Tobias Nipkow, Jasmin Blanchette, Manuel Eberl, Alejandro Gómez-Londoño, Peter Lammich, Christian Sternagel, Simon Wimmer, Bohua Zhan

Functional Algorithms, Verified!

\[
\text{datatype} \ 'a\ \text{tree} = \text{Leaf} \mid \text{Node} \ ('a \ \text{tree}) \ 'a \ ('a \ \text{tree})
\]

\[
T_{\text{insert}} \ a\ t + \Phi(\text{insert} \ a\ t) - \Phi(t) \leq a \cdot \lg(|t| + 3) + 2
\]
August 7, 2022
This book is an introduction to data structures and algorithms for functional languages, with a focus on proofs. It covers both functional correctness and running time analysis. It does so in a unified manner with inductive proofs about functional programs and their running time functions.

The unique feature of this book is that all the proofs have been machine-checked by the proof assistant Isabelle. That is, in addition to the text in the book, which requires no knowledge of proof assistants, there are the Isabelle definitions and proofs that can be accessed by following (in the PDF file) the links attached to section headings with a \symbol{c} symbol. The structured nature of Isabelle proofs permits even novices to browse them and follow the high-level arguments.

This book has been classroom-tested for a number of years in a course for graduate and advanced undergraduate students. At the same time it is a reference for programmers and researchers who are interested in the details of some algorithm or proof.

Isabelle \symbol{c}

Isabelle [63, 76, 57] is a proof assistant for the logic HOL (= Higher-Order Logic), which is why the system is often called Isabelle/HOL. HOL is a generalization of first-order logic: functions can be passed as parameters and returned as results, just as in functional programming, and they can be quantified over. Isabelle's version of HOL also supports a simple version of Haskell's type classes.

The main strength of Isabelle and other proof assistants is their trustworthiness: all definitions, lemma statements, and inferences are checked. Beyond trustworthiness, formal proofs can also clarify arguments, by exposing and
explaining difficult steps. Most Isabelle users will confirm that their pen-and-paper proofs have become more lucid, and more correct, after they subjected themselves to the discipline of formal proof.

As emphasized above, the reader need not be familiar with Isabelle or HOL in order to read this book. However, to take full advantage of our proof assistant approach, readers are encouraged to learn how to write Isabelle definitions and proofs themselves — and to solve some of the exercises in this book. To this end we recommend the tutorial *Programming and Proving in Isabelle/HOL* [52], which is also Part I of the book *Concrete Semantics* [56].

**Prerequisites**

We expect the reader to be familiar with

- the basics of discrete mathematics: propositional and first-order logic, sets and relations, proof principles including induction;
- a typed functional programming language like Haskell [27], OCaml [60] or Standard ML [64];
- simple inductive proofs about functional programs.

**Under Development**

This book is meant to grow. New chapters are meant to be added over time. The list of authors is meant to grow — you could become one of them!

**Colour**

For the quick orientation of the reader, definitions are displayed in coloured boxes:

These boxes display functional programs.

These boxes display auxiliary definitions.

From a logical point of view there is no difference between the two kinds of definitions except that auxiliary definitions need not be executable.

**Acknowledgements** We are obviously indebted to the books by Cormen *et al.* [13] and Okasaki [62]. Mohammad Abdulaziz, Magnus Myreen and Larry Paulson commented on parts of the book. We are very grateful to all of them.
# Contents

1 Basics ................................................................. 7

## Part I Sorting and Selection

2 Sorting ............................................................... 19
3 Selection ............................................................ 35

## Part II Search Trees

4 Binary Trees.......................................................... 49
5 Binary Search Trees .................................................. 61
6 Abstract Data Types ................................................ 75
7 2-3 Trees .............................................................. 81
8 Red-Black Trees ..................................................... 91
9 AVL Trees ............................................................. 101
10 Beyond Insert and Delete: $\cup$, $\cap$ and $-$ ................. 113
11 Arrays via Braun Trees ............................................ 121
12 Tries ................................................................. 143
13 Huffman’s Algorithm ............................................. 151
## Part III Priority Queues

14 Priority Queues .................................................. 167  
15 Leftist Heaps .................................................... 171  
16 Priority Queues via Braun Trees.............................. 175  
17 Binomial Heaps.................................................... 179  

## Part IV Advanced Design and Analysis Techniques

18 Dynamic Programming ......................................... 191  
19 Amortized Analysis .............................................. 213  
20 Queues ............................................................. 219  
21 Splay Trees ......................................................... 231  
22 Skew Heaps ......................................................... 239  
23 Pairing Heaps ....................................................... 243  

## Part V Appendix

A List Library ......................................................... 253  
B Time Functions ................................................... 257  
C Notation ............................................................. 265  
References ............................................................ 271  
Index ................................................................. 277
In this chapter we describe the basic building blocks the book rests on.

**Programs:** The functional programming language we use is merely sketched because of its similarity with other well known functional languages.

**Predefined types and notation:** We introduce the basic predefined types and notations used in the book.

**Inductive proofs:** Although we do not explain proofs in general, we make an exception for certain inductive proofs.

**Running time:** We explain how we model running time by step counting functions.

### 1.1 Programs

The programs in this book are written in Isabelle’s functional programming language which provides recursive algebraic data types (keyword: `datatype`), recursive function definitions and `let`, `if` and `case` expressions. The language is sufficiently close to a number of similar typed functional languages (SML [64], OCaml [60], Haskell [27]) to obviate the need for a detailed explanation. Moreover, Isabelle can generate SML, OCaml, Haskell and Scala code [24]. What distinguishes Isabelle’s functional language from ordinary programming languages is that all functions in Isabelle must terminate. Termination must be proved. For all the functions in this book, termination is not difficult to see and Isabelle can prove it automatically. (If you want to go beyond, consult the function definition tutorial [44].)

Isabelle’s functional language is pure logic. All language elements have precise definitions. However, this book is about algorithms, not programming language semantics. A functional programmer’s intuition suffices for reading it. (If you want to know more about the logical basis see [7, 45, 26].)
A useful bit of notation: any infix operator can be turned into a function by enclosing it in parentheses, e.g. \((+)\).

### 1.2 Types

**Type variables** are denoted by \(a\), \(b\), etc. The function type arrow is \(\Rightarrow\). Type constructor names follow their argument types, e.g. \(\text{a list}\). The notation \(t :\!\!\!\tau\) means that term \(t\) has type \(\tau\). The following types are predefined.

#### Booleans

Type \(\text{bool}\) comes with the constants \(\text{True}\) and \(\text{False}\) and the usual operations.

#### Numbers

There are three numeric types: the natural numbers \(\text{nat}\) (0, 1, …), the integers \(\text{int}\) and the real numbers \(\text{real}\). They correspond to the mathematical sets \(\mathbb{N}\), \(\mathbb{Z}\) and \(\mathbb{R}\) and not to any machine arithmetic. All three types come with the usual (overloaded) operations.

#### Sets

The type \(\text{a set}\) of sets (finite and infinite) over type \(\text{a}\) comes with the standard mathematical operations. The minus sign “\(-\)”, unary or binary, can denote set complement or difference.

#### Lists

The type \(\text{a list}\) of lists whose elements are of type \(\text{a}\) is a recursive data type:

\[
\text{datatype } \text{a list} = \text{Nil} | \text{Cons } \text{a} (\text{a list})
\]

Constant \(\text{Nil}\) represents the empty list and \(\text{Cons } x \ xs\) represents the list with first element \(x\), the **head**, and rest list \(xs\), the **tail**. The following syntactic sugar is sprinkled on top:

\[
\begin{align*}
\text{[]} & \equiv \text{Nil} \\
\text{x \# xs} & \equiv \text{Cons } x \ xs \\
[\text{x}_1, \ldots, \text{x}_n] & \equiv \text{x}_1 \# \ldots \# \text{x}_n \# \text{[]} \\
\end{align*}
\]

The \(\equiv\) symbol means that the left-hand side is merely an abbreviation of the right-hand side.

A library of predefined functions on lists is shown in Appendix A. The length of a list \(xs\) is denoted by \(|xs|\).
1.2 Types

Type `a option

The data type `a option is defined as follows:

```
datatype 'a option = None | Some 'a
```

Pairs and Tuples

Pairs are written `(a, b)`. Functions `fst` and `snd` select the first and second component of a pair: `fst (a, b) = a` and `snd (a, b) = b`. The type `unit` contains only a single element `()`, the empty tuple.

1.2.1 Pattern Matching

Functions are defined by equations and pattern matching, for example over lists. Natural numbers may also be used in pattern-matching definitions:

```
fib (n + 2) = fib (n + 1) + fib n
```

Occasionally we use an extension of pattern matching where patterns can be named. For example, the defining equation

```
f (x # (y # zs := ys)) = ys @ zs
```

introduces a variables `ys` on the left that stands for `y # zs` and can be referred to on the right. Logically it is just an abbreviation of

```
f (x # y # zs) = (let ys = y # zs in ys @ zs)
```

although it is suggestive of a more efficient interpretation. The general format is `pattern ::= variable`.

1.2.2 Numeric Types and Coercions

The numeric types `nat`, `int` and `real` are all distinct. Converting between them requires explicit coercion functions, in particular the inclusion functions `int :: nat ⇒ int` and `real :: nat ⇒ real` that do not lose any information (in contrast to coercions in the other direction). We do not show inclusions unless they make a difference. For example, `(m + n) :: real`, where `m, n :: nat`, is mathematically unambiguous because `real (m + n) = real m + real n`. On the other hand, `(m - n) :: real` is ambiguous because `real (m - n) ≠ real m - real n` because `(0 :: nat) - n = 0`. In some cases we can also drop coercions that are not inclusions, e.g. `nat :: int ⇒ nat`, which coerces negative integers to 0: if we know that `i ≥ 0` then we can drop the `nat` in `nat i`.

We prefer type `nat` over type `real` for ease of (Isabelle) proof. For example, for `m, n :: nat` we prefer `m ≤ 2^n` over `lg m ≤ n`. 
1.2.3 Multisets

Informally, a multiset is a set where elements can occur multiple times. Multisets come with the following operations:

\[
\begin{align*}
\{\} &:: 'a multiset
(c_a) &:: 'a \Rightarrow 'a multiset \Rightarrow bool
add_mset &:: 'a \Rightarrow 'a multiset \Rightarrow 'a multiset
(+1) &:: 'a multiset \Rightarrow 'a multiset \Rightarrow 'a multiset
size &:: 'a multiset \Rightarrow nat
mset &:: 'a list \Rightarrow 'a multiset
set_mset &:: 'a multiset \Rightarrow 'a set
image_mset &:: ('a \Rightarrow 'b) \Rightarrow 'a multiset \Rightarrow 'b multiset
filter_mset &:: ('a \Rightarrow bool) \Rightarrow 'a multiset \Rightarrow 'a multiset
sum_mset &:: 'a multiset \Rightarrow 'a
\end{align*}
\]

Their meaning: \{\} is the empty multiset; \((c_a)\) is the element test; \(add_mset\) adds an element to a multiset; \((+)\) is the sum of two multisets, where multiplicities of elements are added; \(size\) \(M\), written \(\mid M\), is the number of elements in \(M\), taking multiplicities into account; \(mset\) converts a list into a multiset by forgetting about the order of elements; \(set_mset\) converts a multiset into a set; \(image_mset\) applies a function to all elements of a multiset; \(filter_mset\) removes all elements from a multiset that do not satisfy the given predicate; \(sum_mset\) is the sum of the values of a multiset, the iteration of \((+)\) (taking multiplicity into account).

We use some additional suggestive syntax for some of these operations:

\[
\begin{align*}
\{x \in_a M \mid P x\} &\equiv filter_mset P M
\{f x \mid x \in_a M\} &\equiv image_mset f M
\sum_{x \in a} M &\equiv sum_mset M
\sum_{x \in_a M} f x &\equiv sum_mset (image_mset f M)
\end{align*}
\]

See Section C.3 in the appendix for an overview of such syntax.

1.3 Notation

We adopt the following notation:

- The type of functions \('a \Rightarrow 'b\) comes with a predefined pointwise update operation with its own notation:
  \(f(a := b) = (\lambda x. if x = a then b else f x)\)
- Function \(lg\) is the binary logarithm.
We deviate from Isabelle's notation in favour of standard mathematics in a number of points:

- There is only one implication: $\implies$ is printed as $\rightarrow$ and $P \implies Q \implies R$ is printed as $P \land Q \rightarrow R$.
- Multiplication is printed as $x \cdot y$.
- Exponentation is uniformly printed as $x^y$.
- We sweep under the carpet that type $nat$ is defined as a recursive data type: `datatype nat = 0 | Suc nat`. In particular, constructor $Suc$ is hidden: $Suc^k 0$ is printed as $k$ and $Suc^k n$ (where $n$ is not 0) is printed as $n + k$.
- Set comprehension syntax is the canonical $\{x \mid P\}$.

The reader who consults the Isabelle theories referred to in this book should be aware of these discrepancies.

### 1.4 Proofs

Proofs are the raison d'etre of this book. Thus we present them in more detail than is customary in a book on algorithms. However, not all proofs:

- We omit proofs of simple properties of numbers, lists, sets and multisets, our predefined types. Obvious properties (e.g. $|xs \odot ys| = |xs| + |ys|$ or commutativity of $\cup$) are used implicitly without proof.
- With some exceptions, we only state properties if their proofs require induction, in which case we will say so, and we will always indicate which supporting properties were used.
- If a property is simply described as “inductive” or its proof is described by a phrase like “by an easy/automatic induction” it means that in the Isabelle proofs all cases of the induction were automatic, typically by simplification.

As a simple example of an easy induction consider the append function

```plaintext
(\@) :: 'a list \Rightarrow 'a list \Rightarrow 'a list
[] @ ys = ys
(x # xs) @ ys = x # xs @ ys
```

and the proof of $(xs \odot ys) \odot zs = xs \odot ys \odot zs$ by structural induction on $zs$. (Note that $(\@)$ associates to the right.) The base case is trivial by definition: $([], y) \odot zs = [] \odot ys \odot zs$. The induction step is easy:

\[
\begin{align*}
(x \# (xs \odot ys)) \odot zs & = x \# (xs \odot ys) \odot zs \\
& = x \# xs \odot ys \odot zs & \text{by definition of $(\@)$} \\
& \text{by IH}
\end{align*}
\]
Note that IH stands for Induction Hypothesis, in this case \((xs @ ys) @ zs = xs @ ys @ zs\).

### 1.4.1 Computation Induction

Because most of our proofs are about recursive functions, most of them are by induction, and we say so explicitly. If we do not state explicitly what form the induction takes, it is by an obvious structural induction. The alternative and more general induction schema is computation induction where the induction follows the terminating computation, but from the bottom up. For example, the terminating recursive definition for \(gcd :: \text{nat} \Rightarrow \text{nat} \Rightarrow \text{nat}\)

\[
gcd m n = (\text{if } n = 0 \text{ then } m \text{ else } gcd n (m \text{ mod } n))
\]
gives rise to the following induction schema:

If \((n \neq 0 \quad \Rightarrow \quad P n (m \text{ mod } n)) \quad \Rightarrow \quad P m n\) (for all \(m\) and \(n\)), then \(P m n\) (for all \(m\) and \(n\)).

In general, let \(f :: \tau \Rightarrow \tau'\) be a terminating function of, for simplicity, one argument. Proving \(P(x :: \tau)\) by induction on the computation of \(f\) means proving

\[
P r_1 \land \ldots \land P r_n \quad \Rightarrow \quad P e
\]
for every defining equation

\[
f e = \ldots f r_1 \ldots f r_n \ldots
\]
where \(f r_1, \ldots, f r_n\) are all the recursive calls. For simplicity we have ignored the \texttt{if} and \texttt{case} contexts that a recursive call \(f r_i\) occurs in and that should be preconditions of the assumption \(P r_i\) as in the \(gcd\) example. If the defining equations for \(f\) overlap, the above proof obligations are stronger than necessary.

### 1.5 Running Time

Our approach to reasoning about the running time of a function \(f\) is very simple: we explicitly define a function \(T_f\) such that \(T_f x\) models the time the computation of \(f x\) takes. Unless stated otherwise, \(T_f\) counts the number of all function calls in the computation of \(f\). It is not intended that \(T_f\) yields the exact running time but only that the running time of \(f\) is in \(O(T_f)\).

Given a function \(f :: \tau_1 \Rightarrow \ldots \Rightarrow \tau_n \Rightarrow \tau\) we define a (running) time function \(T_f :: \tau_1 \Rightarrow \ldots \Rightarrow \tau_n \Rightarrow \text{nat}\) by translating every defining equation for \(f\) into a defining equation for \(T_f\). The translation is defined by two functions: \(E\) translates defining equations for \(f\) to defining equations for \(T_f\) and \(\mathcal{T}\)
translates expressions that compute some value to expressions that computes the number of function calls. The unusual notation \( E \) and \( T \) emphasizes that they are not functions in the logic.

\[
\begin{align*}
E[f \, p_1 \ldots \, p_n = e] &= (T_f \, p_1 \ldots \, p_n = T[e] + 1) \\
T[g \, e_1 \ldots \, e_k] &= T[e_1] + \ldots + T[e_k] + T_g \, e_1 \ldots \, e_k
\end{align*}
\] (1.1)

This is the general idea. It requires some remarks and clarifications:

- This definition of \( T_f \) is an abstraction of a call-by-value semantics. Thus it is also correct for lazy evaluation but may be a very loose upper bound.
- Definition (1.1) is incomplete: if \( g \) is a variable or constructor function (e.g. \( \texttt{Nil} \) or \( \texttt{Cons} \)), then there is no defining equation and thus no \( T_g \). We simply define \( T_g \) if \( g \) is a variable, constructor function or predefined function on \( \texttt{bool} \) or numbers. These choices are discussed below.
- Conditionals and case expressions are treated specially:

\[
\begin{align*}
T[\text{if } b \text{ then } e_1 \text{ else } e_2] &= T[b] + (\text{if } b \text{ then } T[e_1] \text{ else } T[e_2]) \\
T[\text{case } e \text{ of } p_1 \Rightarrow e_1 \mid \ldots \mid p_k \Rightarrow e_k] &= T[e] + (\text{case } e \text{ of } p_1 \Rightarrow T[e_1] \mid \ldots \mid p_k \Rightarrow T[e_k])
\end{align*}
\]

- let expressions are also special:

\[
T[\text{let } x = e_1 \text{ in } e_2] = T[e_1] + (\lambda x. T[e_2]) \, e_1
\]

Note that \( T[e_2] \) must be evaluated completely before \( (\lambda x. T[e_2]) \, e_1 \) is \( \beta \)-contracted to ensure that \( T[x] = 1 \) (see above) kicks in instead of generating multiple copies of \( T[e_1] \).
- For simplicity we restrict ourselves to a first-order language above. Nevertheless we use a few basic higher-order functions like \texttt{map} in the book. Their running time functions are defined in Appendix B.1.

As an example consider the append function \( @ \) defined above. The defining equations for \( T_{\text{append}} : \ 'a \text{ list} \Rightarrow 'a \text{ list} \Rightarrow \text{nat} \) are easily derived. The first equation translates like this:

\[
\begin{align*}
E[] @ ys &= ys \\
&= (T_{\text{append}} [] \, ys = T[ys] + 1) \\
&= (T_{\text{append}} [] \, ys = 2)
\end{align*}
\]

The right-hand side of the second equation translates like this:
Thus the two defining equations for $T_{\text{append}}$ are

\[
T_{\text{append}} [] \ ys = 2 \\
T_{\text{append}} (x # xs) \ ys = T_{\text{append}} xs ys + 5
\]

The constants 2 and 5 in the above definition are somewhat arbitrary because accessing a variable and calling (#) take different amounts of time, even if we assume they are both constant-time operations. Because we are only interested in $O(T_f)$ it is justified to reduce all additive constants to 1. This is what we will call the canonical running time function. Thus the canonical $T_{\text{append}}$ is this:

\[
T_{\text{append}} [] \ ys = 1 \\
T_{\text{append}} (x # xs) \ ys = T_{\text{append}} xs ys + 1
\]

Unless noted otherwise we always work with canonical running time functions. This may also involve looking inside if/case when reducing additive constants, e.g. reducing (if \(b\) then 2 + \(e\) else 3) + 2 to if \(b\) then 1 + \(e\) else 1.

Occasionally we will apply one more simplification step: when defining some $T_f$, you may drop additive constants when they are added to some term $T_g \ldots$ where $T_g$ has already been defined (i.e. \(g \neq f\) and \(f\) and \(g\) are not mutually recursive). This is because by construction all $T_g$ return positive results and adding a constant does not change $O(T_f)$.

In the main body of the book we initially show the definition of each $T_f$. Once the principles above have been exemplified sufficiently, the time functions are relegated to Appendix B.

### 1.5.1 Example: List Reversal

This section exemplifies not just the definition of time functions but also their analysis. The standard list reversal function $\text{rev}$ is defined in Appendix A. This is the corresponding canonical time function:

\[
T_{\text{rev}} :: 'a \text{ list} \Rightarrow \text{nat} \\
T_{\text{rev}} [] = 1 \\
T_{\text{rev}} (x # xs) = T_{\text{rev}} xs + T_{\text{append}} (\text{rev} \ xs) [x] + 1
\]
The canonical $T_{\text{append}}$ was shown before.

A simple induction shows $T_{\text{append}} \; xs \; ys = |zs| + 1$. The precise formula for $T_{\text{rev}}$ is less immediately obvious (exercise!) but an upper bound is easy to guess and verify by induction:

$$T_{\text{rev}} \; xs \leq (|xs| + 1)^2$$

We will frequently prove upper bounds only.

Of course one can also reverse a list in linear time:

```haskell
itrev :: 'a list ⇒ 'a list ⇒ 'a list
itrev [] ys = ys
itrev (x ≠ xs) ys = itrev xs (x ≠ ys)
```

$$T_{\text{itrev}} :: \text{'a list} ⇒ \text{'a list} ⇒ \text{nat}$$

$$T_{\text{itrev}} [] _- = 1$$

$$T_{\text{itrev}} (x ≠ xs) ys = T_{\text{itrev}} xs (x ≠ ys) + 1$$

Function $\text{itrev}$ has linear running time: $T_{\text{itrev}} \; xs \; ys = |xs| + 1$. A simple induction yields $\text{itrev} \; xs \; ys = \text{rev} \; xs @ ys$. Thus $\text{itrev}$ implements $\text{rev}$:

$$\text{rev} \; xs = \text{itrev} \; xs \; []$$

### 1.5.2 Discussion

Analysing the running time of any program, let alone a functional one, is tricky unless you have two formally specified entities: a machine model that includes a notion of running time and a translation from programs to machines. For imperative programs the standard model is the Random Access Machine (RAM) where each instruction takes one time unit. For functional programs a standard measure is the number of function calls. We adopt this measure and equate it with time although a more honest name would be computation steps.

A full proof that the execution time of our functional programs is in $O(T_f)$ on some actual soft- and hardware is a major undertaking: one would need to formalize the full stack of compiler, runtime system and hardware. We do not offer such a proof. Thus our formalization of “time” should be seen as conditional: given a stack that satisfies our basic assumptions in the definition of $\mathcal{E}$ and $\mathcal{T}$, our analyses are correct for that stack. We argue that these assumptions are not unreasonable (on a RAM) provided we accept that both the address space and numbers have a fixed size and cannot grow arbitrarily.
Of course this means that actual program execution may abort if the resources are exhausted.

Our basic assumption is that function calls take constant time. This is reasonable because we just need to allocate (and later deallocate) a new stack frame of constant size because all parameters are references or numbers and thus of fixed size. We also assumed that variable access takes constant time. This is a standard RAM assumption. Assuming that constructor functions take constant time is reasonable because the memory manager could simply employ a single reference to the first free memory cell and increment that with each constructor function call. How to account for garbage collection is less clear. In the worst case we have to assume that garbage collection is switched off, which simply exhausts memory more quickly. Finally we assume that operations on bool and numbers take constant time. The former is obvious, the latter follows from our assumption that we have fixed-size numbers.

In the end, we are less interested in a specific model of time and more in the principle that time (and other resources) can be analyzed just as formally as functional correctness once the ground rules (e.g. $T$) have been established.

### 1.5.3 Asymptotic Notation

The above approach to running time analysis is nicely concrete and avoids the more sophisticated machinery of asymptotic notation, $O(.)$ and friends. Thus we have intentionally lowered the entry barrier to the book for readers who want to follow the Isabelle formalization: we require no familiarity with Isabelle’s real analysis library and in particular with the existing formalization of and automation for asymptotic notation [17]. Of course this comes at a price: one has to come up with and reason about somewhat arbitrary constants in the analysis of individual functions. Moreover we rarely appeal to the “master theorem” (although Eberl [17] provides a generalized version) but prove solutions to recurrence equations correct by induction. Again, this is merely to reduce the required mathematical basis and to show that it can be done. In informal explanations, typically when considering inessential variations, we do use standard mathematical notation and write, for example, $O(n \lg n)$.
Part I

Sorting and Selection
In this chapter we define and verify the following sorting functions: insertion sort, quicksort, and one top-down and two bottom-up merge sorts. We also analyze their running times, except for quicksort.

Sorting involves an ordering. In this book we assume such an ordering to be provided by comparison operators $\leq$ and $<$ defined on the underlying type.

Sortedness of lists is defined as follows:

\[
\text{sorted} :: ('a::linorder) list \Rightarrow bool
\]
\[
\text{sorted} \ [] = \ True
\]
\[
\text{sorted} \ (x \#\ ys) = ((\forall\ y \in\ \text{set}\ ys.\ x \leq\ y) \land\ \text{sorted}\ ys)
\]

That is, every element is $\leq$ than all elements to the right of it: the list is sorted in increasing order. But what does $'a::\text{linorder}$ mean?

The type variable $'a$ is annotated with $\text{linorder}$, which means that sorted is only applicable if a binary predicate $(\leq) :: 'a \Rightarrow 'a \Rightarrow bool$ is defined and $(\leq)$ is a linear order, i.e. the following properties are satisfied:

- **reflexivity**: $x \leq x$
- **transitivity**: $x \leq y \land y \leq z \Rightarrow x \leq z$
- **antisymmetry**: $a \leq b \land b \leq a \Rightarrow a = b$
- **linearity/totality**: $x \leq y \lor y \leq x$

Moreover, the binary predicate $(<)$ must satisfy

\[x < y \iff x \leq y \land x \neq y.\]

Note that $\text{linorder}$ is a specific predefined example of a type class [25]. We will not explain type classes any further because we do not require the general concept. In fact, we will mostly not even show the $\text{linorder}$ annotation.
in types: you can assume that if you see \( \leq \) or \( < \) on a generic type \('a\) in this book, \('a\) is implicitly annotated with \texttt{linorder}. Note further that \((\leq)\) on the numeric types \texttt{nat}, \texttt{int} and \texttt{real} is a linear order.

### 2.1 Specification of Sorting Functions

A sorting function \(\text{sort} :: \ 'a \ \text{list} \Rightarrow \ 'a \ \text{list} \) (where, as usual, \('a::\text{linorder}\)) must obviously satisfy the following property:

\[
\text{sorted} (\text{sort} \ \text{xs})
\]

However, this is not enough — otherwise, \(\text{wrong\_sort} \ \text{xs} = []\) would be a correct sorting function. The set of elements in the output must be the same as in the input, and each element must occur the same number of times. This is most readily captured with the notion of a multiset (see Section 1.2.3). The second property that a sorting function \(\text{sort}\) must satisfy is

\[
\text{mset} (\text{sort} \ \text{xs}) = \text{mset} \ \text{xs}
\]

where function \(\text{mset}\) converts a list into its corresponding multiset.

### 2.2 Insertion Sort

Insertion sort is well-known for its intellectual simplicity and computational inefficiency. Its simplicity makes it an ideal starting point for this book. Below, it is implemented by the function \(\text{insort}\) with the help of the auxiliary function \(\text{insort1}\) that inserts a single element into an already sorted list.

\[
\begin{align*}
\text{insort1} :: \ 'a \Rightarrow \ 'a \ \text{list} \Rightarrow \ 'a \ \text{list} \\
\text{insort1} \ x \ [] & = [x] \\
\text{insort1} \ x \ (y \ # \ ys) & = (\text{if} \ x \ \leq \ y \ \text{then} \ x \ # \ y \ # \ ys \ \text{else} \ y \ # \ \text{insort1} \ x \ ys)
\end{align*}
\]

\[
\begin{align*}
\text{insort} :: \ 'a \ \text{list} \Rightarrow \ 'a \ \text{list} \\
\text{insort} \ [] & = [] \\
\text{insort} \ (x \ # \ xs) & = \text{insort1} \ x \ (\text{insort} \ xs)
\end{align*}
\]

#### 2.2.1 Functional Correctness

We start by proving the preservation of the multiset of elements:
2.2 Insertion Sort

\[
\begin{align*}
\text{mset} (\text{insort} \ 1 \ x \ xs) &= \{x\} + \text{mset} \ xs \quad (2.1) \\
\text{mset} (\text{insort} \ xs) &= \text{mset} \ xs \quad (2.2)
\end{align*}
\]

Both properties are proved by induction; the proof of (2.2) requires (2.1).

Now we turn to sortedness. Because the definition of sorted involves set, it is frequently helpful to prove multiset preservation first (as we have done above) because that yields preservation of the set of elements. That is, from (2.1) we obtain:

\[
\text{set} (\text{insort} \ 1 \ x \ xs) = \{x\} \cup \text{set} \ xs \quad (2.3)
\]

Two inductions prove

\[
\begin{align*}
\text{sorted} (\text{insort} \ 1 \ a \ xs) &= \text{sorted} \ xs \quad (2.4) \\
\text{sorted} (\text{insort} \ xs) &= \text{sorted} \ xs \quad (2.5)
\end{align*}
\]

where the proof of (2.4) uses (2.3) and the proof of (2.5) uses (2.4).

2.2.2 Running Time Analysis

These are the canonical running time functions (according to Section 1.5):

\[
\begin{align*}
T_{\text{insort} \ 1} :: 'a \Rightarrow 'a \ list \Rightarrow \text{nat} \\
T_{\text{insort} \ 1} \ [] &= 1 \\
T_{\text{insort} \ 1} \ x \ (y \# \ ys) &= (\text{if } x \leq y \ \text{then } 0 \ \text{else } T_{\text{insort} \ 1} \ x \ ys) + 1 \\
T_{\text{insort} :: 'a \ list \Rightarrow \text{nat}} \\
T_{\text{insort} \ []} &= 1 \\
T_{\text{insort}} \ (x \# \ xs) &= T_{\text{insort}} \ xs + T_{\text{insort} \ 1} \ (\text{insort} \ xs) + 1
\end{align*}
\]

A dismal quadratic upper bound for the running time of insertion sort is proved readily:

**Lemma 2.1.** \( T_{\text{insort}} \ xs \leq (|xs| + 1)^2 \)

**Proof.** The following properties are proved by induction on \( xs \):

\[
\begin{align*}
T_{\text{insort} \ 1} \ x \ xs &\leq |xs| + 1 \quad (2.6) \\
|\text{insort} \ 1 \ x \ xs| &= |xs| + 1 \quad (2.7) \\
|\text{insort} \ xs| &= |xs| \quad (2.8)
\end{align*}
\]

The proof of (2.8) needs (2.7). The proof of \( T_{\text{insort}} \ xs \leq (|xs| + 1)^2 \) is also by induction on \( xs \). The base case is trivial. The induction step is easy:
Exercise 2.2 asks you to show that \textit{insort} actually has quadratic running time on all lists \([n, n-1, \ldots, 0]\).

2.2.3 Exercises

Exercise 2.1. Show that any sorting function has the same input/output behaviour as insertion sort:
\[
(\forall xs. mset (f xs) = mset xs) \land (\forall xs. sorted (f xs)) \implies f xs = \text{insort} xs
\]

Exercise 2.2. Show that \(T_{\text{insort}}\) achieves its optimal value of \(2 \cdot n + 1\) for sorted lists, and its worst-case value of \((n+1) \cdot (n+2)\) \(\text{div} \ 2\) for the list \(\text{rev} [0..<n]\).

