop}_{\text{white}} - \texttt{Bishop}_{\text{black}}) \cdot 3;$$

4. In a similar way we would count knights, queens, and pawns. Approximately, the weights of knights are 3, a queen is worth 9 and a pawn is worth 1.

[2] For reference, compare the wikipedia article on the so-called Shannon number. The Shannon number estimates that there are at most 10^{120} different states in chess. However, if we discount those states where any of the player has made an obviously ridiculous move, we are left with approximately 10^{40} different states.

The resulting `sum` can then be used as an approximation of the value of a state. More details about the weights of the pieces can be found in the Wikipedia article "chess piece relative value".

Exercise 9: Read up on the game Connect Four. You can play it online at

`http://www.connectfour.org/connect-4-online.php`

Your task is to implement this game. At the address

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/SetlX/connect-four-frame.stlx

is a frame that can be used to solve this exercise. In this frame, only the following procedures need to be implemented.

1. The procedure `nextStates`(s, p) is called with a state s and a player p. It is supposed to return the set of all states that can be reached from the given state s if player p is next to make a move. In order to implement this function efficiently, you should make use of the function `find_empty` that is described below.

2. The procedure `find_empty`(s, c) is called with a state s and a column c. It searches the column c of the board specified by s bottom up for the first empty square. Once an empty square is found, the corresponding row is returned. If all the squares in the column c are filled, then the number 7 is returned instead.

3. The procedure `utility`(s, p) is called with a state s and a player p. It computes the utility that this state has for player p if player p is the next to move. For efficiency reasons, the implementation of `utility` should make use of the auxiliary functions `all_lines` and `last_line_filled` that are described below.

4. The procedure `all_lines`() takes no arguments. It is supposed to compute all sets of coordinates that form a line of four consecutive squares. For example, one such set would be the set

   $$\{\langle 1,1\rangle, \langle 1,2\rangle, \langle 1,3\rangle, \langle 1,4\rangle\}.$$

   This set would be the vertical line that starts at $\langle 1,1\rangle$. The horizontal line that starts at the location $\langle 1,1\rangle$ is the set

   $$\{\langle 1,1\rangle, \langle 2,1\rangle, \langle 3,1\rangle, \langle 4,1\rangle\},$$

   while the rising diagonal starting at $\langle 1,1\rangle$ is represented by the set

   $$\{\langle 1,1\rangle, \langle 2,2\rangle, \langle 3,3\rangle, \langle 4,4\rangle\}.$$

   Note that the procedure `all_lines` should be implemented as a `cachedProcedure`. This way it is guaranteed that the necessary computation is only performed once.

5. The procedure `last_line_filled`(s) takes a state s. It checks whether the game is drawn. The game is drawn iff the board is completely filled. This can be checked most efficiently by checking the final row of the board. If this row is completely filled, all other rows must have been filled, too.

Once you have a working implementation of Connect Four, try to improve the strength of your program by adding a non-trivial heuristic to evaluate non-terminal states. As an example of a non-trivial heuristic you can define a triple as a set of three marks of either `X`s or `O`s in a row that is followed by a blank space. The blank space could also be between the marks. Now if there is a state s that has a triples of `X`s and b triples of `O`s and the game is not finished, then define

$$\texttt{value}(s, \texttt{X}, \texttt{limit}, \alpha, \beta) = \frac{a-b}{10} \quad \text{if } \texttt{limit} = 0.$$

In order to implement a heuristic of this kind you have to change line 5 in the implementation of the function `value` shown in Figure 4.4 on page 75. ◇

Chapter 5

Linear Regression

A great deal of the current success of artificial intelligence is due to recent advances in machine learning. In order to get a first taste of what machine learning is about, we introduce linear regression in this chapter, since linear regression is one of the most basic algorithms in machine learning. It is also the foundation for more advanced forms of machine learning like logistic regression and neural networks. Furthermore, linear regression is surprisingly powerful. Finally, many of the fundamental problems of machine learning can already be illustrated with linear regression. Therefore it is only natural that we begin our study of machine learning with the study of linear regression.

5.1 Simple Linear Regression

Assume we want to know how the engine displacement of a car engine relates to its fuel consumption. One approach to understand this relation would be to derive a theoretical model that is able to predict the fuel consumption from the engine displacement by using the appropriate laws of physics and chemistry. However, due to our lack of understanding of the underlying theory, this is not an option for us. Instead, we will look at different engines and compare their engine displacement with the corresponding fuel consumption. This way, we will collect a set of m observations of the form $\langle x_1, y_1\rangle, \cdots, \langle x_m, y_m\rangle$ where x_i is the engine displacement of the engine in the i-th car, while y_i is the fuel consumption of the i-th car. We call x the independent variable, while y is the dependent variable. We define the vectors $\mathbf{x}$ and $\mathbf{y}$ as follows:

$$\mathbf{x} := \langle x_1, \cdots, x_m\rangle^\top \quad \text{and} \quad \mathbf{y} := \langle y_1, \cdots, y_m\rangle^\top.$$

Here, the operator $^\top$ is interpreted as the transpose operator, i.e. $\mathbf{x}$ and $\mathbf{y}$ are considered to be column vectors. By using the transpose operator I am able to write these vectors in a single line.

In linear regression, we use a linear hypothesis and assume that the dependent variable y_i is related to the independent variable x_i via a linear equation of the form

$$y_i := \vartheta_1 \cdot x_i + \vartheta_0.$$

We do not expect this equation to hold exactly. The reason is that there are many other factors besides the engine displacement that influence the fuel consumption. For example, both the weight of a car and its aerodynamics certainly influence the fuel consumption. We want to calculate those values ϑ_0 and ϑ_1 such that the mean squared error, which is defined as

$$\texttt{MSE}(\vartheta_0, \vartheta_1) := \frac{1}{m-1} \cdot \sum_{i=1}^{m} \big(\vartheta_1 \cdot x_i + \vartheta_0 - y_i\big)^2, \tag{5.1}$$

is minimized. It can be shown that the solution to this minimization problem is given as follows:

$$\vartheta_1 = r_{x,y} \cdot \frac{s_y}{s_x} \quad \text{and} \quad \vartheta_0 = \bar{\mathbf{y}} - \vartheta_1 \cdot \bar{\mathbf{x}}. \tag{5.2}$$

This solution makes use of the values $r_{x,y}$, s_x, and s_y. In order to define these values, we first define the sample mean values $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ of $\mathbf{x}$ and $\mathbf{y}$ respectively, i.e. we have

$$\bar{\mathbf{x}} = \frac{1}{m} \cdot \sum_{i=1}^{m} x_i \quad \text{and} \quad \bar{\mathbf{y}} = \frac{1}{m} \cdot \sum_{i=1}^{m} y_i.$$

Furthermore, s_x and s_y are the sample standard deviations of $\mathbf{x}$ and y, i.e. we have

$$s_x = \sqrt{\frac{1}{m-1} \cdot \sum_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big)^2} \quad \text{and} \quad s_y = \sqrt{\frac{1}{m-1} \cdot \sum_{i=1}^{m} \big(y_i - \bar{\mathbf{y}}\big)^2}.$$

Next, $\mathrm{Cov}[\mathbf{x}, \mathbf{y}]$ is the sample covariance and is defined as

$$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = \frac{1}{(m-1)} \cdot \sum_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big) \cdot \big(y_i - \bar{\mathbf{y}}\big).$$

Finally, $r_{x,y}$ is the sample correlation coefficient that is defined as

$$r_{x,y} = \frac{1}{(m-1) \cdot s_x \cdot s_y} \cdot \sum_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big) \cdot \big(y_i - \bar{\mathbf{y}}\big) = \frac{\mathrm{Cov}[\mathbf{x}, \mathbf{y}]}{s_x \cdot s_y}.$$

The number $r_{x,y}$ is also known as the Pearson correlation coefficient or Pearson's r. It is named after Karl Pearson (1857 – 1936). Note that the formula for the parameter ϑ_1 can be simplified to

$$\vartheta_1 = \frac{\sum\limits_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big) \cdot \big(y_i - \bar{\mathbf{y}}\big)}{\sum\limits_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big)^2} \tag{5.3}$$

This latter formula should be used to calculate ϑ_1. However, the previous formula is also useful because it shows the the correlation coefficient is identical to the coefficient ϑ_1, provided the variables $\mathbf{x}$ and $\mathbf{y}$ have been normalized so that their standard deviation is 1.

Exercise 10: Prove Equation 5.2 and Equation 5.3.

Hint: Take the partial derivatives of $\texttt{MSE}(\vartheta_0, \vartheta_1)$ with respect to ϑ_0 and ϑ_1. If the expression $\texttt{MSE}(\vartheta_0, \vartheta_1)$ is minimal, then these partial derivatives have to be equal to 0. ◇

5.1.1 Assessing the Quality of Linear Regression

Assume that we have been given a set of m observations of the form $\langle x_1, y_1 \rangle, \cdots, \langle x_m, y_m \rangle$ and that we have calculated the parameters ϑ_0 and ϑ_1 according to Equation 5.2 and Equation 5.3. Provided that not all x_i have the same value, these formulæ will return two numbers for ϑ_0 and ϑ_1 that define a linear model for $\mathbf{y}$ in terms of $\mathbf{x}$. However, at the moment we still lack a number that tells us how good this linear model really is. In order to judge the quality of the linear model given by

$$y = \vartheta_0 + \vartheta_1 \cdot x$$

we can compute the mean squared error according to Equation 5.1. However, the mean squared error is an absolute number that, by itself, is difficult to interpret. The reason is that the variable $\mathbf{y}$ might be inherently noisy and we have to relate this noise to the mean squared error. Now the noise contained in $\mathbf{y}$ can be measured by the sample variance of $\mathbf{y}$ and is given by the formula

$$\texttt{Var}(y) := \frac{1}{m-1} \cdot \sum_{i=1}^{m} \big(y_i - \bar{y}\big)^2. \tag{5.4}$$

If we compare this formula to the formula for the mean squared error

$$\texttt{MSE}(\vartheta_0, \vartheta_1) := \frac{1}{m-1} \cdot \sum_{i=1}^{m} \big(\vartheta_1 \cdot x_i + \vartheta_0 - y_i\big)^2,$$

we see that the sample variance of $\mathbf{y}$ is an upper bound for the mean squared error since we have

$$\texttt{Var}(\mathbf{y}) = \texttt{MSE}(\bar{\mathbf{y}}, 0),$$

i.e. the sample variance is the value that we would get for the mean squared error if we set ϑ_0 to the average value of $\mathbf{y}$ and ϑ_1 to zero. Since ϑ_0 and ϑ_1 are chosen to minimize the mean squared error, we have

$$\texttt{MSE}(\vartheta_0, \vartheta_1) \leq \texttt{MSE}(\bar{\mathbf{y}}, 0) = \texttt{Var}(\mathbf{y}).$$

The mean squared error is an absolute value and, therefore, difficult to interpret. The fraction

$$\frac{\texttt{MSE}(\vartheta_0, \vartheta_1)}{\texttt{Var}(y)}$$

is called the proportion of the unexplained variance because it is the variance that is still left if we use our linear model to predict the values of $\mathbf{y}$ given the values of $\mathbf{x}$. The proportion of the explained variance which is also known as the $\texttt{R}^2$ statistic is defined as

$$\texttt{R}^2 := \frac{\texttt{Var}(\mathbf{y}) - \texttt{MSE}(\vartheta_0, \vartheta_1)}{\texttt{Var}(\mathbf{y})} = 1 - \frac{\texttt{MSE}(\vartheta_0, \vartheta_1)}{\texttt{Var}(\mathbf{y})}. \tag{5.5}$$

Here MSE is short for $\texttt{MSE}(\vartheta_0, \vartheta_1)$ where we have substituted the values from equation 5.2 and 5.3. The statistic $\texttt{R}^2$ measures the quality of our model: If it is small, then our model does not explain the variation of the value of $\mathbf{y}$ when the value of $\mathbf{x}$ changes. On the other hand, if it is near to 100%, then our model does a good job in explaining the variation of $\mathbf{y}$ when $\mathbf{x}$ changes.

Since the formulæ for $\texttt{Var}(\mathbf{y})$ and $\texttt{MSE}(\vartheta_0, \vartheta_1)$ have the same denominator $m - 1$, this denominator can be cancelled when $\texttt{R}^2$ is computed. To this end we define the total sum of squares TSS as

$$\texttt{TSS} := \sum_{i=1}^{m} \big(y_i - \bar{\mathbf{y}}\big)^2 = (m-1) \cdot \texttt{Var}(\mathbf{y})$$

and the residual sum of squares RSS as

$$\texttt{RSS} := \sum_{i=1}^{m} \big(\vartheta_1 \cdot x_i + \vartheta_0 - y_i\big)^2 = (m-1) \cdot \texttt{MSE}(\vartheta_0, \vartheta_1).$$

Then the formula for the $\texttt{R}^2$ statistic can be written as

$$\texttt{R}^2 = 1 - \frac{\texttt{RSS}}{\texttt{TSS}}.$$

This is the formula that we will use when we implement simple linear regression.

It should be noted that R^2 is the square of Pearson's r. The notation is a bit inconsistent since Pearson's r is written in lower case, while R^2 is written in upper case. However, since this is the notation used in most books on statistics, we will use it too. The number R^2 is also known as the coefficient of determination. It tells us to what extend the value of the variable y is determined by the value of x.

5.1.2 Putting the Theory to the Test

In order to get a better feeling for linear regression, we want to test it to investigate the factors that determine the fuel consumption of cars. Figure 5.1 on page 80 shows the head of the data file "`cars.csv`" which I have adapted from the file

`http://www-bcf.usc.edu/~gareth/ISL/Auto.csv`.

Figure 5.1 on page 80 shows the column headers and the first ten data entries contained in this file. Altogether, this file contains data about 392 different car models.

The file "`cars.csv`" is part of the data set accompanying the excellent book Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani [JWHT14]. The file "`cars.csv`" contains the fuel consumption of a number of different cars that were in widespread use during the seventies and early eighties of the last century. The first column of this data set gives the miles per gallonof a car, i.e. the number of miles a car can go with one gallon of gas. Note that this number is inverse to the fuel consumption: If a car A can go twice as many miles per gallon than another car B, then the fuel consumption of A is half of the

```
 mpg, cyl, displacement,    hp, weight,  acc, year, name
18.0,   8,        307.0, 130.0, 3504.0, 12.0,   70, chevrolet chevelle malibu
15.0,   8,        350.0, 165.0, 3693.0, 11.5,   70, buick skylark 320
18.0,   8,        318.0, 150.0, 3436.0, 11.0,   70, plymouth satellite
16.0,   8,        304.0, 150.0, 3433.0, 12.0,   70, amc rebel sst
17.0,   8,        302.0, 140.0, 3449.0, 10.5,   70, ford torino
15.0,   8,        429.0, 198.0, 4341.0, 10.0,   70, ford galaxie 500
14.0,   8,        454.0, 220.0, 4354.0,  9.0,   70, chevrolet impala
14.0,   8,        440.0, 215.0, 4312.0,  8.5,   70, plymouth fury iii
14.0,   8,        455.0, 225.0, 4425.0, 10.0,   70, pontiac catalina
15.0,   8,        390.0, 190.0, 3850.0,  8.5,   70, amc ambassador dpl
```

Figure 5.1: The head of the file `cars.csv`.

fuel consumption of B. Furthermore, besides the miles per gallon, for every car the following other parameters are listed:

1. `cyl` is the number of cylinders,
2. `displacement` is the engine displacement in cubic inches,
3. `hp` is the engine power given in units of horsepower,
4. `weight` is the weight in pounds,
5. `acc` is the acceleration given as the time in seconds needed to accelerate from 0 miles per hour to 60 miles per hour,
6. `year` is the year in which the model was introduced, and
7. `name` is the name of the model.

Our aim is to determine what part of the fuel consumption of a car is explained by its engine displacement. To this end, I have written the function `simple_linear_regression` shown in Figure 5.2 on page 81.
The procedure `simple_linear_regression` takes two arguments:

(a) `X` is a `NumPy` array containing the independent variable.

(b) `Y` is a `NumPy` array containing the dependent variable.

The implementation of the procedure `simple_linear_regression` works as follows:

1. `m` is the number of data that are present in the array `X`.
2. `xMean` is the mean value $\bar{x}$ of the independent variable `x`.
3. `yMean` is the mean value $\bar{\mathbf{y}}$ of the dependent variable `y`.
4. The coefficient `theta1` is computed according to Equation 5.3, which is repeated here for convenience:

$$\vartheta_1 = \frac{\sum_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big) \cdot \big(y_i - \bar{\mathbf{y}}\big)}{\sum_{i=1}^{m} \big(x_i - \bar{\mathbf{x}}\big)^2}.$$

   Note that the expression `(X - xMean)` computes an array of the same shape as `X` by subtracting `xMean` from every entries of `X`. Next, the expression `(X - xMean) * (Y - yMean)` computes the elementwise product of the arrays `X - xMean` and `Y - yMean`. The expression `(X - xMean) ** 2` computes the elementwise squares of the array `X - xMean`. Finally, the function `sum` computes the sum of all the elements of an array.

```
def simple_linear_regression(X, Y):
    """
    This function implements linear regression.

    * X:     explaining variable, numpy array
    * Y:     dependent  variable, numpy array

    Output: The R2 value of the linear regression.
    """
    m      = len(X)
    xMean  = sum(X) / m;
    yMean  = sum(Y) / m;
    theta1 = sum((X - xMean) * (Y - yMean)) / sum((X - xMean) ** 2)
    theta0 = yMean - theta1 * xMean;
    TSS    = sum((Y - yMean) ** 2)
    RSS    = sum((theta1 * X + theta0 - Y) ** 2)
    R2     = 1 - RSS / TSS;
    return R2
```

Figure 5.2: Simple Linear Regression

5. The coefficient `theta0` is computed according to Equation 5.2, which reads

$$\vartheta_0 = \bar{\mathtt{y}} - \vartheta_1 \cdot \bar{\mathtt{x}}.$$

6. `TSS` is the total sum of squares and is computed using the formula

$$\mathtt{TSS} = \sum_{i=1}^{m} \big(y_i - \bar{\mathtt{y}}\big)^2.$$

7. `RSS` is the residual sum of squares and is computed as

$$\mathtt{RSS} := \sum_{i=1}^{m} \big(\vartheta_1 \cdot x_i + \vartheta_0 - y_i\big)^2.$$

8. `R2` is the $\mathtt{R}^2$ statistic and measures the proportion of the explained variance. It is computed using the formula

$$\mathtt{R}^2 = \frac{\mathtt{TSS} - \mathtt{RSS}}{\mathtt{TSS}}.$$

In order to use the function we can use the code that is shown in Figure 5.3 on page 82.

1. We import the module `csv` in order to be able to read the Csv file "`cars.csv`" conveniently.
2. We import the module `numpy` in order to use `NumPy` arrays.
3. We open the file "`cars.csv`".
4. This file is processed as a Csv file where different columns are separated by the character "`,`".
5. `kpl` is a list of the numbers that appear in the first column of the Csv file. The numbers in the Csv file are interpreted as the *miles per gallon* of a car. These numbers are converted into metric units, i.e. how many kilometer a car can run on a litre.
6. `displacement` is a list the numbers appearing in the third column of the CSV file. These numbers are interpreted as the *engine displacement* in cubic inches. These numbers are converted to litres.

```
import csv
import numpy as np

if __name__ == '__main__':
    with open('cars.csv') as input_file:
        reader       = csv.reader(input_file, delimiter=',')
        line_count   = 0
        kpl          = []
        displacement = []
        for row in reader:
            if line_count != 0:  # skip header of file
                kpl .append(float(row[0]) * 0.00425144)
                displacement.append(float(row[2]) * 0.0163871)
            line_count += 1
    m  = len(displacement)
    X  = np.array(displacement)
    Y  = np.array([1 / kpl[i] for i in range(m)])
    R2 = simple_linear_regression(X, Y)
    print(f'The explained variance is {R2}%')
```

Figure 5.3: Calling the procedure `simple_linear_regression`.

dependent variable	explained variance
`displacement`	0.75
`cyl`	0.70
`hp`	0.73
`weight`	0.78
`acc`	0.21
`year`	0.31

Table 5.1: Explained variance for various dependent variables.

7. The first line of the Csv file contains a header. This header is skipped. In order to do so we use the variable `line_count`.
8. `m` is the number of data pairs that have been read.
9. The independent variable `X` is given by the engine displacement. In order to be able to use `NumPy` features later we convert this list into a `NumPy` array.
10. The dependent variable `Y` is given by the inverse of the variable `kph`.
11. Finally, the coefficient of determination R^2 is computed using the function `simple_linear_regression`.

In the same way as we have computed the coefficient of determination that measures how the fuel consumption is influenced by the engine displacement we can also compute the coefficient of determination for other variables like the number of cylinders or the weight of the car. The resulting values are shown in Table 5.1. It seems that, given the data in the file "`cars.csv`", the best indicator for the fuel consumption is the `weight` of a car. The `displacement`, the power `hp` of an engine, and the number of cylinders `cyl` are also good predictors. But notice that the `weight` is the real cause of fuel consumption: If a car has a big weight, it will also need a more powerful engine. Hence the variable `hp` is correlated with the variable `weight` and will therefore also provide a reasonable explanation of the fuel consumption, although the high engine power is not the most important cause of the fuel consumption.

5.2 General Linear Regression

In practise, it is rarely the case that a given observed variable y only depends on a single variable x. To take the example of the fuel consumption of a car further, in general we would expect that the fuel consumption of a car depends not only on the mass of the car but is also related to the other parameters. To this end, we present the theory of general linear regression. In a general regression problem we are given a list of m pairs of the form $\langle \mathbf{x}^{(i)}, y^{(i)} \rangle$ where $\mathbf{x}^{(i)} \in \mathbb{R}^p$ and $y^{(i)} \in \mathbb{R}$ for all $i \in \{1, \cdots, m\}$. The number p is called the number of features, while the pairs are called the training examples. Our goal is to compute a function

$$F : \mathbb{R}^p \to \mathbb{R}$$

such that $F\big(\mathbf{x}^{(i)}\big)$ approximates $y^{(i)}$ as precisely as posssible for all $i \in \{1, \cdots, m\}$, i.e. we want to have

$$F\big(\mathbf{x}^{(i)}\big) \approx y^{(i)} \quad \text{for all } i \in \{1, \cdots, m\}.$$

In order to make the notation $F\big(\mathbf{x}^{(i)}\big) \approx y^{(i)}$ more precise, we define the mean squared error

$$\texttt{MSE} := \frac{1}{m-1} \cdot \sum_{i=1}^{m} \Big(F\big(\mathbf{x}^{(i)}\big) - y^{(i)}\Big)^2. \tag{5.6}$$

Then, given the list of training examples $[\langle \mathbf{x}^{(1)}, y^{(1)} \rangle, \cdots, \langle \mathbf{x}^{n}, y^{(n)} \rangle]$, our goal is to minimize MSE. In order to proceed, we need to have a model for the function F. The simplest model is a linear model, i.e. we assume that F is given as

$$F(\mathbf{x}) = \sum_{j=1}^{p} w_j \cdot x_j + b = \mathbf{x}^\top \cdot \mathbf{w} + b \quad \text{where } \mathbf{w} \in \mathbb{R}^p \text{ and } b \in \mathbb{R}.$$

Here, the expression $\mathbf{x}^\top \cdot \mathbf{w}$ denotes the matrix product of the vector $\mathbf{x}^\top$, which is viewed as a 1-by-m matrix, and the vector $\mathbf{w}$, where $\mathbf{w}$ is viewed as a m-by-1 matrix. Alternatively, this expression could be interpreted as the dot product of the vector $\mathbf{x}$ and the vector $\mathbf{w}$. At this point you might wonder why it is useful to introduce matrix notation here. The reason is that this notation shortens the formula and, furthermore, is more efficient to implement since most programming languages used in machine learning have special library support for matrix operations. Provided the computer is equipped with a graphics card, some programming languages are even able to delegate matrix operations to the graphics unit. This results in a considerable speed-up.

The definition of F given above is the model used in linear regression. Here, $\mathbf{w}$ is called the weight vector and b is called the bias. It turns out that the notation can be simplified if we extend the p-dimensional feature vector $\mathbf{x}$ to an $p+1$-dimensional vector $\mathbf{x}'$ such that

$$x'_j := x_j \quad \text{for all } j \in \{1, \cdots, p\} \quad \text{and} \quad x'_{m+1} := 1.$$

To put it in words, the vector $\mathbf{x}'$ results from the vector $\mathbf{x}$ by appending the number 1:

$$\mathbf{x}'^\top = \langle x_1, \cdots, x_p, 1 \rangle^\top \quad \text{where } \langle x_1, \cdots, x_p \rangle = \mathbf{x}^\top.$$

Furthermore, we define

$$\mathbf{w}'^\top := \langle w_1, \cdots, w_p, b \rangle^\top \quad \text{where } \langle w_1, \cdots, w_p \rangle = \mathbf{w}^\top.$$

Then we have

$$F(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \mathbf{w}' \cdot \mathbf{x}',$$

where $\mathbf{w}' \cdot \mathbf{x}'$ denotes the dot product of the vectors $\mathbf{w}'$ and $\mathbf{x}'$. Hence, the bias has been incorporated into the weight vector at the cost of appending the number 1 at the end of input vector $\mathbf{x}$. As we want to use this simplification, from now on we assume that the input vectors $\mathbf{x}^{(i)}$ have all been extended so that their last component is 1. Using this assumption, we define the function F as

$$F(\mathbf{x}) := \mathbf{x}^\top \cdot \mathbf{w}.$$

Now equation (5.6) can be rewritten as follows:

$$\texttt{MSE}(\mathbf{w}) = \frac{1}{m-1} \cdot \sum_{i=1}^{m} \Big((\mathbf{x}^{(i)})^\top \cdot \mathbf{w} - y^{(i)} \Big)^2. \tag{5.7}$$

Our aim is to rewrite the sum appearing in this equation as a scalar product of a vector with itself. To this end, we first define the vector $\mathbf{y}$ as follows:

$$\mathbf{y}^\top := \langle y^{(1)}, \cdots, y^{(m)} \rangle^\top.$$

Note that $\mathbf{y} \in \mathbb{R}^m$ since it has a component for all of the m training examples. Next, we define the design matrix X as follows:

$$X := \begin{pmatrix} (\mathbf{x}^{(1)})^\top \\ \vdots \\ (\mathbf{x}^{(m)})^\top \end{pmatrix}$$

Defined this way, the row vectors of the matrix X are the vectors $\mathbf{x}^{(i)}$ transposed. Now we have the following:

$$X \cdot \mathbf{w} - \mathbf{y} = \begin{pmatrix} (\mathbf{x}^{(1)})^\top \\ \vdots \\ (\mathbf{x}^{(m)})^\top \end{pmatrix} \cdot \mathbf{w} - \mathbf{y} = \begin{pmatrix} (\mathbf{x}^{(1)})^\top \cdot \mathbf{w} - y_1 \\ \vdots \\ (\mathbf{x}^{(m)})^\top \cdot \mathbf{w} - y_m \end{pmatrix}$$

Taking the square of the vector $X \cdot \mathbf{w} - \mathbf{y}$ we discover that we can rewrite equation (5.7) as follows:

$$\texttt{MSE}(\mathbf{w}) = \frac{1}{m-1} \cdot \big(X \cdot \mathbf{w} - \mathbf{y}\big)^\top \cdot \big(X \cdot \mathbf{w} - \mathbf{y}\big). \tag{5.8}$$

5.2.1 Some Useful Gradients

In the last section, we have computed the mean squared error $\texttt{MSE}(\mathbf{w})$ using equation (5.8). Our goal is to minimize the $\texttt{MSE}(\mathbf{w})$ by choosing the weight vector $\mathbf{w}$ appropriately. A necessary condition for $\texttt{MSE}(\mathbf{w})$ to be minimal is

$$\nabla \texttt{MSE}(\mathbf{w}) = \mathbf{0},$$

i.e. the gradient of $\texttt{MSE}(\mathbf{w})$ with respect to $\mathbf{w}$ needs to be zero. In order to prepare for the computation of $\nabla \texttt{MSE}(\mathbf{w})$, we first compute the gradient of two simpler functions.

Computing the Gradient of $f(\mathbf{x}) = \mathbf{x}^\top \cdot C \cdot \mathbf{x}$

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f(\mathbf{x}) := \mathbf{x}^\top \cdot C \cdot \mathbf{x} \quad \text{where } C \in \mathbb{R}^{n \times n}.$$

If we write the matrix C as $C = (c_{i,j})_{\substack{i=1,\cdots,n \\ j=1,\cdots,n}}$ and the vector $\mathbf{x}$ as $\mathbf{x} = \langle x_1, \cdots, x_n \rangle^\top$, then $f(\mathbf{x})$ can be computed as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{n} c_{i,j} \cdot x_j = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \cdot c_{i,j} \cdot x_j.$$

We compute the partial derivative of f with respect to x_k and use the product rule together with the definition of the Kronecker delta $\delta_{i,j}$, which is defined as 1 if $i = j$ and as 0 otherwise:

$$\begin{aligned}
\frac{\partial f}{\partial x_k} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\frac{\partial x_i}{\partial x_k}\cdot c_{i,j}\cdot x_j + x_i\cdot c_{i,j}\cdot\frac{\partial x_j}{\partial x_k}\Big)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\delta_{i,k}\cdot c_{i,j}\cdot x_j + x_i\cdot c_{i,j}\cdot\delta_{j,k}\Big)\\
&= \sum_{j=1}^{n} c_{k,j}\cdot x_j + \sum_{i=1}^{n} x_i\cdot c_{i,k}\\
&= \big(C\cdot\mathbf{x}\big)_k + \big(C^\top\cdot\mathbf{x}\big)_k
\end{aligned}$$

Hence we have shown that

$$\nabla f(\mathbf{x}) = (C + C^\top)\cdot\mathbf{x}.$$

If the matrix C is symmetric, i.e. if $C = C^\top$, this simplifies to

$$\nabla f(\mathbf{x}) = 2\cdot C\cdot\mathbf{x}.$$

Next, if the function $g:\mathbb{R}^n\to\mathbb{R}$ is defined as

$$g(\mathbf{x}) := \mathbf{b}^\top\cdot A\cdot\mathbf{x},\quad \text{where } \mathbf{b}\in\mathbb{R}^n \text{ and } A\in\mathbb{R}^{n\times n},$$

then a similar calculation shows that

$$\nabla g(\mathbf{x}) = A^\top\cdot\mathbf{b}.$$

Exercise 11: Prove this equation.

5.2.2 Deriving the Normal Equation

Next, we derive the so called normal equation for linear regression. To this end, we first expand the product in equation (5.8):

$$\begin{aligned}
\texttt{MSE}(\mathbf{w}) &= \frac{1}{m-1}\cdot\big(X\cdot\mathbf{w}-\mathbf{y}\big)^\top\cdot\big(X\cdot\mathbf{w}-\mathbf{y}\big)\\
&= \frac{1}{m-1}\cdot\big(\mathbf{w}^\top\cdot X^\top-\mathbf{y}^\top\big)\cdot\big(X\cdot\mathbf{w}-\mathbf{y}\big) && \text{since } (A\cdot B)^\top = B^\top\cdot A^\top\\
&= \frac{1}{m-1}\cdot\big(\mathbf{w}^\top\cdot X^\top\cdot X\cdot\mathbf{w}-\mathbf{y}^\top\cdot X\cdot\mathbf{w}-\mathbf{w}^\top\cdot X^\top\cdot\mathbf{y}+\mathbf{y}^\top\cdot\mathbf{y}\big)\\
&= \frac{1}{m-1}\cdot\big(\mathbf{w}^\top\cdot X^\top\cdot X\cdot\mathbf{w}-2\cdot\mathbf{y}^\top\cdot X\cdot\mathbf{w}+\mathbf{y}^\top\cdot\mathbf{y}\big) && \text{since } \mathbf{w}^\top\cdot X^\top\cdot\mathbf{y} = \mathbf{y}^\top\cdot X\cdot\mathbf{w}
\end{aligned}$$

The fact that

$$\mathbf{w}^\top\cdot X^\top\cdot\mathbf{y} = \mathbf{y}^\top\cdot X\cdot\mathbf{w}$$

might not be immediately obvious. It follows from two facts:

1. For two matrices A and B such that the matrix product $A\cdot B$ is defined we have
$$(A\cdot B)^\top = B^\top\cdot A^\top.$$
2. The matrix product $\mathbf{w}^\top\cdot X^\top\cdot\mathbf{y}$ is a real number. The transpose $r^\top$ of a real number r is the number itself, i.e. $r^\top = r$ for all $r\in\mathbb{R}$. Therefore, we have
$$\mathbf{w}^\top\cdot X^\top\cdot\mathbf{y} = \big(\mathbf{w}^\top\cdot X^\top\cdot\mathbf{y}\big)^\top = \mathbf{y}^\top\cdot X\cdot\mathbf{w}.$$

Hence we have shown that

$$\texttt{MSE}(\mathbf{w}) = \frac{1}{m-1}\cdot\Big(\mathbf{w}^\top\cdot\big(X^\top\cdot X\big)\cdot\mathbf{w}-2\cdot\mathbf{y}^\top\cdot X\cdot\mathbf{w}+\mathbf{y}^\top\cdot\mathbf{y}\Big) \tag{5.9}$$

holds. The matrix $X^\top \cdot X$ used in the first term is symmetric because

$$\big(X^\top \cdot X\big)^\top = X^\top \cdot \big(X^\top\big)^\top = X^\top \cdot X.$$

Using the results from the previous section we can now compute the gradient of $\texttt{MSE}(\mathbf{w})$ with respect to $\mathbf{w}$. The result is

$$\nabla \texttt{MSE}(\mathbf{w}) = \frac{2}{m-1} \cdot \Big(X^\top \cdot X \cdot \mathbf{w} - X^\top \cdot \mathbf{y}\Big).$$

If the squared error $\texttt{MSE}(\mathbf{w})$ has a minimum for the weights $\mathbf{w}$, then we must have

$$\nabla \texttt{MSE}(\mathbf{w}) = \mathbf{0}.$$

This leads to the equation

$$\frac{2}{m-1} \cdot \Big(X^\top \cdot X \cdot \mathbf{w} - X^\top \cdot \mathbf{y}\Big) = \mathbf{0}.$$

This equation can be rewritten as

$$\boxed{\big(X^\top \cdot X\big) \cdot \mathbf{w} = X^\top \cdot \mathbf{y}.} \tag{5.10}$$

This equation is called the normal equation.