2.3 Quicksort

Quicksort is a divide-and-conquer algorithm that sorts a list as follows: pick a pivot element from the list; partition the remaining list into those elements that are smaller and those that are greater than the pivot (equal elements can go into either sublist); sort these sublists recursively and append the results. A particularly simple version of this approach, where the first element is chosen as the pivot, looks like this:

\[
\text{quicksort} :: 'a list \Rightarrow 'a list
\]
\[
\text{quicksort} \ [] = []
\]
\[
\text{quicksort} \ (x \# xs) = \text{quicksort} \ (\text{filter} \ (\lambda y. y < x) \ xs) \ @ [x] \ @ \text{quicksort} \ (\text{filter} \ (\lambda y. y \geq x) \ xs)
\]

2.3.1 Functional Correctness

Preservation of the multiset of elements
\[
mset (\text{quicksort} \ xs) = mset \ xs \quad (2.9)
\]
is proved by computation induction using these lemmas:

\[
\text{mset} \left( \text{filter } P \ xs \right) = \text{filter}_\text{mset} P \left( \text{mset } xs \right)
\]

\[
(\forall x. \ P \ x = (\neg \ Q \ x)) \rightarrow \text{filter}_\text{mset} P \ M + \text{filter}_\text{mset} Q \ M = M
\]

A second computation induction proves sortedness

\[
\text{sorted} \left( \text{quicksort } xs \right)
\]

using the lemmas

\[
\text{sorted} \left( xs \ @ \ ys \right) = \left( \text{sorted } xs \land \text{sorted } ys \land (\forall x \in \text{set } xs. \forall y \in \text{set } ys. \ x \leq y) \right)
\]

\[
\text{set} \left( \text{quicksort } xs \right) = \text{set } xs
\]

where the latter one is an easy consequence of (2.9).

We do not analyze the running time of \textit{quicksort}. It is well known that in the worst case it is quadratic (exercise!) but that the average-case running time (in a certain sense) is \(O(n \lg n)\). If the pivot is chosen randomly instead of always choosing the first element, the \textit{expected} running time is also \(O(n \lg n)\). The necessary probabilistic analysis is beyond the scope of this text but can be found elsewhere [16, 18].

2.3.2 Exercises

\textbf{Exercise 2.3.} Function \textit{quicksort} appends the lists returned from the recursive calls. This is expensive because the running time of \((@)\) is linear in the length of its first argument. Define a function \textit{quicksort2} :: \textit{a list => 'a list => 'a list} that avoids \((@)\) but accumulates the result in its second parameter via \((#)\) only. Prove \textit{quicksort2} \(xs \ @ \ ys = \text{quicksort} \ xs \ @ \ ys\).

\textbf{Exercise 2.4.} There is one obvious optimisation to the version of quicksort that we studied before: instead of partitioning the list into those elements that are smaller than the pivot and those that are at least as big as the pivot, we can use three-way partitioning:

\[
\text{partition3} :: \textit{a list => 'a list => 'a list => 'a list}
\]

\[
\text{partition3} \ x \ xs = (\text{filter } (\lambda y. \ y < x) \ xs, \text{filter } (\lambda y. \ y = x) \ xs, \text{filter } (\lambda y. \ y > x) \ xs)
\]

\[
\text{quicksort3} :: \textit{a list => 'a list}
\]

\[
\text{quicksort3} [] = []
\]
Prove that this version of quicksort also produces the correct results.

**Exercise 2.5.** In this exercise, we will examine the worst-case behaviour of Quicksort, which is e.g. achieved if the input list is already sorted. Consider the time function for Quicksort:

\[
T_{\text{quicksort}} :: \text{'a list} \Rightarrow \text{nat} \\
T_{\text{quicksort}} [] = 1 \\
T_{\text{quicksort}} (x \# xs) = T_{\text{quicksort}} (\text{filter} (\lambda y. y < x) xs) + T_{\text{quicksort}} (\text{filter} (\lambda y. y \geq x) xs) + 2 \cdot T_{\text{filter}} (\lambda \_. 1) xs + 1
\]

1. Show that Quicksort takes quadratic time on sorted lists by proving

\[
\text{sorted} \; xs \rightarrow T_{\text{quicksort}} \; xs = a \cdot |xs|^2 + b \cdot |xs| + c
\]

for suitable values \(a, b, c\).

2. Show that this is the worst-case running time by proving

\[
T_{\text{quicksort}} \; xs \leq a \cdot |xs|^2 + b \cdot |xs| + c
\]

for the values of \(a, b, c\) you determined in the previous step.

### 2.4 Top-Down Merge Sort

Merge sort is another prime example of a divide-and-conquer algorithm, and one whose running time is guaranteed to be \(O(n \lg n)\). We will consider three variants and start with the simplest one: split the list into two halves, sort the halves separately and merge the results.

\[
\text{merge} :: \text{'a list} \Rightarrow \text{'a list} \Rightarrow \text{'a list} \\
\text{merge} [] \; ys = ys
\]
2.4 Top-Down Merge Sort

merge xs [] = xs
merge (x # xs) (y # ys) = (if x ≤ y then x # merge xs (y # ys) else y # merge (x # xs) ys)

msort :: 'a list ⇒ 'a list
msort xs = (let n = |xs|
in if n ≤ 1 then xs
else merge (msort (take (n div 2) xs)) (msort (drop (n div 2) xs)))

2.4.1 Functional Correctness

We start off with multisets and sets of elements:

\[
\begin{align*}
\text{mset} \ (\text{merge} \ xs \ ys) & = \text{mset} \ xs + \text{mset} \ ys \\
\text{set} \ (\text{merge} \ xs \ ys) & = \text{set} \ xs ∪ \text{set} \ ys
\end{align*}
\]

(2.10) (2.11)

Proposition (2.10) is proved by induction on the computation of merge and (2.11) is an easy consequence.

Lemma 2.2. mset (msort xs) = mset xs

Proof. The proof is by induction on the computation of msort. Let \( n = |xs| \). The base case \( n ≤ 1 \) is trivial. Now assume \( n > 1 \) and let \( ys = \text{take} \ (n \ \text{div} \ 2) \ xs \) and \( zs = \text{drop} \ (n \ \text{div} \ 2) \ xs \).

\[
\begin{align*}
\text{mset} \ (\text{msort} \ xs) & = \text{mset} \ (\text{msort} \ ys) + \text{mset} \ (\text{msort} \ zs) & \text{by (2.10)} \\
& = \text{mset} \ ys + \text{mset} \ zs & \text{by IH} \\
& = \text{mset} \ (ys @ zs) \\
& = \text{mset} \ xs
\end{align*}
\]

Now we turn to sortedness. An induction on the computation of merge, using (2.11), yields

\[
\text{sorted} \ (\text{merge} \ xs \ ys) = (\text{sorted} \ xs ∧ \text{sorted} \ ys)
\]

(2.12)

Lemma 2.3. sorted (msort xs)

The proof is an easy induction on the computation of msort. The base case \( n ≤ 1 \) follows because every list of length \( ≤ 1 \) is sorted. The induction step follows with the help of (2.12).
2.4.2 Running Time Analysis

To simplify the analysis, and in line with the literature, we only count the number of comparisons:

\[
C_{\text{merge}} :: \text{a list } \Rightarrow \text{a list } \Rightarrow \text{nat}
\]
\[
C_{\text{merge}} \text{[]} = 0
\]
\[
C_{\text{merge}} \text{[]} = 0
\]
\[
C_{\text{merge}} (x \# x) (y \# y) = 1 + (\text{if } x < y \text{ then } C_{\text{merge}} x y \text{ else } C_{\text{merge}} x y)
\]

\[
C_{\text{msort}} :: \text{a list } \Rightarrow \text{nat}
\]
\[
C_{\text{msort}} x = \begin{cases} 
\text{let } n = |x|; \\
\quad y = \text{take} (n \div 2) x; \\
\quad z = \text{drop} (n \div 2) x \\
\text{in if } n \leq 1 \text{ then } 0 \\
\text{else } C_{\text{msort}} y + C_{\text{msort}} z + C_{\text{merge}} (\text{msort} y) (\text{msort} z)
\end{cases}
\]

By computation inductions we obtain:

\[
|\text{merge} x y| = |x| + |y|
\] (2.13)
\[
|\text{msort} x| = |x|
\] (2.14)
\[
C_{\text{merge}} x y \leq |x| + |y|
\] (2.15)

where the proof of (2.14) uses (2.13).

To simplify technicalities, we prove the \( n \cdot \lg n \) bound on the number of comparisons in \( \text{msort} \) only for \( n = 2^k \), in which case the bound becomes \( k \cdot 2^k \).

**Lemma 2.4.** \( |xs| = 2^k \rightarrow C_{\text{msort}} x s \leq k \cdot 2^k \)

**Proof.** The proof is by induction on \( k \). The base case is trivial and we concentrate on the step. Let \( n = |xs|, y = \text{take} (n \div 2) x \) and \( z = \text{drop} (n \div 2) x \). The case \( n \leq 1 \) is trivial. Now assume \( n > 1 \).

\[
C_{\text{msort}} x s
\]
\[
= C_{\text{msort}} y + C_{\text{msort}} z + C_{\text{merge}} (\text{msort} y) (\text{msort} z)
\]
\[
\leq C_{\text{msort}} y + C_{\text{msort}} z + |y| + |z| \quad \text{using (2.15) and (2.14)}
\]
\[
\leq k \cdot 2^k + k \cdot 2^k + |y| + |z| \quad \text{by IH}
\]
\[
= k \cdot 2^k + k \cdot 2^k + |z|
\]
\[
= (k + 1) \cdot 2^k + 1 \quad \text{by assumption } |xs| = 2^k + 1 \quad \Box
\]
Exercise 2.6. The definition of \texttt{msort} is inefficient in that it calls \texttt{length}, \texttt{take} and \texttt{drop} for each list. Instead we can split the list into two halves by traversing it only once and putting its elements alternately on two piles, for example \texttt{halve [2, 3, 4]} ([0], [1]) = ([4, 2, 0], [3, 1]). Define \texttt{halve} and \texttt{msort2}

\begin{verbatim}
msort2 [] = []
msort2 [x] = [x]
msort2 xs = (let (ys1, ys2) = halve xs ([], []) in merge (msort2 ys1) (msort2 ys2))
\end{verbatim}

and prove \texttt{mset (msort2 xs)} = \texttt{mset xs} and \texttt{sorted (msort2 xs)}. (Hint for Isabelle users: The definition of \texttt{msort2} is tricky because its termination relies on suitable properties of \texttt{halve}.)

2.5 Bottom-Up Merge Sort

Bottom-up merge sort starts by turning the input \([x_1, \ldots, x_n]\) into the list \([[[x_1], \ldots, [x_n]]]\). Then it passes over this list of lists repeatedly, merging pairs of adjacent lists on every pass until at most one list is left.

\begin{verbatim}
merge_adj :: 'a list list ⇒ 'a list list
merge_adj [] = []
merge_adj [xs] = [xs]
merge_adj (xs # ys # zss) = merge xs ys # merge_adj zss

merge_all :: 'a list list ⇒ 'a list
merge_all [] = []
merge_all [xs] = xs
merge_all xss = merge_all (merge_adj xss)

msort_bu :: 'a list ⇒ 'a list
msort_bu xs = merge_all (map (λx. [x]) xs)
\end{verbatim}

Termination of \texttt{merge_all} relies on the fact that \texttt{merge_adj} halves the length of the list (rounding up). Computation induction proves

\[ |merge_adj xs| = (|xs| + 1) \div 2 \] (2.16)

2.5.1 Functional Correctness

We introduce the abbreviation \texttt{mset_mset :: 'a list list ⇒ 'a multiset}:
These are the key properties of the functions involved:

\[ \text{mset}_\text{mset} (\text{merge}_\text{adj} \ xss) = \text{mset}_\text{mset} \ xss \]
\[ \text{mset} (\text{merge}_\text{all} \ xss) = \text{mset}_\text{mset} \ xss \]  \hspace{1cm} (2.17)
\[ \text{mset} (\text{msort}_\text{bu} \ xs) = \text{mset} \ xss \]
\[ (\forall x \in \text{set} \ xss. \ \text{sorted} \ x) \rightarrow (\forall x \in \text{set} (\text{merge}_\text{adj} \ xss). \ \text{sorted} \ x) \]
\[ (\forall x \in \text{set} \ xss. \ \text{sorted} \ x) \rightarrow \text{sorted} (\text{merge}_\text{all} \ xss) \]  \hspace{1cm} (2.18)
\[ \text{sorted} (\text{msort}_\text{bu} \ xs) \]

The third and the last property prove functional correctness of \text{msort}_\text{bu}. The proof of each proposition may use the preceding proposition and the propositions (2.10) and (2.12). The propositions about \text{merge}_\text{adj} and \text{merge}_\text{all} are proved by computation inductions.

### 2.5.2 Running Time Analysis

Again, we count only comparisons:

\[ C_{\text{merge}_\text{adj}} :: \text{nat} \]
\[ C_{\text{merge}_\text{adj}} [] = 0 \]
\[ C_{\text{merge}_\text{adj}} [\_\_] = 0 \]
\[ C_{\text{merge}_\text{adj}} (x \# y \# zss) = C_{\text{merge}} \ xss + C_{\text{merge}_\text{adj}} \ zss \]

\[ C_{\text{merge}_\text{all}} :: \text{nat} \]
\[ C_{\text{merge}_\text{all}} [] = 0 \]
\[ C_{\text{merge}_\text{all}} [\_\_] = 0 \]
\[ C_{\text{merge}_\text{all}} \ xss = C_{\text{merge}_\text{adj}} \ xss + C_{\text{merge}_\text{all}} (\text{merge}_\text{adj} \ xss) \]

\[ C_{\text{msort}_\text{bu}} :: \text{nat} \]
\[ C_{\text{msort}_\text{bu}} \ xs = C_{\text{merge}_\text{all}} (\text{map} (\lambda x. \ [x]) \ xs) \]

By simple computation inductions we obtain:

\[ \text{even} \ |\ xss| \wedge (\forall x \in \text{set} \ xss. \ |x| = m) \rightarrow \]
\[ (\forall x \in \text{set} (\text{merge}_\text{adj} \ xss). \ |x| = 2 \cdot m) \]  \hspace{1cm} (2.19)
\[ (\forall x \in \text{set} \ xss. \ |x| = m) \rightarrow C_{\text{merge}_\text{adj}} \ xss \leq m \cdot |\ xss| \]  \hspace{1cm} (2.20)

using (2.13) for (2.19) and (2.15) for (2.20).
Lemma 2.5. \((\forall xss \in \text{set. } |xss| = m) \land |xss| = 2^k \rightarrow C_{merge\_all\ xss} \leq m \cdot k \cdot 2^k\)

Proof by induction on the computation of merge\_all. We concentrate on the nontrivial recursive case arising from the third equation. We assume \(|xss| > 1, \forall xss \in \text{set xss}. |xss| = m\) and \(|xss| = 2^k\). Clearly \(k \geq 1\) and thus even \(|xss|\).

Thus (2.19) implies \(\forall xss \in \text{set (merge\_adj xss). } |xss| = 2 \cdot m\). Also note

\[
|\text{merge\_adj xss}|
= (|xss| + 1) \div 2 \\
= 2^k - 1 
\]

using \(|xss| = 2^k\) and \(k \geq 1\) by arithmetic

Let \(yss = \text{merge\_adj xss}\). We can now prove the lemma:

\[
C_{merge\_all\ xss} = C_{merge\_adj\ xss} + C_{merge\_all\ yss}
\leq m \cdot 2^k + C_{merge\_all\ yss}
\leq m \cdot 2^k + 2 \cdot m \cdot (k - 1) \cdot 2^k - 1
\]

by IH using \(\forall xss \in \text{set yss. } |xss| = 2 \cdot m\) and \(|yss| = 2^k - 1\)

\[
= m \cdot k \cdot 2^k 
\]

\(\square\)

Setting \(m = 1\) we obtain the same upper bound as for top-down merge sort in Lemma 2.4:

Corollary 2.6. \(|xs| = 2^k \rightarrow C_{msort\_bu\ xs} \leq k \cdot 2^k\)

2.6 Natural Merge Sort

A disadvantage of all the sorting functions we have seen so far (except insertion sort) is that even in the best case they do not improve upon the \(n \cdot \lg n\) bound. For example, given the sorted input \([1, 2, 3, 4, 5]\), \(msort\_bu\) will, as a first step, create \([[[1], [2], [3], [4], [5]]]\) and then merge this list of lists recursively.

A slight variation of bottom-up merge sort, sometimes referred to as "natural merge sort," first partitions the input into its constituent ascending and descending subsequences (collectively referred to as \textit{runs}) and only then starts merging. In the above example we would get merge\_all \([[[1], [2, 3, 4, 5]]]\), which returns immediately with the result \([1, 2, 3, 4, 5]\). Assuming that obtaining \textit{runs} is of linear complexity, this yields a best-case performance that is linear in the number of list elements.

Function \textit{runs} computes the initial list of lists; it is defined mutually recursively with \textit{asc} and \textit{desc}, which gather ascending and descending runs in accumulating parameters:
Function `desc` needs to reverse the descending run it collects. Therefore a natural choice for the type of its accumulator `as` is `list`, since recursively prepending elements (using `(##)`) ultimately yields a reversed list.

Function `asc` collects an ascending run and is slightly more complicated than `desc`. If we used lists, we could accumulate the elements similarly to `desc` but using `as @ [a]` instead of `a @ as`. This would take quadratic time in the number of appended elements. Therefore the “standard” solution is to accumulate elements using `(##)` and to reverse the accumulator in linear time (as shown in Section 1.5.1) at the end. However, another interesting option (that yields better performance for some functional languages, like Haskell) is to use difference lists. This is the option we chose for `asc`.

In the functional programming world, difference lists are a well-known trick to append lists in constant time by representing lists as functions of type `'a list ⇒ 'a list`. For difference lists, we have the following correspondences: empty list `[] ≈ λx. x`, singleton list `[x] ≈ (##) x`, and list append `xs @ ys ≈ xs o ys` (taking constant time). Moreover, transforming a difference list `xs` into a normal list is as easy as `xs []` (taking linear time).

Note that, due to the mutually recursive definitions of `runs`, `asc`, and `desc`, whenever we prove a property of `runs`, we simultaneously have to prove suitable properties of `asc` and `desc` using mutual induction.

Natural merge sort is the composition of `merge_all` and `runs`:
2.6 Natural Merge Sort

\[ \text{nmsort :: 'a list \rightarrow 'a list} \]
\[ \text{nmsort xs = merge\_all (runs xs)} \]

2.6.1 Functional Correctness

We have

\[
(\forall \mathbf{xs} \ \mathbf{ys}. \ \mathbf{f} \ (\mathbf{xs} @ \mathbf{ys}) = \mathbf{f} \mathbf{xs} @ \mathbf{ys}) \longrightarrow \\
\text{mset\_mset (asc } \mathbf{x} \ f \ \mathbf{ys}) = \{x\} + \text{mset} \ (f \ []) + \text{mset} \mathbf{ys} \quad (2.21) \\
\text{mset\_mset (desc } \mathbf{x} \ \mathbf{xs} \ \mathbf{ys}) = \{x\} + \text{mset} \mathbf{xs} + \text{mset} \mathbf{ys} \quad (2.22) \\
\text{mset\_mset (runs} \mathbf{xs}) = \text{mset} \mathbf{xs} \quad (2.23) \\
\text{mset} \ (\text{nmsort } \mathbf{xs}) = \text{mset} \mathbf{xs} \quad (2.24)
\]

where (2.23), (2.21), and (2.22) are proved simultaneously. The assumption of (2.21) on \( \mathbf{f} \) ensures that \( \mathbf{f} \) is a difference list. We use (2.23) together with (2.17) in order to show (2.24). Moreover, we have

\[
\forall \mathbf{x} \in \text{set} \ (\text{runs } \mathbf{xs}). \ \text{sorted} \ \mathbf{x} \quad (2.25) \\
\text{sorted} \ (\text{nmsort } \mathbf{xs}) \quad (2.26)
\]

where we use (2.25) together with (2.18) to obtain (2.26).

2.6.2 Running Time Analysis

Once more, we only count comparisons:

\[ \text{C\_runs :: 'a list \Rightarrow nat} \]
\[ \text{C\_runs } (\mathbf{a} \ # \ b \ # \ \mathbf{xs}) = 1 + (\text{if } b < a \ \text{then } C\_\text{desc } b \ \mathbf{xs} \ \text{else } C\_\text{asc } b \ \mathbf{xs}) \]
\[ \text{C\_runs } [] = 0 \]
\[ \text{C\_runs } [\_] = 0 \]

\[ \text{C\_asc :: 'a \Rightarrow 'a list \Rightarrow nat} \]
\[ \text{C\_asc } a \ (b \ # \ \mathbf{bs}) = 1 + (\text{if } \neg \ b < a \ \text{then } C\_\text{asc } b \ \mathbf{bs} \ \text{else } C\_\text{runs } (b \ # \ \mathbf{bs}) \]
\[ \text{C\_asc } \_ [] = 0 \]

\[ \text{C\_desc :: 'a \Rightarrow 'a list \Rightarrow nat} \]
\[ \text{C\_desc } a \ (b \ # \ \mathbf{bs}) = 1 + (\text{if } b < a \ \text{then } C\_\text{desc } b \ \mathbf{bs} \ \text{else } C\_\text{runs } (b \ # \ \mathbf{bs}) \]
\[ \text{C\_desc } \_ [] = 0 \]
\[ \text{C}_{\text{nsort}} :: 'a \text{ list} \Rightarrow \text{nat} \]
\[ \text{C}_{\text{nsort}} \; \text{xs} = \text{C}_{\text{runs}} \; \text{xs} + \text{C}_{\text{merge\_all}} \left( \text{runs} \; \text{xs} \right) \]

Again note the mutually recursive definitions of \( \text{C}_{\text{runs}}, \text{C}_{\text{asc}}, \) and \( \text{C}_{\text{desc}}. \) Hence the remark on proofs about \( \text{runs} \) also applies to proofs about \( \text{C}_{\text{runs}}. \)

Before talking about \( \text{C}_{\text{nsort}}, \) we need a variant of Lemma 2.5 that also works for lists whose lengths are not powers of two (since the result of \( \text{runs} \) will usually not satisfy this property).

To this end, we will need the following two results, which we prove by two simple computation inductions using (2.15) and (2.13):

\[ \text{C}_{\text{merge\_adj}} \; \text{xss} \leq |\text{concat} \; \text{xss}| \tag{2.27} \]
\[ |\text{concat} \left( \text{merge\_adj} \; \text{xss} \right)| = |\text{concat} \; \text{xss}| \tag{2.28} \]

**Lemma 2.7.** \( \text{C}_{\text{merge\_all}} \; \text{xss} \leq |\text{concat} \; \text{xss}| \cdot |\lg |\text{xss}|| \)

*Proof* by induction on the computation of \( \text{C}_{\text{merge\_all}}. \) We concentrate on the nontrivial recursive case arising from the third equation. It follows that \( \text{xss} \) is of the form \( \text{xs} \neq \text{ys} \neq \text{zss}. \) Further note that for all \( n :: \text{nat}: \)

\[ 2 \leq n \rightarrow |\lg n| = |\lg \left( \left( n - 1 \right) \div 2 + 1 \right) \div 2 + 1 \right) + 1 \tag{2.29} \]

Now, let \( m = |\text{concat} \; \text{xss}|. \) Then we have

\[
\begin{align*}
\text{C}_{\text{merge\_all}} \; \text{xss} &= \text{C}_{\text{merge\_adj}} \; \text{xss} + \text{C}_{\text{merge\_all}} \left( \text{merge\_adj} \; \text{xss} \right) \\
&\leq m + \text{C}_{\text{merge\_all}} \left( \text{merge\_adj} \; \text{xss} \right) \quad \text{using (2.27)} \\
&\leq m + |\text{concat} \left( \text{merge\_adj} \; \text{xss} \right)| \cdot |\lg |\text{merge\_adj} \; \text{xss}|| \quad \text{by IH} \\
&= m + m \cdot |\lg |\text{merge\_adj} \; \text{xss}|| \\
&= m + m \cdot |\lg \left( \left( |\text{xss}| + 1 \right) \div 2 \right) + 2 + 1 \right) | \\
&= m \cdot \left( |\lg \left( \left( |\text{xss}| + 1 \right) \div 2 + 1 \right) \div 2 + 1 \right) + 1) \\
&= m \cdot |\lg \left( \left( |\text{xss}| + 2 \right) \right) | \\
&= m \cdot |\lg |\text{xss}|| \quad \Box
\end{align*}
\]

Three simple computation inductions, each performed simultaneously for the corresponding mutually recursive definitions, yield:

\[
(\forall \text{xs} \; \text{ys}. \; f \; (\text{xs} @ \text{ys}) = f \; \text{xs} @ \text{ys}) \rightarrow \\
|\text{concat} \left( \text{asc} \; a \; f \; \text{ys} \right)| = 1 + |\text{f}|| + |\text{ys}|, \\
|\text{concat} \left( \text{desc} \; a \; \text{xs} \; \text{ys} \right)| = 1 + |\text{xs}| + |\text{ys}|, \\
|\text{concat} \left( \text{runs} \; \text{xs} \right)| = |\text{xs}| \\
\tag{2.30} \]

\[
(\forall \text{xs} \; \text{ys}. \; f \; (\text{xs} @ \text{ys}) = f \; \text{xs} @ \text{ys}) \rightarrow |\text{asc} \; a \; f \; \text{ys}| \leq 1 + |\text{ys}|, \\
\]
2.7 Stability

A sorting function is called stable if the order of equal elements is preserved. However, this only makes a difference if elements are not identified with their keys, as we have done so far. Let us assume instead that sorting is parameterized with a key function \( f : : a \rightarrow k \) that maps an element to its key and that the keys \( k \) are linearly ordered, not the elements. This is the specification of a sorting function \( \text{sort\_key} \):

\[
\text{mset} (\text{sort\_key } f \, x) = \text{mset} \, x \\
\text{sorted} (\text{map } f (\text{sort\_key } f \, x))
\]

Assuming (for simplicity) we are sorting pairs of keys and some attached information, stability means that sorting \([(2, x), (1, z), (1, y)]\) yields \([(1, z), (1, y), (2, x)]\) and not \([(1, y), (1, z), (2, x)]\). That is, if we extract all elements with the same key after sorting \( x \), they should be in the same order as in \( x \):

\[
\text{filter} (\lambda y. \, f \, y = k) (\text{sort\_key } f \, x) = \text{filter} (\lambda y. \, f \, y = k) \, x
\]

We will now define insertion sort adapted to keys and verify its correctness and stability.
The proofs of the functional correctness properties
\[\text{mset} (\text{insort} \_\text{key} f \ xs) = \text{mset} \ xs \] (2.33)
\[\text{sorted} (\text{map} f (\text{insort} \_\text{key} f \ xs))\]
are completely analogous to their counterparts for plain \text{insort}.

The proof of stability uses three auxiliary properties:
\[\forall x \in \text{set} \ xs. \ f \ a \leq f \ x \rightarrow \text{insort} \_\text{key} f \ a \ xs = a \# xs\] (2.34)
\[\neg P \ x \rightarrow \text{filter} P (\text{insort} \_\text{key} f \ x \ xs) = \text{filter} P \ xs\] (2.35)
\[\text{sorted} (\text{map} f \ xs) \land P \ x \rightarrow \text{filter} P (\text{insort} \_\text{key} f \ x \ xs) = \text{insort} \_\text{key} f \ x \ (\text{filter} P \ xs)\] (2.36)

The first one is proved by a case analysis on \(xs\). The other two are proved by induction on \(xs\), using (2.34) in the proof of (2.36).

**Lemma 2.9 (Stability of \text{insort} \_\text{key}).**
\[\text{filter} (\lambda y. f \ y = k) (\text{insort} \_\text{key} f \ xs) = \text{filter} (\lambda y. f \ y = k) \ xs\]

**Proof.** The proof is by induction on \(xs\). The base case is trivial. In the induction step we consider the list \(a \# xs\) and perform a case analysis. If \(f \ a \neq k\) the claim follows by IH using (2.35). Now assume \(f \ a = k\):
\[
\text{filter} (\lambda y. f \ y = k) (\text{insort} \_\text{key} f (a \# xs)) = \text{filter} (\lambda y. f \ y = k) (\text{insort} \_\text{key} f a (\text{insort} \_\text{key} f xs)) = \text{insort} \_\text{key} f a (\text{filter} (\lambda y. f \ y = k) (\text{insort} \_\text{key} f xs))
\]
using \(f \ a = k\), (2.36), (2.33)
\[
= \text{insort} \_\text{key} f a (\text{filter} (\lambda y. f \ y = k) \ xs) \quad \text{by IH}
\]
\[
= a \# \text{filter} (\lambda y. f \ y = k) \ xs \quad \text{using } f \ a = k \text{ and (2.34)}
\]
\[
= \text{filter} (\lambda y. f \ y = k) (a \# xs) \quad \text{using } f \ a = k \quad \Box
\]

As exercises we recommend to adapt some of the other sorting algorithms above to sorting with keys and to prove their correctness and stability.
Selection

A topic that is somewhat related to that of sorting is selection: given a list \(xs\) of length \(n\) with some linear ordering defined on it and a natural number \(k < n\), return the \(k\)-th smallest number in the list. If \(xs\) is sorted, this is exactly the \(k\)-th element of the list.

The defining properties of the selection operation are as follows:

\[
k < |xs| \rightarrow \left| \{ y \in_s mset xs \mid y < \text{select } k \text{ } xs \} \right| \leq k
\]
\[
\wedge \left| \{ y \in_s mset xs \mid y > \text{select } k \text{ } xs \} \right| < |xs| - k
\] (3.1)

These properties uniquely define the selection property, as shown by the following theorem:

**Theorem 3.1 (Uniqueness of the selection operation).**

*If* \(k < |xs|\) *and*

\[
\left| \{ z \in_s mset xs \mid z < x \} \right| \leq k
\]
\[
\left| \{ z \in_s mset xs \mid z > y \} \right| \leq k
\]

*then* \(x = y\).

**Proof.** Suppose \(x \neq y\) and then w.l.o.g. \(x < y\). This implies:

\[
\{ z \in_s mset xs \mid z \leq x \} \subseteq \{ z \in_s mset xs \mid z < y \}
\] (3.2)

From this we can prove the contradiction \(|xs| < |xs|\):

\[
|xs| = \left| \{ z \in_s mset xs \mid z \leq x \} \right| + \left| \{ z \in_s mset xs \mid z > x \} \right|
\]
\[
\leq \left| \{ z \in_s mset xs \mid z < y \} \right| + \left| \{ z \in_s mset xs \mid z > x \} \right|
\]
\[
< k + (|xs| - k)
\]
\[
= |xs|
\] using (3.1)
An equivalent, more concrete definition is the following:

\[
\begin{align*}
\text{select} & :: \text{nat} \Rightarrow \text{a list} \Rightarrow \text{a} \\
\text{select} \ k \ \text{xs} & = \text{sort} \ \text{xs} ! k
\end{align*}
\] (3.3)

**Theorem 3.2.** Our definition of \( \text{select} \) satisfies the conditions (3.1).

**Proof.** If \( \text{ys} \) is a sorted list, the following can be proved by a straightforward induction on \( \text{ys} \):

\[
\{ x \in_\ast \text{mset} \ \text{ys} \mid x < \text{ys} ! k \} \subseteq_\ast \text{mset} \ (\text{take} \ k \ \text{ys}) \{ x \in_\ast \text{mset} \ \text{ys} \mid x > \text{ys} ! k \} \subseteq_\ast \text{mset} \ (\text{drop} \ (k + 1) \ \text{ys})
\]

Taking the size of these multisets on both sides, we obtain:

\[
|\{ x \in_\ast \text{mset} \ \text{ys} \mid x < \text{ys} ! k \}| \leq k |\{ x \in_\ast \text{mset} \ \text{ys} \mid x > \text{ys} ! k \}| < |\text{ys}| - k
\]

By setting \( \text{ys} := \text{sort} \ \text{ys} \) here, we have

\[
\begin{align*}
 k & \geq |\{ x \in_\ast \text{mset} \ (\text{sort} \ \text{xs}) \mid x < \text{sort} \ \text{xs} ! k \}| \\
 &= |\{ x \in_\ast \text{mset} \ \text{xs} \mid x < \text{sort} \ \text{xs} ! k \}| \quad \text{using} \ \text{mset} \ (\text{sort} \ \text{xs}) = \text{mset} \ \text{xs} \\
 &= |\{ x \in_\ast \text{mset} \ \text{xs} \mid x < \text{select} \ k \ \text{xs} \}| \quad \text{using} \ (3.3)
\end{align*}
\]

and analogously for the elements greater than \( \text{select} \ k \ \text{xs} \). \( \square \)

We will frequently need two more important lemmas about \( \text{sort} \) and \( \text{select} \), namely that they are invariant under permutation of the input list:

\[
\text{mset} \ \text{xs} = \text{mset} \ \text{ys} \longrightarrow \text{sort} \ \text{xs} = \text{sort} \ \text{ys} \ (3.4) \ \
\text{select} \ k \ \text{xs} = \text{select} \ k \ \text{ys} \ \ (3.5)
\]

Equation (3.4) follows directly from the uniqueness of the \( \text{sort} \) operation, and (3.5) then follows from (3.4) and our definition of \( \text{select} \).

The definition using \( \text{sort} \ \text{xs} ! k \) already gives us a straightforward \( O(n \lg n) \) algorithm for the selection operation: sort the list with one of our \( O(n \lg n) \) sorting algorithms and then return the \( k \)-th element of the resulting sorted list. It is also fairly easy to come up with an algorithm that has running time \( O(kn) \), i.e. that runs in linear-time in \( n \) for any fixed \( k \) (see Exercise 3.3).

In the remainder of this chapter, we will look at a selection algorithm due to Blum et al. [10] that achieves \( O(n) \) running time unconditionally (i.e. for any \( k < n \)). Since a selection algorithm must, in general, inspect each element at least once (see Exercise 3.4), this running time is asymptotically optimal.
3.0.1 Exercises

Exercise 3.1. A simple special case of selection is \( \text{select 0 } xs \), i.e. the minimum. Implement a linear-time function \( \text{select0} \) such that

\[ xs \neq [] \rightarrow \text{select0 } xs = \text{select 0 } xs \]

and prove this. This function should be tail-recursive and traverse the list exactly once. You need not prove the linear running time (it should be obvious).

Exercise 3.2. How can your \( \text{select0} \) algorithm be modified to obtain an analogous algorithm \( \text{select1} \) such that

\[ |xs| > 1 \rightarrow \text{select1 } xs = \text{select 1 } xs \]

Do not try to prove the correctness yet; it gets somewhat tedious and you will be able to prove it more easily after the next exercise.

Exercise 3.3.

1. Based on the previous two exercises, implement and prove correct an algorithm \( \text{select\_fixed} \) that fulfils

\[ k < |xs| \rightarrow \text{select\_fixed } k xs = \text{select } k xs \]

The algorithm must be tail-recursive with running time \( O(kn) \) and traverse the list exactly once.

Hint: one approach is to first define a function \( \text{take\_sort} \) that computes \( \text{take } m (\text{sort } xs) \) in time \( O(mn) \).

2. Prove your \( \text{select1} \) from the previous exercise correct by showing that it is equivalent to \( \text{select\_fixed} 1 \).

3. Define a suitable time function for your \( \text{select\_fixed} \). Prove that this time function is \( O(kn) \), i.e. that

\[ T_{\text{select\_fixed} } k xs \leq C_1 \cdot k \cdot |xs| + C_2 \cdot |zs| + C_3 \cdot k + C_4 \]

for all \( k < |zs| \) for some constants \( C_1 \) to \( C_4 \).

If you have trouble finding the concrete values for these constants, try proving the result with symbolic constants first and observe what conditions need to be fulfilled in order to make the induction step go through.

Exercise 3.4. Show that if \( xs \) is a list of integers with no repeated elements, an algorithm computing the result of \( \text{select } k xs \) must examine every single element, i.e. for any index \( i < |zs| \), the \( i \)-th element can be replaced by some other number such that the result changes. Formally:

\[ k < |zs| \land i < |xs| \land \text{distinct } xs \rightarrow (\exists z. \text{distinct } (xs[i := z]) \land \text{select } k (xs[i := z]) \neq \text{select } k xs) \]

Here, the notation \( xs[i := z] \) denotes the list \( xs \) where the \( i \)-th element has been replaced with \( z \) (the first list element, as always, having index 0).

Hint: a lemma you might find useful is that \( \lambda k. \text{select } k xs \) is injective if \( xs \) has no repeated elements.
3.1 A Divide-and-Conquer Approach

As a first step in our attempt to derive an efficient algorithm for selection, recall what we did with the function \texttt{partition3} in the threeway quicksort algorithm in Exercise 2.4: we picked some pivot \(x\) from \(xs\) and partitioned the input list \(xs\) into the sublists \(ls\), \(es\), and \(gs\) that were smaller, equal, and greater than \(x\), respectively.

If we do the same for \(select\ k\ xs\), there are three possible cases:

- If \(k < |ls|\), the element we are looking for is located in \(ls\). To be more precise, it is the \(k\)-th smallest element of \(ls\), i.e. \(select\ k\ ls\).
- If \(k < |ls| + |es|\), the element we are looking for is located in \(es\) and must therefore be \(x\) itself.
- Otherwise, the element we are looking for must be located in \(gs\). More precisely, it is the \(k'\)-th smallest element of \(gs\) where \(k' = k - |ls| - |es|\).

This gives us a straightforward recursive divide-and-conquer algorithm for selection. To prove this correct, we first prove the following lemma about \(select\):

\[
select k (ys @ zs) = (\text{if } k < |ys| \text{ then } select k ys \text{ else } select (k - |ys|) zs)
\]  

(3.6)

\textbf{Proof.} The assumptions imply that \(sort\ xs \mathbin{@} sort\ ys\) is sorted, so that due to the uniqueness of the \(sort\) operation, we have:

\[
sort (xs \mathbin{@} ys) = sort xs \mathbin{@} sort ys
\]  

(3.7)

Then:

\[
select k (zs @ ys) = sort (zs @ ys) ! k
\]  

using (3.3)

\[
= (sort zs @ sort ys) ! k
\]  

using (3.7)

\[
= \text{if } k < |zs| \text{ then } sort zs ! k \text{ else } sort ys ! (k - |zs|)
\]

\[
= \text{if } k < |zs| \text{ then } select k zs \text{ else } select (k - |zs|) ys
\]  

using (3.3)

\[\square\]

Now the recurrence outlined before is a direct consequence:

\textbf{Theorem 3.3 (A recurrence for select).}

\[
k < |xs| \rightarrow
\]

\[
select k xs
\]

\[
= (\text{let } (ls, es, gs) = \text{partition3} x xs
\]

\[\text{in } \text{if } k < |ls| \text{ then } select k ls
\]

\[\text{else if } k < |ls| + |es| \text{ then } x \text{ else } select (k - |ls| - |es|) gs)\]
Proof. We clearly have \( \text{mset} \; \text{xs} = \text{mset} \; \text{ls} + \text{mset} \; \text{es} + \text{mset} \; \text{gs} \) and \( |\text{zs}| \)  

\[ |\text{ls}| + |\text{es}| + |\text{gs}|. \]  

Then:

\[
\text{select} \; k \; \text{xs} = \text{select} \; (k \; \text{ls} \; @ \; \text{es} \; @ \; \text{gs}) \tag{3.5}
\]

- if \( k < |\text{ls}| \) then \( \text{select} \; k \; \text{ls} \)
- else if \( k < |\text{ls}| < |\text{es}| \) then \( \text{select} \; (k - |\text{ls}|) \; \text{es} \) \tag{3.6}
- else \( \text{select} \; (k - |\text{ls}| - |\text{es}|) \; \text{gs} \)

Clearly, \( k - |\text{ls}| < |\text{es}| \leftrightarrow k < |\text{ls}| + |\text{es}| \) and \( \text{select} \; (k - |\text{ls}|) \; \text{es} = x \) since \( \text{select} \; (k - |\text{ls}|) \; \text{es} \in \text{set es} \) and \( \text{set es} = \{x\} \) by construction. \( \square \)

Note that this holds for any pivot \( x \). Indeed, \( x \) need not even be in the list itself. Therefore, the algorithm is (partially) correct no matter what pivot we choose. However, the number of recursive calls (and thereby the running time) depends strongly on the pivot choice:

- If we always choose a pivot that is smaller than any element in the list, the algorithm does not terminate at all.
- If we choose the smallest element in the list as a pivot every time, only one element is removed from the list in every recursion step so that we get \( n \) recursive calls in total. Since we do a linear amount of work in every step, this leads to a running time of \( \Theta(n^2) \).
- If we choose pivots from the list at random, the worst-case running time is again \( \Theta(n^2) \), but the expected running time can be shown to be \( \Theta(n) \), similarly to the situation in Quicksort. Indeed, it can also be shown that it is very unlikely that the running time is significantly "worse than linear" [39, Section 2.5].
- If we choose a pivot that cuts the list in half every time, we get roughly \( \log n \) recursive steps and, by the Master Theorem, a running time of \( \Theta(n) \) (assuming we can find such a pivot in linear time).

Clearly, the last case is the most desirable one. The element that cuts the list in half is called the median (a term widely used in statistics, albeit with a slightly different meaning). For lists of odd length, there is a unique element that achieves this, whereas for lists of even length there are two such elements (e.g. for the list \([1 \; 2 \; 3 \; 4]\), both 2 and 3 work). For our purposes, we always use the smaller one and define:

\[
\text{median} :: 'a list \Rightarrow 'a \\
\text{median} \; \text{xs} = \text{select} \; (|\text{xs}| - 1 \; \text{div} \; 2) \; \text{xs}
\]

Unfortunately, computing the median of a list is no easier than selection (see Exercise 3.5), so it seems that, for now, this does not really help us.
Exercise 3.5. Show that select \( k \) \( xs \) can be reduced to computing the median of a list in linear time, i.e. give a linear-time function \( \text{reduce\_select\_median} \) such that

\[
\begin{align*}
xs \neq [] \land k < |xs| &\rightarrow \\
\text{reduce\_select\_median} k \ xs \neq [] \land \\
\text{median} (\text{reduce\_select\_median} k \ xs) = \text{select} \ k \ xs
\end{align*}
\]

3.2 The Median of Medians

Since, as we have seen, computing a true median in every recursive step is too expensive, the next natural question is: is there something that is almost as good as a median but easier to compute?

This is indeed the case, and this is where the ingenuity of the algorithm is hidden: instead of computing the median of all the list elements, we use another list element \( M \), which is computed as follows:

- chop the list into groups of 5 elements each (possibly with one smaller group at the end if \( n \) is not a multiple of 5)
- compute the median of each of the \( \lceil \frac{n}{5} \rceil \) groups (which can be done in constant time for each group using e.g. insertion sort, since their sizes are bounded by 5)
- compute the median \( M \) of these \( \lceil \frac{n}{5} \rceil \) elements (which can be done by a recursive call to the selection algorithm)

We call \( M \) the median of medians. \( M \) is not quite as good a pivot as the true median, but it is still fairly decent:

**Theorem 3.4 (Pivoting bounds for the median of medians).**

Let \( xs \) be a list and let \( \prec \) be either \( < \) or \( > \). Define

\[
M := \text{median} (\text{map} \ \text{median} (\text{chop} \ 5 \ xs))
\]

where the \text{chop} function performs the chopping described earlier. Then:

\[
|\{ y \in_{\neq} \text{mset} \ xs \mid y \prec M \}| \leq \lfloor 0.7 \cdot n + 3 \rfloor
\]

**Proof.** The result of \( \text{chop} \ 5 \ xs \) is a list of \( \lceil \frac{n}{5} \rceil \) chunks, each of size at most 5, i.e. \( |\text{chop} \ 5 \ xs| = \lceil \frac{n}{5} \rceil \). Let us split these chunks into two groups according to whether their median is \( < M \) or \( \geq M \):

\[
Y_\prec := \{ ys \in_{\neq} \text{mset} (\text{chop} \ 5 \ xs) \mid \text{median} \ ys \prec M \}
\]

\[
Y_\geq := \{ ys \in_{\neq} \text{mset} (\text{chop} \ 5 \ xs) \mid \text{median} \ ys \geq M \}
\]

We clearly have
3.2 The Median of Medians

\[ mset \; xs = (\sum_{ys \leftarrow \text{chop}\; 5\; \text{xs}} mset\; ys) \quad (3.8) \]

\[ mset \; (\text{chop}\; 5\; \text{xs}) = Y_\prec + Y_\succ \quad (3.9) \]

\[ \lceil n / 5 \rceil = |Y_\prec| + |Y_\succ| \quad (3.10) \]

and since \( M \) is the median of the medians of the groups, we also know that:

\[ |Y_\prec| < \frac{1}{2} \cdot \lceil n / 5 \rceil \quad (3.11) \]

The core idea of the proof is that any group \( ys \in_s Y_\succ \) can have at most 2 elements that are \( \prec M \):

\[
\left\{ y \in_s mset\; ys \mid y \prec M \right\} \\
\leq \left\{ y \in_s mset\; ys \mid y \prec \text{median}\; ys \right\} \quad \text{because median}\; ys \preceq M \\
\leq |ys| \text{ div } 2 \\
\leq 5 \text{ div } 2 = 2
\]

And of course, since each group has size at most 5, any group in \( ys \in_s Y_\prec \) can contribute at most 5 elements. In summary, we have:

\[
\forall ys \in_s Y_\prec. \left\{ y \in_s mset\; ys \mid y \prec M \right\} \leq 2 \\
\forall ys \in_s Y_\succ. \left\{ y \in_s mset\; ys \mid y \prec M \right\} \leq 5 
\]

(3.12)

With this, we can begin our estimation of the number of elements \( \prec M \):

\[
\left\{ y \in_s mset\; xs \mid y \prec M \right\} \\
= \left\{ y \in_s (\sum_{ys \leftarrow \text{chop}\; 5\; \text{xs}} mset\; ys) \mid y \prec M \right\} \quad \text{using}(3.8) \\
= \sum_{ys \leftarrow \text{chop}\; 5\; \text{xs}} \left\{ y \in_s mset\; ys \mid y \prec M \right\} \\
= \sum_{ys \in_s (Y_\prec + Y_\succ)} \left\{ y \in_s mset\; ys \mid y \prec M \right\} \quad \text{using} \; (3.9)
\]

Taking the size of both sides, we have

\[
\left\{ y \in_s mset\; xs \mid y \prec M \right\} \\
\leq \sum_{ys \in_s (Y_\prec + Y_\succ)} \left\{ y \in_s mset\; ys \mid y \prec M \right\} \\
= \sum_{ys \in_s Y_\prec} \left\{ y \in_s mset\; ys \mid y \prec M \right\} + \sum_{ys \in_s Y_\succ} \left\{ y \in_s mset\; ys \mid y \prec M \right\} \\
\leq (\sum_{ys \in_s Y_\prec} 5) + (\sum_{ys \in_s Y_\succ} 2) \quad \text{using} \; (3.12) \\
= 5 \cdot |Y_\prec| + 2 \cdot |Y_\succ| \\
= 2 \cdot \lceil n / 5 \rceil + 3 \cdot |Y_\succ| \quad \text{using} \; (3.10) \\
\leq 3.5 \cdot \lceil n / 5 \rceil \quad \text{using} \; (3.11) \\
= 3.5 \cdot \lceil n / 5 \rceil \leq [0.7 \cdot n + 3]
\]

The delicate arithmetic reasoning about rounding in the end can thankfully be done fully automatically by Isabelle’s linarith tactic.
3.3 Selection in Linear Time

We now have all the ingredients to write down our algorithm: the base cases (i.e., sufficiently short lists) can be handled using the naive approach of performing insertion sort and then returning the \( k \)-th element. For bigger lists, we perform the divide-and-conquer approach outlined in Theorem 3.3 using \( M \) as a pivot. We have two recursive calls: one on a list with exactly \([0.2 \cdot n]\) elements to compute \( M \), and one on a list with at most \([0.7 \cdot n + 3]\). We will still need to show later that this actually leads to a linear-time algorithm, but the fact that \( 0.7 + 0.2 < 1 \) is at least encouraging.

The full algorithm looks like this:

```haskell
chop :: nat \Rightarrow 'a list \Rightarrow 'a list list
chop 0 _ = []
chop _ [] = []
chop n xs = take n xs \# chop n (drop n xs)

slow_select :: nat \Rightarrow 'a list \Rightarrow 'a
slow_select k xs = insort xs ! k

slow_median :: 'a list \Rightarrow 'a
slow_median xs = slow_select ((|xs| - 1) div 2) xs

mom_select :: nat \Rightarrow 'a list \Rightarrow 'a
mom_select k xs
= (if |xs| \leq 20 then slow_select k xs
   else let M = mom_select ((|xs| / 5) - 1) div 2)
           (map slow_median (chop 5 xs));
       (ls, es, gs) = partition3 M xs
       in if k < |ls| then mom_select k ls
          else if k < |ls| + |es| then M
          else mom_select (k - |ls| - |es|) gs)
```

The (partial) correctness statement is simple:

**Theorem 3.5 (Partial Correctness of \( \text{mom\_select} \)).** Let \( xs \) be a list and \( k < |xs| \). Then if \( \text{mom\_select} \) \( k \) \( xs \) terminates, we have

\[
\text{mom\_select} \ k \ \text{xs} = \text{select} \ k \ \text{xs}.
\]

**Proof.** Straightforward computation induction using Theorem 3.3. \( \square \)
Theorem 3.6 (Termination of `mom_select`). Let \(xs\) be a list and \(k < |xs|\). Then `mom_select k xs` terminates.

Proof. We prove termination by using the length of the list \(xs\) as a termination measure. We perform a computation induction and need to show that this length decreases in every recursive call if \(20 < |xs|\). This is easy to see:

- The list in the first recursive call has length \(|xs| / 5\), which is strictly less than \(|xs|\) if \(1 < |xs|\).
- The length of the list in the second recursive call is at most \(|xs| - 1\): by induction hypothesis, the first recursive call terminates, so by Theorem 3.5 we know that \(M = \text{median} (\text{map median} (\text{chop} \ 5 \ xs))\) and thus:

\[
M \in \text{set} (\text{map median} (\text{chop} \ 5 \ xs))
\]

\[
= \{\text{median} \ ys \mid ys \in \text{set} (\text{chop} \ 5 \ xs)\}
\]

\[
\subseteq \bigcup_{ys \in \text{set} (\text{chop} \ 5 \ xs)} \text{set} \ ys = \text{set} \ xs
\]

Hence, \(M \in \text{set} \ xs\) but \(M \notin \text{set} \ ls\) and \(M \notin \text{set} \ gs\) by construction.

\[\square\]

Exercise 3.6. The recursive definition of `mom_select` handles the cases \(|xs| \leq 20\) through the naive algorithm using insertion sort. The constant 20 here seems somewhat arbitrary. Find the smallest constant \(n_0\) for which the algorithm still works. Why do you think 20 was chosen?

3.4 Time Functions

It remains to show now that this indeed leads to a linear-time algorithm. The time function for our selection algorithm is as follows:

\[
T_{mom\_select} :: \text{nat} \rightarrow \text{'a list} \rightarrow \text{nat}
\]

\[
T_{mom\_select} \ k \ xs
\]

\[
= \begin{cases} T_{slow\_select} \ k \ xs & \text{if } |xs| \leq 20 \\ \\
\text{else let } xss = \text{chop} \ 5 \ xs; \\
ms = \text{map slow\_median} \ xss; \\
idx = (\lceil |xs| + 4 \rceil / 5 - 1) / 2; \\
x = \text{mom\_select} \ idx \ ms; \\
(ls, es, gs) = \text{partition3} \ x \ xs \\
in T_{mom\_select} \ idx \ ms + T_{chop} \ 5 \ xs + T_{map} \ T_{slow\_median} \ xss + T_{partition3} \ x \ xs + T_{length} \ ls + T_{length} \ es + 1 + \\
\text{(if } k < |ls| \text{ then } T_{mom\_select} \ k \ ls \\
\text{else if } k < |ls| + |es| \text{ then } 0 \\
\text{else } T_{mom\_select} \ (k - |ls| - |es|) \ gs) \\
\end{cases}
\]
We can then prove
\[
T_{\text{mom\_select}} k \text{xs} \leq T'_{\text{mom\_select}} |\text{xs}|
\]
where the upper bound \( T'_{\text{mom\_select}} \) is defined as follows:

\[
T'_{\text{mom\_select}} :: \text{nat} \Rightarrow \text{nat}
\]
\[
T'_{\text{mom\_select}} n = \begin{cases} 
463 & \text{if } n \leq 20 \\
T'_{\text{mom\_select}} [0.2 \cdot n] + T'_{\text{mom\_select}} [0.7 \cdot n + 3] + 17 \cdot n + 50 & \text{else}
\end{cases}
\]

The time functions of the auxiliary functions used here can be found in Section B.2 in the appendix. The proof is a simple computation induction using Theorem 3.4 and the time bounds for the auxiliary functions from Chapter B in the appendix.

The next section will be dedicated to showing that \( T'_{\text{mom\_select}} \in O(n) \).

**Exercise 3.7.** Show that the upper bound \([0.7 \cdot n + 3]\) is fairly tight by giving an infinite family \((\text{xs}_i)_{i \in \mathbb{N}}\) of lists with increasing lengths for which more than 70\% of the elements are larger than the median of medians (with chopping size 5). In Isabelle terms: define a function \( f :: \text{nat} \Rightarrow \text{nat list} \) such that \( \forall n. |f n| < |f (n + 1)| \) and
\[
\frac{|\{ y \in \text{mset} (f n) \mid y > \text{mom} (f n) \}|}{|f n|} > 0.7
\]
where \( \text{mom } \text{xs} = \text{median} (\text{map} \text{ median} (\text{chop} 5 \text{xs})) \).

### 3.5 Akra–Bazzi Light

The function \( T_{\text{mom\_select}} \) (let us write it as \( f \) for now) fulfils the recurrence
\[
n > 20 \rightarrow f n = f [0.2 \cdot n] + f [0.7 \cdot n + 3] + 17 \cdot n + 50
\]
Such divide-and-conquer recurrences are beyond the “normal” Master Theorem, but a generalisation, the Akra–Bazzi Theorem [3, 48, 17], does apply to them. Let us first abstract the situation a bit and consider the recurrence
\[
n > 20 \rightarrow f n = f [a \cdot n + b] + f [c \cdot n + d] + C_1 \cdot n + C_2
\]
where \( 0 < a, b < 1 \) and \( C_1, C_2 > 0 \). The Akra–Bazzi Theorem then tells us that such a function is \( O(n) \) if (and only if) \( a + b < 1 \). We will prove one direction of this (the one we need) now – Akra–Bazzi Light, so to say.
Instead of presenting the full theorem statement and its proof right away, let us take a more explorative approach. What we want to prove in the end is that there are real constants \( C_3 > 0 \) and \( C_4 \) such that \( f(n) \leq C_3 \cdot n + C_4 \) for all \( n \). Suppose these constants are fixed already and now we want to prove that the inequality holds for them. For simplicity of the presentation, we assume \( b, d \geq 0 \), but note that these assumptions are unnecessary and the proof still works for negative \( b \) and \( d \) if we replace \( b \) and \( d \) with \( \max 0 \, b \) and \( \max 0 \, d \).

The obvious approach to show this is by induction on \( n \), following the structure of the recurrence above. To do this, we use complete induction (i.e. the induction hypothesis holds for all \( n' < n \))\(^1\) and a case analysis on \( n > n_0 \) (where \( n_0 \) is some constant we will determine later).

The two cases we have to show in the induction are then:

**Base case:** \( \forall n \leq n_0. \ f(n) \leq C_3 \cdot n + C_4 \)

**Step:** \( \forall n > n_0. \ (\forall m < n. \ f(m) \leq C_3 \cdot m + C_4) \rightarrow f(n) \leq C_3 \cdot n + C_4 \)

We can see that in order to even be able to apply the induction hypothesis in the induction step, we need \( |a \cdot n + b| < n \). We can make the estimate\(^2\)

\[
|a \cdot n + b| \leq a \cdot n + b + 1 < n
\]

and then solve for \( n \), which gives us \( n \geq \frac{b + 1}{1 - a} \). If we do the same for \( c \) and \( d \) as well, we get the conditions

\[
n_0 \geq \frac{b + 1}{1 - a} \quad \text{and} \quad n_0 \geq \frac{d + 1}{1 - c}
\]

However, it will later turn out that these are implied by the other conditions we will have accumulated anyway.

Now that we have ensured that the basic structure of our induction will work out, let us continue with the two cases.

The base cases \( (n \leq n_0) \) is fairly uninteresting: we can simply choose \( C_4 \) to be large enough to satisfy all of them.

In the recursive step, unfolding one step of the recurrence and applying the induction hypotheses leaves us with the following proof obligation

\[
C_3 \cdot [a \cdot n + b] + C_4 + (C_3 \cdot [c \cdot n + d] + C_4) + C_1 \cdot n + C_2 \leq C_3 \cdot n + C_4
\]

\(^1\) In Isabelle, the corresponding rule is called `less_induct`.

\(^2\) The notation \(<\) stands for “must be less than”. It emphasises that this inequality is not a consequence of what we have shown so far, but something that we still need to show, or in this case something that we need to ensure by adding suitable preconditions.
or, equivalently,

\[ C_3 \cdot (a \cdot n + b) + (c \cdot n + d) - n) + C_1 \cdot n + C_2 + C_4 \leq 0 , \]

We estimate the left-hand side like this:

\[ C_3 \cdot (a \cdot n + b + 1 + (c \cdot n + d + 1) - n) + C_1 \cdot n + C_2 + C_4 \]
\[ = C_3 \cdot (b + d + 2) + C_2 + C_4 - C_3 \cdot (1 - a - c) - C_1 \cdot n \]
\[ \leq C_3 \cdot (b + d + 2) + C_2 + C_4 - C_3 \cdot (1 - a - c) - C_1 \cdot n_0 \]
\[ \leq 0 \]

The step from (*) to (†) uses the fact that \( n > n_0 \) and requires the factor \( C_3 \cdot (1 - a - c) - C_1 \) in front of the \( n \) to be positive, i.e. we need to add the assumption

\[ C_3 > \frac{C_1}{1 - a - c} . \]  

(3.14)

The term (†) (which we want to be \( \leq 0 \)) is now a constant. If we solve that inequality for \( C_3 \), we get the following two additional conditions:

\[ n_0 > \frac{b + d + 2}{1 - a - c} \quad \text{and} \quad C_3 \geq \frac{C_1 \cdot n_0 + C_2 + C_4}{(1 - a - c) \cdot n_0 - b - d - 2} \]  

(3.15)

The former of these directly implies our earlier conditions (3.13), so we can safely discard those now.

Now all we have to do is to find a combination of \( n_0, C_3, \) and \( C_4 \) that satisfy (3.14) and (3.15). This is straightforward:

\[ n_1 := \max n_0 \left( \frac{b + d + 2}{1 - a - c} + 1 \right) \]
\[ C_4 := \max \{ f \mid n \leq 20 \} \]
\[ C_3 := \max \left( \frac{C_1}{1 - a - c} \right) \left( \frac{C_1 \cdot n_1 + C_2 + C_4}{(1 - a - c) \cdot n_1 - b - d - 2} \right) \]

And with that, the induction goes through and we get the following theorem:

**Theorem 3.7 (Akra Bazzi Light).**

\[ a > 0 \land c > 0 \land a + c < 1 \land C_1 > 0 \land \]
\[ (\forall n > n_0. \ f(n) = f \left( \left[ a \cdot n + b \right] + \left[ c \cdot n + d \right] + C_1 \cdot n + C_2 \right) \rightarrow \]
\[ \left( \exists C_3, C_4, n. \ f(n) \leq C_3 \cdot n + C_4 \right) \]

**Exercise 3.8.**

1. Suppose that instead of groups of 5, we now chop into groups of size \( l \geq 1 \).
   Prove a corresponding generalisation of Theorem 3.4.

2. Examine (on paper only): how does this affect correctness and running time of our selection algorithm? Why do you think \( l = 5 \) was chosen?
Part II

Search Trees
Binary Trees

Binary trees are defined as a recursive data type:

```
datatype 'a tree = Leaf | Node ('a tree) 'a ('a tree)
```

The following syntactic sugar is sprinkled on top:

```
  () ≡ Leaf
  (l, x, r) ≡ Node l x r
```

The trees l and r are the left and right children of the node (l, x, r).

Because most of our tree will be binary trees, we drop the “binary” most of the time and have also called the type merely tree.

When displaying a tree in the usual graphical manner we show only the nodes. For example, ([[]], 3, [], 9, [{}], 7, []) is displayed like this:

```
  9
 / \
3   7
```

The (label of the) root node is 9. The depth (or level) of some node (or leaf) in a tree is the distance from the root. We use these concepts only informally.

4.1 Basic Functions

Two canonical functions on data types are `set` and `map`:
set_tree :: 'a tree ⇒ 'a set
set_tree () = {}
set_tree (l, x, r) = set_tree l ∪ {x} ∪ set_tree r

map_tree :: ('a ⇒ 'b) ⇒ 'a tree ⇒ 'b tree
map_tree f () = ()
map_tree f (l, x, r) = (map_tree f l, f x, map_tree f r)

The inorder, preorder and postorder traversals (we omit the latter) list the elements in a tree in a particular order:

inorder :: 'a tree ⇒ 'a list
inorder () = []
inorder (l, x, r) = inorder l @ [x] @ inorder r

preorder :: 'a tree ⇒ 'a list
preorder () = []
preorder (l, x, r) = x # preorder l @ preorder r

These two size functions count the number of Nodes and Leafs in a tree:

size :: 'a tree ⇒ nat
|()| = 0
|⟨l, __, r⟩| = |l| + |r| + 1

sizel :: 'a tree ⇒ nat
|()|₁ = 1
|⟨l, __, r⟩|₁ = |l|₁ + |r|₁

The syntactic sugar $|t|$ for size $t$ and $|t|₁$ for sizel $t$ is only used in this text, not in the Isabelle theories.

Induction proves a convenient fact that explains the name sizel:

$|t|₁ = |t| + 1$

The height and the minimal height of a tree are defined as follows:
4.2 Complete Trees

A complete tree is one where all the leaves are on the same level. An example is shown in Figure 4.1. The predicate complete is defined recursively:

Exercise 4.1. Function inorder has quadratic complexity because the running time of (@) is linear in the length of its first argument. Define a function inorder2 :: 'a tree ⇒ 'a list ⇒ 'a list that avoids (@) but accumulates the result in its second parameter via (#) only. Its running time should be linear in the size of the tree. Prove inorder2 t xs = inorder t @ xs.

You can think of them as the longest and shortest (cycle-free) path from the root to a leaf. The real names of these functions are height and min_height. The abbreviations h and mh are only used in this text, not in the Isabelle theories.

The obvious properties \( h t \leq |t| \) and \( mh t \leq h t \) and the following classical properties have easy inductive proofs:

\[
2^{mh t} \leq |t|_1 \quad |t|_1 \leq 2^{h t}
\]

We will simply use these fundamental properties without referring to them by a name or number.

The set of subtrees of a tree is defined as follows:

Exercise 4.1. Function inorder has quadratic complexity because the running time of (@) is linear in the length of its first argument. Define a function inorder2 :: 'a tree ⇒ 'a list ⇒ 'a list that avoids (@) but accumulates the result in its second parameter via (#) only. Its running time should be linear in the size of the tree. Prove inorder2 t xs = inorder t @ xs.