Remark: Although the matrix $X^\top \cdot X$ will often be invertible, for numerical reasons it is not advisable to rewrite the normal equation as

$$\mathbf{w} = \big(X^\top \cdot X\big)^{-1} \cdot X^\top \cdot \mathbf{y}.$$

Instead, when solving the normal equation we will use the *Python* function $\texttt{numpy.linalg.solve}(A, b)$, which takes a matrix $A \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$ and solves the equation

$$A \cdot \mathbf{x} = \mathbf{b}.$$ ◇

5.2.3 Implementation

Figure 5.4 on page 87 shows an implementation of general linear regression. The procedure

```
linear_regression(fileName, target, explaining, f)
```

takes four arguments:

1. `fileName` is a string that is interpreted as the name of a Csv file containing the data.
2. `target` is an integer that specifies the column that contains the dependent variable.
3. `explaining` is a list of integers. These integers specify the columns of the Csv file that contain the independent variables. These are also called the explaining variables.
4. `f` is a function that takes one floating point argument and outputs one floating point function. This function is used to modify the dependent variable.

   Later, when we call the function `linear_regression` to investigate the fuel consumption, we will use the function

   $$x \mapsto \frac{1}{x}$$

   to transform the variable *miles per gallon* into a variable expressing the fuel consumption. The reason is that there is reciprocal relation between the number of miles that a car drives on one gallon of gasoline and the fuel consumption: If you drive only a few miles with one gallon of gas, then your fuel consumption is high.

```
1  import csv
2  import numpy as np
3
4  def linear_regression(fileName, target, explaining, f):
5      with open(fileName) as input_file:
6          reader     = csv.reader(input_file, delimiter=',')
7          line_count = 0
8          goal       = []
9          Causes     = []
10         for row in reader:
11             if line_count != 0:
12                 goal  .append(f(float(row[target])))
13                 Causes.append([float(row[i]) for i in explaining] + [1.0])
14             line_count += 1
15     m = len(goal)
16     X = np.array(Causes)
17     y = np.array(goal)
18     w = np.linalg.solve(X.T @ X, X.T @ y)
19     RSS   = np.sum((X @ w - y) ** 2)
20     yMean = np.sum(y) / m
21     TSS   = sum((y - yMean) ** 2)
22     R2    = 1 - RSS / TSS
23     return R2
24
25 def main():
26     explaining = [1, 2, 3, 4, 5, 6]
27     R2 = linear_regression("cars.csv", 0, explaining, lambda x: 1/x)
28     print(f'portion of explained variance : {R2}')
```

Figure 5.4: General linear regression.

The function `linear_regression` works as follows:

1. It reads the specified Csv file line by line and stores the data in the variables `goal` and `Causes`. This is done by creating a `csv` reader in line 6. This reader returns the entries in the specified input file line by line in the for loop in line 10.

   (a) `goal` is a list containing the data of the dependent variable that was specified by `target`. This list is initialized in line 8. It is filled with data in line 12. Note that the values stored in `goal` are transformed by the function `f`.

   (b) `Causes` is a list of lists containing the data of the explaining variables. Every row in the Csv file corresponds to one list in the list Causes. Note also that we append the number 1.0 to each of these lists. This corresponds to adding a constant feature to our data and it enables us to use the normal equations as we have derived them.

2. `m` is the number of data pairs and is computed in line 15.

3. `Causes` is transformed into the `NumPy` matrix `X` in line 16.

4. `goal` is transformed into the `NumPy` array `y` in line 17.

5. The normal equation $(X^\top \cdot X) \cdot \mathbf{w} = X^\top \cdot \mathbf{y}$ is formulated and solved using the function `np.linalg.solve` in line 18.

Note that `X.T` is the transpose of the matrix `X`. The operator `@` computes the matrix product. Hence the expression `X.T @ X` is interpreted as $X^\top \cdot X$. Similarly, the expression `X.T @ y` is interpreted as $X^\top \cdot \mathtt{y}$.

6. The expression `(X @ w - y)` is the difference between the predictions of the linear model and the observed values `y`. By squaring it and the summing over all entries of the resulting vector we compute is the residual sum of squares `RSS` in line 19.

7. `yMean` is the mean value of the variable `y`.

8. `TSS` is the total sum of squares.

9. `R2` is the proportion of the explained variance.

When we run the program shown in Figure 5.4 on page 87 with the data stored in `cars.csv`, which had been discussed previously, then the proportion of explained variance is 88%. Considering that our data does not take the aerodynamics of the cars into take account, this seems like a reasonable result. A Jupyter notebook containing a similar program is available at

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/Python/Linear-Regression.ipynb.

Exercise 12: The file "`trees.csv`", which is available at

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/Python/trees.csv,

contains data about 31 lovely cherry trees from the Allegheny National Forest in Pennsylvania that have fallen prey to a chainsaw massacre. I have taken this data from

http://www.statsci.org/data/general/cherry.txt.

1. The first column of this Csv file contains the diameter of these trees at a height of 54 inches above the ground.

2. The second column lists the heights of these trees in inches.

3. The third column list the volume of wood that has been harvested from these trees. This volume is given in cubic inches.

Try to derive a model that estimates the volume of the trees from the diameter and the height. ◇

Exercise 13: The file "`nba.csv`", which is available at

https://github.com/karlstroetmann/Artificial-Intelligence/blob/master/Python/nba.csv,

contains various data about professional basket ball players.

1. The first column gives the name of the player.

2. The second column specifies the position of the player.

3. The third column lists the height of the player.

4. The fourth column contains the weight of the player.

5. The fifth column shows the age of each player.

To what extend can you predict the weight of a player given his height and his age? ◇

Chapter 6

Classification

One of the earliest application of artificial intelligence is classification. A good example of classification is spam detection. A system for spam detection classifies an email as either spam or not spam. To do so, it first computes various features of the email and then uses these features to determine whether the email is likely to be spam. For example, a possible feature would be the number of occurrences of the word "`pharmacy`" in the text of the email.

6.1 Introduction

Formally, the classification problem in machine learning can be stated as follows: We are given a set of objects $S := \{o_1, \cdots, o_n\}$ and a set of classes $C := \{c_1, \cdots, c_k\}$. Furthermore, there exists a function

$$\texttt{classify} : S \to C$$

that assigns a class $\texttt{classify}(o)$ to every object $o \in S$. The set S is called the sample space. In the example of spam detection, the sample space S is the set of all emails that we might receive, i.e. S is the set of all strings, while

$$C = \{\texttt{spam}, \texttt{ham}\}.$$

If $\texttt{classify}(o) = \texttt{spam}$, we consider the email o to be spam, while we would consider it a genuine message otherwise. Our goal is to compute the function `classify`. In order to do this, we use an approach known as supervised learning: We take a subset $S_{Train} \subseteq S$ of emails where we already know whether the emails are spam or not. This set S_{Train} is called the training set. Next, we define a set of D features for every $o \in S$. These features have to be computable, i.e. we must have a function

$$\texttt{feature} : S \times \{1, \cdots, D\} \to \mathbb{R}$$

such that $\texttt{feature}(o, j)$ computes the j-th feature and we have to be able to implement this function with reasonable efficiency. In general, the values of the features are real values. However, there are cases where these values are just Booleans. If

$$\texttt{feature}(o, j) \in \mathbb{B} \quad \text{for all } o \in S,$$

then the j-th feature is called a binary feature. If we encode `false` as -1 and `true` as $+1$, then the set of Boolean values $\mathbb{B}$ can be considered a subset of $\mathbb{R}$ and hence Boolean features can be considered as real numbers. For example, in the case of spam detection, the first feature could be the occurrence of the string "`pharmacy`". In this case, we would have

$$\texttt{feature}(o, 1) := \begin{cases} +1 & \text{if } \texttt{pharmacy} \in o, \\ -1 & \text{if } \texttt{pharmacy} \notin o, \end{cases}$$

i.e. the first feature would be to check whether the email o contains the string "`pharmacy`". If we want to be more precise, we can instead define the first feature as

$$\texttt{feature}(o,1) := \texttt{count}(\texttt{"pharmacy"}, o),$$

i.e. we would count the number of occurrences of the string "pharmacy" in our email o. As the value of

$$\texttt{count}(\texttt{"pharmacy"}, o)$$

is always a natural number, in this case the first feature would be a discrete feature. However, we can be even more precise than just counting the number of occurrences of "pharmacy". After all, there is a difference if the string "pharmacy" occurs once in an email containing but a hundred characters or whether is occurs once in an email with a length of several thousand characters. To this end, we would then define the first feature as

$$\texttt{feature}(o,1) := \frac{\texttt{count}(\texttt{"pharmacy"}, o)}{\#o},$$

where $\#o$ defines the number of characters in the string o. In this case, the first feature would be a continuous feature and as this is the most general case, unless stated otherwise, we deal with the continuous case.

Having defined the features, we next need a model of the function `classify` that tries to approximate the function `classify` via the features. This model is given by a function

$$\texttt{model} : \mathbb{R}^D \to C$$

such that

$$\texttt{model}\big(\texttt{feature}(o,1), \cdots, \texttt{feature}(o,D)\big) \approx \texttt{classify}(o).$$

Using the function `model`, we can than approximate the function classify using a function `guess` that is defined as

$$\texttt{guess}(o) := \texttt{model}\big(\texttt{feature}(o,1), \cdots, \texttt{feature}(o,D)\big)$$

Most of the time, the function `guess` will only approximate the function `classify`, i.e. we will have

$$\texttt{guess}(o) = \texttt{classify}(o)$$

for most objects of $o \in S$ but not for all of them. The accuracy of our model is then defined as the fraction of those objects that are classified correctly, i.e.

$$\texttt{accuracy} := \frac{\#\{o \in S \mid \texttt{guess}(o) = \texttt{classify}(o)\}}{\#S}.$$

If the set S is infinite, this equation has to be interpreted as a limit, i.e. we can define S_n as the set of strings that have a length of at most n. Then, the accuracy can be defined as follows:

$$\texttt{accuracy} := \lim_{n\to\infty} \frac{\#\{o \in S_n \mid \texttt{guess}(o) = \texttt{classify}(o)\}}{\#S_n}.$$

The function `model` is usually determined by a set of parameters or weights $\mathbf{w}$. In this case, we have

$$\texttt{model}(\mathbf{x}) = \texttt{model}(\mathbf{x}; \mathbf{w})$$

where $\mathbf{x}$ is the vector of features, while $\mathbf{w}$ is the vector of weights. Later, when we introduce logistic regression, we will assume that the number of weights is the same as the number of features. Then, the weights specify the relative importance of the different features.

When it comes to the choice of model, it is important to understand that, at least in practical applications, all models are wrong. Nevertheless, some models are useful. There are two reasons for this:

1. We do not fully understand the function `classify` that we want to approximate by the function `model`.
2. The function `classify` is so complex, that even if we could compute it exactly, the resulting model would be much too complicated.

The situation is similar in physics: Let us assume that we intend to model the fall of an object. A model that is a hundred percent accurate would have to include the following forces:

(a) gravitational acceleration,

(b) air friction,

(c) tidal forces, i.e. the effects that the rotation of the earth has on moving objects,

(d) celestial forces, i.e. the gravitational acceleration caused by celestial objects like the moon or the sun.

(e) In case we have a metallic object we have to be aware of the magnetic forces caused by the geomagnetic field.

(f) To be fully accurate, we might have to include corrections from relativistic physics and even quantum physics.

(g) As physics is not a closed subject, there might be other forces at work which we still do not know of.

Hence, a correct model would be so complicated that it would be unmanageable and therefore useless.

Let us summarize our introductory discussion of machine learning in general and classification in particular. A set S of objects and a set C of classes are given. Our goal is to approximate a function

$$\texttt{classify} : S \to C$$

using certain features of our objects. The function `classify` is then approximated using a function `model` as follows:

$$\texttt{model}\big(\texttt{feature}(o,1), \cdots, \texttt{feature}(o,D); \mathbf{w}\big) \approx \texttt{classify}(o).$$

The model depends on a vector of parameters $\mathbf{w}$. In order to learn these parameters, we are given a training set S_{Train} that is a subset of S. As we are dealing with supervised learning, the function `classify` is known for all objects $o \in S_{Train}$. Our goal is to determine the parameters $\mathbf{w}$ such that the number of mistakes we make on the training set is minimized.

6.1.1 Notation

We conclude this introductory section by fixing some notation. Let us assume that the objects $o \in S_{Train}$ are numbered from 1 to N, while the features are numbered from 1 to D. Then we define

1. $\mathbf{x}_i := [\texttt{feature}(o_i, 1), \cdots, \texttt{feature}(o_i, D)]$ for all $i \in \{1, \cdots, N\}$.
   i.e. $\mathbf{x}_i$ is a D-dimensional vector that collects the features of the i-th training object.
2. $x_{i,j} := \texttt{feature}(o_i, j)$ for all $i \in \{1, \cdots, N\}$ and $j \in \{1, \cdots, D\}$.
   i.e. $x_{i,j}$ is the j-th feature of the i-th object.
3. $y_i := \texttt{classify}(o_i)$ for all $i \in \{1, \cdots, N\}$
   i.e. y_i is the class of the i-th object.

Mathematically, our goal is now to maximize the accuracy of our model as a function of the parameters $\mathbf{w}$.

6.1.2 Applications of Classification

Besides spam detection, there are many other classification problems that can be solved using machine learning. To give just one more example, imagine a general practitioner that receives a patient and examines his symptoms. In this case, the symptoms can be seen as the features of the patient. For example, these features could be

(a) body temperature,

(b) blood pressure,

(c) heart rate,

(d) body weight,

(e) breathing difficulties,

(f) age,

to name but a few of the possible features. Based on these symptoms, the general practitioner would then decide on an illness, i.e. the set of classes for the classification problem would be

$$\{\texttt{commonCold}, \texttt{pneumonia}, \texttt{asthma}, \texttt{flu}, \cdots, \texttt{unknown}\}.$$

Hence, the task of disease diagnosis is a classification problem. This was one of the earliest problem that was tackled by artificial intelligence. As of today, computer-aided diagnosis has been used for more than 40 years in many hospitals. There are a number of diseases that today can be diagnosed more accurately by a computer than by a specialist. One such example is the diagnosis of heart disease. Other examples of classifications are the following:

1. image recognition,
2. speech recognition,
3. credit card fraud detection,
4. credit approval.

6.2 Digression: The Method of Gradient Ascent

In machine learning, it is often the case that we have to find either the maximum or the minimum of a function

$$f : \mathbb{R}^n \to \mathbb{R}.$$

For example, when we discuss logistic regression in the next section, we will have to find the maximum of a function. First, let us introduce the arg max function. The idea is that

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

is that value of $\mathbf{x} \in \mathbb{R}^n$ that maximizes $f(\mathbf{x})$. Formally, we have

$$\forall \mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f\Big(\arg\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})\Big).$$

Of course, the function argmax is only defined when the maximum is unique. If the function f is differentiable, we know that a necessary condition for a vector $\widehat{\mathbf{x}} \in \mathbb{R}^n$ to satisfy

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{is that we must have} \quad \nabla f(\widehat{\mathbf{x}}) = \mathbf{0},$$

i.e. the gradient of f, which we will write as ∇f, vanishes at the maximum. Remember that the gradient of f is defined as

$$\nabla f := \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{pmatrix}.$$

Unfortunately, in many cases the equation

$$\nabla f(\widehat{\mathbf{x}}) = \mathbf{0}$$

cannot be solved explicitly. This is already true in the one-dimensional case, i.e. if $n = 1$. For example, consider the function $f : \mathbb{R} \to \mathbb{R}$ that is defined as

$$f(x) := \sin(x) - \frac{1}{2} \cdot x^2.$$

This function is shown in Figure 6.1 on page 93. From the graph of the function it is obvious that this function has a maximum somewhere between 0.6 and 0.8. In order to compute this maximum, we can compute the

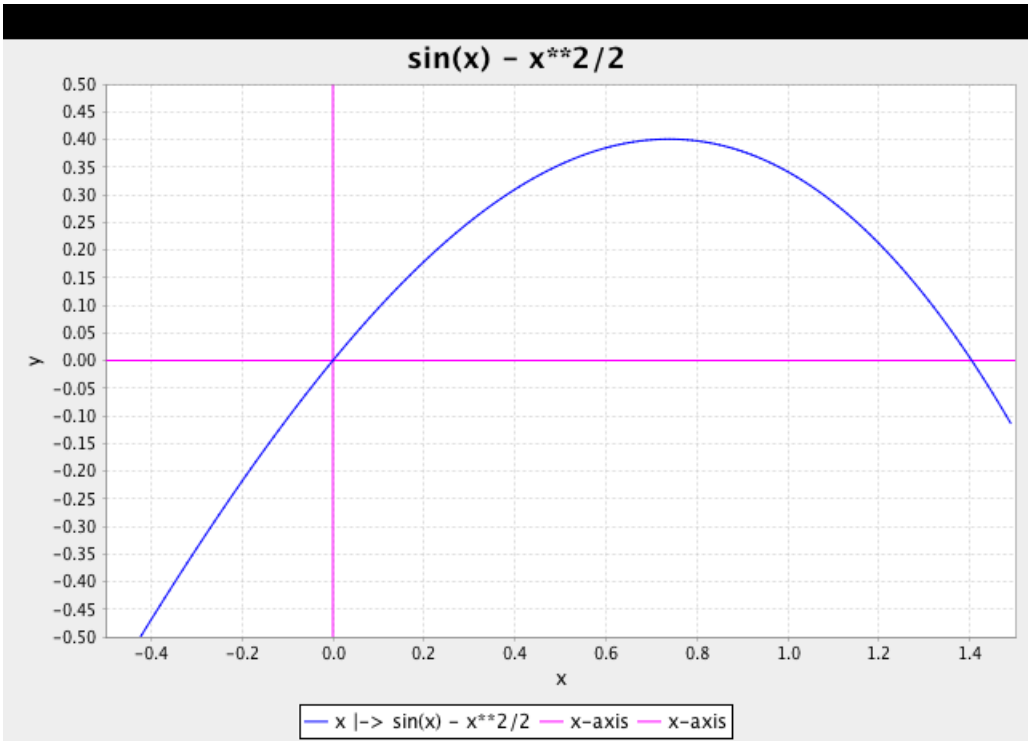

Figure 6.1: The function $x \mapsto \sin(x) - \frac{1}{2} \cdot x^2$.