A complete tree is one where all the leaves are on the same level. An example is shown in Figure 4.1. The predicate complete is defined recursively:

\[
\text{height} :: \text{'}a\text{ tree }\Rightarrow\text{ nat}
\]

\[
h \langle \rangle = 0
\]

\[
h \langle l, _, r \rangle = \max (h l) (h r) + 1
\]

\[
\text{min_height} :: \text{'}a\text{ tree }\Rightarrow\text{ nat}
\]

\[
mh \langle \rangle = 0
\]

\[
mh \langle l, _, r \rangle = \min (mh l) (mh r) + 1
\]

Note that every tree is a subtree of itself.

Exercise 4.1. Function inorder has quadratic complexity because the running time of (@) is linear in the length of its first argument. Define a function inorder2 :: 'a tree ⇒ 'a list ⇒ 'a list that avoids (@) but accumulates the result in its second parameter via (#) only. Its running time should be linear in the size of the tree. Prove inorder2 t xs = inorder t @ xs.

4.2 Complete Trees

A complete tree is one where all the leaves are on the same level. An example is shown in Figure 4.1. The predicate complete is defined recursively:

\[
\text{height} :: \text{'}a\text{ tree }\Rightarrow\text{ nat}
\]

\[
h \langle \rangle = 0
\]

\[
h \langle l, _, r \rangle = \max (h l) (h r) + 1
\]

\[
\text{min_height} :: \text{'}a\text{ tree }\Rightarrow\text{ nat}
\]

\[
mh \langle \rangle = 0
\]

\[
mh \langle l, _, r \rangle = \min (mh l) (mh r) + 1
\]
This recursive definition is equivalent with the above definition that all leaves must have the same distance from the root. Formally:

**Lemma 4.1.** `complete t` $\iff$ $mh t = h t$

*Proof.* The proof is by induction and case analyses on $min$ and $max$. $\Box$

The following classic property of complete trees is easily proved by induction:

**Lemma 4.2.** `complete t` $\rightarrow$ $|t|_1 = 2^h t$

It turns out below that this is in fact a defining property of complete trees.

For complete trees we have $(2::')^mh t \leq |t|_1 = 2^h t$. For incomplete trees the equalities the $=$ sign becomes a $<$ as the following two lemmas prove:

**Lemma 4.3.** $\neg complete t$ $\rightarrow$ $|t|_1 < 2^h t$

*Proof.* The proof is by induction. We focus in the induction step where $t = (l, x, r)$. If $t$ is incomplete, there are a number of cases and we prove $|t|_1 < 2^h t$ in each case. If $h l \neq h r$, consider the case $h l < h r$ (the case $h r < h l$ is symmetric). From $2^h l < 2^h r$, $|l|_1 \leq 2^h l$ and $|r|_1 \leq 2^h r$ the claim follows: $|t|_1 = |l|_1 + |r|_1 \leq 2^h l + 2^h r < 2 \cdot 2^h r = 2^h t$. If $h l = h r$, then either $l$ or $r$ must be incomplete. We consider the case $\neg complete l$ (the case $\neg complete r$ is symmetric). From the IH $|l|_1 < 2^h l$, $|r|_1 \leq 2^h r$ and $h l = h r$ the claim follows: $|t|_1 = |l|_1 + |r|_1 < 2^h l + 2^h r = 2 \cdot 2^h r = 2^h t$. $\Box$

**Lemma 4.4.** $\neg complete t$ $\rightarrow$ $2mh t < |t|_1$

The proof of this lemma is completely analogous to the previous proof except that one also needs to use Lemma 4.1.

From the contrapositive of Lemma 4.3 one obtains $|t|_1 = 2^h t$ $\rightarrow$ `complete t`, the converse of Lemma 4.2. Thus we arrive at:
Corollary 4.5. \( \text{complete } t \iff |t|_1 = 2^h t \)

The complete trees are precisely the ones where the height is exactly the logarithm of the number of leaves.

Exercise 4.2. Define a function \( \text{mcs} \) that computes a maximal complete subtree of some given tree. You are allowed only one traversal of the input but you may freely compute the height of trees and may even compare trees for equality. You are not allowed to use \text{complete} or \text{subtrees}.

Prove that \( \text{mcs} \) returns a complete subtree (which should be easy) and that it is maximal in height:

\[ u \in \text{subtrees } t \land \text{complete } u \rightarrow h u \leq h (\text{mcs } t) \]

Bonus: get rid of any tree equality tests in \( \text{mcs} \).

4.3 Almost Complete Trees

An almost complete tree is one where the leaves may occur not just at the lowest level but also one level above:

\[
\text{acomplete :: } \forall a \text{ tree } \Rightarrow \text{ bool }
\]
\[
\text{acomplete } t = (h t - mh t \leq 1)
\]

An example of an almost complete tree is shown in Figure 4.2. You can think of an almost complete tree as a complete tree with (possibly) some additional nodes one level below the last full level.

![Almost complete tree](image)

Fig. 4.2. Almost complete tree

Almost complete trees are important because among all the trees with the same number of nodes they have minimal height:
Lemma 4.6. $\text{acomp}utes s \land |s| \leq |t| \rightarrow h_s \leq h_t$

Proof. The proof is by cases. If $\text{comp}utes s$ then, by Lemma 4.2, $2^h s = |s| \leq |t| \leq 2^h t$ and thus $h_s \leq h_t$. Now assume $\neg \text{comp}utes s$. Then Lemma 4.4 yields $2^{mh} s < |s| \leq |t| \leq 2^h t$ and thus $mh s < h_t$. Furthermore we have $h s = mh s \leq 1$ (from $\text{acomp}utes s$), $h s \neq mh s$ (from Lemma 4.1) and $mh s \leq h s$, which together imply $mh s + 1 = h s$. With $mh s < h_t$ this implies $h_s \leq h_t$.

This is relevant for search trees because their height determines the worst case running time. Almost complete trees are optimal in that sense.

The following lemma yields an explicit formula for the height of an almost complete tree:

Lemma 4.7. $\text{acomp}utes t \rightarrow h_t = \lfloor \log |t| \rfloor$

Proof. The proof is by cases. If $t$ is complete, the claim follows from Lemma 4.2. Now assume $t$ is incomplete. Then $h_t = mh t + 1$ because $\text{acomp}utes t$, $mh t \leq h_t$ and $\text{comp}utes t \leftrightarrow mh t = h t$ (Lemma 4.1). Together with $|t| \leq 2^h t$ this yields $|t| \leq 2^{mh t + 1}$ and thus $\log |t| \leq mh t + 1$. By Lemma 4.4 we obtain $mh t < \log |t|$. These two bounds for $\log |t|$ together imply the claimed $h_t = \lfloor \log |t| \rfloor$.

In the same manner we also obtain:

Lemma 4.8. $\text{acomp}utes t \rightarrow mh t = \lfloor \log |t| \rfloor$

4.3.1 Converting a List into an Almost Complete Tree

We will now see how to convert a list $xs$ into an almost complete tree $t$ such that $\text{inorder } t = xs$. If the list is sorted, the result is an almost complete binary search tree (see the next chapter). The basic idea is to cut the list in two halves, turn them into almost complete trees recursively and combine them. But cutting up the list in two halves explicitly would lead to an $n \cdot \log n$ algorithm but we want a linear one. Therefore we use an additional $\text{nat}$ parameter to tell us how much of the input list should be turned into a tree.

The remaining list is returned with the tree:

\[
\begin{align*}
\text{bal} :: \text{nat} & \Rightarrow \text{a list} \Rightarrow \text{a tree} \times \text{a list} \\
\text{bal} n xs &= (\text{if } n = 0 \text{ then } ((), xs) \\
\text{else let } m &= n \div 2; \\
(l, ys) &= \text{bal } m \\ & \quad \text{xs};
\end{align*}
\]
\[(r, zs) = bal (n - 1 - m) (tl ys) \]
\[in ((l, hd ys, r), zs))\]

The trick is not to chop \(xs\) in half but \(n\) because we assume that arithmetic is constant-time. Hence \(bal\) runs in linear time (see Exercise 4.4). Figure 4.3 shows the result of \(bal 10 \ [0..9]\).

Fig. 4.3. Balancing \([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]\)

Balancing some prefix or all of a list or tree is easily derived:

\[
bal_list :: nat \Rightarrow 'a list \Rightarrow 'a tree
bal_list n xs = fst (bal n xs)
\]

\[
balance_list :: 'a list \Rightarrow 'a tree
balance_list xs = bal_list |xs| xs
\]

\[
bal_tree :: nat \Rightarrow 'a tree \Rightarrow 'a tree
bal_tree n t = bal_list n (inorder t)
\]

\[
balance_tree :: 'a tree \Rightarrow 'a tree
balance_tree t = bal_tree |t| t
\]

Correctness

The following lemma clearly expresses that \(bal n xs\) turns the prefix of length \(n\) of \(xs\) into a tree and returns the corresponding suffix of \(xs\):
Lemma 4.9. \( n \leq |xs| \land bal n xs = (t, zs) \rightarrow xs = inorder t @ zs \land |t| = n \)

Proof by complete induction on \( n \), assuming that the property holds for all values below \( n \). If \( n = 0 \) the claim is trivial. Now assume \( n \neq 0 \) and let \( m = n \div 2 \) and \( m' = n - 1 - m \) (and thus \( m, m' < n \)). From \( bal n zs = (t, zs) \) we obtain \( l, r \) and \( ys \) such that \( bal m xs = (l, ys), bal m' (tl ys) = (r, zs) \) and \( t = \langle l, hd ys, r \rangle \). Because \( m < n \leq |xs| \) the induction hypothesis implies \( xs = inorder l @ ys \land |l| = m \) (1). This in turn implies \( m' \leq |tl ys| \) and thus the induction hypothesis implies \( tl ys = inorder r @ zs \land |r| = m' \) (2). Properties (1) and (2) together with \( t = \langle l, hd ys, r \rangle \) imply the claim \( xs = inorder t @ zs \land |t| = n \) because \( ys \neq [] \).

The corresponding correctness properties of the derived functions are easy consequences:

\[
\begin{align*}
\text{Lemma 4.10. } n \leq |xs| \land bal n xs = (t, zs) & \rightarrow h t = \lfloor \log (n + 1) \rfloor \\
\text{Proof. } & \text{The proof structure is the same as for Lemma 4.9 and we reuse the variable names introduced there. In the induction step we obtain the simplified} \\
& \text{induction hypothesis } h l = \lfloor \log (m + 1) \rfloor \text{ and } h r = \lfloor \log (m' + 1) \rfloor. \text{ This} \\
& \text{leads to} \\
& h t = \max (h l) (h r) + 1 \\
& = h l + 1 \quad \text{because } m' \leq m \\
& = \lfloor \log (m + 1) + 1 \rfloor \\
& = \lfloor \log (n + 1) \rfloor \quad \text{by (2.29)} \quad \Box 
\end{align*}
\]

The following complementary lemma is proved in the same way:

Lemma 4.11. \( n \leq |xs| \land bal n xs = (t, zs) \rightarrow mh t = \lfloor \log (n + 1) \rfloor \)

By definition of \( acomplete \) and because \( |x| - |y| \leq 1 \) we obtain that \( bal \) (and consequently the functions that build on it) returns an almost complete tree:

Corollary 4.12. \( n \leq |xs| \land bal n xs = (t, ys) \rightarrow acomplete t \)
4.3.2 Exercises

Exercise 4.3. Find a formula $B$ such that \( \text{acomplete} \langle l, x, r \rangle = B \) where $B$ may only contain the functions \( \text{acomplete}, \text{complete}, \text{height} \), arithmetic and boolean operations, $l$ and $r$, but in particular not \( \text{min\_height} \) or \( \text{Node} \) (\( = \langle _, _, _ \rangle \)). Prove \( \text{acomplete} \langle l, x, r \rangle = B \).

Exercise 4.4. Prove that the running time of function \( \text{bal} \) is linear in its first argument. (Isabelle hint: you need to refer to \( \text{bal} \) as \( \text{Balance.bal} \)).

4.4 Augmented Trees

A tree of type \( 'a \text{ tree} \) only stores elements of type \( 'a \). However, it is frequently necessary to store some additional information of type \( 'b \) in each node too, usually for efficiency reasons. Typical examples are:

- The size or the height of the tree. Because recomputing them requires traversing the whole tree.
- Lookup tables where each key of type \( 'a \) is associated with a value of type \( 'b \).

In this case we simply work with trees of type \( ('a \times 'b) \text{ tree} \) and call them augmented trees. As a result we need to redefine a few functions that should ignore the additional information. For example, function \( \text{inorder} \), when applied to an augmented tree, should return an \( 'a \text{ list} \). Thus we redefine it in the obvious way:

\[
\text{inorder} :: ('a \times 'b) \text{ tree} \Rightarrow 'a \text{ list} \\
\text{inorder} () = [] \\
\text{inorder} \langle l, (a, _), r \rangle = \text{inorder} l @ a \# \text{inorder} r
\]

Another example is \( \text{set\_tree} :: ('a \times 'b) \text{ tree} \Rightarrow 'a \text{ set} \). In general, if a function \( f \) is originally defined on type \( 'a \text{ tree} \) but should ignore the \( 'b \)-values in an \( ('a \times 'b) \text{ tree} \) then we assume that there is a corresponding revised definition of \( f \) on augmented trees that focuses on the \( 'a \)-values just like \( \text{inorder} \) above does. Of course functions that do not depend on the information in the nodes, e.g. size and height, stay unchanged.

Note that there are two alternative redefinitions of \( \text{inorder} \) (and similar functions): \( \text{map \_fst} \circ \text{Tree.inorder} \) or \( \text{Tree.inorder} \circ \text{map\_tree \_fst} \) where \( \text{Tree.inorder} \) is the original function.
4.4.1 Maintaining Augmented Trees

Maintaining the ‘b-values in a tree can be hidden inside a suitable smart version of Node that has only a constant time overhead. Take the example of augmentation by size:

\[
\text{sz} :: (\text{'a} \times \text{nat}) \text{ tree} \Rightarrow \text{nat}
\]
\[
\text{sz} \langle \rangle = 0
\]
\[
\text{sz} \langle \_ , (\_ , n) , \_ \rangle = n
\]

\[
\text{node\_sz} :: (\text{'a} \times \text{nat}) \text{ tree} \Rightarrow (\text{'a} \times \text{nat}) \text{ tree} \Rightarrow (\text{'a} \times \text{nat}) \text{ tree}
\]
\[
\text{node\_sz} \ l \ a \ r = \langle l , (a , \text{sz} \ l + \text{sz} \ r + 1) , r \rangle
\]

A (\text{'a} \times \text{nat}) \text{ tree} satisfies \text{invar\_sz} if the size annotation of every node is computed from its children as specified in \text{node\_sz}:

\[
\text{invar\_sz} :: (\text{'a} \times \text{nat}) \text{ tree} \Rightarrow \text{bool}
\]
\[
\text{invar\_sz} \langle \rangle = \text{True}
\]
\[
\text{invar\_sz} \langle \_ , (\_ , n) , r \rangle
\]
\[
= (\text{n} = \text{sz} \ l + \text{sz} \ r + 1 \land \text{invar\_sz} \ l \land \text{invar\_sz} \ r)
\]

This property is preserved by \text{node\_sz} (i.e. \text{invar\_sz} \ l \land \text{invar\_sz} \ r \rightarrow \text{invar\_sz} \ (\text{node\_sz} \ l \ a \ r)) and it guarantees that \text{sz} returns the size:

\[
\text{invar\_sz} \ t \rightarrow \text{sz} \ t = |t|
\]

We can generalize this example easily. Assume we have a constant \text{zero} :: \text{'b} and a function \text{f} :: \text{'b} \Rightarrow \text{'a} \Rightarrow \text{'b} \Rightarrow \text{'b} which we iterate over the tree:

\[
\text{F} :: (\text{'a} \times \text{'b}) \text{ tree} \Rightarrow \text{'b}
\]
\[
\text{F} \langle \rangle = \text{zero}
\]
\[
\text{F} \langle l , (a , \_ ) , r \rangle = \text{f} \ (\text{F} \ l) \ a \ (\text{F} \ r)
\]

This generalizes the definition of size. Let \text{node\_f} compute the \text{'b}-value from the \text{'b}-value of its children via \text{f}:

\[
\text{b\_val} :: (\text{'a} \times \text{'b}) \text{ tree} \Rightarrow \text{'b}
\]
\[
\text{b\_val} \langle \rangle = \text{zero}
\]
4.4 Augmented Trees

\[ b\_val (\_, (\_, b), \_) = b \]

\[ \text{node}_f :: \ ('a \times 'b) \ \text{tree} \Rightarrow 'a \Rightarrow ('a \times 'b) \ \text{tree} \Rightarrow ('a \times 'b) \ \text{tree} \]

\[ \text{node}_f l a r = \langle l, (a, f (b\_val l) a (b\_val r)), r \rangle \]

If all \( 'b\)-values are computed as in \text{node}_f

\[
\begin{align*}
\text{invar}_f :: (&'a \times 'b) \ \text{tree} & \Rightarrow \text{bool} \\
\text{invar}_f \langle \rangle & = \text{True} \\
\text{invar}_f (l, (a, b), r) & = (b = f (b\_val l) a (b\_val r) \land \text{invar}_f l \land \text{invar}_f r)
\end{align*}
\]

then all \( 'b\)-values equal \( F \): \text{invar}_f t \rightarrow b\_val t = F t. 

4.4.2 Exercises

Exercise 4.5. Augment trees by a pair of a boolean and something else where the boolean indicates whether the tree is complete or not. Define \text{ch}, \text{node}_\text{ch} and \text{invar}_\text{ch} as in Section 4.4.1 and prove the following properties:

\[
\begin{align*}
\text{invar}_\text{ch} t & \rightarrow \text{ch} t = (\text{complete} t, ?) \\
\text{invar}_\text{ch} l \land \text{invar}_\text{ch} r & \rightarrow \text{invar}_\text{ch} (\text{node}_\text{ch} l a r)
\end{align*}
\]

Exercise 4.6. Assume type \( 'a \) is of class \text{linorder} and augment each \text{Node} with the maximum value in that tree. Following Section 4.4.1 (but mind the \text{option} type!) define \text{mx} :: ('a \times 'b) \ \text{tree} \Rightarrow 'b \ \text{option}, \text{node}_\text{mx} and \text{invar}_\text{mx} and prove

\[
\begin{align*}
\text{invar}_\text{mx} t & \rightarrow \\
\text{mx} t & = \langle \text{if} t = \langle \rangle \text{ then } \text{None else } \text{Some} (\text{Max (set_tree} t)) \rangle
\end{align*}
\]

where \text{Max} is the predefined maximum operator on finite, non-empty sets.
The purpose of this chapter is threefold: to introduce binary search trees (BSTs), to discuss their correctness proofs, and to provide a first example of an abstract data type, a notion discussed in more detail in the next chapter.

Search trees are a means for storing and accessing collections of elements efficiently. In particular they can support sets and maps. We concentrate on sets. We have already seen function `set_tree` that maps a tree to the set of its elements. This is an example of an abstraction function that maps concrete data structures to the abstract values that they represent.

BSTs require a linear ordering on the elements in the tree (as in Chapter Sorting). For each node, the elements in the left child are smaller than the root and the elements in the right child are bigger:

\[
\text{bst} :: (\text{linorder}) \rightarrow \text{bool} \\
\text{bst} \langle \langle \rangle \rangle = \text{True} \\
\text{bst} \langle l, a, r \rangle = (\forall x \in \text{set_tree} l. x < a) \land (\forall x \in \text{set_tree} r. a < x) \land \text{bst} l \land \text{bst} r
\]

This is an example of a (coincidentally almost complete) BST:

![Binary Search Tree Diagram]

It is obvious how to search for an element in a BST by comparing the element with the root and descending into one of the two children if you have
not found it yet. In the worst case this takes time proportional to the height of the tree. In later chapters we discuss a number of methods for ensuring that the height of the tree is logarithmic in its size. For now we ignore all efficiency considerations and permit our BSTs to degenerate.

Exercise 5.1. The above recursive definition of \( bst \) is not a direct translation of the description “For each node” given in the text. For a more direct translation define a function

\[
\text{nodes} :: 'a \text{ tree} \Rightarrow (\langle 'a \times 'a \times 'a \text{ tree} \rangle) \text{ set}
\]

that collects all the nodes as triples \((l, a, r)\). Now define \( \text{bst}_\text{nodes} \) as \( \text{bst}_\text{nodes} t = (\forall (l,a,r) \in \text{nodes} t. \ldots) \) and prove \( \text{bst}_\text{nodes} t = \text{bst} t \).

5.1 Interfaces

Trees are concrete data types that provide the building blocks for realizing abstract data types like sets. The abstract type has a fixed interface, i.e. set of operations, through which the values of the abstract type can be manipulated. The interface hides all implementation detail. In the Search Tree part of the book we focus on the abstract type of sets with the following interface:

\[
\begin{align*}
\text{empty} & :: 's \\
\text{insert} & :: 'a \Rightarrow 's \Rightarrow 's \\
\text{delete} & :: 'a \Rightarrow 's \Rightarrow 's \\
\text{isin} & :: 's \Rightarrow 'a \Rightarrow \text{bool}
\end{align*}
\]

where \( 's \) is the type of sets of elements of type \( 'a \). Most of our implementations of sets will be based on variants of BSTs and will require a linear order on \( 'a \), but the general interface does not require this. The correctness of an implementation of this interface will be proved by relating it back to HOL’s type \( 'a \text{ set} \) via an abstraction function, e.g. \( \text{set\_tree} \).

5.2 Implementing Sets via unbalanced BSTs

So far we have compared elements via \( =, \leq \) and \( < \). Now we switch to a comparator-based approach:

\[
\text{datatype cmp\_val} = LT \mid EQ \mid GT
\]

\[
\text{cmp} :: ('a::linorder) \Rightarrow 'a \Rightarrow \text{cmp\_val}
\]

\[
cmp x y = \begin{cases} 
LT & \text{if } x < y \\
EQ & \text{if } x = y \\
GT & \text{if } x > y
\end{cases}
\]
We will frequently phrase algorithms in terms of \( \text{cmp} \), \( LT \), \( EQ \) and \( GT \) instead of \(<\), \( =\) and \( >\). This leads to more symmetric code. If some type comes with its own primitive \( \text{cmp} \) function this can yield a speed-up over the above generic \( \text{cmp} \) function.

Below you find an implementation of the set interface in terms of BSTs. Functions \( \text{isin} \) and \( \text{insert} \) are self-explanatory. Deletion is more interesting.

```haskell
empty :: 'a tree
empty = ()

isin :: 'a tree \( \Rightarrow \) 'a \( \Rightarrow \) bool
isin () _ = False
isin (l, a, r) x
= (case \( \text{cmp} \) x a of \( LT \) \\Rightarrow \) isin l x | \( EQ \) \Rightarrow \) True | \( GT \) \Rightarrow \) isin r x)

insert :: 'a \( \Rightarrow \) 'a tree \( \Rightarrow \) 'a tree
insert x () = ((), x, ())
insert x (l, a, r) = (case \( \text{cmp} \) x a of
  \( LT \) \Rightarrow (insert x l, a, r) |
  \( EQ \) \Rightarrow (l, a, r) |
  \( GT \) \Rightarrow (l, a, insert x r))

delete :: 'a \( \Rightarrow \) 'a tree \( \Rightarrow \) 'a tree
delete x () = ()
delete x (l, a, r)
= (case \( \text{cmp} \) x a of
  \( LT \) \Rightarrow (delete x l, a, r) |
  \( EQ \) \Rightarrow if r = () then l
  else let (a', r') = split_min r in (l, a', r') |
  \( GT \) \Rightarrow (l, a, delete x r))

split_min :: 'a tree \( \Rightarrow \) 'a \times 'a tree
split_min (l, a, r)
= (if l = () then (a, r) else let (x, l') = split_min l in (x, (l', a, r)))
```

5.2.1 Deletion

Function \( \text{delete} \) deletes \( a \) from \( (l, a, r) \) (where \( r \neq () \)) by replacing \( a \) with \( a' \) and \( r \) by \( r' \) where
$a'$ is the leftmost (least) element of $r$, also called the inorder successor of $a$, $r'$ is the remainder of $r$ after removing $a'$.

We call this deletion by replacing. Of course one can also obtain $a'$ as the inorder predecessor of $a$ in $l$.

An alternative is to delete $a$ from $(l, a, r)$ by “joining” $l$ and $r$:

\[
delete2 :: 'a ⇒ 'a tree ⇒ 'a tree
\]
\[
delete2 _ _ () = ()
delete2 _ _ (x, a, r) = (\text{case } \text{cmp } x a \text{ of}
LT ⇒ (\delete2 _ _ x, a, r) |
EQ ⇒ join _ _ l r |
GT ⇒ (l, a, delete2 _ _ r))
\]

\[
join :: 'a tree ⇒ 'a tree ⇒ 'a tree
\]
\[
join t () = t
join () t = t
join (t1, a, t2) (t3, b, t4) = (\text{case } join _ _ t2 _ _ of
⟨⟩ ⇒ (t1, a, ⟨⟩, b, t4) |
⟨u2, x, u3⟩ ⇒ (⟨t1, a, u2⟩, x, ⟨u3, b, t4⟩))
\]

We call this deletion by joining. The characteristic property of $join$ is $\text{inorder } (\text{join } l r) = \text{inorder } l \mathbin{@} \text{inorder } r$.

The definition of $join$ may appear needlessly complicated. Why not this much simpler version:

\[
join0 :: 'a tree ⇒ 'a tree ⇒ 'a tree
\]
\[
join0 t () = t
join0 () t = t
join0 (t1, a, t2) (t3, b, t4) = (t1, a, ⟨join0 _ _ t2 _ _⟩, b, t4)
\]

Because with this version of $join$, deletion may almost double the height of the tree, in contrast to $join$ and also deletion by replacing, where the height cannot increase:

**Exercise 5.2.** First prove that $join$ behaves well:

\[
h (\text{join } l r) ≤ \max (h l) (h r) + 1
\]

Now show that $join0$ behaves badly: find an upper bound $ub$ of $h (\text{join0 } l r)$ such that $ub$ is a function of $h l$ and $h r$. Prove $h (\text{join0 } l r) ≤ ub$ and prove that $ub$ is a tight upper bound if $l$ and $r$ are complete trees.

We focus on $delete$, deletion by replacing, in the rest of the chapter.
5.3 Correctness

Why is the above implementation correct? Roughly speaking, because the implementations of `empty`, `insert`, `delete` and `isin` on type `'a tree` simulate the behaviour of `{}`, `∪`, `−` and `∈` on type `'a set`. Taking the abstraction function into account we can formulate the simulation precisely:

\[
\begin{align*}
\text{set_tree} \; \text{empty} &= \emptyset \\
\text{set_tree} \; (\text{insert} \; x \; t) &= \text{set_tree} \; t \cup \{x\} \\
\text{set_tree} \; (\text{delete} \; x \; t) &= \text{set_tree} \; t \setminus \{x\} \\
\text{isin} \; t \; x &= (x \in \text{set_tree} \; t)
\end{align*}
\]

However, the implementation only works correctly on BSTs. Therefore we need to add the precondition `bst t` to all but the first property. But why are we permitted to assume this precondition? Only because `bst` is an invariant of this implementation: `bst` holds for `empty`, and both `insert` and `delete` preserve `bst`. Therefore every tree that can be manufactured through the interface is a BST. Of course this adds another set of proof obligations for correctness, invariant preservation:

\[
\begin{align*}
\text{bst} \; \text{empty} \\
\text{bst} \; t \quad \text{⇒} \quad \text{bst} \; (\text{insert} \; x \; t) \\
\text{bst} \; t \quad \text{⇒} \quad \text{bst} \; (\text{delete} \; x \; t)
\end{align*}
\]

When looking at the abstract data type of sets from the user (or 'client') perspective, we would call the collection of all proof obligations for the correctness of an implementation the specification of the abstract type.

Exercise 5.3. Verify the implementation in Section 5.2 by showing all the proof obligations above, without the detour via sorted lists explained below.

5.4 Correctness Proofs

It turns out that direct proofs of the properties in the previous section are cumbersome — at least for `delete` and for proof assistants like Isabelle. Yet the correctness of the implementation is quite obvious to most (functional) programmers. Which is why most algorithm texts do not spend any time on functional correctness of search trees and concentrate on non-obvious structural properties that imply the logarithmic height of the trees — of course our simple BSTs do not guarantee the latter.

We will now present how the vague notion of “obvious” can be concretized and automated to such a degree that we do not need to discuss functional correctness of search tree implementations again in this book. This is because our
5 Binary Search Trees

approach is quite generic: it works not only for the BSTs in this chapter but also for the more efficient variants discussed in later chapters. The remainder of this section can be skipped if one is not interested in proof automation.

5.4.1 The Idea

The key idea [54] is to express bst and set_tree via inorder:

\[
\begin{align*}
\text{bst } t &= \text{sorted } (\text{inorder } t) \quad \text{and} \quad \text{set_tree } t = \text{set } (\text{inorder } t)
\end{align*}
\]

where

\[
\begin{align*}
sorted :: \text{'}a \text{ list } \Rightarrow \text{ bool} \\
sorted [] &= \text{True} \\
sorted [\_] &= \text{True} \\
sorted (x \# y \# zs) &= (x < y \land \text{sorted } (y \# zs))
\end{align*}
\]

Note that this is “sorted w.r.t. (\text{<})” whereas in the chapter on sorting sorted was defined as “sorted w.r.t. (\leq)”. Instead of showing directly that BSTs implement sets, we show that they implement an intermediate specification based on lists (and later that the list-based specification implies the set-based one). We can assume that the lists are sorted because they are abstractions of BSTs. Insertion and deletion on sorted lists can be defined as follows:

\[
\begin{align*}
\text{ins_list } :: \text{'}a \Rightarrow \text{'}a \text{ list } \Rightarrow \text{'}a \text{ list} \\
\text{ins_list } x [] &= [x] \\
\text{ins_list } x (a \# xs) &= \begin{cases} 
\text{if } x < a \text{ then } x \# a \# xs \\
\text{else if } x = a \text{ then } a \# xs \\
\text{else } a \# \text{ins_list } x xs 
\end{cases}
\end{align*}
\]

\[
\begin{align*}
\text{del_list } :: \text{'}a \Rightarrow \text{'}a \text{ list } \Rightarrow \text{'}a \text{ list} \\
\text{del_list } _[] &= [] \\
\text{del_list } x (a \# xs) &= \begin{cases} 
\text{if } x = a \text{ then } xs \\
\text{else } a \# \text{del_list } x xs 
\end{cases}
\end{align*}
\]

The abstraction function from trees to lists is function inorder. The specification in Figure 5.1 expresses that empty, insert, delete and isin implement [], ins_list, del_list and \( \lambda xs \ x . \ x \in \text{set } xs \). One nice aspect of this specification is that it does not require us to prove invariant preservation explicitly: it follows from the fact (proved below) that ins_list and del_list preserve sorted.
inorder empty = []
sorted (inorder t) \rightarrow \text{inorder} (insert x t) = \text{ins_list} x (\text{inorder} t)
sorted (inorder t) \rightarrow \text{inorder} (delete x t) = \text{del_list} x (\text{inorder} t)
bst t \rightarrow \text{isin} t x = (x \in \text{set_tree} t)

Fig. 5.1. List-based Specification of BSTs

5.4.2 BSTs Implement Sorted Lists — A Framework

We present a library of lemmas that automate the functional correctness proofs for the BSTs in this chapter and the more efficient variants in later chapters. This library is motivated by general considerations concerning the shape of formulas that arise during verification.

As a motivating example we examine how to prove
\[ \text{sorted} (\text{inorder} t) \rightarrow \text{inorder} (\text{insert} x t) = \text{ins_list} x (\text{inorder} t) \]
The proof is by induction on \(t\) and we consider the case \(t = \langle l, a, r \rangle\) such that \(x < a\). Ideally the proof looks like this:
\[
\begin{align*}
inorder (\text{insert} x t) &= \text{inorder} (\text{insert} x l) @ a \# \text{inorder} r \\
&= \text{ins_list} x (\text{inorder} l) @ a \# \text{inorder} r \\
&= \text{ins_list} x (\text{inorder} l @ a \# \text{inorder} r) = \text{ins_list} x t
\end{align*}
\]
The first and last step are by definition, the second step by induction hypothesis, and the third step by lemmas in Figure 5.2: (5.1) rewrites the assumption \(\text{sorted} (\text{inorder} t)\) to \(\text{sorted} (\text{inorder} l @ [a]) \land \text{sorted} (a \# \text{inorder} r)\), thus allowing (5.5) to rewrite the term \(\text{ins_list} x (\text{inorder} l @ a \# \text{inorder} r)\) to \(\text{ins_list} x (\text{inorder} l) @ a \# \text{inorder} r\).