derivative of f. This derivative is given as

$$f'(x) = \cos(x) - x$$

As it happens, the equation $\cos(x) - x = 0$ does not seem to have a solution in closed form. Hence, we can only approximate the solution numerically via a sequence of numbers $(x_n)_{n\in\mathbb{N}}$ such that the limit $\lim\limits_{n\to\infty} x_n$ exists and is a solution of the equation $\cos(x) - x = 0$, i.e. we want to have

$$\cos\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} x_n.$$

The method of gradient ascent is a numerical method that can be used to find the maximum of a function

$$f : \mathbb{R}^n \to \mathbb{R}$$

numerically. The basic idea is to take a vector $\mathbf{x}_0 \in \mathbb{R}^n$ as the start value and define a sequence of vectors $(\mathbf{x}_n)_{n\in\mathbb{N}}$ such that we have

$$f(\mathbf{x}_{n+1}) \geq f(\mathbf{x}_n) \quad \text{for all } n \in \mathbb{N}.$$

Hopefully, this sequence will converge against $\widehat{\mathbf{x}} = \arg\max\limits_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. If we do not really know where to start our search, we define $\mathbf{x}_0 := \mathbf{0}$. In order to compute $\mathbf{x}_{n+1}$ given $\mathbf{x}_n$, the idea is to move from $\mathbf{x}_n$ in that direction where we have the biggest change in the values of f. This direction happens to be the gradient of f at $\mathbf{x}_n$. Therefore, the definition of $\mathbf{x}_{n+1}$ is given as follows:

$$\mathbf{x}_{n+1} := \mathbf{x}_n + \alpha \cdot \nabla f(\mathbf{x}_n) \quad \text{for all } n \in \mathbb{N}_0.$$

Here, α is called the step size also known as the learning rate. It determines by how much we move in the direction of the gradient. In practise, it is best to adapt the step size dynamically during the iterations. Figure 6.2 on page 95 shows how this is done. The function `findMaximum` takes four arguments:

1. `f` is the function that is to be maximized. It is assumed that `f` takes a vector `x` $\in \mathbb{R}^n$ as its input and that it returns a real number.

2. `gradF` is the gradient of `f`. It takes a vector $\mathtt{x} \in \mathbb{R}^n$ as its input and returns the vector $\nabla \mathtt{f}(\mathtt{x})$.
3. `start` is a vector from $\mathbb{R}^n$ that is used as the value of $\mathbf{x}_0$. In practice, we will often use $\mathbf{0} \in \mathbb{R}^n$ as the start vector.
4. `eps` is the precision that we need for the maximum. We will have to say more on how `eps` is related to the precision later. As we are using double precision floating point arithmetic, it won't make sense to use a value for `eps` that is smaller than 10^{-15}.

Next, let us discuss the implementation of gradient ascent.

1. `x` is initialized with the parameter `start`. Hence, `start` is really the same as $\mathbf{x}_0$.
2. `fx` is the value that the function f takes for the argument `x`.
3. `alpha` is the learning rate. We initialize `alpha` as 1.0. The learning rate will be adapted dynamically.
4. The body of the `while` loop starting in line 5 executes one iteration of gradient ascent.
5. In each iteration, we store the values of $\mathbf{x}_n$ and $f(\mathbf{x}_n)$ in the variables `xOld` and `fOld`. This is needed since we need to ensure that the values of $f(\mathbf{x}_n)$ are increasing.
6. Next, we compute $\mathbf{x}_{n+1}$ in line 7 using the formula
$$\mathbf{x}_{n+1} := \mathbf{x}_n + \alpha \cdot \nabla f(\mathbf{x}_n).$$
7. The corresponding value $f(\mathbf{x}_{n+1})$ is computed in line 8.
8. If we are unlucky, $f(\mathbf{x}_{n+1})$ is smaller than $f(\mathbf{x}_n)$. This happens if the step size α is too large. Hence, in this case we decrease the value of α, discard both $\mathbf{x}_{n+1}$ and $f(\mathbf{x}_{n+1})$ and start over again.
9. Otherwise, $\mathbf{x}_{n+1}$ is a better approximation of the maximum than $\mathbf{x}_n$. In order to increase the speed of the convergence of our algorithm we will then increase the learning rate α by 20%.
10. The idea of our implementation is to stop the iteration when the relative difference of $f(\mathbf{x}_{n+1})$ and $f(\mathbf{x}_n)$ is less than ε or, to be more precise, if
$$f(\mathbf{x}_{n+1}) < f(\mathbf{x}_n) \cdot (1 + \varepsilon).$$
As the sequence $\big(f(\mathbf{x}_n)\big)_{n \in \mathbb{N}}$ will be monotonically increasing, i.e. we have
$$f(\mathbf{x}_{n+1}) \geq f(\mathbf{x}_n) \quad \text{for all } n \in \mathbb{N},$$
the condition given above is sufficient. Now, if the increase of $f(\mathbf{x}_{n+1})$ is less than $f(\mathbf{x}_n) \cdot (1 + \varepsilon)$ we assume that we have reached the maximum with the required precision. In this case we return both the value of `x` and the corresponding function value $f(\mathtt{x})$.

The implementation of gradient ascent given above is not the most sophisticated variant of this algorithm. Furthermore, there are algorithms that are more powerful than gradient ascent. The first of these methods is the conjugate gradient method. A refinement of this method is the BFGS-algorithm that has been invented by Broyden, Fletcher, Goldfarb, and Shanno. Unfortunately, we do not have the time to discuss this algorithm. However, our implementation of gradient ascent is sufficient for our applications and as this is not a course on numerical analysis but rather on artificial intelligence we will not delve deeper into this topic but, instead, we refer readers interested in more efficient algorithms to the literature [Sny05]. If you ever need to find the maximum of a function numerically, you should try to use a predefined library routine that implements a state of the art algorithm. For example, in Python the method minimize from the package `scipy.optimize` offers various algorithms for minimization.

```
findMaximum := procedure(f, gradF, start, eps) {
    x     := start;
    fx    := f(x);
    alpha := 1.0;
    while (true) {
        [xOld, fOld] := [x, fx];
        x  += alpha * gradF(x);
        fx := f(x);
        if (fx < fOld) {
            alpha   *= 0.5;
            [x, fx] := [xOld, fOld];  // reset values and
            continue;                 // start over
        } else {
            alpha *= 1.2;
        }
        if (abs(fx - fOld) <= abs(fx) * eps) {
            return [x, fx];
        }
    }
};
```

Figure 6.2: The gradient ascent algorithm.

6.3 Logistic Regression

If we have a model such that

$$\texttt{model}(\mathbf{x}, \mathbf{w}) \approx \texttt{classify}(\mathbf{x})$$

we want to choose the weight vector $\mathbf{w}$ in a way such that the accuracy

$$\texttt{accuracy}(\mathbf{w}) := \frac{\#\big\{\mathbf{o} \in S \;\big|\; \texttt{model}\big(\texttt{feature}(\mathbf{o}), \mathbf{w}\big) = \texttt{classify}(\mathbf{o})\big\}}{\#S}$$

is maximized. However, there is a snag: The accuracy is not a smooth function of the weight vector $\mathbf{w}$. It can't be because the number of errors of our model is a natural number and not a real number that could change smoothly when the weight vector $\mathbf{w}$ is changed. Hence, the accuracy is not differentiable as a function of the weight vector. The way to proceed is to work with probabilities instead. Instead of assigning a class to an object $\mathbf{o}$ we rather assign a probability p to the object $\mathbf{o}$ that measures how probable it is that object $\mathbf{o}$ has a given class c. Then we try to maximize this probability. In logistic regression we use a linear model that is combined with the sigmoid function. Before we can discuss the details of logistic regression we need to define this function and state some of its properties.

6.3.1 The Sigmoid Function

Definition 6 (Sigmoid Function) The sigmoid function $S : \mathbb{R} \to [0, 1]$ is defined as

$$S(t) = \frac{1}{1 + \exp(-t)}.$$

Figure 6.3 on page 96 shows the sigmoid function. ◇

Let us note some immediate consequences of the definition of the sigmoid function. As we have

$$\lim_{x \to -\infty} \exp(-x) = \infty, \quad \lim_{x \to +\infty} \exp(-x) = 0, \quad \text{and} \quad \lim_{x \to \infty} \frac{1}{x} = 0,$$

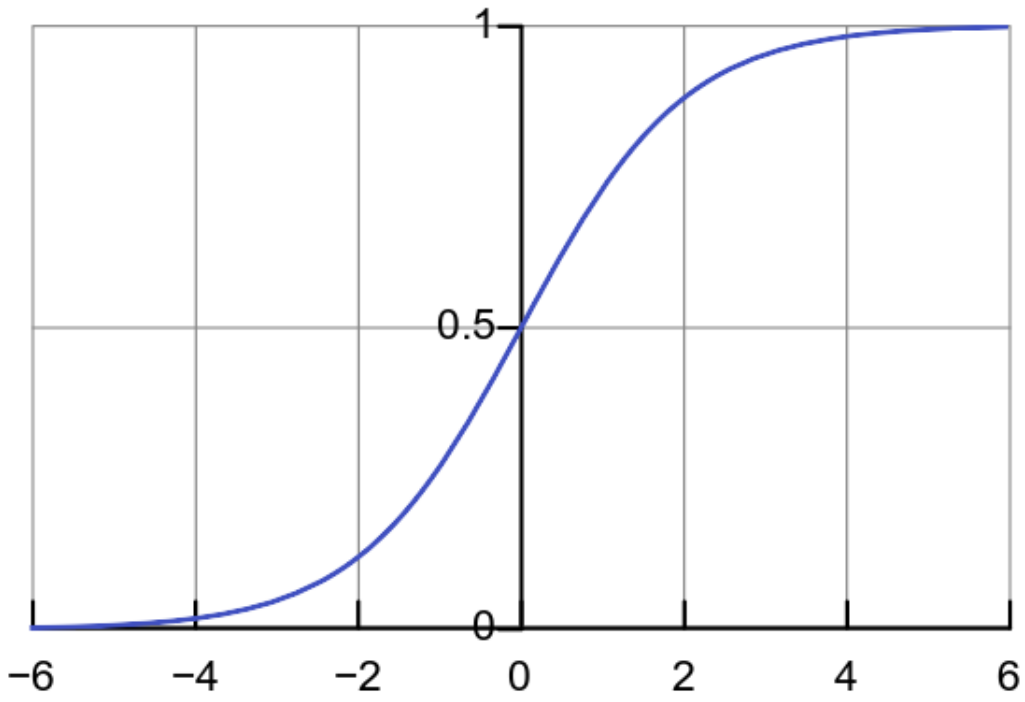

Figure 6.3: The sigmoid function.

the sigmoid function has the following properties:

$$\lim_{t\to-\infty} S(t) = 0 \quad \text{and} \quad \lim_{t\to+\infty} S(t) = 1.$$

As the sigmoid function is monotonically increasing, this shows that indeed

$$0 \leq S(t) \leq 1 \quad \text{for all } t \in \mathbb{R}.$$

Therefore, the value of the sigmoid function can be interpreted as a probability. Another important property of the sigmoid function is its symmetry. Figure 6.3 shows that if the sigmoid function is shifted down by $\frac{1}{2}$, the resulting function is centrally symmetric, i.e. we have

$$S(-t) - \frac{1}{2} = -\Big(S(t) - \frac{1}{2}\Big).$$

Adding $\frac{1}{2}$ on both sides of this equation shows that this is equivalent to the equation

$$\boxed{S(-t) = 1 - S(t),}$$

The proof of this fact runs as follows:

$$
\begin{aligned}
1 - S(t) &= 1 - \frac{1}{1+\exp(-t)} && \text{by definition of } S(t) \\
&= \frac{1+\exp(-t)-1}{1+\exp(-t)} && \text{common denominator} \\
&= \frac{\exp(-t)}{1+\exp(-t)} && \text{simplify} \\
&= \frac{1}{\exp(+t)+1} && \text{expand fraction by } \exp(t) \\
&= \frac{1}{1+\exp(+t)} \\
&= S(-t). \quad \square && \text{by definition of } S(-t)
\end{aligned}
$$

The exponential function can be expressed via the sigmoid function. Let us start with the definition of the sigmoid function.

$$S(t) = \frac{1}{1+\exp(-t)}$$

Multiplying this equation with the denominator yields

$$S(t) \cdot \big(1+\exp(-t)\big) = 1.$$

Dividing both sides by $S(t)$ gives:

$$
\begin{aligned}
& 1+\exp(-t) = \frac{1}{S(t)} \\
\Leftrightarrow \quad & \exp(-t) = \frac{1}{S(t)} - 1 \\
\Leftrightarrow \quad & \exp(-t) = \frac{1-S(t)}{S(t)}
\end{aligned}
$$

We highlight this formula, as we need it later

$$\boxed{\exp(-t) = \frac{1-S(t)}{S(t)}.}$$

If we take the reciprocal of both sides of this equation, we have

$$\exp(t) = \frac{S(t)}{1-S(t)}.$$

Applying the natural logarithm on both sides of this equation yields

$$t = \ln\left(\frac{S(t)}{1-S(t)}\right).$$

This shows that the inverse of the sigmoid function is given as

$$\boxed{S^{-1}(y) = \ln\left(\frac{y}{1-y}\right).}$$

This function is known as the logit function. Next, let us compute the derivative of $S(t)$, i.e. $S'(t) = \frac{\mathrm{d}S}{\mathrm{d}t}$. We have

$$\begin{aligned} S'(t) &= -\frac{-\exp(-t)}{\big(1+\exp(-t)\big)^2} \\ &= \exp(-t)\cdot S(t)^2 \\ &= \frac{1-S(t)}{S(t)}\cdot S(t)^2 \\ &= \big(1-S(t)\big)\cdot S(t) \end{aligned}$$

We have shown

$$\boxed{S'(t) = \big(1-S(t)\big)\cdot S(t).}$$

We will later need the derivative of the natural logarithm of the logistic function. We define

$$L(t) := \ln\big(S(t)\big).$$

Then we have

$$\begin{aligned} L'(t) &= \frac{S'(t)}{S(t)} && \text{by the chain rule} \\ &= \frac{(1-S(t))\cdot S(t)}{S(t)} \\ &= 1-S(t) && \text{simplify} \\ &= S(-t) && \text{by the symmetry} \end{aligned}$$

As this is our most important result, we highlight it:

$$\boxed{L'(t) = S(-t) \quad \text{where} \quad L(t) := \ln\big(S(t)\big).}$$

6.3.2 The Model of Logistic Regression

In logistic regression we deal with binary classification, i.e. we assume that we just need to decide whether a given object has a fixed class or not. We use the following model to compute the probability that an object with features $\mathbf{x}$ will be of the given class:

$$P(y=+1 \mid \mathbf{x};\mathbf{w}) = S(\mathbf{x}\cdot\mathbf{w}).$$

Here $\mathbf{x}\cdot\mathbf{w}$ denotes the dot product of the vectors $\mathbf{x}$ and $\mathbf{w}$, i.e. we have

$$\mathbf{x}\cdot\mathbf{w} = \sum_{i=1}^{D} x_i\cdot w_i.$$

To simplify calculations, it is assumed that $\mathbf{x}$ contains a constant feature that always takes the value of 1. Seeing this model the first time you might think that this model is not very general and that it can only be applied in very special circumstances. However, the fact is that the features x_i can be functions of arbitrary complexity and hence this model is much more general than it appears on first sight.

We assume that y can only take the values $+1$ or -1, e.g. in the example of spam detection $y=1$ if the email is spam and $y=-1$ otherwise. Since complementary probabilities add up to 1, we have

$$P(y=-1 \mid \mathbf{x};\mathbf{w}) = 1 - P(y=+1 \mid \mathbf{x};\mathbf{w}) = 1 - S(\mathbf{x}\cdot\mathbf{w}) = S(-\mathbf{x}\cdot\mathbf{w}).$$

Hence, we can combine the equations for $P(y=-1 \mid \mathbf{x};\mathbf{w})$ and $P(y=+1 \mid \mathbf{x};\mathbf{w})$ into a single equation

$$\boxed{P(y \mid \mathbf{x};\mathbf{w}) = S\big(y\cdot(\mathbf{x}\cdot\mathbf{w})\big).}$$

Given N objects $o_1,\cdots,o_n$ with feature vectors $\mathbf{x}_1,\cdots,\mathbf{x}_n$ we want to determine the weight vector $\mathbf{w}$ such

that the likelihood $\ell(\mathbf{X}, \mathbf{y})$ of all of our observations is maximized. This approach is called the maximum likelihood estimation of the weights. As we assume the probabilities of different observations are independent, the individual probabilities have to be multiplied to compute the overall likelihood $\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ of a given training set:

$$\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) = \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i; \mathbf{w}).$$

Here, we have combined the different attribute vectors $\mathbf{x}_i$ into the matrix $\mathbf{X}$ such that $\mathbf{x}_i$ is the i-th row of the matrix $\mathbf{X}$. Since it is easier to work with sums than with products, instead of maximizing the function $\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ we maximize the function

$$\ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) := \ln\big(\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})\big).$$

As the natural logarithm is a monotonically increasing function, the functions $\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ and $\ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ take their maximum at the same value of $\mathbf{w}$. As we have

$$\ln(a \cdot b) = \ln(a) + \ln(b),$$

the natural logarithm of the likelihood is

$$\ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) = \sum_{i=1}^{N} \ln\Big(S\big(y_i \cdot (\mathbf{x}_i \cdot \mathbf{w})\big)\Big) = \sum_{i=1}^{N} L\big(y_i \cdot (\mathbf{x}_i \cdot \mathbf{w})\big).$$

Our goal is to maximize the likelihood. Since this is the same as maximizing the log-likelihood, we need to determine those values of the coefficients $\mathbf{w}$ that satisfy

$$\frac{\partial}{\partial\, w_j} \ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) = 0.$$

In order to compute the partial derivative of $\ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ with respect to the coefficients $\mathbf{w}$ we need to compute the partial derivative of the dot product $\mathbf{x}_i \cdot \mathbf{w}$ with respect to the weights w_j. We define

$$h(\mathbf{w}) := \mathbf{x}_i \cdot \mathbf{w} = \sum_{k=1}^{D} x_{i,k} \cdot w_k.$$

Then we have

$$\frac{\partial}{\partial\, w_j} h(\mathbf{w}) = \frac{\partial}{\partial\, w_j} \sum_{k=1}^{D} x_{i,k} \cdot w_k = \sum_{k=1}^{D} x_{i,k} \cdot \frac{\partial}{\partial\, w_j} w_k = \sum_{k=1}^{D} x_{i,k} \cdot \delta_{j,k} = x_{i,j}.$$

Now we are ready to compute the partial derivative of $\ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w})$ with respect to $\mathbf{w}$:

$$\begin{aligned}
&\frac{\partial}{\partial\, w_j} \ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) \\
= \;& \frac{\partial}{\partial\, w_j} \sum_{i=1}^{N} L\big(y_i \cdot (\mathbf{x}_i \cdot \mathbf{w})\big) \\
= \;& \sum_{i=1}^{N} y_i \cdot x_{i,j} \cdot S\big(-y_i \cdot (\mathbf{x}_i \cdot \mathbf{w})\big), \quad \text{since} \quad \frac{\mathrm{d}L(x)}{\mathrm{d}x} = S(-x).
\end{aligned}$$

Hence, the partial derivative of the log-likelihood function is given as follows:

$$\frac{\partial}{\partial\, w_j} \ell\ell(\mathbf{X}, \mathbf{y}; \mathbf{w}) = \sum_{i=1}^{N} y_i \cdot x_{i,j} \cdot S(-y_i \cdot \mathbf{x}_i \cdot \mathbf{w})$$

Next, we have to find the value of $\mathbf{w}$ such that

$$\sum_{i=1}^{N} y_i \cdot x_{i,j} \cdot S(-y_i \cdot \mathbf{x}_i \cdot \mathbf{w}) = 0 \quad \text{for all } j \in \{1, \cdots, D\}.$$

These are D equations for the D variables $w_1, \cdots, w_D$. Due to the occurrence of the sigmoid function, these equations are nonlinear. We can not solve these equations explicitly. Nevertheless, our computation of the gradient of the log-likelihood was not for nought: We will use gradient ascent in order to find the value of $\mathbf{w}$ that maximizes the log-likelihood. This method has been outlined in the previous section.

6.3.3 Implementing Logistic Regression

In order to implement logistic regression we need a data structure for tabular data. Figure 6.4 on page 100 shows the class table that can be used to administer this kind of data. Figure 6.4 shows an example of tabular data that is stored in a Csv file. In this case, the data stores the hours a student has learned for a particular exam and the fact whether the student has passed of failed. The first column stores pass or fail, where a pass is coded using the number 1, while a fail is coded as 0. The second column stores the number of hours that the student has learned in order to pass the exam.

```
class table(columnNames, types, data) {
    mColumnNames := columnNames;
    mTypes       := types;
    mData        := data;

  static {
      getColumnNames := [ ] |-> mColumnNames;
      getTypes       := [ ] |-> mTypes;
      getData        := [ ] |-> mData;
      getRow         := [r] |-> mData[r];
      getLength      := []  |-> #mData;

      head := procedure(limit := 10) {
          print(mColumnNames);
          print(mTypes);
          for (i in [1 .. limit]) {
              print(mData[i]);
          }
      };
  }
}
readTable := procedure(fileName, types) {
    all         := readFile(fileName);
    columnNames := split(all[1], ',\s*');
    data        := [];
    for (i in [2 .. #all]) {
        row := split(all[i], ',\s*');
        data[i-1] := [eval("$type$($s$)") : [type, s] in types >< row];
    }
    return table(columnNames, types, data);
};
```

Figure 6.4: A class to represent tabular data.

There is no need for us to discuss every detail of the implementation of the class `table`. The important thing to note is that the data is stored as a list of lists in the member variable `mData`. Each of the inner lists corresponds to one row of the Csv file. This member variable can be accessed using the function getData. The function `readTable` has the responsibility to read a Csv file and to convert it into an object of class `table`. In order to do this, it has to be called with two arguments. The first argument is the file name, the second argument is a list of the types of each column in the Csv file. For example, to read the file "`exam.csv`" we would call `readTable` as follows:

```
readTable("exam.csv", ["int", "double"]).
```

```
Pass, Hours
0,    0.50
0,    0.75
0,    1.00
0,    1.25
0,    1.50
0,    1.75
1,    1.75
0,    2.00
1,    2.25
0,    2.50
1,    2.75
0,    3.00
1,    3.25
0,    3.50
1,    4.00
1,    4.25
1,    4.50
1,    4.75
1,    5.00
1,    5.50
```

Figure 6.5: Results of an exam.

The program shown in Figure 6.6 on page 102 implements logistic regression. As there are a number of subtle points that might easily be overlooked otherwise, we proceed to discuss this program line by line.

1. First, we have to load both the class `table` and our implementation of gradient ascent that has already been discussed in Section 6.2.

2. Line 4 implements the sigmoid function

$$S(x) = \frac{1}{1 + \exp(-x)}.$$

3. Line 5 starts the implementation of the natural logarithm of the sigmoid function, i.e. we implement

$$L(x) = \ln\bigl(S(X)\bigr) = \ln\left(\frac{1}{1 + \exp(-x)}\right) = -\ln\bigl(1 + \exp(-x)\bigr).$$

The implementation is more complicated than you might expect. The reason has to do with overflow. Consider values of x that are smaller than, say, -1000. The problem is that the expression `exp`$(1000))$ evaluates to `Infinity`, which represents the mathematical value ∞. But then $1 +$ `exp`$(1000))$ is also `Infinity` and finally `log(1 + exp(1000))` is `Infinity`. However, in reality we have

$$\ln\bigl(1 + \exp(1000)\bigr) \approx 1000.$$

The argument works as follows:

```
load("table.stlx");
load("gradient-ascent.stlx");

sigmoid := procedure(x) { return 1.0 / (1.0 + exp(-x)); };
logSigmoid := procedure(x) {
    if (x >= -100) {
        return -log(1.0 + exp(-x));
    } else {
        return x;
    }
};
ll := procedure(X, y, w) {
    return +/ [ logSigmoid(y[i] * (X[i] * w)) : i in [1 .. #X] ];
};
gradLL := procedure(X, y, w) {
    result := [];
    for (j in [1 .. #X[1]]) {
        result[j] := +/ [y[i]*X[i][j]*sigmoid(-y[i]*(X[i]*w)) : i in [1..#X]];
    }
    return la_vector(result);
};
logisticRegressionFile := procedure(fileName, types) {
    csv    := readTable(fileName, types);
    data   := csv.getData();
    number := #data;
    dmnsn  := #data[1];
    yList  := [];
    xList  := [];
    for (i in [1 .. number]) {
        yList[i] := data[i][1];
        xList[i] := la_vector([1.0] + data[i][2..]);
    }
    X := la_matrix(xList);
    y := la_vector([2 * y - 1 : y in yList]);
    start := la_vector([0.0 : i in [1 .. dmnsn]]);
    eps   := 10 ** -15;
    f     := w |=> ll(X, y, w);
    gradF := w |=> gradLL(X, y, w);
    return findMaximum(f, gradF, start, eps)[1];
};
```

Figure 6.6: An implementation of logistic regression.

$$
\begin{array}{rcll}
\ln\bigl(1+\exp(x)\bigr) & = & \ln\bigl(\exp(x)\cdot(1+\exp(-x))\bigr) & \\
& = & \ln\bigl(\exp(x)\bigr)+\ln\bigl(1+\exp(-x)\bigr) & \\
& = & x+\ln\bigl(1+\exp(-x)\bigr) & \\
& \approx & x+\ln(1)+\exp(-x) & \text{Taylor expansion of } \ln(1+x) \\
& = & x+0+\exp(-x) & \\
& \approx & x & \text{since } \exp(-x)\approx 0 \text{ for large } x
\end{array}
$$

This is the reason that `logSigmoid` returns `x` if the value of `x` is less than -100.

4. The function $\texttt{ll}(\mathbf{x},\mathbf{y},\mathbf{w})$ computes the log-lokelihood

$$\ell\ell(\mathbf{X},\mathbf{y};\mathbf{w}) = \sum_{i=1}^{N} L\bigl(y_i\cdot(\mathbf{x}_i\cdot\mathbf{w})\bigr).$$

Here L denotes the natural logarithm of the sigmoid of the argument. It is assumed that $\mathbf{X}$ is a matrix. Every observation corresponds to a row in this matrix, i.e. the vector $\mathbf{x}_i$ is the feature vector containing the features of the i-th observation. $\mathbf{y}$ is a vector describing the outcomes, i.e. the elements of this vector are either $+1$ or -1. Finally, $\mathbf{w}$ is the vector of coefficients.

5. The function $\texttt{gradLL}(\mathbf{x},\mathbf{y},\mathbf{w})$ computes the gradient of the log-lokelihood according to the formula

$$\frac{\partial}{\partial\, w_j}\ell\ell(\mathbf{X},\mathbf{y};\mathbf{w}) = \sum_{i=1}^{N} y_i\cdot x_{i,j}\cdot S(-y_i\cdot\mathbf{x}_i\cdot\mathbf{w}).$$

The different components of this gradient are combined into a vector. The arguments are the same as the arguments to the log-lokelihood.

6. Finally, the function `logisticRegressionFile` takes two arguments. The first argument is the name of the `.csv` file containing the data, while the second argument is a list specifying the types of the columns. The elements of this list have to be either `"int"` or `"double"`. The task of this function is to read the `.csv` file, convert the data in the matrix $\mathbf{X}$ and the vector $\mathbf{y}$, and then use the method of gradient ascent to find the weight vector $\mathbf{w}$ that maximize the likelihood.

As we assume that the entries of the vector $\mathbf{y}$ are either $+1$ or -1 but the data provided in the file are instead 1 or 0, we need to apply a function that maps 1 to $+1$ and 0 to -1. The function

$$y \mapsto 2\cdot y - 1$$

is used for this mapping in the construction of the vector $\mathbf{y}$.

Next, we define the functions `f` and `gradF` as closures using the operator "`|=>`". This operator ensures that the values of the matrix `X` and the vector `y` are available on invocations of `f` and `gradF`.

Finally, the function `findMaximum` computes the maximum of the log-likelihood using gradient ascent.

If we run the function `logisticRegressionFile` using the data shown in Figure 6.5 via the command

```
logisticRegressionFile("exam.csv", ["int", "double"]);
```

the resulting coefficients are:

```
<<-4.077649741107752 1.5046211108850898>>
```

This shows that the probability $P(h)$ that a student who has studied for h hours will pass the exam is given approximately as follows:

$$P(h) \approx \frac{1}{1+\exp(4.1-1.5\cdot h)}$$

Figure 6.7 shows a plot of the probability $P(x)$. This figure has been taken from the Wikipedia article on logistic regression. It has been created by Michaelg2015.

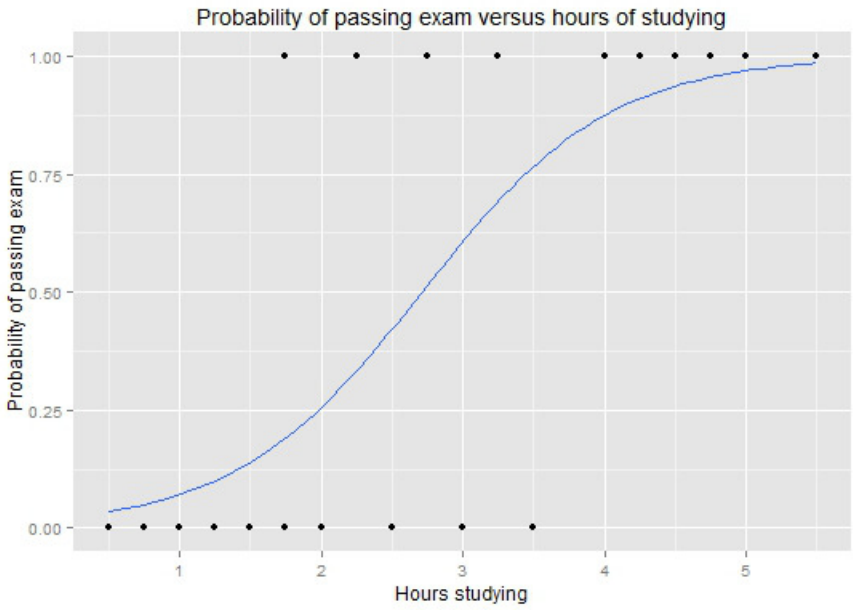

Figure 6.7: Probability of passing an exam versus hours of studying.

Chapter 7

Neural Networks

In this chapter, we discuss neural networks. Many of the most visible breakthroughs in artificial intelligence have been achieved through the use of neural networks:

1. The current system used by Google to automatically translate web pages is called "Google Neural Machine Translation" and, as the name suggests, is based on neural networks.
2. DeepL is another translator that is based on deep neural networks.
3. AlphaGo uses neural networks together with tree search [SHM+16]. It has beaten beaten the world champion Ke Jie in the game of Go.
4. Image recognition is best done via neural networks.
5. Autonomous driving makes heavy use of neural networks.

The list given above is far from being complete. In this chapter, we will only discuss feedforward neural networks. Although recently both convolutional neural networks and recurrent neural networks have gotten a lot of attention, these type of neural networks are more difficult to train and are therefore beyond the scope of this introduction. The rest of this chapter is strongly influenced by the online book

http://neuralnetworksanddeeplearning.com/index.html

that has been written by Michael Nielsen [Nie15]. This book is easy to read, carefully written, and free to access. I recommend this book to anybody who wants to dive deeper into the fascinating topic of neural networks. Furthermore, I can recommend the specialization Deep Learning that is offered by Coursera.

7.1 Feedforward Neural Networks

A neural network is built from neurons. Neural networks are inspired by biological neurons. However, in order to understand artificial neural networks it is not necessary to know how biological neurons work and it is definitely not necessary to understand how networks of biological neurons, i.e. brains, work[1]. Instead, we develop a mathematical abstraction of neurons that will serve as the foundation of the theory developed in this chapter. At the abstraction level that we are looking at neural networks, a single neuron with n inputs is specified by a pair $\langle \mathbf{w}, b \rangle$ where the vector $\mathbf{w} \in \mathbb{R}^m$ is called the weight vector and the number $b \in \mathbb{R}$ is called the bias. Conceptually, a neuron is a function that maps an input vector $\mathbf{x} \in \mathbb{R}^m$ into the interval $[0, 1]$. This function is defined as follows:

$$\mathbf{x} \mapsto a(\mathbf{x} \cdot \mathbf{w} + b).$$

Here, a is called the activation function. In our applications, we will always use the sigmoid function (Definition 6) as our activation function, i.e. we have

[1] Actually, when it comes to brains, although there are many speculations, surprisingly little is known for a fact.

$$a(t) := S(t) = \frac{1}{1+\exp(-t)}.$$

The function modelling the neuron can be written more explicitly using index notation. If

$$\mathbf{w} = \langle w_1, \cdots, w_m \rangle^\top$$

is the weight vector and

$$\mathbf{x} = \langle x_1, \cdots, x_m \rangle^\top$$

is the input vector, then we have

$$\mathbf{x} \mapsto S\left(\left(\sum_{i=1}^{m} x_i \cdot w_i\right) + b\right).$$

If we compare this function to a similar function appearing in the last chapter, you will notice that so far a neuron works just like logistic regression. The only difference is that the bias b is now explicit in our notation. In logistic regression, we had assumed that the first component x_1 of our feature vector $\mathbf{x}$ was always equal to 1. This assumption enabled us to incorporate the bias b into the weight vector $\mathbf{w}$.

A feedforward neural network is a layered network of neurons. Formally, the topology of a neural network is given by a number $L \in \mathbb{N}$ and a list $[m(1), \cdots, m(L)]$ of L natural numbers. The number L is called the number of layers and for $i \in \{2, \cdots, L\}$ the number $m(i)$ is the number of neurons in the l-th layer. The first layer is called the input layer. The input layer does not contain neurons but instead just contains input nodes. The last layer (i.e. the layer with index L) is called the output layer and the remaining layers are called hidden layers. If there is more than one hidden layer, the neural network is called a deep neural network. Figure 7.1 on page 106 shows a small neural network with two hidden layers.