The lemma library in Figure 5.2 helps to prove the properties in Figure 5.1. These proofs are by induction on \(t\) and lead to (possibly nested) tree constructor terms like \(\langle t_1, a_1, t_2 \rangle\), \(a_2, t_3 \rangle\) where the \(t_i\) and \(a_i\) are variables. Evaluating \(\text{inorder}\) of such a tree leads to a list of the following form:
\[ \text{inorder} t_1 @ a_1 \# \text{inorder} t_2 @ a_2 \# \ldots \# \text{inorder} t_n \]
Now we discuss the lemmas in Figure 5.2 that simplify the application of \(\text{sorted}, \text{ins_list}\) and \(\text{del_list}\) to such terms.

Terms of the form \(\text{sorted} (xs @ a_1 \# xs_2 @ a_2 \# \ldots \# xs_n)\) are decomposed into the following basic formulas
\[
\begin{align*}
\text{sorted} (xs @ [a]) &\quad (\text{simulating } \forall x \in \text{set} xs. x < a) \\
\text{sorted} (a \# xs) &\quad (\text{simulating } \forall x \in \text{set} xs. a < x) \\
a < b
\end{align*}
\]
by the rewrite rules (5.1)–(5.2). Lemmas (5.3)–(5.4) enable deductions from basic formulas.
58 5 Binary Search Trees

\[
\text{sorted } \left(\mathit{xs} \oplus \mathit{ys}\right) = \left(\text{sorted } \left(\mathit{xs} \oplus [y]\right) \land \text{sorted } \left(y \not\in \mathit{ys}\right)\right) \quad (5.1)
\]

\[
\text{sorted } \left(x \not\in \mathit{xs}\right) \iff \text{sorted } \mathit{xs}
\]

\[
\text{sorted } \left(\mathit{xs} \oplus [a]\right) \implies \text{ins_list } x \left(\mathit{xs} \oplus a \not\in \mathit{ys}\right) = \quad (5.5)
\]

\[
\begin{align*}
\text{if } x < a & \text{ then } \text{ins_list } x \left(\mathit{xs} \oplus a \not\in \mathit{ys}\right) \text{ else } \mathit{xs} \oplus \text{ins_list } x \left(a \not\in \mathit{ys}\right) \\
\text{sorted } \left(\mathit{xs} \oplus a \not\in \mathit{ys}\right) & \implies \text{del_list } x \left(\mathit{xs} \oplus a \not\in \mathit{ys}\right) = \\
\text{if } x < a & \text{ then } \text{del_list } x \left(\mathit{xs} \oplus a \not\in \mathit{ys}\right) \text{ else } \mathit{xs} \oplus \text{del_list } x \left(a \not\in \mathit{ys}\right) \\
\text{sorted } \left(\mathit{xs} \oplus [x]\right) & = (\text{sorted } \mathit{xs} \land (\forall y \in \mathit{set} \mathit{xs}. \ y < x))
\end{align*}
\quad (5.6)
\]

Fig. 5.2. Lemmas for sorted, ins_list, del_list

Terms of the form \(\text{ins_list } x \left(\mathit{xs}_1 \oplus a_1 \not\in \mathit{xs}_2 \oplus a_2 \not\in \ldots \not\in \mathit{xs}_n\right)\) are rewritten with equation (5.5) (and the defining equations for \(\text{ins_list}\)) to push \(\text{ins_list}\) inwards. Terms of the form \(\text{del_list } x \left(\mathit{xs}_1 \oplus a_1 \not\in \mathit{xs}_2 \oplus a_2 \not\in \ldots \not\in \mathit{xs}_n\right)\) are rewritten with equation (5.6) (and the defining equations for \(\text{del_list}\)) to push \(\text{del_list}\) inwards.

The \(\text{isin}\) property in Figure 5.1 can be proved with the help of (5.1), (5.7) and (5.8).

The lemmas in Figure 5.2 form the complete set of basic lemmas on which the automatic proofs of almost all search trees in the book rest; only splay trees (see Chapter 21) need additional lemmas.

5.4.3 Sorted Lists Implement Sets

It remains to be shown that the list-based specification (Figure 5.1) implies the set-based correctness properties in Section 5.3. Remember that \(\text{bst } t = \text{sorted } \left(\text{inorder } t\right)\). The correctness properties

\[
\text{set_tree empty} = \{
\}
\]

\[
\begin{align*}
\text{sorted } \left(\text{inorder } t\right) & \iff \text{set_tree } \left(\text{insert } x \ t\right) = \text{set_tree } t \cup \{x\} \\
\text{sorted } \left(\text{inorder } t\right) & \iff \text{set_tree } \left(\text{delete } x \ t\right) = \text{set_tree } t \setminus \{x\} \\
\text{sorted } \left(\text{inorder } t\right) & \iff \text{isin } t \ x = (x \in \text{set_tree } t)
\end{align*}
\]

are a consequence of \(\text{set_tree } t = \text{set } \left(\text{inorder } t\right)\) (a basic tree lemma), the properties in Figure 5.1 and the inductive correctness properties

\[
\begin{align*}
\text{set } \left(\text{ins_list } x \ \mathit{xs}\right) & = \text{set } \mathit{xs} \cup \{x\} \\
\text{sorted } \mathit{xs} & \iff \text{set } \left(\text{del_list } x \ \mathit{xs}\right) = \text{set } \mathit{xs} \setminus \{x\}
\end{align*}
\]
Preservation of the invariant

\[
\begin{align*}
\text{sorted (inorder empty)} \\
\text{sorted (inorder t) } \rightarrow \text{sorted (inorder (insert x t))} \\
\text{sorted (inorder t) } \rightarrow \text{sorted (inorder (delete x t))}
\end{align*}
\]

are trivial or consequences of the properties in Figure 5.1 and the preservation of \text{sorted} by \text{ins_list} and \text{del_list}:

\[
\begin{align*}
\text{sorted xs } \rightarrow \text{sorted (ins_list x xs)} \\
\text{sorted xs } \rightarrow \text{sorted (del_list x xs)}
\end{align*}
\]

This means in particular that preservation of \text{sorted} \circ \text{inorder} is guaranteed for any implementation of \text{empty}, \text{insert} and \text{delete} that satisfies the list-based specification in Figure 5.1.

### 5.5 Interval Trees

In this section we study binary trees for representing a set of intervals, called interval trees. In addition to the usual insertion and deletion functions of standard BSTs, interval trees support a function for determining whether a given interval overlaps with some interval in the tree.

#### 5.5.1 Augmented BSTs

The efficient implementation of the search for an overlapping interval relies on an additional piece of information in each node. Thus interval trees are another example of augmented trees as introduced in Section 4.4. We reuse the modified definitions of \text{set_tree} and \text{inorder} from that section. Moreover we use a slightly adjusted version of \text{isin}:

\[
\text{isin} :: (\text{a} \times \text{b}) \rightarrow \text{bool}
\]

\[
\text{isin} \text{'} = \text{False}
\]

\[
\text{isin}\ (l, (a, _), r) x = (\text{case } \text{cmp} x a \text{ of } \text{LT} \Rightarrow \text{isin} l x | \text{EQ} \Rightarrow \text{True} | \text{GT} \Rightarrow \text{isin} r x)
\]

This works for any kind of augmented BST, not just interval trees.

#### 5.5.2 Intervals

An interval \text{a ivl} is simply a pair of lower and upper bound, accessed by functions \text{low} and \text{high}, respectively. Intuitively, an interval represents the
closed set between \( low \) and \( high \). The standard mathematical notation is \([l, h]\) ,
the Isabelle notation is \( \{l..h\} \). We restrict ourselves to non-empty intervals:

\[ low \ p \leq high \ p \]

Type \( 'a \) can be any linearly ordered type with a minimum element \( \bot \) (for
example, the natural numbers or the real numbers extended with \(-\infty\)). In-
trvals can be linearly ordered by first comparing \( low \), then comparing \( high \).
The definitions are as follows:

\[
\begin{align*}
(x < y) &= (low x < low y \lor low x = low y \land high x < high y) \\
(x \leq y) &= (low x < low y \lor low x = low y \land high x \leq high y)
\end{align*}
\]

Two intervals overlap if they have at least one point in common:

\( overlap \ x \ y = (low y \leq high x \land low x \leq high y) \)

The readers should convince themselves that \( overlap \) does what it is supposed
to do: \( overlap \ x \ y = \{\{low x..high x\} \cap \{low y..high y\} \neq \{\}\} \)

We also define the concept of an interval overlapping with some interval
in a set:

\[ has\_overlap \ S \ y = (\exists x \in S. \ overlap \ x \ y) \]

### 5.5.3 Interval Trees

An interval tree associates to each node a number \( max\_hi \), which records the
maximum \( high \) value of all intervals in the subtrees. This value is updated
during insert and delete operations, and will be crucial for enabling efficient
determination of overlap with some interval in the tree.

\[
\begin{align*}
\text{type_s\_ynonym} \quad & 'a \ ivl\_tree = ('a \ ivl \times 'a) \ tree \\
max\_hi :: \quad & 'a \ ivl\_tree \Rightarrow 'a \\
\text{max\_hi} \ \bot &= \bot \\
\text{max\_hi} \ (\_, (\_, m), \_) &= m
\end{align*}
\]

If the \( max\_hi \) value of every node in a tree agrees with \( max3 \)
it follows by induction that $\text{max}_hi$ is the maximum value of $\text{high}$ in the tree and comes from some node in the tree:

**Lemma 5.1.** $\text{inv\_max\_hi} \ t \land \ a \in \text{set\_tree} \ t \rightarrow \text{high} \ a \leq \text{max}_hi \ t$

**Lemma 5.2.** $\text{inv\_max\_hi} \ t \land \ t \neq \emptyset \rightarrow \exists \ a \in \text{set\_tree} \ t. \ \text{high} \ a = \text{max}_hi \ t$

### 5.5.4 Implementing Sets of Intervals via Interval Trees

Interval trees can implement sets of intervals via unbalanced BSTs as in Section 5.2. Of course $\text{empty} = \emptyset$. Function $\text{isin}$ was already defined in Section 5.5.1 Insertion and deletion are also very close to the versions in Section 5.2, but the value of $\text{max}_hi$ must be computed (by $\text{max3}$) for each new node. We follow Section 4.4 and introduce a smart constructor $\text{node}$ for that purpose and replace $\langle l, a, r \rangle$ by $\text{node} \ l \ a \ r$ (on the right-hand side):

\[
\text{node} :: \text{`a ivl\_tree} \Rightarrow \text{`a ivl} \Rightarrow \text{`a ivl\_tree} \Rightarrow \text{`a ivl\_tree}
\]

\[
\text{node} \ l \ a \ r = \langle l, (a, \text{max3} \ a \ (\text{max}_hi \ l \ (\text{max}_hi \ r)) \ (\text{max}_hi \ r)), \ r \rangle
\]

\[
\text{insert} :: \text{`a ivl} \Rightarrow \text{`a ivl\_tree} \Rightarrow \text{`a ivl\_tree}
\]

\[
\text{insert} \ x \emptyset = \langle\emptyset, (x, \text{high} \ x), \emptyset\rangle
\]

\[
\text{insert} \ x \langle l, (a, m), r \rangle = \text{\textbf{case} \ cmp \ x \ a \ \textbf{of}}
\]

\[
\LT \Rightarrow \text{node} \ (\text{insert} \ x \ l) \ a \ r \ |
\]

\[
\EQ \Rightarrow \langle l, (a, m), r \rangle |
\]

\[
\GT \Rightarrow \text{node} \ l \ a \ (\text{insert} \ x \ r)\)
\]

\[
\text{split\_min} :: \text{`a ivl\_tree} \Rightarrow \text{`a ivl} \times \text{`a ivl\_tree}
\]

\[
\text{split\_min} \langle l, (a, \_), r \rangle
\]

\[
= \text{\textbf{if} } l = \langle\emptyset \rangle \text{ \textbf{then} } (a, r)
\]
else let (x, l') = split_min l in (x, node l' a r))

delete :: 'a ivl ⇒ 'a ivl_tree ⇒ 'a ivl_tree

delete _ () = ()
delete x (l, (a, _), r) = (case cmp x a of
LT ⇒ node (delete x l) a r |
EQ ⇒ if r = () then l else let (x, y) = split_min r in node l x y |
GT ⇒ node l a (delete x r))

The correctness proofs for insertion and deletion cover two aspects. Functional correctness and preservation of the invariant \textit{sorted} \circ \textit{inorder} (the BST property) are proved exactly as in Section 5.3 for ordinary BSTs. Preservation of the invariant \textit{inv_max_hi} can be proved by a sequence of simple inductive properties. In the end the main correctness properties are

\[
\text{sorted (inorder t)} \longrightarrow \text{inorder (insert x t)} = \text{ins_list x (inorder t)} \\
\text{sorted (inorder t)} \longrightarrow \text{inorder (delete x t)} = \text{del_list x (inorder t)} \\
\text{inv_max_hi t} \longrightarrow \text{inv_max_hi (insert x t)} \\
\text{inv_max_hi t} \longrightarrow \text{inv_max_hi (delete x t)}
\]

Defining \textit{invar} \( t = (\text{inv_max_hi} t \land \text{sorted (inorder t)}) \) we obtain the following top-level correctness corollaries:

\[
\text{invar s} \longrightarrow \text{set_tree (insert x s)} = \text{set_tree s} \cup \{x\} \\
\text{invar s} \longrightarrow \text{set_tree (delete x s)} = \text{set_tree s} - \{x\} \\
\text{invar s} \longrightarrow \text{invar (insert x s)} \\
\text{invar s} \longrightarrow \text{invar (delete x s)}
\]

The above insertion function allows overlapping intervals to be added into the tree and deletion supports only deletion of whole intervals. This is appropriate for the computational geometry application sketched below in Subsection 5.5.6. Other applications may require a different design.

5.5.5 Searching for an Overlapping Interval

The added functionality of interval trees over ordinary BSTs is function \textit{search} that searches for an overlapping rather than identical interval:
The following theorem expresses the correctness of \texttt{search} assuming the same invariants as before; \texttt{bst t} would work just as well as \texttt{sorted (inorder t)}.

\textbf{Theorem 5.3.} \texttt{inv_max_hi t \land sorted (inorder t) \rightarrow search t x = has_overlap (set_tree t) x}

\textbf{Proof.} The result is clear when \( t = \langle \rangle \). Now suppose \( t \) is in the form \( \langle l, (a, m), r \rangle \), where \( m \) is the value of \texttt{max_hi} at root. If \( a \) overlaps with \( x \), search returns \texttt{True} as expected. Otherwise, there are two cases.

\begin{itemize}
  \item If \( l \neq \langle \rangle \) and \( low x \leq max_hi l \), the search goes to the left child. If there is an interval in the left child overlapping with \( x \), then the search returns \texttt{True} as expected. Otherwise, we show there is also no interval in the right child overlapping with \( x \). Since \( l \neq \langle \rangle \), Lemma 5.2 yields a node \( p \) in the left child such that \( high p = max_hi l \). Since \( low x \leq max_hi l \), we have \( low x \leq high p \). Since \( p \) does not overlap with \( x \), we must have \( high x < low p \). But then, for every interval \( rp \) in the right child, \( low p \leq low rp \), so that \( high x < low rp \), which implies that \( rp \) does not overlap with \( x \).
  \item Now we consider the case where either \( l = \langle \rangle \) or \( max_hi l < low x \). In this case, the search goes to the right. We show there is no interval in the left child that overlaps with \( x \). This is clear if \( l = \langle \rangle \). Otherwise, for each interval \( lp \), we have \( high lp \leq max_hi l \) by Lemma 5.1, so that \( high lp < low x \), which means \( lp \) does not overlap with \( x \).
\end{itemize}

\textbf{Exercise 5.4.} Define a function that determines if a given point is in some interval in a given interval tree. Starting with

\texttt{in_ivl :: 'a ivl \Rightarrow bool}

\texttt{in_ivl x iv = (low iv \leq x \land x \leq high iv)}

write a recursive function

\texttt{search1 :: 'a ivl_tree \Rightarrow 'a \Rightarrow bool}

(without using \texttt{search}) such that \texttt{search1 x t} is \texttt{True} iff there is some interval \texttt{iv} in \( t \) such that \texttt{in_ivl x iv}. Prove

\texttt{inv_max_hi t \land bst t \rightarrow search1 t x = (\exists iv \in set_tree t. in_ivl x iv)}
5.5.6 Application

While this section demonstrated how to augment an ordinary binary tree with intervals, any of the balanced binary trees (such as red-black tree) can be augmented in a similar manner. We leave this as exercises.

Interval trees have many applications in computational geometry. A basic example is as follows. Consider a set of rectangles whose sides are aligned to the x and y-axes. We wish to efficiently determine whether any pair of rectangles in the set intersect each other (i.e. sharing a point, including boundaries). This can be done using a "sweep line" algorithm as follows. For each rectangle $[x_l, x_h] \times [y_l, y_h]$, we create two events: insert interval $[x_l, x_h]$ at y-coordinate $y_l$ and delete interval $[x_l, x_h]$ at y-coordinate $y_h$. Perform the events, starting from an empty interval tree, in ascending order of y-coordinates, with insertion events performed before deletion events. At each insertion, check whether the interval to be inserted overlaps with any of the existing intervals in the tree. If yes, we have found an intersection between two rectangles. If no overlap of intervals is detected throughout the process, then no pair of rectangles intersect. When using an interval tree based on a balanced binary tree, the time complexity of this procedure is $O(n \lg n)$, where $n$ is the number of rectangles.

We refer to Cormen et al. [13, Section 14.3] for another exposition on interval trees and their applications. Interval trees, together with the application of finding rectangle intersection, have been formalized by Zhan [81].
6

Abstract Data Types

In the previous chapter we looked at a very specific example of an abstract data type, namely sets. In this chapter we consider abstract data types in general and in particular the model-oriented approach to the specification of abstract data types. This will lead to a generic format for such specifications.

6.1 Abstract Data Types

Abstract data types (ADTs) can be summarized by the following slogan:

\[
\text{ADT} = \text{interface} + \text{specification}
\]

where the interface lists the operations supported by the ADT and the specification describes the behaviour of these operations. For example, our set ADT has the following interface:

- `empty :: 's`
- `insert :: 'a ⇒ 's ⇒ 's`
- `delete :: 'a ⇒ 's ⇒ 's`
- `isin :: 's ⇒ 'a ⇒ bool`

The purpose of an ADT is to be able to write applications based on this ADT that will work with any implementation of the ADT. To this end one can prove properties of the application that are solely based on the specification of the ADT. That is, one can write generic algorithms and prove generic correctness theorems about them in the context of the ADT specification.

6.2 Model-Oriented Specifications

We follow the model-oriented style of specification [35]. In that style, an abstract type is specified by giving an abstract model for it. For simplicity we
assume that each ADT describes one type of interest \( T \). In the set interface \( T \) is \( 's \). This type \( T \) must be specified by some existing HOL type \( A \), the abstract model. In the case of sets this is straightforward: the model for sets is simply the HOL type \( 'a \) set. The motto is that \( T \) should behave like \( A \). In order to bridge the gap between the two types, the specification needs an

- **abstraction function** \( \alpha :: T \Rightarrow A \)

that maps concrete values to their abstract counterparts. Moreover, in general only some elements of \( T \) represent elements of \( A \). For example, in the set implementation in the previous chapter not all trees but only BSTs represent sets. Thus the specification should also take into account an

- **invariant** \( \text{invar} :: T \Rightarrow \text{bool} \)

Note that the abstraction function and the invariant are not part of the interface, but they are essential for specification and verification purposes.

As an example, the ADT of sets is shown in Figure 6.1 with suggestive keywords and a fixed mnemonic naming schema for the labels in the specification. This is the template for ADTs that we follow throughout the book.

**ADT Set =**

**interface**

\[
\begin{aligned}
\text{empty} & :: \ 's \\
\text{insert} & :: \ 'a \Rightarrow \ 's \Rightarrow \ 's \\
\text{delete} & :: \ 'a \Rightarrow \ 's \Rightarrow \ 's \\
\text{isin} & :: \ 's \Rightarrow \ 'a \Rightarrow \text{bool}
\end{aligned}
\]

**abstraction** \( \text{set} :: \ 's \Rightarrow \ 'a \) set

**invariant** \( \text{invar} :: \ 's \Rightarrow \text{bool} \)

**specification**

\[
\begin{aligned}
\text{invar} \ \text{empty} & \quad (\text{empty-inv}) \\
\text{set} \ \text{empty} = & \quad \{} \quad (\text{empty}) \\
\text{invar} \ s \Rightarrow \text{invar} \ (\text{insert} \ x \ s) & \quad (\text{insert-inv}) \\
\text{invar} \ s \Rightarrow \text{set} (\text{insert} \ x \ s) = \text{set} s \cup \{x\} & \quad (\text{insert}) \\
\text{invar} \ s \Rightarrow \text{invar} \ (\text{delete} \ x \ s) & \quad (\text{delete-inv}) \\
\text{invar} \ s \Rightarrow \text{set} (\text{delete} \ x \ s) = \text{set} s \setminus \{x\} & \quad (\text{delete}) \\
\text{invar} \ s \Rightarrow \text{isin} \ s \ x = (x \in \text{set} s) & \quad (\text{isin})
\end{aligned}
\]

Fig. 6.1. ADT Set

We have intentionally refrained from showing the Isabelle formalization using so-called locales and have opted for a more intuitive textual format that is not Isabelle-specific, in accordance with the general philosophy of this book.
The actual Isabelle text can of course be found in the source files, and locales are explained in a dedicated manual [5].

We conclude this section by explaining what the specification of an arbitrary ADT looks like. We assume that for each function \( f \) of the interface there is a corresponding function \( f_A \) in the abstract model, i.e. defined on \( A \). For a uniform treatment we extend \( \alpha \) and \( \text{invar} \) to arbitrary types by setting \( \alpha x = x \) and \( \text{invar} x = \text{True} \) for all types other than \( T \). Each function \( f \) of the interface gives rise to two properties in the specification: preservation of the invariant and simulation of \( f_A \). The precondition is shared:

\[
\begin{align*}
\text{invar} x_1 \land \ldots \land \text{invar} x_n & \implies \\
\text{invar}(f x_1 \ldots x_n) & \implies (f-\text{inv}) \\
\alpha(f x_1 \ldots x_n) = f_A(\alpha x_1) \ldots (\alpha x_n) & \implies (f)
\end{align*}
\]

To understand how the specification of ADT \( \text{Set} \) is the result of this uniform schema one has to take two things into account:

- Precisely which abstract operations on type ‘a set model the functions in the interface of the ADT \( \text{Set} \)? This correspondence is implicit in the specification: \text{empty} is modeled by \{\}, \text{insert} is modeled by \( \lambda x \, s \, s' = \{x\} \), \text{delete} is modeled by \( \lambda x \, s \, s' = s \setminus \{x\} \) and \text{isin} is modeled by \( \lambda s \, x \, \in s \).
- Because of the artificial extension of \( \alpha \) and \text{invar} the above uniform format often collapses to something simpler where some \( \alpha \)'s and \text{invar}'s disappear.

### 6.3 Maps

An even more versatile data structure than sets are (efficient) maps from ‘a to ‘b. In fact, sets can be viewed as maps from ‘a to bool. Conversely, many data structures for sets also support maps, e.g. BSTs. Although, for simplicity, we focus on sets in this book, the ADT of maps should at least be introduced. It is shown in Figure 6.2. Type ‘m is the type of maps from ‘a to ‘b. The ADT \( \text{Map} \) is very similar to the ADT \( \text{Set} \) except that the abstraction function \text{lookup} is also part of the interface: it abstracts a map to a function of type ‘a \( \Rightarrow \) ‘b \( \text{option} \). This implies that the equations are between functions of that type. We use the function update notation (see Section 1.3) to explain \text{update} and \text{delete}: \text{update} is modeled by \( \lambda m \, a \, b \, m(a := b) \) and \text{delete} by \( \lambda m \, a \, m(a := \text{None}) \).
ADT Map =

interface
empty :: 'm
update :: 'a ⇒ 'b ⇒ 'm ⇒ 'm
delete :: 'a ⇒ 'm ⇒ 'm
lookup :: 'm ⇒ 'a ⇒ 'b option

abstraction lookup
invariant invar :: 'm ⇒ bool

specification
invar empty                  (empty-inv)
lookup empty = (λa. None)     (empty)
invar m → invar (update a b m) (update-inv)
invar m → lookup (update a b m) = (lookup m)(a := Some b) (update)
invar m → invar (delete a m)   (delete-inv)
invar m → lookup (delete a m) = (lookup m)(a := None)     (delete)

Fig. 6.2. ADT Map

6.4 Implementing ADTs

An implementation of an ADT consists of definitions for all the functions in the interface. If you want to verify the correctness of the implementation, you also need to provide definitions for the abstraction function and the invariant. The latter two need not be executable unless they also occur in the interface and the implementation is meant to be executable. The abstraction function need not be surjective. For example, implementations of the ADT Set will normally only represent finite sets, e.g. by BSTs.

For Isabelle users: because ADTs are formalized as locales, an implementation of an ADT is an interpretation of the corresponding locale.

6.5 Exercises

Exercise 6.1. Modify the ADT Set specification by making isin the abstraction function (from 's to 'a ⇒ bool). Follow the example of the ADT Map specification.

Exercise 6.2. In the ADT Map specification, following the general schema, there should be a property labeled (lookup), but it is missing. The reason is that given the correct abstract model of lookup, the equation becomes trivial: invar m → lookup m a = lookup m a. Why is that, which function models lookup?
Exercise 6.3. Implement ADT Map via unbalanced BSTs (like Set in Chapter 5) using augmented trees. Verify the implementation by proving all the correctness properties in the specification of ADT Map directly, without any detour via sorted lists as in Section 5.4.
This is the first in a series of chapters examining balanced search trees where the height of the tree is logarithmic in its size and which can therefore be searched in logarithmic time.

The most popular first example of balanced search trees are red-black trees. We start with 2-3 trees, where nodes can have 2 or 3 children, because red-black trees are best understood as an implementation of (a variant of) 2-3 trees. We introduce red-black trees in the next chapter. The type of 2-3 trees is similar to binary trees but with an additional constructor `Node3`:

```
datatype 'a tree23 =
  Leaf |
  Node2 ('a tree23) 'a ('a tree23) |
  Node3 ('a tree23) 'a ('a tree23) 'a ('a tree23)
```

The familiar syntactic sugar is sprinkled on top:

```
⟨⟩ ≡ Leaf
⟨l, a, r⟩ ≡ Node2 l a r
⟨l, a, m, b, r⟩ ≡ Node3 l a m b r
```

The size, height and the completeness of a 2-3 tree are defined by adding an equation for `Node3` to the corresponding definitions on binary trees:

```
|⟨l, __, m, __, r⟩| = |l| + |m| + |r| + 1
```
\[ h (l, _, m, _, r) = \max (h l) (\max (h m) (h r)) + 1 \]

\[
\text{complete } (l, _, m, _, r) = (h l = h m \land h m = h r \land \text{complete } l \land \text{complete } m \land \\
\text{complete } r)
\]

A trivial induction yields complete \( t \rightarrow 2^h t \leq |t| + 1 \): thus all operations on complete 2-3 trees have logarithmic complexity if they descend along a single branch and take constant time per node. This is the case and we will not discuss complexity in any more detail.

A nice property of 2-3 trees is that for every \( n \) there is a complete 2-3 tree of size \( n \). As we will see below, completeness can be maintained under insertion and deletion in logarithmic time.

Exercise 7.1. Define a function \( \text{maxt} :: \text{nat} \Rightarrow \text{unit} \) tree23 that creates the tree with the largest number of nodes given the height of the tree. We use type \( \text{unit} \) because we are not interested in the elements in the tree. Prove \( |\text{maxt } n| = (3^n - 1) \div 2 \) and that no tree of the given height can be larger: \( |t| \leq (3^h t - 1) \div 2 \). Note that both subtraction and division on type \( \text{nat} \) can be tedious to work with. You may want to prove the two properties as corollaries of subtraction- and division-free properties. Alternatively, work with \( \text{real} \) instead of \( \text{nat} \) by replacing \( \div \) by \( / \).

### 7.1 Implementation of ADT Set

The implementation will maintain the usual ordering invariant and additionally completeness. When we speak of a 2-3 tree we will implicitly assume these two invariants now.

Searching a 2-3 tree is like searching a binary tree (see Section 5.2) but with one more defining equation:

\[
\text{isin } (l, a, m, b, r) x = (\text{case } \text{cmp } x a \text{ of } LT \Rightarrow \text{isin } l x \mid EQ \Rightarrow True \\
| GT \Rightarrow \text{case } \text{cmp } x b \text{ of } LT \Rightarrow \text{isin } m x \mid EQ \Rightarrow True \\
| GT \Rightarrow \text{isin } r x)
\]

Insertion into a 2-3 tree must preserve the completeness invariant. Thus recursive calls must report back to the caller if the child has “overflowed”, i.e. increased in height. Therefore insertion returns a result of type ‘a upI:
7.1 Implementation of ADT Set

\[
\text{datatype } \text{'a upI} = \text{TI ('}\text{a tree23}) \mid \text{OF ('}\text{a tree23}) \text{'a ('}\text{a tree23})
\]

This is the idea: If insertion into \( t \) returns
\( \text{TI } t' \) then \( t \) and \( t' \) should have the same height,
\( \text{OF } l \times r \) then \( t \) and \( l \) and \( r \) should have the same height.

The insertion functions are shown in Figure 7.1. The actual work is performed by the recursive function \text{ins}. The element to be inserted is propagated down to a leaf, which causes an overflow of the leaf. If an overflow is returned from a recursive call it can be absorbed into a Node2 but in a Node3 it causes another overflow. At the root of the tree, function treeI converts values of type 'a upI back into trees:

\[
\text{treeI} :: \text{'a upI} \Rightarrow \text{'a tree23}
\]
\[
\text{treeI} (\text{TI } t) = t
\]
\[
\text{treeI} (\text{OF } l \times a \times r) = \langle l, a, r \rangle
\]

Deletion is dual. Recursive calls must report back to the caller if the child has "underflown", i.e. decreased in height. Therefore deletion returns a result of type 'a upD:

\[
\text{datatype } \text{'a upD} = \text{TD ('}\text{a tree23}) \mid \text{UF ('}\text{a tree23})
\]

This is the idea: If deletion from \( t \) returns
\( \text{TD } t' \) then \( t \) and \( t' \) should have the same height,
\( \text{UF } t' \) then \( t \) should be one level higher than \( t' \).