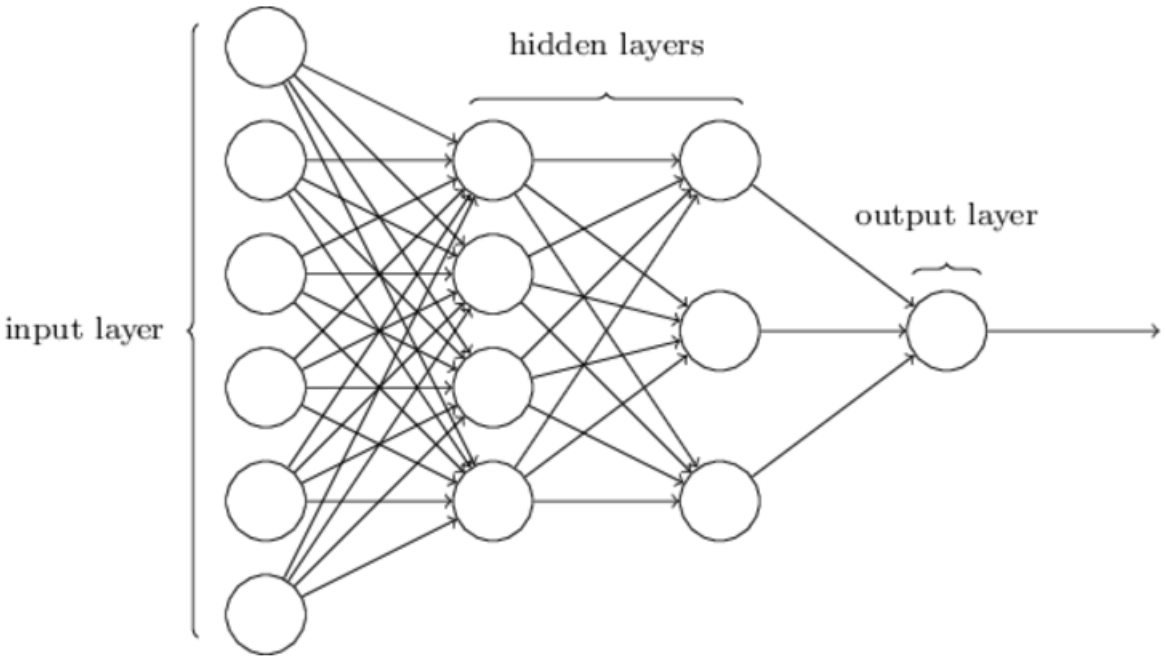

Figure 7.1: A neural network taken from the book by Michael Nielsen[Nie15].

As the first layer is the input layer, the input dimension is defined as $m(1)$. Similarly, the output dimension is defined as $m(L)$. Every node in the l-th layer is connected to every node in the $(l+1)$-th layer via a weight. The weight $w_{j,k}^{(l)}$ is the weight of the connection from the k-th neuron in layer $l-1$ to the j-th neuron in layer l. The weights in layer l are combined into the weight matrix $W^{(l)}$ of the layer l: This matrix is defined as

$$W^{(l)} := \big(w^{(l)}_{j,k}\big).$$

Note that $W^{(l)}$ is an $m(l) \times m(l-1)$ matrix, i.e. we have

$$W^{(l)} \in \mathbb{R}^{m(l)\times m(l-1)}.$$

The j-th neuron in layer l has the bias $b^{(l)}_j$. These biases of layer l are combined into the bias vector

$$\mathbf{b}^{(l)} := \langle b^{(l)}_1, \cdots, b^{(l)}_{m(l)}\rangle^\top.$$

Then, the activation of the j-th neuron in layer l is denoted as $a^{(l)}_j$ and is defined recursively as follows:

1. For the input layer we have

$$a^{(1)}_j(\mathbf{x}) := x_j. \tag{FF1}$$

   To put it differently, the input vector $\mathbf{x}$ is the activation of the input nodes.

2. For all other layers we have

$$a^{(l)}_j(\mathbf{x}) := S\left(\left(\sum_{k=1}^{m(l-1)} w^{(l)}_{j,k} \cdot a^{(l-1)}_k(\mathbf{x})\right) + b^{(l)}_j\right) \quad \text{for all } l \in \{2, \cdots, L\}. \tag{FF2}$$

The activation vector of layer l is defined as

$$\mathbf{a}^{(l)}(\mathbf{x}) := \langle a^{(l)}_1(\mathbf{x}), \cdots, a^{(l)}_{m(l)}(\mathbf{x})\rangle^\top.$$

Using vector notation, the feed forward equations (FF1) and (FF2) can be written as follows:

$$\mathbf{a}^{(1)}(\mathbf{x}) := \mathbf{x}, \tag{FF1v}$$

$$\mathbf{a}^{(l)}(\mathbf{x}) := S\Big(W^{(l)} \cdot \mathbf{a}^{(l-1)}(\mathbf{x})\Big) + \mathbf{b}^{(l)}\Big) \quad \text{for all } l \in \{2, \cdots, L\}. \tag{FF2v}$$

The output of our neural network for an input $\mathbf{x}$ is given by the neurons in the output layer, i.e. the output vector $\mathbf{o}(\mathbf{x}) \in \mathbb{R}^{m(L)}$ is defined as

$$\mathbf{o}(\mathbf{x}) := \langle a^{(L)}_1(\mathbf{x}), \cdots, a^{(L)}_{m(L)}(\mathbf{x})\rangle^\top = \mathbf{a}^{(L)}(\mathbf{x}).$$

Note that the equations (FF1) and (FF2) describe how information propagates through the neural network:

1. Initially, the input vector $\mathbf{x}$ is given and stored in the input layer of the neural network:

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

2. The first layer of neurons, which is the second layer of nodes, is activated and computes the activation vector $\mathbf{a}^{(2)}$ according to the formula

$$\mathbf{a}^{(2)}(\mathbf{x}) := S\big(W^{(2)} \cdot \mathbf{a}^{(1)}(\mathbf{x}) + \mathbf{b}^{(2)}\big) = S\big(W^{(2)} \cdot \mathbf{x} + \mathbf{b}^{(2)}\big).$$

3. The second layer of neurons, which is the third layer of nodes, is activated and computes the activation vector $\mathbf{a}^{(3)}(\mathbf{x})$ according to the formula

$$\mathbf{a}^{(3)}(\mathbf{x}) := S\big(W^{(3)} \cdot \mathbf{a}^{(2)}(\mathbf{x}) + \mathbf{b}^{(3)}\big) = S\Big(W^{(3)} \cdot S\big(W^{(2)} \cdot \mathbf{x} + \mathbf{b}^{(2)}\big) + \mathbf{b}^{(3)}\Big)$$

4. This proceeds until the output layer is reached and the output

$$\mathbf{o}(\mathbf{x}) := \mathbf{a}^{(L)}(\mathbf{x})$$

   has been computed. Note that every neuron of the neural network performs logistic regression.

Next, we assume that we have n training examples

$$\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle \quad \text{for } i = 1, \cdots, n$$

such that

$$\mathbf{x}^{(i)} \in \mathbb{R}^{m(1)} \text{ and } \mathbf{y}^{(i)} \in \mathbb{R}^{m(L)}.$$

Our goal is to choose the weight matrices $W^{(l)}$ and the bias vectors $\mathbf{b}^{(l)}$ in a way such that

$$\mathbf{o}\big(\mathbf{x}^{(i)}\big) = \mathbf{y}^{(i)} \quad \text{for all } i \in \{1, \cdots, n\}.$$

Unfortunately, in general we will not be able to achieve equality for all $i \in \{1, \cdots, n\}$. Therefore, our goal is to minimize the error instead. To be more precise, the quadratic error cost function is defined as

$$C\Big(W^{(2)}, \cdots, W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}; \mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \cdots, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}\Big) := \frac{1}{2 \cdot n} \cdot \sum_{i=1}^{n} \Big(\mathbf{o}\big(\mathbf{x}^{(i)}\big) - \mathbf{y}^{(i)}\Big)^2.$$

Note that this cost function is additive in the training examples $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$. In order to simplify the notation we define

$$C_{\mathbf{x},\mathbf{y}}\Big(W^{(2)}, \cdots, W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}\Big) := \frac{1}{2} \cdot \Big(\mathbf{a}^{(L)}(\mathbf{x}) - \mathbf{y}\Big)^2,$$

i.e. $C_{\mathbf{x},\mathbf{y}}$ is the part of the cost function that is associated with a single training example $\langle \mathbf{x}, \mathbf{y} \rangle$. Then, we have

$$C\Big(W^{(2)}, \cdots, W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}; \mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \cdots, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}\Big) := \frac{1}{n} \cdot \sum_{i=1}^{n} C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}\Big(W^{(2)}, \cdots W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}\Big).$$

As the notation

$$C_{\mathbf{x},\mathbf{y}}\Big(W^{(2)}, \cdots, W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}\Big)$$

is far too heavy, we will abbreviate this term as $C_{\mathbf{x},\mathbf{y}}$ in the following discussion of the backpropagation algorithm. Similarly, we abbreviate the quadratic error cost function as C. Our goal is to choose the weight matrices $W^{(l)}$ and the bias vectors $\mathbf{b}^{(l)}$ such that the quadratic error cost function C is minimized. We will use a variation of gradient descent to find this minimum[2]. Unfortunately, the cost function C when regarded as a function of the weights and biases has many local minima. Hence, all we can hope for is to find a local minimum that is good enough.

7.2 Backpropagation

There are three reasons for the recent success of neural networks.

1. The computing power that is available today has vastly increased in the last 20 years. For example, today the AMD RX Vega 64 Liquid graphic card offers about 13.7 teraflops in single precision performance. It consumes about 350 watt. Contrast this with ASCI White, which was the most powerful supercomputer in 2000: In 2000 when it topped the rankings of the supercomputers, it offered a performance of 7.2 teraflops and needed 6 megawatt to operate. The cost to build ASCI White was about $110,000,000\,\$$. On the contrary, the AMD RX Vega 64 Liquid costs $699\,\$$.

   The Nvidia Titan V comes at $2,999\,\$$ and offers a stunning 110 teraflops for deep learning via the CUDA API.

2. The breakthrough in the theory of neural networks was the rediscovering of the backpropagation algorithm by David Rumelhart, Geoffrey Hinton, and Ronald Williams [RHW86] in 1986. The backpropagation algorithm had first been discovered by Arthur E. Bryson, Jr. and Yu-Chi Ho [BH69]. In recent years, there have been a number of other theoretical advances that have helped in speeding up the learning algorithms for neural networks.

[2] In logistic regression we have tried to *maximize* the log-likelihood. Here, instead we *minimize* the quadratic error cost function. Hence, instead of gradient *ascent* we use gradient *descent*.

3. Lastly, as neural networks have large sets of parameters, they need large sets of training examples. The recent digitization of our society has made these large data sets available.

Essentially, the backpropagation algorithm is an efficient way to compute the partial derivatives of the cost function C with respect to the weights $w_{j,k}^{(l)}$ and the biases $b_j^{(l)}$. Before we can proceed to compute these partial derivatives, we need to define some auxiliary variables.

7.2.1 Definition of some Auxiliary Variables

We start by defining the auxiliary variables $z_j^{(l)}$. The expressions $z_j^{(l)}$ are defined as the inputs of the activation function S of the j-th neuron in layer l:

$$z_j^{(l)} := \left(\sum_{k=1}^{m(l-1)} w_{j,k}^{(l)} \cdot a_k^{(l-1)} \right) + b_j^{(l)} \quad \text{for all } j \in \{1, \cdots, m(l)\} \text{ and } l \in \{2, \cdots, L\}.$$

Of course, the term $a_k^{(l-1)}$ really is a function of the input vector $\mathbf{x}$. However, it is better to suppress this dependence in the notation since otherwise the formulæ get too cluttered. Essentially, $z_j^{(l)}$ is the input to the sigmoid function when the activation $a_j^{(l)}$ is computed, i.e. we have

$$a_j^{(l)} = S\Big(z_j^{(l)}\Big).$$

We will later see that the partial derivatives of the cost function $C_{\mathbf{x},\mathbf{y}}$ with respect to both the weights $w_{j,k}^{(l)}$ and the biases $b_j^{(l)}$ can be computed easily if we first compute the partial derivatives of $C_{\mathbf{x},\mathbf{y}}$ with respect to $z_j^{(l)}$. Therefore we define

$$\varepsilon_j^{(l)} := \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(l)}} \quad \text{for all } j \in \{1, \cdots, m(l)\} \text{ and } l \in \{2, \cdots, L\},$$

that is we regard $C_{\mathbf{x},\mathbf{y}}$ as a function of the $z_j^{(l)}$ and take the partial derivatives according to these variables. Note that $\varepsilon_j^{(l)}$ does depend on both $\mathbf{x}$ and $\mathbf{y}$. Since the notation would get too cumbersome if we would write $\varepsilon(\mathbf{x},\mathbf{y})_j^{(l)}$, we regard the training example $\langle \mathbf{x}, \mathbf{y} \rangle$ as fixed for now. Next, the quantities $\varepsilon_j^{(l)}$ are combined into a vector:

$$\boldsymbol{\varepsilon}^{(l)} := \begin{pmatrix} \varepsilon_1^{(l)} \\ \vdots \\ \varepsilon_{m(l)}^{(l)} \end{pmatrix}.$$

The quantity $\boldsymbol{\varepsilon}^{(l)}$ is called the error in layer l.

7.2.2 The Hadamard Product

Later, we will have need of the Hadamard product of two vectors. Assume that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The Hadamard product of $\mathbf{x}$ and $\mathbf{y}$ is a vector that is defined by multiplying the vectors $\mathbf{x}$ and $\mathbf{y}$ elementwise:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \odot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} := \begin{pmatrix} x_1 \cdot y_1 \\ x_2 \cdot y_2 \\ \vdots \\ x_n \cdot y_n \end{pmatrix},$$

i.e. the i-th component of the Hadamard product $\mathbf{x} \odot \mathbf{y}$ is the product of the i-th component of $\mathbf{x}$ with the i-th component of $\mathbf{y}$. Do not confuse the Hadamard product with the dot product! Although both multiply the vectors componentwise, the Hadamard product returns a vector, while the dot product returns a number.

7.2.3 Backpropagation: The Equations

Now we are ready to state the backpropagation equations. The first of these four equations reads as follows:

$$\varepsilon_j^{(L)} = (a_j^{(L)} - y_j) \cdot S'\big(z_j^{(L)}\big) \quad \text{for all } j \in \{1, \cdots, m(L)\}, \tag{BP1}$$

where $S'(x)$ denotes the derivative of the sigmoid function. We have shown in Chapter 6 that

$$S'(t) = \big(1 - S(t)\big) \cdot S(t)$$

holds. The equation (BP1) can also be written in vectorized form using the Hadamard product:

$$\boldsymbol{\varepsilon}^{(L)} = (\mathbf{a}^{(L)} - \mathbf{y}) \odot S'\big(\mathbf{z}^{(L)}\big) \tag{BP1v}$$

Here, we have vectorized the application of the function S' to the vector $\mathbf{z}^{(L)}$, i.e. the expression $S'\big(\mathbf{z}^{(L)}\big)$ is defined as follows:

$$S'\begin{pmatrix} z_1^{(L)} \\ \vdots \\ z_{m(L)}^{(L)} \end{pmatrix} := \begin{pmatrix} S'\big(z_1^{(L)}\big) \\ \vdots \\ S'\big(z_{m(L)}^{(L)}\big) \end{pmatrix}.$$

The next equation computes $\varepsilon_j^{(l)}$ for $l < L$.

$$\varepsilon_j^{(l)} = \sum_{i=1}^{m(l+1)} w_{i,j}^{(l+1)} \cdot \varepsilon_i^{(l+1)} \cdot S'\big(z_j^{(l)}\big) \quad \text{for all } j \in \{1, \cdots, m(l)\} \text{ and } l \in \{2, \cdots, L-1\}. \tag{BP2}$$

This equation is more succinct in vectorized notation:

$$\boldsymbol{\varepsilon}^{(l)} = \Big(\big(W^{(l+1)}\big)^\top \cdot \boldsymbol{\varepsilon}^{(l+1)}\Big) \odot S'\big(\mathbf{z}^{(l)}\big) \quad \text{for all } l \in \{2, \cdots, L-1\}. \tag{BP2v}$$

Note that this equation computes $\boldsymbol{\varepsilon}^{(l)}$ in terms of $\boldsymbol{\varepsilon}^{(l+1)}$: The error $\boldsymbol{\varepsilon}^{(l+1)}$ at layer $l+1$ is propagated backwards through the neural network to produce the error $\boldsymbol{\varepsilon}^{(l)}$ at layer l. This is the reason for calling the algorithm backpropagation.

Next, we have to compute the partial derivative of $C_{\mathbf{x},\mathbf{y}}$ with respect to the bias of the j-th neuron in layer l, which is denoted as $b_j^{(l)}$. We have

$$\frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial b_j^{(l)}} = \varepsilon_j^{(l)} \quad \text{for all } j \in \{1, \cdots, m(l)\} \text{ and } l \in \{2, \cdots, L\}. \tag{BP3}$$

In vectorized notation, this equation takes the following form:

$$\nabla_{\mathbf{b}^{(l)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(l)} \quad \text{for all } l \in \{2, \cdots, L\}. \tag{BP3v}$$

Here, $\nabla_{\mathbf{b}^{(l)}} C_{\mathbf{x},\mathbf{y}}$ denotes the gradient of $C_{\mathbf{x},\mathbf{y}}$ with respect to the bias $\mathbf{b}^{(l)}$. Finally, we can compute the partial derivative of $C_{\mathbf{x},\mathbf{y}}$ with respect to the weights:

$$\frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial w_{j,k}^{(l)}} = a_k^{(l-1)} \cdot \varepsilon_j^{(l)} \quad \text{for all } j \in \{1, \cdots, m(l)\},\ k \in \{1, \cdots, m(l-1)\}, \text{ and } l \in \{2, \cdots, L\}. \tag{BP4}$$

In vectorized notation, this equation can be written as:

$$\nabla_{W^{(l)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(l)} \cdot \big(\mathbf{a}^{(l-1)}\big)^\top \quad \text{for all } l \in \{2, \cdots, L\}. \tag{BP4v}$$

Here, the expression $\boldsymbol{\varepsilon}^{(l)} \cdot \big(\mathbf{a}^{(l-1)}\big)^\top$ denotes the matrix product of the column vector $\boldsymbol{\varepsilon}^{(l)}$ that is regarded as an $m(l) \times 1$ matrix and the row vector $\big(\mathbf{a}^{(l-1)}\big)^\top$ that is regarded as an $1 \times m(l-1)$ matrix.

As the backpropagation equations are at the very core of the theory of neural networks, we highlight them:

$$\boldsymbol{\varepsilon}^{(L)} = (\mathbf{a}^{(L)} - \mathbf{y}) \odot S'(\mathbf{z}^{(L)}) \tag{BP1v}$$

$$\boldsymbol{\varepsilon}^{(l)} = \Big((W^{(l+1)})^\top \cdot \boldsymbol{\varepsilon}^{(l+1)}\Big) \odot S'(\mathbf{z}^{(l)}) \quad \text{for all } l \in \{2, \cdots, L-1\} \tag{BP2v}$$

$$\nabla_{\mathbf{b}^{(l)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(l)} \quad \text{for all } l \in \{2, \cdots, L\} \tag{BP3v}$$

$$\nabla_{W^{(l)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(l)} \cdot (\mathbf{a}^{(l-1)})^\top \quad \text{for all } l \in \{2, \cdots, L\} \tag{BP4v}$$

The equations (BP3) and (BP4) show why it was useful to introduce the numbers $\varepsilon_j^{(l)}$: These numbers enable us to compute the partial derivatives of the cost function with respect to both the biases and the weights. Furthermore, the equations (BP1) and (BP2) show how these numbers can be computed. An implementation of backpropagation should use the vectorized versions of these equations since this is more efficient for two reasons:

1. Interpreted languages like SetlX, *Python*, or *Octave* take much more time to execute a loop than to execute a simple matrix-vector multiplication. The reason is that in a loop, in addition to executing the statement a given number of times, the statement has to be interpreted every time it is executed.

2. Languages that are optimized for machine learning often take care to delegate the execution of matrix operations to the graphical coprocessor which is optimized for these kinds of operations.

7.2.4 The Proof of the Backpropagation Equations

Next, we prove the backpropagation equations. Although the proof is a bit tedious, it should be accessible: The chain rule of multivariate calculus is all that is needed to understand why the backpropagation equations hold. As a reminder, the chain rule in multivariate calculus works as follows: Assume that the functions $f = f(\mathbf{y})$ and $g = g(\mathbf{x})$ where $\mathbf{y} \in \mathbb{R}^k$, $\mathbf{x} \in \mathbb{R}^n$, $g(\mathbf{x}) \in \mathbb{R}^k$, and $f(\mathbf{y}) \in \mathbb{R}$ are differentiable[3]. So we have

$$f : \mathbb{R}^k \to \mathbb{R} \quad \text{and} \quad g : \mathbb{R}^n \to \mathbb{R}^k.$$

If the function $h : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$h(\mathbf{x}) := f\big(g(\mathbf{x})\big) \quad \text{for all } \mathbf{x} \in \mathbb{R}^n,$$

then the partial derivative of h with respect to x_j satisfies

$$\frac{\partial h}{\partial x_j} = \sum_{i=1}^{k} \frac{\partial f}{\partial y_i} \cdot \frac{\partial g_i}{\partial x_j}.$$

Remember that we have defined the numbers $\varepsilon_j^{(l)}$ as

$$\varepsilon_j^{(l)} = \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(l)}},$$

while the numbers $z_j^{(l)}$ have been defined as

$$z_j^{(l)} := \left(\sum_{k=1}^{m(l-1)} w_{j,k}^{(l)} \cdot a_k^{(l-1)}(\mathbf{x}) \right) + b_j^{(l)}.$$

Since the quadratic error cost function $C_{\mathbf{x},\mathbf{y}}$ for the training example $\langle \mathbf{x}, \mathbf{y} \rangle$ has been defined in terms of the activation $\mathbf{a}^{(L)}(\mathbf{x})$ as

$$C_{\mathbf{x},\mathbf{y}} = \frac{1}{2} \cdot \big(\mathbf{a}^{(L)}(\mathbf{x}) - \mathbf{y}\big)^2$$

and we have $\mathbf{a}^{(L)}(\mathbf{x}) = S\big(\mathbf{z}^{(L)}\big)$, the chain rule tells us that $\varepsilon_j^{(L)}$ can be computed as follows:

[3] If this text had been written in German, I would have said that f and g are "total differenzierbar".

$$
\begin{aligned}
\varepsilon_j^{(L)} &= \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(L)}} \\
&= \frac{\partial}{\partial z_j^{(L)}} \frac{1}{2} \cdot \left(\mathbf{a}^{(L)}(\mathbf{x}) - \mathbf{y}\right)^2 \\
&= \frac{1}{2} \cdot \frac{\partial}{\partial z_j^{(L)}} \sum_{i=1}^{m(L)} \left(a_i^{(L)}(\mathbf{x}) - y_i\right)^2 \\
&= \frac{1}{2} \cdot \frac{\partial}{\partial z_j^{(L)}} \sum_{i=1}^{m(L)} \left(S\big(z_i^{(L)}\big) - y_i\right)^2 \\
&= \frac{1}{2} \cdot \sum_{i=1}^{m(L)} 2 \cdot \left(S\big(z_i^{(L)}\big) - y_i\right) \cdot \frac{\partial}{\partial z_j^{(L)}} S\big(z_i^{(L)}\big) \\
&= \sum_{i=1}^{m(L)} \left(S\big(z_i^{(L)}\big) - y_i\right) \cdot S'\big(z_i^{(L)}\big) \cdot \frac{\partial z_i^{(L)}}{\partial z_j^{(L)}} \\
&= \sum_{i=1}^{m(L)} \left(S\big(z_i^{(L)}\big) - y_i\right) \cdot S'\big(z_i^{(L)}\big) \cdot \delta_{i,j} \qquad \delta_{i,j} \text{ denotes the Kronecker delta} \\
&= \left(S\big(z_j^{(L)}\big) - y_j\right) \cdot S'\big(z_j^{(L)}\big) \\
&= \left(a_j^{(L)} - y_j\right) \cdot S'\big(z_j^{(L)}\big)
\end{aligned}
$$

Thus we have proved equation BP1. Next, let us compute $\varepsilon_j^{(l)}$ for $l < L$. We have

$$
\begin{aligned}
\varepsilon_j^{(l)} &= \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(l)}} \\
&= \sum_{i=1}^{m(l+1)} \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_i^{(l+1)}} \cdot \frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} \qquad \text{using the chain rule} \\
&= \sum_{i=1}^{m(l+1)} \varepsilon_i^{(l+1)} \cdot \frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} \qquad \text{using the definition of } \varepsilon_i^{(l+1)}
\end{aligned}
$$

In order to proceed, we have to remember the definition of $z_i^{(l+1)}$. We have

$$
z_i^{(l+1)} = \left(\sum_{k=1}^{m(l)} w_{i,k}^{(l+1)} \cdot S\big(z_k^{(l)}\big)\right) + b_i^{(l+1)}
$$

Therefore, the partial derivatives $\dfrac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}}$ can be computed as follows:

$$
\begin{aligned}
\frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} &= \sum_{k=1}^{m(l)} w_{i,k}^{(l+1)} \cdot S'\big(z_k^{(l)}\big) \cdot \frac{\partial z_k^{(l)}}{\partial z_j^{(l)}} \\
&= \sum_{k=1}^{m(l)} w_{i,k}^{(l+1)} \cdot S'\big(z_k^{(l)}\big) \cdot \delta_{k,j} \\
&= w_{i,j}^{(l+1)} \cdot S'\big(z_j^{(l)}\big)
\end{aligned}
$$

If we substitute this expression back into the result we got for $\varepsilon_j^{(l)}$ we have shown the following:

$$\begin{aligned}\varepsilon_j^{(l)} &= \sum_{i=1}^{m(l+1)} \varepsilon_i^{(l+1)} \cdot \frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} \\ &= \sum_{i=1}^{m(l+1)} \varepsilon_i^{(l+1)} \cdot w_{i,j}^{(l+1)} \cdot S'\big(z_j^{(l)}\big) \\ &= \sum_{i=1}^{m(l+1)} w_{i,j}^{(l+1)} \cdot \varepsilon_i^{(l+1)} \cdot S'\big(z_j^{(l)}\big)\end{aligned}$$

Therefore, we have now proven equation (BP2). We proceed to prove equation (BP4).

According to the chain rule we have

$$\frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial w_{j,k}^{(l)}} = \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial w_{j,k}^{(l)}}$$

Now by definition of $\varepsilon_j^{(l)}$, the first factor on the right hand side of this equation is equal to $\varepsilon_j^{(l)}$:

$$\varepsilon_j^{(l)} = \frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial z_j^{(l)}}.$$

In order to proceed, we need to evaluate the partial derivative $\dfrac{\partial z_j^{(L)}}{\partial w_{j,k}^{(l)}}$. The term $z_j^{(l)}$ has been defined as follows:

$$z_j^{(l)} = \left(\sum_{k=1}^{m(l)} w_{j,k}^{(l)} \cdot S\big(z_k^{(l-1)}\big)\right) + b_j^{(l)}.$$

Hence we have

$$\frac{\partial z_j^{(l)}}{\partial w_{j,k}^{(l)}} = S\big(z_k^{(l-1)}\big) = a_k^{(l-1)}.$$

Combining these equations we arrive at

$$\frac{\partial C_{\mathbf{x},\mathbf{y}}}{\partial w_{j,k}^{(l)}} = \varepsilon_j^{(l)} \cdot a_k^{(l-1)}.$$

Therefore, equation (BP4) has been verified.

Exercise 14: Prove equation (BP3). ◇

7.3 Stochastic Gradient Descent

The equations describing backpropagation describe the gradient of the cost function for a single training example $\langle \mathbf{x}, \mathbf{y} \rangle$. However, when we train a neural network, we need to take all training examples into account. If we have n training examples

$$\langle \mathbf{x}^{(1)}, \mathbf{y}^{(1)} \rangle\rangle, \langle \mathbf{x}^{(2)}, \mathbf{y}^{(2)} \rangle\rangle, \cdots, \langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\rangle,$$

then the quadratic error cost function has been previously defined as the sum

$$C\Big(W^{(2)}, \cdots, W^{(L)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(L)}; \mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \cdots, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}\Big) := \frac{1}{2 \cdot n} \cdot \sum_{i=1}^{n} \Big(\mathbf{o}\big(\mathbf{x}^{(i)}\big) - \mathbf{y}^{(i)}\Big)^2.$$

In practical applications of neural networks, the number of training examples is usually big. For example, when we later develop a neural network to classify handwritten digits, we will have $60,000$ training examples. More ambitious projects that use neural networks to classify objects in images use millions of training examples. When we compute the gradient of the quadratic error function with respect to a weight matrix $W^{(l)}$ or a bias

vector $b^{(l)}$ we have to compute the sums

$$\frac{1}{2\cdot n}\cdot\sum_{i=1}^{n}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial w_{j,k}^{(l)}} \quad \text{and} \quad \frac{1}{2\cdot n}\cdot\sum_{i=1}^{n}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial b_{j}^{(l)}}$$

over all training examples in order to perform a single step of gradient descent. If n is large, this is computationally costly. Note that these sums can be regarded as computing average values. In stochastic gradient descent, we approximate these sums by randomly choosing a small subset of the training examples. In order to formulate this approximation in a convenient notation, let us assume that instead of using all n training examples, we just use the first m training examples. Then we approximate the sums shown above as follows:

$$\frac{1}{2\cdot n}\cdot\sum_{i=1}^{n}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial w_{j,k}^{(l)}} \approx \frac{1}{2\cdot m}\cdot\sum_{i=1}^{m}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial w_{j,k}^{(l)}} \quad \text{and} \quad \frac{1}{2\cdot n}\cdot\sum_{i=1}^{n}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial b_{j}^{(l)}} \approx \frac{1}{2\cdot m}\cdot\sum_{i=1}^{m}\frac{\partial C_{\mathbf{x}^{(i)},\mathbf{y}^{(i)}}}{\partial b_{j}^{(l)}},$$

i.e. we approximate these sums by the average value of their first m training examples. Of course, in general we will not choose the first m training examples but rather we will choose m random training examples. The randomness of this choice is the reason this algorithm is called stochastic gradient descent. It turns out that if we take care that eventually all training examples are used during gradient descent, then the approximations given above can speed up the learning of neural networks substantially.

7.4 Implementation

Next, we will take a look at a neural network that is able to recognize handwritten digits. The Mnist database of handwritten digits contains 70 000 images of handwritten digits. These images have a size of 28×28 pixels. Figure 7.2 shows the first 10 images visualized.

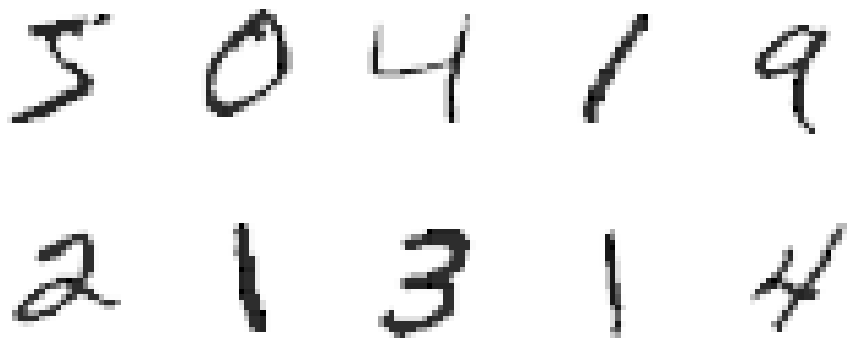

Figure 7.2: The first 10 images of the Mnist dataset visualized

We will use the first 60 000 of these images to train a neural network, while the remaining 10 000 images will be used to check the accuracy of the trained network. Since SetlX lacks the ability to read image files, the images have been converted into two large `.csv` files. You can download the 10 000 images from the test set at

https://github.com/karlstroetmann/Artificial-Intelligence/raw/master/SetlX/Neural-Networks/mnist_test.csv.gz

while the images from the training set are found at

https://github.com/karlstroetmann/Artificial-Intelligence/raw/master/SetlX/Neural-Networks/mnist_train.csv.gz

Note that these files are compressed. In order to train the neural network with these files, these files first have to be decompressed. Figure 7.3 on page 115 contains the code that is used to read these files.

The function that is called to load both the training data and the test data is the function `load_data`.

```
load_data := procedure(train_length, test_length) {
    test_data  := parse_csv("mnist_test.csv" , test_length);
    train_data := parse_csv("mnist_train.csv", train_length);
    train_data := [ [image, encode(label)] : [image, label] in train_data];
    return [train_data, test_data];
};
parse_csv := procedure(file, length) {
    Pairs := [0] * length;
    csv := readFile(file);
    for (i in [1..length]) {
        Line  := split(csv[i], ",");
        label := int(Line[1]);
        image := la_vector([double(x)/255: x in Line[2..]]);
        Pairs[i] := [image, label];
    }
    return Pairs;
};
encode := procedure(digit) {
    result          := la_vector([0] * 10);
    result[digit+1] := 1; // digit zero has index 1
    return result;
};
```

Figure 7.3: SetlX code to load the image files.

1. The function `load_data` receives two parameters.

   (a) `train_length` is the number of images used for training.

   (b) `test_length` is the number of images used for training.

   Usually, we will have `train_length` = 60 000 and `test_length` = 10 000. However, since training the neural network takes about 15 minutes to complete when all data are used, these parameters can be set to smaller values to speed up experiments.

   On completion, the function `load_data` will return a pair of the form

   `[train_data, test_data]`.

2. Both the training data and the test data are read via the function `parse_csv`. This function returns a list of pairs. The first component of these pairs is an image of a handwritten digit, while the second component is a number d from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. This number specifies the digit that is shown in the image.

3. For the training data, the number d specifying the digit needs to be encoded as a vector of dimension 10. The idea is that this vector contains nine occurrences of the digit `0` and one occurrence of the digit `1`. The position of the digit `1` is given as $d + 1$. For example, $d = 0$ is encoded as the vector

   $$\langle 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 \rangle^\top,$$

   while $d = 3$ is encoded as

   $$\langle 0, 0, 0, 1, 0, 0, 0, 0, 0, 0 \rangle^\top.$$

   The encoding of digits as vectors is done via the function `encode`.

The function `parse_csv` is responsible for reading a given `.csv` file.

1. This function is called with two parameters.

   (a) `file` is the name of the `.csv` file that is to be read.

   (b) `length` specifies the number of lines that should be read.