The main deletion functions are shown in Figure 7.2. The actual work is performed by the recursive function \text{del}. If the element to be deleted is in a child, the result of a recursive call is reintegrated into the node via the auxiliary functions \text{nodeij} from Figure 7.3 that create a node with \( i \) children where child \( j \) is given as an 'a upD value. If the element to be deleted is in the node itself, a replacement is obtained and deleted from a child via \text{split_min}. At the root of the tree, values of type 'a upD are converted back into trees:

\[
\text{treeD} :: \text{'a upD} \Rightarrow \text{'a tree23}
\]
\[
\text{treeD} (\text{TD } t) = t
\]
\[
\text{treeD} (\text{UF } t) = t
\]
\[ \text{insert } x \ t = \text{treeI} \ (\text{ins } x \ t) \]

\[ \text{ins } :: \ 'a \Rightarrow 'a \ \text{treeI23} \Rightarrow 'a \ \text{upI} \]

\[ \text{ins } \emptyset = \text{OF} \ \emptyset \ x \ \emptyset \]

\[ \text{ins } \{l, a, r\} = (\text{case } \text{cmp } x \ a \text{ of} \]

\[ \begin{align*}
LT & \Rightarrow \text{case } \text{ins } x \ l \text{ of } \text{TI } \ l' \Rightarrow \text{TI} \ \{l', a, r\} \\
| & \quad \text{OF} \ l_1 \ b \ l_2 \Rightarrow \text{TI} \ \{l_1, b, l_2, a, r\} \\
| & \quad \text{EQ} \Rightarrow \text{TI} \ \{l, a, r\} \\
| & \quad \text{GT} \Rightarrow \text{case } \text{ins } x \ r \text{ of } \text{TI } r' \Rightarrow \text{TI} \ \{l, a, r'\} \\
| & \quad \text{OF} \ r_1 \ b \ r_2 \Rightarrow \text{TI} \ \{l, a, r_1, b, r_2\} \\
\end{align*} \]

\[ \text{ins } \{l, a, m, b, r\} = (\text{case } \text{cmp } x \ a \text{ of} \]

\[ \begin{align*}
LT & \Rightarrow \text{case } \text{ins } x \ l \text{ of } \text{TI } \ l' \Rightarrow \text{TI} \ \{l', a, m, b, r\} \\
| & \quad \text{OF} \ l_1 \ c \ l_2 \Rightarrow \text{OF} \ \{l_1, c, l_2\} \ a \ \{m, b, r\} \\
| & \quad \text{EQ} \Rightarrow \text{TI} \ \{l, a, m, b, r\} \\
| & \quad \text{GT} \Rightarrow \text{case } \text{cmp } x \ b \text{ of} \\
LT & \Rightarrow \text{case } \text{ins } x \ b \text{ of } \text{TI } m' \Rightarrow \text{TI} \ \{l, a, m', b, r\} \\
| & \quad \text{OF} \ m_1 \ c \ m_2 \Rightarrow \text{OF} \ \{l, a, m_1\} \ c \ \{m_2, b, r\} \\
| & \quad \text{EQ} \Rightarrow \text{TI} \ \{l, a, m, b, r\} \\
| & \quad \text{GT} \Rightarrow \text{case } \text{ins } x \ r \text{ of } \text{TI } r' \Rightarrow \text{TI} \ \{l, a, m, b, r'\} \\
| & \quad \text{OF} \ r_1 \ c \ r_2 \Rightarrow \text{OF} \ \{l, a, m\} \ b \ \{r_1, c, r_2\} \\
\end{align*} \]

Fig. 7.1. Insertion into 2-3 tree

### 7.2 Preservation of Completeness

As explained in Section 5.4, we do not go into the automatic functional correctness proofs but concentrate on invariant preservation. To express the relationship between the height of a tree before and after insertion we define a height function \( hI \) on \( 'a \ \text{upI} \) as follows:

\[ hI \ (\text{TI } t) = h \ t \]

\[ hI \ (\text{OF } l \ a \ r) = h \ l \]

The intuition is that \( hI \) is the height of the tree before insertion. A routine induction proves

\[ \text{complete } t \rightarrow \text{complete } (\text{treeI } (\text{ins } a \ t)) \lor hI \ (\text{ins } a \ t) = h \ t \]

which implies by definition that

\[ \text{complete } t \rightarrow \text{complete } (\text{insert } a \ t) \]
Fig. 7.2. Deletion from 2-3 tree: main functions

The fact that deletion preserves completeness can be proved by a sequence of small lemmas. To express the relationship between the height of a tree before and after deletion we define a height function \( hD \) on \('a \ upD \) as follows:

\[

delete x t = treeD (del x t)
\]

\[
del :: 'a \Rightarrow 'a \ tree23 \Rightarrow 'a \ upD
\]

\[
del x \emptyset = TD \emptyset
\]

\[
del x \langle i, a, \emptyset \rangle = (if \ x = a \ then \ UF \langle i, a \rangle \ else \ TD \langle i, a, \emptyset \rangle)
\]

\[
del x \langle i, a, \emptyset, b, \emptyset \rangle = TD (if \ x = a \ then \ \langle i, b, \emptyset \rangle \ else \ if \ x = b \ then \ \langle i, a, \emptyset \rangle \ else \ \langle \emptyset, a, \emptyset \rangle)
\]

\[
del x \langle i, a, b, r \rangle = (case \ cmp x a of \ LT \Rightarrow \ node21 (del x l) a r
\ | \ EQ \Rightarrow \ \langle i, a', r' \rangle = split_min r \ in \ node22 l a' r'
\ | \ GT \Rightarrow \ node22 l a \ (del x r))
\]

\[
split_min :: 'a \ tree23 \Rightarrow 'a \ x \ 'a \ upD
\]

\[
split_min \langle i, a, \emptyset \rangle = (a, UF \langle i \rangle)
\]

\[
split_min \langle i, a, \emptyset, b, \emptyset \rangle = (a, TD \langle i, b, \emptyset \rangle)
\]

\[
split_min \langle l, a, r \rangle = (let \ (x, l') = split_min l \ in \ node21 l' a r)
\]

\[
split_min \langle l, a, m, b, r \rangle = (let \ (x, l') = split_min l \ in \ node31 l' a m b r)
\]

\[
hD (TD t) = h t
\]

\[
hD (UF t) = h t + 1
\]

The intuition is that \( hD \) is the height of the tree before deletion. We now list a sequence of properties that build on each other and culminate in completeness preservation of \( \text{delete} \):

\[
complete r \land complete (treeD l') \land h r = hD l' \lor \rightarrow \ complete (treeD (node21 l' a r))
\]

\[
0 < h r \rightarrow hD (node21 l' a r) = \max (hD l') (h r) + 1
\]
For each property of nodeij there are analogues properties for the other nodeij functions which we omit.
7.3 Converting a List into a 2-3 Tree

The naive method of converting a list of elements into a 2-3 tree is to insert them one by one starting from the empty tree. However, that takes time $\Theta(n \lg n)$. This holds for any data structure where insertion takes time proportional to $\lg n$. In that case inserting $n$ elements one by one takes time proportional to $\frac{1}{2} n \lg n \leq \lg(n!)$. Thus $\lg(n!) \in \Theta(n \lg n)$ (which also follows from Stirling’s formula). We have intentionally proved a $\Theta$ property because the $O$ property is obvious but one might hope that $\lg 1 + \cdots + \lg n$ has a lower order of growth than $n \lg n$.

Luckily we can convert a whole list into a 2-3 tree in linear time. The bottom-up algorithm is particularly intuitive. It repeatedly passes over an alternating list $t_1; e_1; t_2; e_2; \ldots$ of trees and elements, combining trees and elements into new trees. Given elements $a_1, \ldots, a_n$ we start with the alternating list $() ; a_1 ; () ; a_2 ; () ; \ldots ; a_n ; ()$. On every pass over this list, we replace adjacent triples $t; a; t_0$ by $h_{t; a; t_0}$, possibly creating a 3-node instead of a 2-node at the end of the list. Once a single tree is left over, we terminate.

We define this type of alternating (and non-empty) list as a new data type:

```
datatype 'a tree23s = T ('a tree23) | TTs ('a tree23) 'a ('a tree23s)
```

The following examples demonstrate the encoding of alternating lists as terms of type `'a tree23s`:

Alternating list: $t_1; e_1; t_2; e_2; \ldots; t_k$

Encoding: $T t_1; TTs t_1; e_1; (T t_2); TTs t_1; e_1; (TTs t_2; e_2; ts)$

We also need the following auxiliary functions:

```
len :: 'a tree23s => nat
len (T _) = 1
len (TTs _ _ ts) = len ts + 1

trees :: 'a tree23s => 'a tree23 set
trees (T t) = {t}
trees (TTs t _ ts) = {t} \cup trees ts

inorder2 :: 'a tree23s => 'a list
inorder2 (T t) = inorder t
inorder2 (TTs t a ts) = inorder t @ a @ inorder2 ts
```
Repeatedly passing over the alternating list until only a single tree remains is expressed by the following functions:

\[
\begin{align*}
\text{join}_\text{all} & : \text{'a tree23s} \Rightarrow \text{'a tree23} \\
\text{join}_\text{all} (\ T \ t) & = t \\
\text{join}_\text{all} \ ts & = \text{join}_\text{all} (\text{join}_\text{adj} \ ts)
\end{align*}
\]

\[
\begin{align*}
\text{join}_\text{adj} & : \text{'a tree23s} \Rightarrow \text{'a tree23s} \\
\text{join}_\text{adj} (\ TTs \ t_1 \ a \ (T \ t_2)) & = T \ (t_1, \ a, \ t_2) \\
\text{join}_\text{adj} (\ TTs \ t_1 \ a \ (TTs \ t_2 \ b \ (T \ t_3))) & = T \ (t_1, \ a, \ t_2, \ b, \ t_3) \\
\text{join}_\text{adj} (\ TTs \ t_1 \ a \ (TTs \ t_2 \ b \ ts)) & = TTs \ (t_1, \ a, \ t_2) \ b \ (\text{join}_\text{adj} \ ts)
\end{align*}
\]

Note that \text{join}_\text{adj} is not and does not need to be defined on single trees. We express this precondition with an abbreviation:

\[
\text{not}_T \ ts \equiv \forall \ t. \ ts \neq T \ t
\]

Also note that \text{join}_\text{all} terminates only because \text{join}_\text{adj} shortens the list:

\[
\text{not}_T \ ts \rightarrow \text{len} (\text{join}_\text{adj} \ ts) < \text{len} \ ts
\]

In fact, it reduces the length at least by a factor of 2:

\[
\text{not}_T \ ts \rightarrow \text{len} (\text{join}_\text{adj} \ ts) \leq \text{len} \ ts \ \text{div} \ 2 \quad (7.1)
\]

The whole process starts with a list of alternating leaves and elements:

\[
\begin{align*}
\text{tree23}_\text{of_list} & : \text{'a list} \Rightarrow \text{'a tree23} \\
\text{tree23}_\text{of_list} \ as & = \text{join}_\text{all} (\text{leaves} \ as)
\end{align*}
\]

\[
\begin{align*}
\text{leaves} & : \text{'a list} \Rightarrow \text{'a tree23s} \\
\text{leaves} [\ ] & = T \ () \\
\text{leaves} (a \ # \ as) & = TTs \ () \ a \ (\text{leaves} \ as)
\end{align*}
\]

### 7.3.1 Functional Correctness

Functional correctness is easily established. The \text{inorder} and the completeness properties are proved independently by the following inductive lemmas:

\[
\begin{align*}
\text{not}_T \ ts \rightarrow \text{inorder2} (\text{join}_\text{adj} \ ts) & = \text{inorder2} \ ts \\
\text{inorder} (\text{join}_\text{all} \ ts) & = \text{inorder2} \ ts \\
\text{inorder} (\text{tree23}_\text{of_list} \ as) & = as
\end{align*}
\]
7.3 Converting a List into a 2-3 Tree

\((\forall t \in \text{trees } ts. \text{ complete } t \land h t = n) \land \text{not}_\text{T } ts \rightarrow\)
\((\forall t \in \text{trees } (\text{join}_\text{adj} ts). \text{ complete } t \land h t = n + 1)\)
\((\forall t \in \text{trees } ts. \text{ complete } t \land h t = n) \rightarrow \text{ complete } (\text{join}_\text{all} ts)\)
\(t \in \text{trees } (\text{leaves } \text{as}) \rightarrow \text{ complete } t \land h t = 0\)
\(\text{complete } (\text{tree}_\text{23}_\text{of_list } \text{as})\)

7.3.2 Running Time Analysis

Why does the conversion take linear time? Because the first pass over an alternating list of length \(n\) takes \(n\) steps, the next pass \(n/2\) steps, the next pass \(n/4\) steps, etc, and this sums up to \(2n\). The canonical time functions for the formal proof are shown in Appendix B.3. The following upper bound is easily proved by induction on the computation of \(\text{join}_\text{adj}\):

\[
\text{not}_\text{T } ts \rightarrow T_{\text{join}_\text{adj} } ts \leq \text{len } ts \div 2
\]  
(7.2)

An upper bound \(T_{\text{join}_\text{all} } ts \leq 2 \cdot \text{len } ts\) follows by induction on the computation of \(\text{join}_\text{adj}\). We focus on the induction step:

\[
T_{\text{join}_\text{all} } ts
= T_{\text{join}_\text{adj} } ts + T_{\text{join}_\text{all} } (\text{join}_\text{adj} ts) + 1
\leq \text{len } ts \div 2 + 2 \cdot \text{len } (\text{join}_\text{adj} ts) + 1
\leq \text{len } ts \div 2 + 2 \cdot (\text{len } ts \div 2 + 1) + 1
\leq 2 \cdot \text{len } ts
\]

by (7.1)

Now it is routine to derive

\[
T_{\text{tree}_\text{23}_\text{of_list } \text{as}} \leq 3 \cdot |\text{as}| + 4
\]

Bibliographic Remarks

The invention of 2-3 trees is credited to Hopcroft in 1970 [13, p. 337]. Equational definitions were given by Hoffmann and O’Donnell [29] (only insertion) and Reade [66]. Our formalisation is based on the teaching material by Turbak [74] and the article by Hinze [28].
Red-black trees are a popular implementation technique for BSTs: they guarantee logarithmic height just like 2-3 trees but the code is arguably simpler. The nodes are colored either red or black. Abstractly, red-black trees encode 2-3-4 trees where nodes have between 2 and 4 children. Each 2-3-4 node is encoded by a group of between 2 and 4 colored binary nodes as follows:

\[
\begin{align*}
\langle & \rangle \approx \langle \rangle \\
\langle A, a, B \rangle \approx \langle A, a, B \rangle \\
\langle A, a, B, b, C \rangle & \approx \langle \langle A, a, B \rangle, b, C \rangle \text{ or } \langle A, a, \langle B, b, C \rangle \rangle \\
\langle A, a, B, b, C, c, D \rangle & \approx \langle \langle A, a, B \rangle, b, \langle C, c, D \rangle \rangle
\end{align*}
\]

Color expresses grouping: a black node is the root of a 2-3-4 node, a red node is part of a bigger 2-3-4 node. Thus a red-black tree needs to satisfy the following properties or invariants:

1. The root is black.
2. Every \( \langle \rangle \) is considered black.
3. If a node is red, its children are black.
4. All paths from a node to a leaf have the same number of black nodes.

The final property expresses that the corresponding 2-3-4 tree is complete. The last two properties imply that the tree has logarithmic height (see below).

We implement red-black trees as binary trees augmented (see Section 4.4) with a color tag:

```plaintext
datatype color = Red | Black

type_synonym 'a rbt = ('a × color) tree
```

Some new syntactic sugar is sprinkled on top:
The following functions get and set the color of a node:

\[
\begin{align*}
\text{color} & : 'a rbt \Rightarrow color \\
\text{color} \emptyset &= \text{Black} \\
\text{color} \langle \_, \_, c, \_ \rangle &= c
\end{align*}
\]

\[
\begin{align*}
\text{paint} & : color \Rightarrow 'a rbt \Rightarrow 'a rbt \\
\text{paint} \emptyset &= \emptyset \\
\text{paint} c \langle l, (a, \_), r \rangle &= \langle l, (a, c), r \rangle
\end{align*}
\]

Note that the color of a leaf is by definition black.

### 8.1 Invariants

The above informal description of the red-black tree invariants is formalized as the predicate \( \text{rbt} \) which (for reasons of modularity) is split into a color and a height invariant \( \text{invc} \) and \( \text{invh} \):

\[
\begin{align*}
\text{rbt} & : 'a rbt \Rightarrow bool \\
\text{rbt} t &= (\text{invc} t \land \text{invh} t \land \text{color} t = \text{Black})
\end{align*}
\]

The color invariant expresses that red nodes must have black children:

\[
\begin{align*}
\text{invc} & : 'a rbt \Rightarrow bool \\
\text{invc} \emptyset &= \text{True} \\
\text{invc} \langle l, \_, c, r \rangle &= ((c = \text{Red} \Rightarrow \text{color} l = \text{Black} \land \text{color} r = \text{Black}) \land \\
& \quad \text{invc} l \land \text{invc} r)
\end{align*}
\]

The height invariant expresses (via the black height \( bh \)) that all paths from the root to a leaf have the same number of black nodes:
Note that although $bh$ traverses only the left spine of the tree, the fact that $invh$ traverses the complete tree ensures that all paths from the root to a leaf are considered. (See Exercise 8.2)

The split of the invariant into $invc$ and $invh$ improves modularity: frequently one can prove preservation of $invc$ and $invh$ separately, which facilitates proof search. For compactness we will mostly present the combined invariance properties.

### 8.1.1 Logarithmic Height

In a red-black tree, i.e. $rbt$ $t$, every path from the root to a leaf has the same number of black nodes, and no such path has two red nodes in a row. Thus each leaf is at most twice as deep as any other leaf, and therefore $h t \leq 2 \cdot \lg |t|_1$. The detailed proof starts with the key inductive relationship between height and black height

\[
invc t \land invh t \rightarrow h t \leq 2 \cdot bh t + (\text{if } color t = \text{Black} \text{ then } 0 \text{ else } 1)
\]

which has the easy corollary $rbt t \rightarrow h t / 2 \leq bh t$. Together with the easy inductive property

\[
invc t \land invh t \rightarrow 2^{bh t} \leq |t|_1
\]

this implies $2^{h t / 2} \leq 2^{bh t} \leq |t|_1$ and thus $h t \leq 2 \cdot \lg |t|_1$ if $rbt t$.

### 8.2 Implementation of ADT Set

We implement sets by red-black trees that are also BSTs. As usual, we only discuss the proofs of preservation of the $rbt$ invariant.

Function $isin$ is implemented as for all augmented BSTs (see Section 5.5.1).
8 Red-Black Trees

8.2.1 Insertion

Insertion is shown in Figure 8.1. The workhorse is function ins. It descends to the leaf where the element is inserted and it adjusts the colors on the way back up. The adjustment is performed by baliL/baliR. They combine arguments l a r into a tree. If there is a red-red conflict in l/r, they rebalance and replace it by red-black. Inserting into a red node needs no immediate balancing because that will happen at the black node above it:

\[
\begin{align*}
\text{ins } x \ t &= \text{paint Black } (\text{ins } x \ t) \\
\text{ins } :: \ 'a \Rightarrow 'a \ rbt \Rightarrow 'a \ rbt \\
\text{ins } x \ () &= R \ () \ x \ () \\
\text{ins } x \ (B \ l \ a \ r) &= \text{case cmp } x \ a \ of \\
& \quad LT \Rightarrow \text{baliL } (\text{ins } x \ l) \ a \ r \\
& \quad EQ \Rightarrow B \ l \ a \ r \\
& \quad GT \Rightarrow \text{baliR } l \ a \ (\text{ins } x \ r) \\
\text{ins } x \ (R \ l \ a \ r) &= \text{case cmp } x \ a \ of \\
& \quad LT \Rightarrow R \ (\text{ins } x \ l) \ a \ r \\
& \quad EQ \Rightarrow R \ l \ a \ r \\
& \quad GT \Rightarrow R \ l \ a \ (\text{ins } x \ r) \\
\text{baliL } :: \ 'a \ rbt \Rightarrow 'a \ rbt \Rightarrow 'a \ rbt \\
\text{baliL } (R \ (R \ t_1 \ a \ t_2) \ b \ t_3) \ c \ t_4 &= R \ (B \ t_1 \ a \ t_2) \ b \ (B \ t_3 \ c \ t_4) \\
\text{baliL } (R \ t_1 \ a \ (R \ t_2 \ b \ t_3)) \ c \ t_4 &= R \ (B \ t_1 \ a \ t_2) \ b \ (B \ t_3 \ c \ t_4) \\
\text{baliL } t_1 \ a \ t_2 &= B \ t_1 \ a \ t_2 \\
\text{baliR } :: \ 'a \ rbt \Rightarrow 'a \ rbt \Rightarrow 'a \ rbt \\
\text{baliR } t_1 \ a \ (R \ t_2 \ b \ (R \ t_3 \ c \ t_4)) &= R \ (B \ t_1 \ a \ t_2) \ b \ (B \ t_3 \ c \ t_4) \\
\text{baliR } t_1 \ a \ (R \ (R \ t_2 \ b \ t_3) \ c \ t_4) &= R \ (B \ t_1 \ a \ t_2) \ b \ (B \ t_3 \ c \ t_4) \\
\text{baliR } t_1 \ a \ t_2 &= B \ t_1 \ a \ t_2 \\
\end{align*}
\]

Fig. 8.1. Insertion into red-black tree
Function \texttt{ins} preserves \texttt{invh} but not \texttt{invc}: it may return a tree with a red root node (see example above). However, once the root node is colored black, everything is fine again. Thus we introduce the weaker invariant \texttt{invc2} as an abbreviation:

\[
invc2 \, t \equiv \texttt{invc}\, (\texttt{paint Black } t)
\]

It is easy to prove that \texttt{baliL} and \texttt{baliR} preserve \texttt{invh} and upgrade from \texttt{invc2} to \texttt{invc}:

\[
\begin{align*}
\texttt{invh } l \land \texttt{invh } r \land \texttt{invc2 } l \land \texttt{invc } r \land \texttt{bh } l = \texttt{bh } r & \longrightarrow \\
\texttt{invc } (\texttt{baliL } l \texttt{ a } r) \land \texttt{invh } (\texttt{baliL } l \texttt{ a } r) \land \texttt{bh } (\texttt{baliL } l \texttt{ a } r) = \texttt{bh } l + 1 \\
\texttt{invh } l \land \texttt{invh } r \land \texttt{invc } l \land \texttt{invc2 } r \land \texttt{bh } l = \texttt{bh } r & \longrightarrow \\
\texttt{invc } (\texttt{baliR } l \texttt{ a } r) \land \texttt{invh } (\texttt{baliR } l \texttt{ a } r) \land \texttt{bh } (\texttt{baliR } l \texttt{ a } r) = \texttt{bh } l + 1
\end{align*}
\]

Another easy induction yields

\[
\begin{align*}
\texttt{invc } t \land \texttt{invh } t \longrightarrow \\
\texttt{invc2 } (\texttt{ins } x \texttt{ t}) \land (\texttt{color } t = \texttt{Black} \longrightarrow \texttt{invc } (\texttt{ins } x \texttt{ t})) \land \\
\texttt{invh } (\texttt{ins } x \texttt{ t}) \land \texttt{bh } (\texttt{ins } x \texttt{ t}) = \texttt{bh } t
\end{align*}
\]

The corollary \(\texttt{rbt } t \longrightarrow \texttt{rbt } (\texttt{insert } x \texttt{ t})\) is immediate.

### 8.2.2 Deletion

Deletion from a red-black tree is shown in Figure 8.2. It follows the deletion-by-replacing approach (Section 5.2.1). The tricky bit is how to maintain the invariants. As before, intermediate trees may only satisfy the weaker invariant \texttt{invc2}. Functions \texttt{del} and \texttt{split_min} decrease the black height of a tree with a black root node and leave the black height unchanged otherwise. To see that this makes sense, consider deletion from a singleton black or red node. The case that the element to be removed is not in the black tree can be dealt with by coloring the root node red. These are the precise input/output relations:

**Lemma 8.1.** \(\texttt{split_min } t = (x, t') \land t \neq () \land \texttt{invh } t \land \texttt{invc } t \longrightarrow \)

\[
\begin{align*}
\texttt{invh } t' \land (\texttt{color } t = \texttt{Red} \longrightarrow \texttt{bh } t' = \texttt{bh } t \land \texttt{invc } t') \land \\
(\texttt{color } t = \texttt{Black} \longrightarrow \texttt{bh } t' = \texttt{bh } t - 1 \land \texttt{invc2 } t')
\end{align*}
\]

**Lemma 8.2.** \(\texttt{invh } t \land \texttt{invc } t \land t' = \texttt{del } x \texttt{ t} \longrightarrow \)

\[
\begin{align*}
\texttt{invh } t' \land (\texttt{color } t = \texttt{Red} \longrightarrow \texttt{bh } t' = \texttt{bh } t \land \texttt{invc } t') \land \\
(\texttt{color } t = \texttt{Black} \longrightarrow \texttt{bh } t' = \texttt{bh } t - 1 \land \texttt{invc2 } t')
\end{align*}
\]

It is easy to see that the \texttt{del}-Lemma implies correctness of \texttt{delete}:

**Corollary 8.3.** \(\texttt{rbt } t \longrightarrow \texttt{rbt } (\texttt{delete } x \texttt{ t})\)
delete \( x \ t = \text{paint} \ \text{Black} \ (\text{del} \ x \ t) \)

del :: 'a ⇒ 'a rbt ⇒ 'a rbt

del _\{
  _\} = _\{
  del z \{\}, (a, _), r\} = (\text{case} \ \text{cmp} \ z \ a \ \text{of}
  \{\}
  \{\}
  \{\}
  \{\}
  \{\}
  \{\}

LT ⇒ let \( l' = \text{del} \ x \ l \)
  in if \( l \neq \{\} \land \text{color} \ l = \text{Black} \) then \( \text{baldL} \ l' \ a \ r \) else \( \text{R} \ l' \ a \ r \) |

EQ ⇒ if \( r = \{\} \) then \( l \)
  else let \( (a', \ r') = \text{split}\_\text{min} \ r \)
  in if \( \text{color} \ r = \text{Black} \) then \( \text{baldR} \ l \ a' \ r' \) else \( \text{R} \ l \ a' \ r' \) |

GT ⇒ let \( r' = \text{del} \ x \ r \)
  in if \( r \neq \{\} \land \text{color} \ r = \text{Black} \) then \( \text{baldR} \ l \ a' \ r \) else \( \text{R} \ l \ a' \ r \)

split\_\text{min} :: 'a rbt ⇒ 'a × 'a rbt

split\_\text{min} (l, (a, _), r)
  = (if \( l = \{\} \) then (a, r)
  else let \( (x, l') = \text{split}\_\text{min} \ l \)
  in (x, if \( \text{color} \ l = \text{Black} \) then \( \text{baldL} \ l' \ a \ r \) else \( \text{R} \ l' \ a \ r \)))

\text{baldL} :: 'a rbt ⇒ 'a ⇒ 'a rbt

\text{baldL} (R \ t_1 \ a \ t_2) \ b \ t_3 = R (B \ t_1 \ a \ t_2) \ b \ t_3

\text{baldL} (R \ a) \ (B \ b \ t_3) = \text{baliL} \ a \ (R \ t_1 \ a \ t_2)

\text{baldL} (R \ (B \ t_2 \ b \ t_3) \ c \ t_4) = R \ (B \ t_1 \ a \ t_2) \ b \ (\text{baliR} \ t_3 \ c \ (\text{paint} \ \text{Red} \ t_4))

\text{baldL} \ t_1 \ a \ t_2 = R \ t_1 \ a \ t_2

\text{baldR} :: 'a rbt ⇒ 'a ⇒ 'a rbt

\text{baldR} \ a \ R \ t_2 \ b \ t_3) = \text{R} \ t_1 \ a \ (B \ t_2 \ b \ t_3)

\text{baldR} (B \ t_1 \ a \ t_2) \ b \ t_3 = \text{baliR} \ (R \ t_1 \ a \ t_2) \ b \ t_3

\text{baldR} (R \ t_1 \ a \ (B \ t_2 \ b \ t_3) \ c \ t_4) = \text{R} \ (\text{baliL} \ (\text{paint} \ \text{Red} \ t_1) \ a \ t_2) \ b \ (B \ t_3 \ c \ t_4)

\text{baldR} \ t_1 \ a \ t_2 = R \ t_1 \ a \ t_2

\text{Fig. 8.2. Deletion from red-black tree}

The proofs of the two lemmas need the following precise characterizations of \text{baldL} and \text{baldR}, the counterparts of \text{baliL} and \text{baliR}:

\text{Lemma 8.4.} \ invh \ l \land \ invh \ r \land \ bh \ l + 1 = bh \ r \land \ invc2 \ l \land \ invc \ r \land \ t' = \text{baldL} \ l \ a \ r \Land 

\text{invh} \ t' \land \ bh \ t' = bh \ r \land \ invc2 \ t' \land \ (\text{color} \ r = \text{Black} \longrightarrow \ invc \ t')

\text{Lemma 8.5.} \ invh \ l \land \ invh \ r \land \ bh \ l = bh \ r + 1 \land \ invc \ l \land \ invc2 \ r \land \ t' = \text{baldR} \ l \ a \ r \Land 

\text{invh} \ t' \land \ bh \ t' = bh \ l \land \ invc2 \ t' \land \ (\text{color} \ l = \text{Black} \longrightarrow \ invc \ t')
The proofs of these lemmas are by case analyses over the defining equations using the characteristic properties of \(baliL\) and \(baliR\) given above.

**Proof.** Lemma 8.2 is proved by induction on the computation of \(del\ x\ t\). The base case is trivial. In the induction step \(t = \langle l, (a, c), r \rangle\) if \(x < a\) then we distinguish three subcases. If \(l = \langle \rangle\) the claim is trivial. Otherwise the claim follows from the IH: if \(color\ l = Red\) then the claim follows directly, if \(color\ l = Black\) then it follows with the help of Lemma 8.4 (with \(l = del\ x\ l\)). The case \(a < x\) is dual and the case \(x = a\) is similar (using Lemma 8.1). We do not show the details because they are tedious but routine.

The proof of Lemma 8.1 is similar but simpler.

### 8.2.3 Deletion by Joining

As an alternative to deletion by replacement we also consider deletion by joining (see Section 5.2.1). The code for red-black trees is shown in Figure 8.3: compared to Figure 8.2, the \(EQ\) case of \(del\) has changed and \(join\) is new.

Invariant preservation is proved much like before except that instead of \(split\_min\) we now have \(join\) to take care of. The characteristic lemma is proved by induction on the computation of \(join\):

**Lemma 8.6.** \(invh\ l \land invh\ r \land bh\ l = bh\ r \land invc\ l \land invc\ r \land\)
\[ t' = join\ l\ r \rightarrow\]
\[ invh\ t' \land bh\ t' = bh\ l \land invc2\ t' \land\]
\[(color\ l = Black \land color\ r = Black \rightarrow invc\ t')\]

### 8.3 Exercises

**Exercise 8.1.** Show that the logarithmic height of red-black trees is already guaranteed by the color and height invariants:

\[ invc\ t \land invh\ t \rightarrow h\ t \leq 2 \cdot \log_2 |t| + 2\]

**Exercise 8.2.** We already discussed informally why the definition of \(invh\) captures “all paths from the root to a leaf have the same number of black nodes” although \(bh\) only traverses the left spine. This exercises formalizes that discussion. The following function computes the set of black heights (number of black nodes) of all paths:

\[ bhs :: \text{`a rbt} \Rightarrow \text{nat set} \]
\[ bhs\ \langle \rangle = \{0\} \]
\[ bhs\ \langle l, (_\_,\ c), r \rangle = (\text{let} \ H = bhs\ l \cup bhs\ r\ \text{in if} \ c = Black\ \text{then} \ Suc\ \text{'} H\ \text{else} \ H)\]
\[
\begin{align*}
\text{del} &:: 'a \Rightarrow 'a \text{ rbt} \Rightarrow 'a \text{ rbt} \\
\text{del} \_\_ () &= () \\
\text{del} \_ (l, (a, \_), r) &= \begin{cases} \\
\text{LT} &\Rightarrow \text{if } l \neq () \land \text{color } l = \text{Black} \text{ then } \text{baldL} (\text{del } x \_ l) \_ a \_ r \text{ else } \text{R} (\text{del } x \_ l) \_ a \_ r \\
\text{EQ} &\Rightarrow \text{join } l \_ r \\
\text{GT} &\Rightarrow \text{if } r \neq () \land \text{color } r = \text{Black} \text{ then } \text{baldR} \_ a \_ (\text{del } x \_ r) \text{ else } \text{R} \_ l \_ a \_ (\text{del } x \_ r) \\
\end{cases}
\end{align*}
\]

\[
\begin{align*}
\text{join} &:: 'a \text{ rbt} \Rightarrow 'a \text{ rbt} \Rightarrow 'a \text{ rbt} \\
\text{join} \_ (t) &= t \\
\text{join} \_ (t1 \_ t2) &= \begin{cases} \\
\text{R} \_ u \_ b \_ u3 &\Rightarrow \text{R} (\text{R} \_ t1 \_ a \_ u2) \_ b \_ (\text{R} \_ u3 \_ c \_ t4) \\
\_ t23 &\Rightarrow \text{R} \_ t1 \_ a \_ (\text{R} \_ t23 \_ c \_ t4) \\
\end{cases}
\end{align*}
\]

Fig. 8.3. Deletion from red-black tree by combining children

where the infix operator (‘) is predefined as \( f ' A = \{ y \mid \exists x \in A. y = f x \} \).