   The `.csv` files containing the images of handwritten digits contain one image per row. Every row contains 785 numbers.

   (a) The first number specifies the digit that is shown in the image.

   (b) The remaining $784 = 28 \times 28$ numbers specify the darkness of the pixels. These numbers are elements of the set $\{0, \cdots, 255\}$. The number 0 specifies a pixel that is white, while the number 255 specifies a pixel that is black. Numbers in between are interpreted as grey values.

   The function returns a list of pairs of the form

   ```
   [image, label]
   ```

   where `image` is a vector of dimension 784, while `label` is a digit from the set $\{0, \cdots, 9\}$.

2. The function `readFile` returns a list of strings. Each string in this list corresponds to one line in the `.csv` file and therefore it corresponds to one image and the corresponding label.

3. The function `split` splits the string at the occurrences of the commas and returns a list of strings of length 785.

4. The first entry of this list is the digit that is shown in the image. This entry is converted to an integer and is then stored in the variable `label`.

5. The remaining 784 entries are the pixels of the image encoded in this line. The grey values of these pixels are elements from the set $\{0, \cdots, 255\}$. In order to improve the convergence speed of stochastic gradient descent, these values are transformed into real values in the range $[0, 1]$.

6. The image and its label are combined into a pair and stored in the list `Pairs`. After `length` lines have been read, the list `Pairs` is returned.

The function `encode` is used to encode a digit into a vector. First, the function creates a vector `result` of dimension 10 that is initialized to all zeros. Then, it inserts a `1` at index `digit` + 1 and returns the resulting vector. This encode technique is called one-hot encoding. We use one hot encoding since the output layer of the neural network that we are going to construct contains 10 output nodes, one node to recognize each digit. The first output node will recognize the digit "0", the second output node will recognize the digit "1", and in general the i-th output node will recognize the i+1-th digit.

Figure 7.4 on page 117 show some auxiliary functions.

1. `rndVector` creates a vector of random numbers that are uniformly distributed in the interval $[-0.5, 0.5]$. The argument `length` specifies the dimension of this vector.

   This function is used later to initialize the biases of the neurons of the neural network.

2. `my_rnd` returns a random number that is uniformly distributed in the interval

   $$\left[-\frac{1}{2} \cdot \frac{1}{28}, \frac{1}{2} \cdot \frac{1}{28}\right].$$

   This function is used to initialize the weights of the neurons. The reason for the upper bound of $\frac{1}{28}$ is the fact that we have to avoid saturation of the neurons. A neuron is said to be saturated if the input argument a to the logistic function S of the neuron is either very low or very high. In that case, if a is small, $S(a)$ would be very small, while if a is big, $S(a)$ would be very close to 1. The reason is that

   $$\lim_{a \to -\infty} S(a) = 0 \quad \text{and} \quad \lim_{a \to \infty} S(a) = 1.$$

   In any case, the derivative of the logistic function, which has the form $S'(a) = S(a) \cdot \big(1 - S(a)\big)$ would

```
rndVector := length   |-> la_vector([random() - 0.5: i in [1..length]]);
my_rnd    := []       |-> (random() - 1/2) / 28; // 28 = sqrt(784)
rndMatrix := [rs, cs] |->
             la_matrix([[my_rnd(): c in [1..cs]]: r in [1..rs]]);
sigmoid   := [x] |-> la_vector([1 / (1 + exp(-a)) : a in x]);
sigmoid_prime := procedure(x) {
    s := sigmoid(x);
    return la_vector([a * (1 - a): a in s]);
};
hadamard := [x, y] |-> la_vector([x[i] * y[i]: i in [1 .. #x]]);
argmax := procedure(x) {
    [maxValue, maxIndex] := [x[1], 1];
    for (i in [2 .. #x] | x[i] > maxValue) {
        [maxValue, maxIndex] := [x[i], i];
    }
    return maxIndex;
};
```

Figure 7.4: Auxiliary functions.

be close to 0 because either $S(a)$ is small or $1 - S(a)$ is small. But then the gradient that is used in gradient descent would be very small also and hence gradient descent would require a very large number of iterations to converge. In order to prevent this from happening we need to ensure that the initial inputs of the logistic functions are are not to big. In the hidden layer, the weight matrices have a size of $28 \times 28 = 784$. If we add n random numbers that are uniformly distributed in the interval $[-1, 1]$, the expected value is 0 but the variance is 784 times the variance of a random variable that is $\text{Unif}(-1, 1)$. As the standard deviation is the square root of the variance and $28 = \sqrt{784}$, choosing the weights from the interval $\left[-\frac{1}{2} \cdot \frac{1}{28},\ \frac{1}{2} \cdot \frac{1}{28}\right]$ will ensure that the neurons are not saturated when initialized.

3. `rndMatrix(rs, cs)` creates a matrix of `rs` rows and `cs` columns of random numbers uniformly distributed in the interval $\left[-\frac{1}{2} \cdot \frac{1}{28},\ \frac{1}{2} \cdot \frac{1}{28}\right]$.

4. `sigmoid(x)` takes a vector `x` and computes the element-wise sigmoid function of this vector.

5. `sigmoid_prime(x)` takes a vector `x` and computes the element-wise derivative of the sigmoid function of this vector. Remember that the derivative $S'(t)$ of the sigmoid function $S(t)$ satisfies the equation

$$S'(t) = S(t) \cdot \big(1 - S(t)\big).$$

```
class network(inputSize, hiddenSize, outputSize) {
    mInputSize  := inputSize;   //        784
    mHiddenSize := hiddenSize;  //  30 .. 100
    mOutputSize := outputSize;  //         10
    mBiasesH  := rndVector(mHiddenSize);  // biases hidden layer
    mBiasesO  := rndVector(mOutputSize);  // biases output layer
    mWeightsH := rndMatrix(mHiddenSize, mInputSize);  // weights hidden layer
    mWeightsO := rndMatrix(mOutputSize, mHiddenSize); // weights output layer
    ...
}
```

Figure 7.5: The constructor of the class `network`.

Figure 7.7 on page 120 shows the outline of the class `network` and its constructor. This class represents a neural network with one hidden layer.

1. The argument `inputSize` specifies the number of input nodes. The neural network for the recognition of handwritten digits has 784 inputs. These inputs are the grey values of the 28×28 pixels that constitute the image of the handwritten digit.

2. The argument `hiddenSize` specifies the number of neurons in the hidden layer. We assume that there is only one hidden layer. I have experimented with 30 neurons, 60 neurons and a hundred neurons.

   (a) For 30 neurons, the computation took about 15 minutes and the trained neural network achieved an accuracy of 96.3%.

   (b) For 60 neurons, the computation took 25 minutes and the network achieved an accuracy of 97.3%.

   (c) If there are 100 neurons in the hidden layer, the computation took 41 minutes and achieved an accuracy of 97.9%.

   For 100 neurons, the number of weights in the hidden layer is $784 \cdot 100 = 78\,400$. Therefore, the number of weights is greater than the number of training examples. Hence, we should really use regularization in order to further increase the accuracy of the network.

3. The argument `mOutputSize` specifies the number of output neurons. For the neural network recognizing handwritten digits this number is 10 since there is an output neuron for every digit.

4. Besides storing the topology of the neural network, the class `network` stores the biases and weights of all the neurons. The biases and weights are initialized as random numbers. As explained when discussing the auxiliary functions in Figure 7.4 on page 117 we have to take care that these random numbers are not to big for otherwise the neurons would saturate and the learning would be very slow.

Figure 7.6 on page 119 shows the method `sgd` of the class `network`. This method implements stochastic gradient descent. It receives 5 arguments.

1. `training_data` is a list of pairs of the form $[\mathbf{x}_i, \mathbf{y}_i]$. Here, $\mathbf{x}_i$ is a vector of dimension 784. This vector contains the pixels of an image showing one of the handwritten digits from the training set. $\mathbf{y}_i$ is the one-hot encoding of the digit that is shown in the image $\mathbf{x}_i$.

2. `epochs` is the number of iterations of gradient descent. In order to train the neural network to recognize handwritten digits, we will use 30 iterations.

3. `mbs` is the size of the mini-batches that are used in stochastic gradient descent. I have achieved the fastest learning when I have used a mini-batch size of 10. Using a mini-batch size of 20 was sightly slower, but this parameter seems to be quite uncritical.

4. `eta` is the learning rate.

5. `test_data` is the list of test data. These data are only used to check the accuracy, they are not used to determine the weights or biases.

The implementation of stochastic gradient descent executes a `for`-loop that runs `epoch` number of times. At the beginning of each iteration, the training data are shuffled randomly. Next, the data is chopped up into chunks of size `mbs`. These chunks are called mini-batches. The inner `for`-loop iterates over all mini-batches and executes one step of gradient descent that only uses the data form the mini-batch. At the end of each iteration of the outer `for`-loop, the accuracy of the current version of the neural net is printed.

The method `update_mini_batch` performs one step of gradient descent for the data from one mini-batch. It receives two arguments.

1. `mini_batch` is the list of training data that constitute one mini-batch.

2. `eta` is the learning rate.

The implementation of `update_mini_batch` works as follows:

```
sgd := procedure(training_data, epochs, mbs, eta, test_data) {
    n_test := #test_data;
    n      := #training_data;
    for (j in [1 .. epochs]) {
        training_data := shuffle(training_data);
        mini_batches  := [training_data[k .. k+mbs-1]: k in [1, mbs..n]];
        for (mini_batch in mini_batches) {
            update_mini_batch(mini_batch, eta);
        }
        print("Epoch $j$: $evaluate(test_data)$ / $n_test$");
    }
};
update_mini_batch := procedure(mini_batch, eta) {
    nabla_BH := 0 * mBiasesH;  // gradient biases hidden layer
    nabla_BO := 0 * mBiasesO;  // gradient biases output layer
    nabla_WH := 0 * mWeightsH; // gradient weights hidden layer
    nabla_WO := 0 * mWeightsO; // gradient weights output layer
    for([x,y] in mini_batch) {
        [dltNbl_BH, dltNbl_BO, dltNbl_WH, dltNbl_WO] := backprop(x, y);
        nabla_BH += dltNbl_BH;
        nabla_BO += dltNbl_BO;
        nabla_WH += dltNbl_WH;
        nabla_WO += dltNbl_WO;
    }
    alpha := eta / #mini_batch;
    this.mBiasesH  -= alpha * nabla_BH;
    this.mBiasesO  -= alpha * nabla_BO;
    this.mWeightsH -= alpha * nabla_WH;
    this.mWeightsO -= alpha * nabla_WO;
};
```

Figure 7.6: Stochastic gradient descent.

1. First, we initialize the vectors `nabla_BH`, `nabla_BO` and the matrices `nabla_WH`, `nabla_WO` to contain only zeros.
   (a) `nabla_BH` will store the gradient of the bias vector of the hidden layer.
   (b) `nabla_BO` will store the gradient of the bias vector of the output layer.
   (c) `nabla_WH` will store the gradient of the weight matrix of the hidden layer.
   (d) `nabla_WO` will store the gradient of the weight matrix of the output layer.
2. Next, we iterate of all training examples in the mini-batch and for every training example `[x,y]` we compute the contribution of this training example to the gradients of the cost function C, i.e. we compute
   $$\nabla_{\mathbf{b}^{(l)}} C_{\mathbf{x},\mathbf{y}} \quad \text{and} \quad \nabla_{W^{(l)}} C_{\mathbf{x},\mathbf{y}}$$
   for the hidden layer and the output layer. These gradients are computed by the function `backprop`.
3. Finally, the bias vectors and the weight matrices are updated according to the learning rate and the computed gradients.

The method `backprop` computes the gradients of the bias vectors and the weight matrices with respect to a single training example $\langle \mathtt{x}, \mathtt{y} \rangle$. The implementation of `backprop` proceeds as follows:

```
backprop := procedure(x, y) {
    ZH := mWeightsH * x + mBiasesH;
    AH := sigmoid(ZH);
    ZO := mWeightsO * AH + mBiasesO;
    AO := sigmoid(ZO);
    epsilonO := hadamard(AO - y, sigmoid_prime(ZO));
    nabla_BO := epsilonO;
    nabla_WO := la_matrix(epsilonO) * la_matrix(AH)!;
    epsilonH := hadamard(mWeightsO! * epsilonO, sigmoid_prime(ZH));
    nabla_BH := epsilonH;
    nabla_WH := la_matrix(epsilonH) * la_matrix(x)!;
    return [nabla_BH, nabla_BO, nabla_WH, nabla_WO];
};
```

Figure 7.7: Implementation of backpropagation.

1. First, the vector `ZH` is computed according to the formula
$$\mathbf{z}^{(2)} = W^{(2)} \cdot \mathbf{x} + \mathbf{b}^{(2)}.$$
Here, $W^{(2)}$ is the weight matrix of the hidden layer that is stored in `mWeightsH`, while $\mathbf{b}^{(2)}$ is the bias vector of the hidden layer. This vector is stored in `mBiasesH`.

2. The activation of the neurons in the hidden layer `AH` is computed by applying the sigmoid function to the vector $\mathbf{z}^{(2)}$.

3. Next, the vector `ZH` is computed according to the formula
$$\mathbf{z}^{(3)} = W^{(3)} \cdot \mathbf{x} + \mathbf{b}^{(3)}.$$
Here, $W^{(3)}$ is the weight matrix of the output layer that is stored in `mWeightsO`, while $\mathbf{b}^{(3)}$ is the bias vector of the output layer. This vector is stored in `mBiasesO`.

4. The activation of the neurons in the output layer `AO` is computed by applying the sigmoid function to the vector $\mathbf{z}^{(3)}$.

   These four step constitute the forward pass of backpropagation.

5. Next, the error in the output layer `epsilonO` is computed using the backpropagation equation (BP1v)
$$\boldsymbol{\varepsilon}^{(3)} = (\mathbf{a}^{(3)} - \mathbf{y}) \odot S'(\mathbf{z}^{(3)}).$$

6. According to equation (BP3v), the gradient of the cost function with respect to the bias vector of the output layer is given as
$$\nabla_{\mathbf{b}^{(3)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(3)}.$$
This gradient is stored in the variable `nabla_BO`.

7. According to equation (BP4v), the gradient of the cost function with respect to the weight matrix of the output layer is given as
$$\nabla_{W^{(3)}} C_{\mathbf{x},\mathbf{y}} = \boldsymbol{\varepsilon}^{(3)} \cdot \left(\mathbf{a}^{(2)}\right)^{\top}.$$
This gradient is stored in the variable `nabla_WO`.

8. Next, the error in the hidden layer `epsilonH` is computed using the backpropagation equation (BP2v)
$$\boldsymbol{\varepsilon}^{(2)} = \left(\left(W^{(3)}\right)^{\top} \cdot \boldsymbol{\varepsilon}^{(3)}\right) \odot S'(\mathbf{z}^{(2)}).$$

9. Finally, the gradients of the cost function with respect to the bias vector and the weight matrix of the hidden layer are computed. This is completely to the computation of the corresponding gradients of the output layer.

```
1  feedforward := procedure(x) {
2      AH := sigmoid(mWeightsH * x  + mBiasesH);
3      AO := sigmoid(mWeightsO * AH + mBiasesO);
4      return AO;
5  };
6  evaluate := procedure(test_data) {
7      test_results := [[argmax(feedforward(x)) - 1, y]: [x, y] in test_data];
8      return #[1 : [a, b] in test_results | a == b];
9  };
```

Figure 7.8: Evaluation functions.

Figure 7.8 shows the implementation of the function `feedforward`. This function receives an image **x** stored a s a vector of dimension 784 and computes the output of the neural network for this image. The code is a straightforward implementation of the equations (FF1) and (FF2). The function `evaluate` is used to evaluate the accuracy of the neural network on the test data. Finally, Figure 7.9 shows how the computation can be started.

```
1  load("nn-loader.stlx");
2  load("nn.stlx");
3
4  main := procedure() {
5      resetRandom();
6      train_length := 60000;
7      test_length  := 10000;
8      inputSize    :=   784;
9      hiddenSize   :=    30;
10     outputSize   :=    10;
11     epochs       :=    30;
12     mbs          :=    10;
13     eta          :=   1.0;
14     [training_data, test_data] := load_data(train_length, test_length);
15     print("Create Network");
16     net := network(inputSize, hiddenSize, outputSize);
17     print("Start SGD");
18     start := now();
19     net.sgd(training_data, epochs, mbs, eta, test_data);
20     stop := now();
21     print("Time needed:\t $stop - start$");
22 };
23 main();
```

Figure 7.9: How to start the training.

Bibliography

[BH69] Arthur Earl Bryson, Jr. and Yu-Chi Ho. *Applied Optimal Control: Optimization, Estimation, and Control*. Blaisdell Pub., Waltham, Massachusetts, 1969.

[HNR68] Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics SSC4*, 4(2):100–107, 1968.

[HNR72] Peter Hart, Nils Nilsson, and Bertram Raphael. Correction to "A formal basis for the heuristic determination of minimum cost paths". *SIGART Newsletter*, 37:28–29, 1972.

[JWHT14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2014.

[KM75] Donald E. Knuth and Ronald W. Moore. An analysis of alpha-beta pruning. *Artificial Intelligence*, 6(4):293–326, 1975.

[Kor85] Richard Korf. Depth-first iterative-deepening: An optimal admissible tree search. *Artificial Intelligence*, 27:97–109, 1985.

[Nie15] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, 1986.

[RN09] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 3rd edition, 2009.

[SHM+16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484, 2016.

[Sny05] Jan A. Snyman. *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*. Springer Publishing, 2005.