Prove \( \text{invh} t \leftarrow bhs t = \{ bh \ t \} \). The \( \leftarrow \) direction should be easy, the other direction should need some lemmas.

Exercise 8.3. Following Section 7.3, define a linear time function \( \text{rbt_of_list} :: 'a \text{ list} \Rightarrow 'a \text{ rbt} \) and prove both \( \text{inorder} (\text{rbt_of_list as}) = \text{as} \) and \( \text{rbt} (\text{rbt_of_list as}) \).

Bibliographic Remarks

Red-Black trees were invented by Bayer [6] who called them “symmetric binary B-trees”. The red-black color convention was introduced by Guibas and Sedgewick [23] who studied their properties in greater depth. The first functional version of red-black trees (without deletion) is due to Okasaki [62] and everybody follows his code. A functional version of deletion was first given
by Kahrs [36]\footnote{The code for deletion is not in the article but can be retrieved from this URL: http://www.cs.ukc.ac.uk/people/staff/smk/redblack/rb.html} and Section 8.2.3 is based on it. Germane [22] presents a function for deletion by replacement that is quite different from the one in Section 8.2.2. Our starting point were Isabelle proofs by Reiter and Krauss (based on Kahrs). Other verifications of red-black trees are reported by Filliâtre and Letouzey [19] (using their own deletion function) and Appel [4] (based on Kahrs).
AVL Trees

The AVL tree [2] (named after its inventors Adelson-Velsky and Landis) is the granddaddy of efficient binary search trees. Its logarithmic height is maintained by rotating subtrees based on their height. For efficiency reasons the height of each subtree is stored in its root node. That is, the underlying data structure is a height-augmented tree (see Section 4.4):

```
type_synonym 'a tree_ht = ('a × nat) tree
```

Function $ht$ extracts the height field and $node$ is a smart constructor that sets the height field:

```
ht :: 'a tree_ht ⇒ nat
ht () = 0
ht (_, (_, n), _) = n

node :: 'a tree_ht ⇒ 'a ⇒ 'a tree_ht ⇒ 'a tree_ht
node l a r = (l, (a, max (ht l) (ht r) + 1), r)
```

An AVL tree is a tree that satisfies the AVL invariant: the height of the left and right child of any node differ by at most 1

```
avl :: 'a tree_ht ⇒ bool
avl () = True
avl (l, (_, n), r)
  = (|int (ht l) - int (ht r)| ≤ 1 ∧
    n = max (ht l) (ht r) + 1 ∧ avl l ∧ avl r)
```
and the height field contains the correct value. The conversion function \( \text{int} :\to \text{nat} \Rightarrow \text{int} \) is required because on natural numbers \( 0 - n = 0 \).

### 9.1 Logarithmic Height

AVL trees have logarithmic height. The key insight for the proof is that \( M \), the minimal number of leaves of an AVL tree of height \( n \), satisfies the recurrence relation \( M \, (n + 2) = M \, (n + 1) + M \, n \). Instead of formalizing this function \( M \) we prove directly that an AVL tree of height \( n \) has at least \( \text{fib} \, (n + 2) \) leaves where \( \text{fib} \) is the Fibonacci function:

\[
\text{fib} :\to \text{nat} \Rightarrow \text{nat} \\
\text{fib} \, 0 = 0 \\
\text{fib} \, 1 = 1 \\
\text{fib} \, (n + 2) = \text{fib} \, (n + 1) + \text{fib} \, n
\]

**Lemma 9.1.** \( \text{avl} \, t \rightarrow \text{fib} \, (h \, t + 2) \leq |t|_1 \)

*Proof.* The proof is by induction on \( t \). We focus on the induction step \( t = \langle l, (a, n), r \rangle \) and assume \( \text{avl} \, t \). Thus the IHs reduce to \( \text{fib} \, (h \, l + 2) \leq |l|_1 \) and \( \text{fib} \, (h \, r + 2) \leq |r|_1 \). We prove \( \text{fib} \, (\text{max} \, (h \, l) \, (h \, r) + 3) \leq |l|_1 + |r|_1 \), from which \( \text{avl} \, t \rightarrow \text{fib} \, (h \, t + 2) \leq |t|_1 \) follows directly. There are two cases. We focus on \( h \, l \geq h \, r \), \( h \, l < h \, r \) is dual.

\[
\text{fib} \, \text{max} \, (h \, l) \, (h \, r) + 3 = \text{fib} \, (h \, l + 3) \\
= \text{fib} \, (h \, l + 2) + \text{fib} \, (h \, l + 1) \\
\leq |l|_1 + \text{fib} \, (h \, l + 1) \quad \text{by } \text{fib} \, (h \, l + 2) \leq |l|_1 \\
\leq |l|_1 + |r|_1 \quad \text{by } \text{fib} \, (h \, r + 2) \leq |r|_1
\]

The last step is justified because \( h \, l + 1 \leq h \, r + 2 \) (which follows from \( \text{avl} \, t \)) and \( \text{fib} \) is monotone. \( \square \)

Now we prove a well-known exponential lower bound for \( \text{fib} \) where \( \varphi \equiv (1 + \sqrt{5}) / 2 \):

**Lemma 9.2.** \( \varphi^n \leq \text{fib} \, (n + 2) \)

*Proof.* The proof is by induction on \( n \) by \( \text{fib} \) computation induction. The case \( n = 0 \) is trivial and the case \( n = 1 \) is easy. Now consider the induction step:
9.2 Implementation of ADT Set

9.2.1 Insertion

Insertion follows the standard approach: insert the element as usual and reestablish the AVL invariant on the way back up.

```plaintext
insert :: 'a ⇒ 'a tree_ht ⇒ 'a tree_ht
insert x () = ((), (x, 1), ()
insert x (l, (a, n), r) = (case cmp x a of
  LT ⇒ balL (insert x l) a r |
  EQ ⇒ (l, (a, n), r) |
  GT ⇒ balR l a (insert x r))
```

Functions balL/balR readjust the tree after an insertion into the left/right child. The AVL invariant has been lost if the difference in height has become 2 — it cannot become more because the height can only increase by 1. Consider the definition of balL in Figure 9.1 (balR in Figure 9.2 is dual). If the AVL invariant has not been lost, i.e. if \( \text{ht AB} \neq \text{ht C} + 2 \), then we can just return the AVL tree node AB c C. But if \( \text{ht AB} = \text{ht C} + 2 \), we need to “rotate” the subtrees suitably. Clearly \( AB \) must be of the form \( \langle A, (a, \_), B \rangle \). There are two cases, which are illustrated in Figure 9.1. Rectangles denote trees. Rectangles of the same height denote trees of the same height. Rectangles with a +1 denote the additional level due to insertion of the new element.

If \( \text{ht B} \leq \text{ht A} \) then balL performs what is known as a single rotation.

If \( \text{ht A} < \text{ht B} \) then \( B \) must be of the form \( \langle B_1, (b, \_), B_2 \rangle \) (where either \( B_1 \) or \( B_2 \) has increased in height) and balL performs what is known as a double rotation.

It is easy to check that in both cases the tree on the right satisfies the AVL invariant.
avlL :: 'a tree_ht ⇒ 'a ⇒ 'a tree_ht ⇒ 'a tree_ht
avlL AB c C
= (if ht AB = ht C + 2
  then case AB of
    ⟨A, ⟨a, x⟩, B⟩ ⇒
      if ht B ≤ ht A then node A a (node B c C)
      else case B of
        ⟨B1, ⟨b, _⟩, B2⟩ ⇒ node (node A a B1) b (node B2 c C)
    else node AB c C)

Single rotation:

Double rotation:

Fig. 9.1. Function avlL
9.2 Implementation of ADT Set

\[
balR :: 'a tree_ht \Rightarrow 'a \Rightarrow 'a tree_ht \Rightarrow 'a tree_ht
\]

\[
balR A a BC = (if \ \text{ht} \ BC = \text{ht} \ A + 2 \ \text{then case} \ BC \ \text{of}
\begin{align*}
\langle B, (c, x), C \rangle \Rightarrow \\
\text{if} \ \text{ht} \ B \leq \text{ht} \ C \text{ then node (node} \ A \ a \ B) \ c \ C \\
\text{else case} \ B \ \text{of}
\langle B_1, (b, \_), B_2 \rangle \Rightarrow \text{node (node} \ A \ a \ B_1) \ b \ (\text{node} \ B_2 \ c \ C)
\end{align*}
\]

\text{else node} \ A \ a \ BC)
\]

Fig. 9.2. Function \(balR\)

Preservation of \(avl\) by \(insert\) cannot be proved in isolation but needs to be proved simultaneously with how \(insert\) changes the height (because \(avl\) depends on the height and \(insert\) requires \(avl\) for correct behaviour):

**Theorem 9.4.** \(avl \ t \rightarrow avl \ (insert \ x \ t) \land h \ (insert \ x \ t) \in \{h \ t, h \ t + 1\}\)

The proof is by induction on \(t\) followed by a complete case analysis (which Isabelle automates).

9.2.2 Deletion

\[
delete :: 'a \Rightarrow 'a tree_ht \Rightarrow 'a tree_ht
\]

\[
delete \ _ \ () = ()
\]

\[
delete \ x \ h \ (a, \_, r) = \ (\text{case} \ \text{cmp} \ x \ a \ \text{of}
\begin{align*}
\text{LT} & \Rightarrow \text{balR} \ (\text{delete} \ x \ l) \ a \ r \\
\text{EQ} & \Rightarrow \text{if} \ l = () \ \text{then} \ r \ \text{else let} \ (l', a') = \text{split_max} \ l \ \text{in} \ \text{balR} \ l' \ a' \ r \\
\text{GT} & \Rightarrow \text{balL} \ l \ a \ (\text{delete} \ x \ r)
\end{align*}
\]

\[
split_max :: 'a tree_ht \Rightarrow 'a tree_ht \times 'a
\]

\[
split_max \ l \ (a, \_, r) = \ (\text{if} \ r = () \ \text{then} \ (l, a) \ \text{else let} \ (r', a') = \text{split_max} \ r \ \text{in} \ \text{balL} \ l \ a \ r' \ a'))
\]

Fig. 9.3. Deletion from AVL tree

Figure 9.3 shows deletion-by-replacing (see 5.2.1). The recursive calls are dual to insertion: in terms of the difference in height, deletion of some element
from one child is the same as insertion of some element into the other child. Thus functions \( \text{balR} / \text{balL} \) can again be employed to restore the invariant.

An element is deleted from a node by replacing it with the maximal element of the left child (the minimal element of the right child would work just as well). Function \( \text{split\_max} \) performs that extraction and uses \( \text{balL} \) to restore the invariant after splitting an element off the right child.

The fact that \( \text{balR} / \text{balL} \) can be reused for deletion can be illustrated by drawing the corresponding rotation diagrams. We look at how the code for \( \text{balL} \) behaves when an element has been deleted from \( C \). Dashed rectangles indicate a single additional level that may or may not be there. The label -1 indicates that the level has disappeared due to deletion.

Single rotation in \( \text{balL} \) after deletion in \( C \):

Double rotation in \( \text{balL} \) after deletion in \( C \):

At least one of \( B_1 \) and \( B_2 \) must have the same height as \( A \).

Preservation of \text{avl} by \text{delete} can be proved in the same manner as for \text{insert} but we provide more of the details (partly because our Isabelle proof is less automatic). The following lemmas express that the auxiliary functions preserve \text{avl}:

\[
\text{avl } l \land \text{avl } r \land h_r - 1 \leq h_l \land h_l \leq h_r + 2 \implies \text{avl } (\text{balL } l \ a \ r)
\]
avl l ∧ avl r ∧ h l - 1 ≤ h r ∧ h r ≤ h l + 2 → avl (balR l a r)
avl t ∧ t ≠ () →
avl (fst (split_max t)) ∧
  h t ∈ {h (fst (split_max t)), h (fst (split_max t)) + 1}

The first two are proved by the obvious cases analyses, the last one also requires induction.

As for insert, preservation of avl by delete needs to be proved simultaneously with how delete changes the height:

**Theorem 9.5.** \(avl t \land t' = delete x t \rightarrow avl t' \land h t \in \{h t', h t' + 1\}\)

**Proof.** The proof is by induction on \(t\) followed by the case analyses dictated by the code for delete. We sketch the induction step. Let \(t = \langle l, (a, n), r \rangle\) and \(t' = delete x t\) and assume the IHs and \(avl t\). The claim \(avl t'\) follows from the preservation of \(avl\) by \(balL, balR\) and \(split_max\) as shown above. The claim \(h t \in \{h t', h t' + 1\}\) follows directly from the definitions of \(balL\) and \(balR\).

9.3 Exercises

**Exercise 9.1.** The logarithmic height of AVL trees can be proved directly. Prove

\[avl t ∧ h t = n \rightarrow 2^n \div 2 \leq |t|_1\]

by \(fib\) computation induction on \(n\). This implies \(avl t \rightarrow h t \leq 2 \cdot \lg |t|_1\).

**Exercise 9.2.** Fibonacci trees are defined in analogy to Fibonacci numbers:

\[
fibt :: \text{nat} \Rightarrow \text{unit tree}
fibt 0 = ()
fibt 1 = ()
fibt (n + 2) = (fibt (n + 1), (), fibt n)
\]

We are only interested in the shape of these trees. Therefore the nodes just contain dummy unit values (). Hence we need to define the AVL invariant again for trees without annotations:

\[
avl0 :: \text{a tree} \Rightarrow \text{bool}
avl0 () = True
avl0 (l, _, r) = (|\text{int} (h l) - \text{int} (h r)| \leq 1 ∧ avl0 l ∧ avl0 r)
\]

Prove the following properties of Fibonacci trees:

\[avl0 (fibt n) \Rightarrow |fibt n|_1 = fib (n + 2)\]
Conclude that the Fibonacci trees are minimal (w.r.t. their size) among all AVL trees of a given height:
\[ \text{avl } t \rightarrow |\text{fibt } (h \ t)|_1 \leq |t|_1 \]

Exercise 9.3. Show that every almost complete tree is an AVL tree:
\[ \text{acomp}lete \ t \rightarrow \text{avl } 0 \ t \]
As in the previous exercise we consider trees without height annotations.

Exercise 9.4. Generalize AVL trees to height-balanced trees where the condition
\[ |\text{int } (h \ l) - \text{int } (h \ r)| \leq 1 \]
in the invariant is replaced by
\[ |\text{int } (h \ l) - \text{int } (h \ r)| \leq m \]
where \( m \geq 1 \) is some fixed integer. Modify the invariant and the insertion and deletion functions and prove that the latter fulfill the same correctness theorems as before. You do not need to prove the logarithmic height of height-balanced trees.

Exercise 9.5. Following Section 7.3, define a linear-time function \( \text{avl}_\text{of}_\text{list} :: \text{'a list} \Rightarrow \text{'a tree } \text{ht} \) and prove both \( \text{inorder } (\text{avl}_\text{of}_\text{list} \ as) = \ as \) and \( \text{avl } (\text{avl}_\text{of}_\text{list} \ as) \).

9.4 An Optimization

Instead of recording the height of the tree in each node, it suffices to record the balance factor, i.e. the difference in height of its two children. Rather than the three integers \(-1, 0, 1\) we utilize a new data type:

```
datatype bal = Lh | Bal | Rh

type_synonym 'a tree_bal = ('a × bal) tree
```

The names \( Lh \) and \( Rh \) stand for “left-heavy” and “right-heavy”. The AVL invariant for these trees reflect these names:
The code for insertion (and deletion) is similar to the height-based version. The key difference is that the test if the AVL invariant as been lost cannot be based on the height anymore. We need to detect if the tree has increased in height upon insertion based on the balance factors. The key insight is that a height increase is coupled with a change from `Bal` to `Lh` or `Rh`. Except when we transition from `()` to `(Lh, (a, Bal), ())`. This insight is encoded in the test `incr`:

```haskell
is_bal :: 'a tree_bal ⇒ bool
is_bal (l, (_, b), _) = (b = Bal)

incr :: 'a tree_bal ⇒ 'b tree_bal ⇒ bool
incr t t' = (t = ()) ∨ is_bal t ∧ ¬ is_bal t'
```

The test for a height increase compares the trees before and after insertion. Therefore it has been pulled out of the balance functions into insertion:

```haskell
insert :: 'a ⇒ 'a tree_bal ⇒ 'a tree_bal
insert x () = (()), (x, Bal), ()
insert x (l, (a, b), r)
  = (case cmp x a of
      LT ⇒ let l' = insert x l
        in if incr l l' then balL l' a b r else (l', (a, b), r) |
      EQ ⇒ (l, (a, b), r) |
      GT ⇒ let r' = insert x r
        in if incr r r' then balR l a b r' else (l, (a, b), r'))
```

The balance functions are shown in Figure 9.4. Function `rot2` implements double rotations. Function `balL` is called if the left child `AB` has increased in height. If the tree was `Lh` then single or double rotations are necessary.
to restore balance. Otherwise we simply need to adjust the balance factors. Function \texttt{balR} is dual to \texttt{balL}.

For deletion we need to test if the height has decreased and \texttt{decr} implements this test:

\begin{verbatim}
decr :: 'a tree_bal ⇒ 'b tree_bal ⇒ bool
decl t t′ = (t ≠ ∅) ∧ (t′ = ∅) ∨ is_bal t ∧ is_bal t′)
\end{verbatim}

The functions \texttt{incr} and \texttt{decr} are almost dual except that \texttt{incr} implicitly assumes \(t′ ≠ ∅\) because insertion is guaranteed to return a \textit{Node}. Thus we could use \texttt{decr} instead of \texttt{incr} but not the other way around.

Deletion and \texttt{split_max} change in the same manner as insertion:
delete :: 'a ⇒ 'a tree_bal ⇒ 'a tree_bal

delete _ ⟨⟩ = ⟨⟩
delete x ⟨l, (a, ba), r⟩ = (case cmp x a of
  LT ⇒ let l' = delete x l
         in if decr l l' then balR l' a ba r else ⟨l', (a, ba), r⟩
  | EQ ⇒ if l = ⟨⟩ then r
         else let (l', a') = split_max l
                in if decr l l' then balR l' a' ba r
                     else ⟨l', (a', ba), r⟩
  | GT ⇒ let r' = delete x r
         in if decr r r' then balL l a ba r' else ⟨l, (a, ba), r'⟩)

split_max :: 'a tree_bal ⇒ 'a tree_bal 

split_max ⟨l, (a, ba), r⟩ = (if r = ⟨⟩ then ⟨l, a⟩
   else let (r', a') = split_max r;
         t' = if decr r r' then balL l a ba r' else ⟨l, (a, ba), r'⟩
         in (t', a'))

In the end we have the following correctness theorems:

**Theorem 9.6.** \(avl \ t \land t' = insert \ x \ t \rightarrow\)
\(avl \ t' \land h \ t' = h \ t + (if \ incr \ t \ t' \ then \ 1 \ else \ 0)\)

This theorem tells us not only that \(avl\) is preserved but also that \(incr\) indicates correctly if the height has increased or not.

Similarly for deletion and \(decr\):

**Theorem 9.7.** \(avl \ t \land t' = delete \ x \ t \rightarrow\)
\(avl \ t' \land h \ t = h \ t' + (if \ decr \ t \ t' \ then \ 1 \ else \ 0)\)

The proofs of both theorems follow the standard pattern of induction followed by an exhaustive (automatic) cases analysis. The proof for \(delete\) requires an analogous lemma for \(split_max\):

\(split_max \ t = (t', a) \land avl \ t \land t \neq ⟨⟩ \rightarrow\)
\(avl \ t' \land h \ t = h \ t' + (if \ decr \ t \ t' \ then \ 1 \ else \ 0)\)

9.5 Exercises

**Exercise 9.6.** We map type \('a \ tree_bal\) back to type \('a \times\ nat\) \ tree called \('a \ tree_ht\) in the beginning of the chapter:
debal :: 'a tree_bal ⇒ ('a × nat) tree

debal () = ()
debal (t, (a, _, r)) = (debal l, (a, max (h l) (h r) + 1), debal r)

Prove that the AVL property is preserved: avl t → avl_ht (debal t) where avl_ht is the avl predicate on type 'a tree_ht from the beginning of the chapter.

Define a function debal2 of the same type that traverses the tree only once and in particular does not use function h. Prove avl t → debal2 t = debal t.

Exercise 9.7. To realize the full space savings potential of balance factors we encode them directly into the node constructors and work with the following special tree type:

datatype 'a tree4 = Leaf
  | Lh ('a 'a tree4) 'a ('a 'a tree4)
  | Bal ('a 'a tree4) 'a ('a 'a tree4)
  | Rh ('a 'a tree4) 'a ('a 'a tree4)

On this type define the AVL invariant, insertion, deletion and all necessary auxiliary functions. Prove theorems 9.6 and 9.7. Hint: modify the theory underlying Section 9.4.
Beyond Insert and Delete: UNION, INTERSECTION AND DIFFERENCE

So far we looked almost exclusively at insertion and deletion of single elements, with the exception of the conversion of whole lists of elements into search trees (see Section 7.3 and Exercises 8.3 and 9.5). This chapter is dedicated to operations that combine two sets (implemented by search trees) by union, intersection and difference. We denote set difference by $-$ rather than \.

Let us focus on set union for a moment and assume that insertion into a set of size $s$ takes time proportional to $\lg s$. Consider two sets $A$ and $B$ of size $m$ and $n$ where $m \leq n$. The naive approach is to insert the elements from one set one by one into the other set. This takes time proportional to $\lg n + \cdots + \lg (n + m - 1)$ or $\lg m + \cdots + \lg (m + n - 1)$ depending on whether the smaller set is inserted into the larger one or the other way around. Of course the former sum is less than or equal to the latter sum. To estimate the growth of $\lg n + \cdots + \lg (n + m - 1) = \lg (n \cdots (n + m - 1))$ we can easily generalize the derivation of $\lg (n!) \in \Theta (n \lg n)$ in the initial paragraph of Section 7.3. The result is $\lg (n \cdots (n + m - 1)) \in \Theta (m \lg n)$. That is, inserting $m$ elements into an $n$ element set one by one takes time $\Theta (m \lg n)$.

There is a second, possibly naive sounding algorithm for computing the union: flatten both trees to ordered lists (using function \textit{inorder2} from Exercise 4.1), merge both lists and convert the resulting list back into a suitably balanced search tree. All three steps take linear time. The last step is the only slightly nontrivial one but has been dealt with before (see Section 7.3 and Exercises 8.3 and 9.5). This algorithm takes time $O(m + n)$ which is significantly better than $O(m \lg n)$ if $m \approx n$ but significantly worse if $m \ll n$.

This chapter is concerned with a third approach that has the following salient features:

- Union, intersection and difference take time $O(m \lg (\frac{n}{m} + 1))$
- It works for a whole class of balanced trees, including AVL, red-black and weight-balanced trees.
10 Beyond Insert and Delete: \( \cup, \cap \) and \(-\)

- It is based on a single function for joining two balanced trees to form a new balanced tree.

We call it the join approach. It is easily and efficiently parallelizable, a property we will not explore here.

The join approach is at least as fast as the one-by-one approach: from \( m + n \leq mn \) it follows that \( \frac{n}{m} + 1 \leq n \) (if \( m, n \geq 2 \)). The join approach is also at least as fast as the tree-to-list-to-tree approach because \( m + n = m(\frac{n}{m} + 1) \) (if \( m \geq 1 \)).

10.1 Specification of Union, Intersection and Difference

Before explaining the join approach we extend the ADT Set by three new functions union, inter and diff. The specification in Figure 10.1 is self-explanatory.

\[
\text{ADT Set2 = Set +}
\]

\text{interface}

\begin{align*}
\text{union} :: 's \Rightarrow 's \Rightarrow 's \\
\text{inter} :: 's \Rightarrow 's \Rightarrow 's \\
\text{diff} :: 's \Rightarrow 's \Rightarrow 's
\end{align*}

\text{specification}

\begin{align*}
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{invar } (\text{union } s_1 s_2) & \quad \text{(union-inv)} \\
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{set } (\text{union } s_1 s_2) = \text{set } s_1 \cup \text{set } s_2 & \quad \text{(union)} \\
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{invar } (\text{inter } s_1 s_2) & \quad \text{(inter-inv)} \\
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{set } (\text{inter } s_1 s_2) = \text{set } s_1 \cap \text{set } s_2 & \quad \text{(inter)} \\
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{invar } (\text{diff } s_1 s_2) & \quad \text{(diff-inv)} \\
i \text{invar } s_1 \land \text{invar } s_2 \rightarrow \text{set } (\text{diff } s_1 s_2) = \text{set } s_1 - \text{set } s_2 & \quad \text{(diff)}
\end{align*}

Fig. 10.1. ADT Set2

10.2 Just Join

Now we come to the heart of the matter, the definition of union, intersection and difference in terms of a single function join. We promised that the algorithms would be generic across a range of balanced trees. Thus we assume that we operate on augmented trees of type ('a \times 'b) tree where 'a is the type of the elements and 'b is the balancing information (which we can ignore...
This enables us to formulate the algorithms via pattern-matching. A more generic approach is the subject of Exercise 10.1.

The whole section is parameterized by the join function and an invariant:

\[
\text{join} :: ('a \times 'b) \text{ tree} \Rightarrow 'a \Rightarrow ('a \times 'b) \text{ tree} \Rightarrow ('a \times 'b) \text{ tree}
\]

\[
\text{inv} :: ('a \times 'b) \text{ tree} \Rightarrow \text{bool}
\]

Function \text{inv} is meant to take care of the balancedness property only, not the BST property. Functions \text{join} and \text{inv} are specified with the help of the standard tree functions \text{set\_tree} and \text{bst} in Figure 10.2. With respect to the set of elements, \text{join} must behave like union. But it need only return a BST if both trees are BSTs and the element \(a\) lies in between the elements of the two trees, i.e. if \(\text{bst} \langle l, (a, _), r \rangle\). The structural invariant \text{inv} must be preserved by formation and destruction of trees. Thus we can see \text{join} as a smart constructor that builds a balanced tree.

To define union and friends we need a number of simple auxiliary functions shown in Figure 10.3. Function \text{split\_min} decomposes a tree into its leftmost (minimal) element and the remaining tree; the remaining tree is reassembled via \text{join}, thus preserving \text{inv}. Function \text{join2} is reduced to \text{join} with the help of \text{split\_min}. Function \text{split} splits a BST w.r.t. a given element \(a\) into a triple \((l, b, r)\) such that \(l\) contains the elements less than \(a\), \(r\) contains the elements greater than \(a\), and \(b\) is true iff \(a\) was in the input tree.

Although insertion and deletion could be defined by means of union and difference, we can define them directly from the auxiliary functions:

\[
\text{insert} :: 'a \Rightarrow ('a \times 'b) \text{ tree} \Rightarrow ('a \times 'b) \text{ tree}
\]

\[
\text{insert} \; x \; t = (\text{let} \; (l, b, r) = \text{split} \; t \; x \; \in \; \text{join} \; l \; x \; r)
\]

\[
\text{delete} :: 'a \Rightarrow ('a \times 'b) \text{ tree} \Rightarrow ('a \times 'b) \text{ tree}
\]

\[
\text{delete} \; x \; t = (\text{let} \; (l, b, r) = \text{split} \; t \; x \; \in \text{join2} \; l \; r)
\]

The efficiency can be improved a little by taking the returned \(b\) into account.
split_min :: ('a \times 'b) tree \Rightarrow 'a \times ('a \times 'b) tree

\[ \text{split_min} \langle l, (a, _), r \rangle = \begin{cases} (a, r) & \text{if } l = () \\ \text{else let } (m, l') = \text{split_min} l \text{ in } (m, \text{join} l' a r) \end{cases} \]

join2 :: ('a \times 'b) tree \Rightarrow ('a \times 'b) tree \Rightarrow ('a \times 'b) tree

\[ \text{join2} \langle \rangle = l \]

\[ \text{join2} \langle r \rangle = (\text{let } (m, r') = \text{split_min} r \text{ in } \text{join} l m r') \]

split :: ('a \times 'b) tree \Rightarrow ('a \times 'b) tree \times \text{bool} \times ('a \times 'b) tree

\[ \text{split} \langle \rangle = (\langle \rangle, \text{False}, \langle \rangle) \]

\[ \text{split} \langle l, (a, _), r \rangle x = \begin{cases} \text{case} \text{cmp} x a \text{ of} \\ \text{LT} \Rightarrow \text{let } (l_1, b, l_2) = \text{split} l x \text{ in } (l_1, b, \text{join} l_2 a r) | \\ \text{EQ} \Rightarrow (l, \text{True}, r) | \\ \text{GT} \Rightarrow \text{let } (r_1, b, r_2) = \text{split} r x \text{ in } (\text{join} l a r_1, b, r_2) \end{cases} \]

Fig. 10.3. Auxiliary functions

But we have bigger functions to fry: union, intersection and difference. They are shown in Figure 10.4. All three are divide-and-conquer algorithms that follow the same schema: both input trees are split at an element \( a \) (by construction or explicitly), the algorithm is applied recursively to the two trees of the elements below \( a \) and to the two trees of the elements above \( a \), and the two results are suitably joined.

10.2.1 Correctness

We need to prove that \textit{union}, \textit{inter} and \textit{diff} satisfy the specification in Figure 10.1 where \textit{set} = \text{set_tree} and \textit{invar} \textit{t} = \text{inv} \textit{t} \land \text{bst} \textit{t}. That is, for each function we show its set-theoretic property and preservation of \textit{inv} and \textit{bst} using the assumptions in Figure 10.2. Most of the proofs in this section are obvious and automatic inductions and we do not discuss them.

First we need to prove suitable properties of the auxiliary functions \textit{split_min}, \textit{join2} and \textit{split}:

\[ \text{split_min} t = (m, t') \land t \neq () \longrightarrow \]

\[ m \in \text{set_tree} t \land \text{set_tree} t' = \{m\} \cup \text{set_tree} t' \]

\[ \text{split_min} t = (m, t') \land \text{bst} t \land t \neq () \longrightarrow \]

\[ \text{bst} t' \land (\forall x \in \text{set_tree} t'. m < x) \]

\[ \text{split_min} t = (m, t') \land \text{inv} t \land t \neq () \longrightarrow \text{inv} t' \]
union :: \( ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree

\[ \text{union } \{ \} \ t = t \]
\[ \text{union } t \{ \} = t \]
\[ \text{union } \langle l_1, (a, _), r_1 \rangle \ t_2 \]
\[ = (\text{let } (l_2, b_2, r_2) = \text{split } t_2 \ a; \]
\[ \quad \text{in } \text{join } \langle \text{union } l_1 \ l_2 \rangle \ a (\text{union } r_1 \ r_2) \]

inter :: \( ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree

\[ \text{inter } \{ \} \ t = \{ \]
\[ \text{inter } t \{ \} = \{ \]
\[ \text{inter } \langle l_1, (a, _), r_1 \rangle \ t_2 \]
\[ = (\text{let } (l_2, b_2, r_2) = \text{split } t_2 \ a; \]
\[ \quad l' = \text{inter } l_1 \ l_2; \ r' = \text{inter } r_1 \ r_2 \]
\[ \quad \text{in } \text{if } b_2 \ \text{then } \text{join } l' \ a \ r' \ \text{else } \text{join2 } l' \ r' \)

diff :: \( ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree \( \Rightarrow ('a \times 'b) \) tree

\[ \text{diff } \{ \} \ t = \{ \]
\[ \text{diff } t \{ \} = t \]
\[ \text{diff } t \ (l_2, (a, _), r_2) \]
\[ = (\text{let } (l_1, b_1, r_1) = \text{split } t_1 \ a; \]
\[ \quad \text{in } \text{join2 } (\text{diff } l_1 \ l_2) (\text{diff } r_1 \ r_2) \)

Fig. 10.4. Union, intersection and difference

\[
\text{set_tree } (\text{join2 } l \ r) = \text{set_tree } l \cup \text{set_tree } r \tag{10.5}
\]
\[
\text{bst } l \land \text{bst } r \land (\forall x \in \text{set_tree } l. \ \forall y \in \text{set_tree } r. \ x < y) \implies \\
\text{bst } (\text{join2 } l \ r)
\]
\[
\text{inv } l \land \text{inv } r \implies \text{inv } (\text{join2 } l \ r)
\]
\[
\text{split } t \ x = (l, b, r) \land \text{bst } t \implies \\
\text{set_tree } l = \{a \in \text{set_tree } t \mid a < x\} \land \\
\text{set_tree } r = \{a \in \text{set_tree } t \mid x < a\} \land \\
\text{bst } l \land \text{bst } r
\]
\[
\text{split } t \ x = (l, b, r) \land \text{inv } t \implies \text{inv } l \land \text{inv } r \tag{10.6}
\]

The correctness properties of \text{insert} and \text{delete} are trivial consequences and are not shown. We move on to \text{union}. Its correctness properties are concretizations of the properties (union) and (union-inv) in Figure 10.1:

\[
\text{bst } t_2 \implies \text{set_tree } (\text{union } t_1 \ t_2) = \text{set_tree } t_1 \cup \text{set_tree } t_2
\]
\[
\text{bst } t_1 \land \text{bst } t_2 \implies \text{bst } (\text{union } t_1 \ t_2)
\]
\[
\text{inv } t_1 \land \text{inv } t_2 \implies \text{inv } (\text{union } t_1 \ t_2)
\]
All three union properties are proved by computation induction. The first property follows easily from assumption (10.1) and (10.6). The assumption \( bst \ t_2 \) (but not \( bst \ t_1 \)) is required because \( t_2 \) is split and (10.6) requires \( bst \). Preservation of \( bst \) follows from assumption (10.2) with the help of the first union property and the preservation of \( bst \) by \( split \). Preservation of \( inv \) follows from assumptions (10.3) and (10.4) with the help of the preservation of \( inv \) by \( split \).

The correctness properties of \( inter \) look similar:

\[
\begin{align*}
 bst \ t_1 \land bst \ t_2 \rightarrow set_tree \ (inter \ t_1 \ t_2) &= set_tree \ t_1 \cap set_tree \ t_2 \\
 bst \ t_1 \land bst \ t_2 \rightarrow bst \ (inter \ t_1 \ t_2) \\
 inv \ t_1 \land inv \ t_2 \rightarrow inv \ (inter \ t_1 \ t_2)
\end{align*}
\]

The proof of the preservation properties are automatic but the proof of the \( set_tree \) property is more involved than the corresponding proof for \( union \) and we take a closer look at the induction. We focus on the case \( t_1 = (l_1, (a, _), r_1) \) and \( t_2 \neq (\) \). Let \( L_1 = set_tree \ l_1 \) and \( R_1 = set_tree \ r_1 \). Let \( (l_2, \ b, \ r_2) = split \ t_2 \ a, \ L_2 = set_tree \ l_2, \ R_2 = set_tree \ r_2 \) and \( A = (if \ b \ then \ {a} \ else \ {}) \). The separation properties

\[
\begin{align*}
 a \notin L_1 \cup R_1 & \quad a \notin L_2 \cup R_2 \\
 L_2 \cap R_2 = \{} & \quad L_1 \cap R_2 = \{} & \quad L_2 \cap R_1 = \{}
\end{align*}
\]

follow from \( bst \ t_1, \ bst \ t_2 \) and (10.6). Now for the main proof:

\[
\begin{align*}
 set_tree \ t_1 \cap set_tree \ t_2 &= \ (L_1 \cup R_1 \cup \{a\}) \cap \ (L_2 \cup R_2 \cup A) & \text{by (10.6), \( bst \ t_2 \)} \\
 &= L_1 \cap L_2 \cup R_1 \cap R_2 \cup A & \text{by the separation properties} \\
 &= set_tree \ (inter \ t_1 \ t_2) & \text{by (10.1), (10.5), IHs, \( bst \ t_1, \ bst \ t_2), (10.6) } \\
\end{align*}
\]

The correctness properties of \( diff \) follow the same pattern and their proofs are similar to the proofs of the \( inter \) properties. This concludes the generic join approach.

### 10.3 Joining Red-Black Trees

This section shows how to implement \( join \) efficiently on red-black trees. The basic idea is simple: descend along the spine of the higher of the two trees until reaching a subtree whose height is the same as the height of the lower tree. With suitable changes this works for other balanced trees as well [9]. The function definitions are shown in Figure 10.5. Function \( join \) calls \( joinR \) (descending along the right spine of \( l \)) if \( l \) is the higher tree, or calls \( joinL \) (descending along the left spine of \( r \)) if \( r \) is the higher tree, or returns \( B \ l x r \).
10.3 Joining Red-Black Trees

\[ \text{joinL} :: 'a rbt \Rightarrow 'a \Rightarrow 'a rbt \Rightarrow 'a rbt \]
\[ \begin{align*}
\text{joinL } l x r &= (\text{if } bh r \leq bh l \text{ then } R l x r \\
&\quad \text{else case } r \text{ of} \\
&\quad\quad (l', (x', \text{Red}), r') \Rightarrow R (\text{joinL } l x l') x' r' \\
&\quad\quad (l', (x', \text{Black}), r') \Rightarrow \text{baliL} (\text{joinL } l x l') x' r')
\end{align*} \]

\[ \text{joinR} :: 'a rbt \Rightarrow 'a \Rightarrow 'a rbt \Rightarrow 'a rbt \]
\[ \begin{align*}
\text{joinR } l x r &= (\text{if } bh l \leq bh r \text{ then } R l x r \\
&\quad \text{else case } l \text{ of} \\
&\quad\quad (l', (x', \text{Red}), r') \Rightarrow R l' x' (\text{joinR } r' x r) \\
&\quad\quad (l', (x', \text{Black}), r') \Rightarrow \text{baliR} l' x' (\text{joinR } r' x r)
\end{align*} \]

\[ \text{join} :: 'a rbt \Rightarrow 'a \Rightarrow 'a rbt \Rightarrow 'a rbt \]
\[ \begin{align*}
\text{join } l x r &= (\text{if } bh r < bh l \text{ then } \text{paint Black} (\text{joinR } l x r) \\
&\quad \text{else if } bh l < bh r \text{ then } \text{paint Black} (\text{joinL } l x r) \text{ else } B l x r)
\end{align*} \]

Fig. 10.5. Function \textit{join} on red-black trees

otherwise. The running time is linear in the black height (and thus logarithmic in the size) if we assume that the black height is stored in each node; our implementation of red-black trees would have to be augmented accordingly. Note that in \text{joinR} (and similarly in \text{joinL}) the comparison is not \( bh l = bh r \) but \( bh l \leq bh r \) to simplify the proofs.

10.3.1 Correctness

We need to prove that \textit{join} on red-black trees (and a suitable \textit{inv}) satisfies its specification in Figure 10.2. We start with properties of \text{joinL}; the properties of function \text{joinR} are completely symmetric. These are the automatically provable inductive properties:

- \( \text{invc } l \land \text{invc } r \land \text{invh } l \land \text{invh } r \land bh l \leq bh r \rightarrow \text{invc2} (\text{joinL } l x r) \land (bh l \neq bh r \land \text{color } r = \text{Black} \rightarrow \text{invc} (\text{joinL } l x r)) \land \text{invh} (\text{joinL } l x r) \land bh (\text{joinL } l x r) = bh r \)
- \( bh l \leq bh r \rightarrow \text{set_tree} (\text{joinL } l x r) = \text{set_tree} l \cup \{x\} \cup \text{set_tree } r \land \text{bst} (l, (a, n), r) \land bh l \leq bh r \rightarrow \text{bst} (\text{joinL } l a r) \)

Because \text{joinL} employs \text{baliL} from the chapter on red-black trees, the proof of the first property makes use of the \text{baliL} property shown in Section 8.2.1.
We define the invariant $\text{inv}$ required for the specification in Figure 10.2 as follows:

$$\text{inv} t = (\text{invc} t \land \text{inv} h t)$$

Although weaker than $r\text{bt}$, it still guarantees logarithmic height (see Exercise 8.1). Note that $r\text{bt}$ itself does not work because it does not satisfy property (10.4). The properties of $\text{join}$ and $\text{inv}$ are now easy consequences of the $\text{joinL}$ (and $\text{joinR}$) properties shown above.

### 10.4 Exercises

**Exercise 10.1.** Define an alternative version $\text{diff1}$ of $\text{diff}$ where in the third equation pattern matching is on $t_1$ and $t_2$ is $\text{split}$. Prove $\text{bst} t_1 \land \text{bst} t_2 \rightarrow \text{set_tree} (\text{diff1} t_1 t_2) = \text{set_tree} t_1 - \text{set_tree} t_2$.

**Exercise 10.2.** Following the general idea of the join function for red-black trees, define a join function for 2-3-trees. Start with two functions $\text{joinL}$, $\text{joinR} :: 'a tree23 \Rightarrow 'a \Rightarrow 'a tree23 \Rightarrow 'a upI$ and combine them into the overall join function:

$$\text{join} :: 'a tree23 \Rightarrow 'a \Rightarrow 'a tree23 \Rightarrow 'a tree23$$

Prove the following correctness properties:

- $\text{complete l} \land \text{complete r} \rightarrow \text{complete (join l x r)}$
- $\text{complete l} \land \text{complete r} \rightarrow \text{inorder (join l x r)} = \text{inorder l} \oplus x \neq \text{inorder r}$

The corresponding (and needed) properties of $\text{joinL}$ and $\text{joinR}$ are slightly more involved.

### Bibliographic Remarks

The join approach goes back to Adams [1]. Blelloch et al. [9] generalized the approach from weight-balanced trees to AVL trees, red-black trees and treaps. In particular they proved the $O(m \lg(\frac{m}{n} + 1))$ bound for the work (and an $O(lg m lg n)$ bound for the span).
Arrays via Braun Trees

Braun trees are a subclass of almost complete trees. In this chapter we explore their use as arrays and in Chapter 16 as priority queues.

11.1 Array

So far we have discussed sets (or maps) over some arbitrary linearly ordered type. Now we specialize that linearly ordered type to \( \text{nat} \) to model arrays. In principle we could model arrays as maps from a subset of natural numbers to the array elements. Because arrays are contiguous, it is more appropriate to model them as lists. The type ‘a list comes with two array-like operations (see Appendix A):

**Indexing:** \( xs ! n \) is the \( n \)th element of the list \( xs \).

**Updating:** \( xs[n := x] \) is \( xs \) with the \( n \)th element replaced by \( x \).

By convention, indexing starts with \( n = 0 \). If \( n \geq \text{length} \) then \( xs ! n \) and \( xs[n := x] \) are underdefined: they are defined terms but we do not know what their value is.

Note that operationally, indexing and updating take time linear in the index, which may appear inappropriate for arrays. However, the type of lists is only an abstract model that specifies the desired functional behaviour of arrays but not their running time complexity.

The ADT of arrays is shown in Figure 11.1. Type ‘a ar is the type of arrays, type ‘a the type of elements in the arrays. The abstraction function list abstracts arrays to lists. It would make perfect sense to include list in the interface as well. In fact, our implementation below comes with a (reasonably efficiently) executable definition of list.

The behaviour of lookup, update, size and array is specified in terms of their counterparts on lists and requires that the invariant is preserved.
ADT Array =

interface
lookup :: 'ar ⇒ nat ⇒ 'a
update :: nat ⇒ 'a ⇒ 'ar ⇒ 'ar
len :: 'ar ⇒ nat
array :: 'a list ⇒ 'ar
abstraction list :: 'ar ⇒ 'a list
invariant invar :: 'ar ⇒ bool

specification
invar ar ∧ n < len ar → lookup ar n = list ar ! n    (lookup)
invar ar ∧ n < len ar → invar (update n x ar)      (update-inv)
invar ar ∧ n < len ar → list (update n x ar) = (list ar)[n := x]   (update)
invar ar → len ar = |list ar|                       (len)
invar (array xs)                                    (array-inv)
list (array xs) = xs                                 (array)

Fig. 11.1. ADT Array

What distinguishes the specifications of lookup and update from the standard schema (see Chapter 6) is that they carry a size precondition because the result of lookup and update is only specified if the index is less than the size of the array.

11.2 Braun Trees

One can implement arrays by any one of the many search trees presented in this book. Instead we take advantage of the fact that the keys are natural numbers and implement arrays by so-called Braun trees which are almost complete and thus have minimal height.

The basic idea is to index a node in a binary tree by the non-zero bit string that leads from the root to that node in the following fashion. Starting from the least significant bit and while we have not reached the leading 1 (which is ignored), we examine the bits one by one. If the current bit is 0, descend into the left child, otherwise into the right child. Instead of bit strings we use the natural numbers \( \geq 1 \) that they represent. The Braun tree with nodes indexed by 1–15 is shown in Figure 11.2. The numbers are the indexes and not the elements stored in the nodes. For example, the index 14 is 0111 in binary (least significant bit first). If you follow the path left-right-right corresponding to 011 in Figure 11.2 you reach node 14.
A tree $t$ is suitable for representing an array if the set of indexes of all its nodes is the interval $\{1..|t|\}$. The following tree is unsuitable because the node indexed by 2 is missing:

\[ \begin{array}{c c c c c c c c c c c c c}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15
\end{array} \]

It turns out that the following invariant guarantees that a tree $t$ contains exactly the nodes indexed by $1, ..., |t|$:

\[
\text{braun} :: \text{a tree} \Rightarrow \text{bool}
\]

\[
\text{braun} \langle \rangle = \text{True}
\]

\[
\text{braun} \langle l, _, r \rangle = (|l| = |r| \lor |l| = |r| + 1) \land \text{braun} \ l \land \text{braun} \ r
\]

The disjunction can alternatively be expresses as $|r| \leq |l| \leq |r| + 1$. We call a tree a Braun tree iff it satisfies predicate `braun`. Although we do not need or prove this here, it is interesting to note that a tree that contains exactly the nodes indexed by $1, ..., |t|$ is a Braun tree.

Let us now prove the earlier claim that Braun trees are almost complete. First, a lemma about the composition of almost complete trees:

**Lemma 11.1.**

$\text{acomplete} \ l \land \text{acomplete} \ r \land |l| = |r| + 1 \rightarrow \text{acomplete} \ <l, x, r>$

**Proof.** Using Lemmas 4.7 and 4.8 and the assumptions we obtain

\[
h \langle l, x, r \rangle = \lfloor \log (|r|_1 + 1) \rfloor + 1 \quad (*)
\]

\[
mh \langle l, x, r \rangle = \lfloor \log |r|_1 \rfloor + 1 \quad (**)\]

Because $1 \leq |r|_1$, there is an $i$ such that $2^i \leq |r|_1 < 2^{i+1}$ and thus $2^i \leq |r|_1 + 1 \leq 2^{i+1}$. This implies $i = \lfloor \log |r|_1 \rfloor$ and $i + 1 = \lfloor \log (|r|_1 + 1) \rfloor$. Together with $(*)$ and $(**)$ this implies $\text{acomplete} \ <l, x, r>$. $\square$
Now we can show that all Braun trees are almost complete:

**Lemma 11.2.** $\text{braun } t \rightarrow \text{acomp} l e t$

Thus we know that Braun trees have optimal height (Lemma 4.6) and can even quantify it (Lemma 4.7).

**Proof.** The proof is by induction. We focus on the induction step where $t = (l, x, r)$. By assumption we have $\text{acomp} l e t$ and $\text{acomp} r e t$. Because of $\text{braun } t$ we can distinguish two cases. First assume $|l| = |r| + 1$. The claim $\text{acomp } t$ follows immediately from the previous lemma. Now assume $|l| = |r|$. By definition, there are four cases to consider when proving $\text{acomp } t$. By symmetry it suffices to consider only two of them. If $h l \leq h r$ and $mh r < mh l$ then $\text{acomp } t$ reduces to $\text{acomp } r$, which is true by assumption. Now assume $h l \leq h r$ and $mh l \leq mh r$. Because $|l| = |r|$, the fact that the height of an almost complete tree is determined uniquely by its size (Lemma 4.7) implies $h l = h r$ and thus $\text{acomp } t$ reduces to $\text{acomp } l$, which is again true by assumption. \qed

Note that the proof does not rely on the fact that it is the left child that is potentially one bigger than the right one; it merely requires that the difference in size between two siblings is at most 1.

### 11.3 Arrays via Braun Trees

In this section we implement arrays by means of Braun trees and verify correctness and complexity. We start by defining array-like functions on Braun trees. After the above explanation of Braun trees the following lookup function will not come as a surprise:

```haskell
lookup1 :: 'a tree ⇒ nat ⇒ 'a
lookup1 (l, x, r) n = (if n = 1 then x else lookup1 (if even n then l else r) (n div 2))
```

The least significant bit is the parity of the index and we advance to the next bit by $\text{div } 2$. The function is called $\text{lookup1}$ rather than $\text{lookup}$ to emphasize that it expects the index to be at least 1. This simplifies the implementation via Braun trees but is in contrast to the $\text{Array}$ interface where by convention indexing starts with 0.

Function $\text{update1}$ descends in the very same manner but also performs an update when reaching 1:
11.3 Arrays via Braun Trees

The second equation updates existing entries in case \( n = 1 \). The first equation, however, creates a new entry and thus supports extending the tree. That is, \( \text{update} 1 (j + 1) \cdot x \cdot t \) extends the tree with a new node \( x \) at index \( j + 1 \).

Function \( \text{adds} \) iterates this process (again expecting \(|t| + 1\) as the index) and thus adds a whole list of elements:

\[
\text{adds} :: \text{'a list } \Rightarrow \text{nat } \Rightarrow \text{'a tree } \Rightarrow \text{'a tree}
\]
\[
\text{adds} \cdot [] \cdot t = t
\]
\[
\text{adds} \cdot (x \not\in \text{xs}) \cdot n \cdot t = \text{adds} \cdot \text{xs} \cdot (n + 1) \cdot (\text{update} 1 (n + 1) \cdot x \cdot t)
\]

The implementation of the \textit{Array} interface in Figure 11.3 is just a thin wrapper around the corresponding functions on Braun trees. An array is represented as a pair of a Braun tree and its size. Note that although \( \text{update} 1 \) can extend the tree, the specification and implementation of the array \textit{update} function does not support that: \( n \) is expected to be below the length of the array. Flexible arrays are specified and implemented in Section 11.4.

\[
\text{lookup} \cdot (t, \_ ) \cdot n = \text{lookup} 1 \cdot t \cdot (n + 1)
\]
\[
\text{update} \cdot n \cdot x \cdot (t, m) = (\text{update} 1 (n + 1) \cdot x \cdot t, m)
\]
\[
\text{len} \cdot (t, m) = m
\]
\[
\text{array} \cdot \text{xs} = (\text{adds} \cdot \text{xs} \cdot 0 \cdot \text{List } \cdot \text{xs})
\]

Fig. 11.3. Array implementation via Braun trees

11.3.1 Functional Correctness

The invariant on arrays is obvious:
invar \( (t, l) = (braun\ t \land l = |t|) \)

The abstraction function \( list \) could be defined in the following intuitive way, where \( [m..<n] \) is the list of natural numbers from \( m \) up to but excluding \( n \) (see Appendix A):

\[
list\ t = map\ (lookup\ t)\ [1..<|t| + 1]
\]

Instead we define \( list \) recursively and derive the above equation later on

\[
list\ ::\ 'a\ tree \Rightarrow \ 'a\ list\\
list\ \emptyset = []\\
list\ (\ell, x, r) = x # splice\ (list\ l)\ (list\ r)
\]

This definition is best explained by looking at Figure 11.2. The subtrees with root 2 and 3 will be mapped to the lists \([2, 4, 6, 8, 10, 12, 14]\) and \([1, 3, 5, 7, 9, 11, 13, 15]\). The obvious way to combine these two lists into \([1, 2, 3, \ldots, 15]\) is to splice them:

\[
splice\ ::\ 'a\ list \Rightarrow \ 'a\ list \Rightarrow \ 'a\ list\\
splice\ \emptyset\ ys = ys\\
splice\ (x # xs)\ ys = x # splice\ ys\ xs
\]

Note that because of this reasonably efficient \( O(n\ lg\ n) \), see Section 11.3.2) implementation of \( list \) we can also regard \( list \) as part of the interface of arrays.

Before we embark on the actual proofs we state a helpful arithmetic truth that is frequently used implicitly below:

\[
braun\ \ell\ x\ r \land n \in \{1..(|\ell|, x, r)|\} \land 1 < n \rightarrow \newline\ (odd\ n \rightarrow n\ div\ 2 \in \{1..|r|\}) \land (even\ n \rightarrow n\ div\ 2 \in \{1..|\ell|\})
\]

where \( \{m..n\} = \{k \mid m \leq k \land k \leq m\} \).

We will now verify that the implementation in Figure 11.3 of the Array interface in Figure 11.1 satisfies the given specification.

We start with property \( (len) \), the correctness of function \( len \). Because of the invariant, \( (len) \) follows directly from

\[
|list\ t| = |t|
\]

which is proved by induction. We will also use this property implicitly in many proofs below.

The following proposition implies the correctness property \( (lookup) \):
The proof is by induction and uses the following proposition that is also proved by induction:

\[ n < |xs| + |ys| \land |ys| \leq |xs| \land |xs| \leq |ys| + 1 \rightarrow \\
splice xs ys ! n = (if ~even~ n then xs else ys) ! (n \ div \ 2) \]

As a corollary to (11.1) we obtain that function \texttt{list} can indeed be expressed via \texttt{lookup}:

\[ braun t \mapsto list = \texttt{map} \ (\texttt{lookup} t) \ [1..<|t| + 1] \tag{11.2} \]

It follows by \texttt{list} \texttt{extensionality}:

\[ xs = ys \iff |xs| = |ys| \land (\forall i<|xs| \cdot xs ! i = ys ! i) \]

Let us now verify \texttt{update} as implemented via \texttt{update1}. The following two preservation properties (proved by induction) prove (\texttt{update-inv}):

\[ braun t \land n \in \{1..|t|\} \rightarrow |update1 n x t| = |t| \\
braun t \land n \in \{1..|t|\} \rightarrow braun (update1 n x t) \tag{11.4} \]

The following property relating \texttt{lookup} and \texttt{update1} is again proved by induction:

\[ braun t \land n \in \{1..|t|\} \rightarrow \\
lookup1 (update1 n x t) m = (if n = m then x else lookup1 t m) \tag{11.5} \]

The last three properties together with (11.2) and \texttt{list} \texttt{extensionality} prove the following proposition, which implies (\texttt{update}):

\[ braun t \land n \in \{1..|t|\} \rightarrow list (update1 n x t) = (list t)[n - 1 := x] \]

Finally we turn to the constructor \texttt{array}. It is implemented in terms of \texttt{adds} and \texttt{update1}. Their correctness is captured by the following properties whose inductive proofs build on each other:

\[ braun t \rightarrow |update1 (|t| + 1) x t| = |t| + 1 \tag{11.3} \]
\[ braun t \rightarrow braun (update1 (|t| + 1) x t) \tag{11.4} \]
\[ braun t \rightarrow list (update1 (|t| + 1) x t) = list t @ [x] \tag{11.5} \]
\[ braun t \rightarrow |adds xs |t| t| = |t| + |xs| \land braun (adds xs |t| t) \]
\[ braun t \rightarrow list (adds xs |t| t) = list t @ xs \]

The last two properties imply the remaining proof obligations (\texttt{array-inv}) and (\texttt{array}). The proof of (11.5) requires the following two properties of \texttt{splice} which are proved by simultaneous induction:

\[ |ys| \leq |xs| \rightarrow splice (xs @ [x]) ys = splice xs ys @ [x] \]
\[ |xs| \leq |ys| + 1 \rightarrow splice xs (ys @ [y]) = splice xs ys @ [y] \]
11.3.2 Running Time Analysis

The running time of \textit{lookup} and \textit{update} is obviously logarithmic because of the logarithmic height of Braun trees. We sketch why \textit{list} and \textit{array} both have running time $O(n \lg n)$. Linear time versions are presented in Section 11.5.

Function \textit{list} is similar to bottom-up merge sort and \textit{splice} is similar to \textit{merge}. We focus on \textit{splice} because it performs almost all the work. Consider calling \textit{list} on a complete tree of height $h$. At each level $k$ (starting with 0 for the root) of the tree, \textit{splice} is called $2^k$ times with lists of size (almost) $2^{h-k-1}$.

The running time of \textit{splice} with lists of the same length is proportional to the size of the lists. Thus the running time at each level is $O(2^k 2^{h-k-1}) = O(2^{h-1}) = O(2^h)$. Thus all the splices together require time $O(h 2^h)$. Because complete trees have size $n = 2^h$, the bound $O(n \lg n)$ follows.

Function \textit{array} is implemented via \textit{adds} and thus via repeated calls of \textit{update}. At the beginning of Section 7.3 we show that because \textit{update} has logarithmic complexity, calling it $n$ times on a growing tree starting with a leaf takes time $\Theta(n \lg n)$.

11.4 Flexible Arrays

Flexible arrays can be grown and shrunk at either end. Figure 11.4 shows the specification of all four operations. (For \textit{tl} and \textit{butlast} see Appendix A.) \textit{Array_Flex} extends the basis specification \textit{Array} in Figure 11.1.

\textbf{ADT} \textit{Array_Flex} = \textit{Array} +

\textbf{interface}
- \textit{add_lo} :: 'a \Rightarrow 'a \Rightarrow 'a
- \textit{del_lo} :: 'a \Rightarrow 'a
- \textit{add_hi} :: 'a \Rightarrow 'a \Rightarrow 'a
- \textit{del_hi} :: 'a \Rightarrow 'a

\textbf{specification}
- $\text{invar ar} \Rightarrow \text{invar (add_lo a ar)}$ (\textit{add_lo-inv})
- $\text{invar ar} \Rightarrow \text{list (add_lo a ar) = a # list ar}$ (\textit{add_lo})
- $\text{invar ar} \Rightarrow \text{invar (del_lo ar)}$ (\textit{del_lo-inv})
- $\text{invar ar} \Rightarrow \text{list (del_lo ar) = tl (list ar)}$ (\textit{del_lo})
- $\text{invar ar} \Rightarrow \text{invar (add_hi a ar)}$ (\textit{add_hi-inv})
- $\text{invar ar} \Rightarrow \text{list (add_hi a ar) = list ar \@ [a]}$ (\textit{add_hi})
- $\text{invar ar} \Rightarrow \text{invar (del_hi ar)}$ (\textit{del_hi-inv})
- $\text{invar ar} \Rightarrow \text{list (del_hi ar) = butlast (list ar)}$ (\textit{del_hi})

\textbf{Fig. 11.4. ADT Array_Flex}
Below we first implement the Array_Flex functions on Braun trees. In a final step an implementation of Array_Flex on (tree,size) pairs is derived.

We have already seen that update1 adds an element at the high end. The inverse operation del_hi removes the high end, assuming that the given index is the size of the tree:

\[
\text{del_hi} :: \text{n \&} \Rightarrow \text{'a tree \&} \Rightarrow \text{'a tree}
\]
\[
\text{del_hi} \ n \ (l, x, r) = (\text{if} \ n = 1 \ \text{then} \ () \ \text{else if} \ \text{even} \ n \ \text{then} \ (\text{del_hi} \ (n \ \text{div} \ 2) \ l, x, r) \ \text{else} \ (l, x, \text{del_hi} \ (n \ \text{div} \ 2) \ r))
\]

This was easy but extending an array at the low end seems hard because one has to shift the existing entries. However, Braun trees support a logarithmic implementation:

\[
\text{add_lo} :: \text{'a \Rightarrow \text{'a tree \&} \Rightarrow \text{'a tree}
\]
\[
\text{add_lo} \ x \ () = ((), x, ())
\]
\[
\text{add_lo} \ x \ (l, a, r) = (\text{add_lo} \ r, x, l)
\]

The intended functionality is list (add_lo x t) = x #: list t. Function add_lo installs the new element x at the root of the tree. Because add_lo needs to shift the indices of the elements already in the tree, the left child (indices 2, 4, \ldots) becomes the new right child (indices 3, 5, \ldots). The old right child becomes the new left child with the old root a added in at index 2 and the remaining elements at indices 4, 6, \ldots. In the following example, add_lo 0 transforms the left tree into the right one. The numbers in the nodes are the actual elements, not their indices.

\[
\begin{array}{c}
\text{add_lo} 0,
\end{array}
\]

Function del_lo simply reverses add_lo by removing the root and merging the children:
**11 Arrays via Braun Trees**

```plaintext
del_lo :: 'a tree \Rightarrow 'a tree
del_lo () = ()
del_lo (l, _, r) = merge l r

merge :: 'a tree \Rightarrow 'a tree \Rightarrow 'a tree
merge () r = r
merge (l, a, r) rr = (rr, a, merge l r)
```

Figure 11.5 shows the obvious implementation of the functions in the Array_Flex specification from Figure 11.4 (on the left-hand side) with the help of the corresponding Braun tree operations (on the right-hand side). It is an extension of the basic array implementation from Figure 11.3. All Array_Flex functions have logarithmic time complexity because the corresponding Braun tree functions do because they descend along one branch of the tree.

```plaintext
add_lo x (t, l) = (add_lo x t, l + 1)
del_lo (t, l) = (del_lo t, l - 1)
add_hi x (t, l) = (update1 (l + 1) x t, l + 1)
del_hi (t, l) = (del_hi t, l - 1)
```

Fig. 11.5. Flexible array implementation via Braun trees

**11.4.1 Functional Correctness**

We now have to prove the properties in Figure 11.4. We have already dealt with `update1` and thus `add_hi` above. Properties `(add_hi-inv)` and `(add_hi)` follow from (11.3), (11.4) and (11.5) stated earlier.

Correctness of `del_hi` on Braun trees is captured by the following two properties proved by induction:

```plaintext
braun t \rightarrow braun (del_hi tl t)
braun t \rightarrow list (del_hi tl t) = butlast (list t) (11.6)
```

They imply `(del_hi-inv)` and `(del_hi)`. The proof of (11.6) requires the simple fact `list t = [] \leftrightarrow t = ()` and the following property of `splice` which is proved by induction:
\begin{verbatim}
butlast (splice xs ys)
  = (if |ys| < |xs| then splice (butlast xs) ys else splice xs (butlast ys))
\end{verbatim}

Correctness of \texttt{add\_lo} on Braun trees is captured by the following two properties proved by induction:

\begin{align*}
\text{braun } t &\rightarrow \text{braun } (\text{add\_lo } x \ t) \\
\text{braun } t &\rightarrow \text{list } (\text{add\_lo } a \ t) = a \# \text{list } t
\end{align*}

Properties (\texttt{add\_lo-inv}) and (\texttt{add\_lo}) follow directly from them.

Finally we turn to \texttt{del\_lo}. Inductions (for \texttt{merge}) and case analyses (for \texttt{del\_lo}) yield the following properties:

\begin{align*}
\text{braun } (l, x, r) &\rightarrow \text{braun } (\text{merge } l \ r) \\
\text{braun } (l, x, r) &\rightarrow \text{list } (\text{merge } l \ r) = \text{splice } (\text{list } l) (\text{list } r) \\
\text{braun } t &\rightarrow \text{braun } (\text{del\_lo } t) \\
\text{braun } t &\rightarrow \text{list } (\text{del\_lo } t) = \text{tl } (\text{list } t)
\end{align*}

The last two properties imply (\texttt{del\_lo-inv}) and (\texttt{del\_lo}).

11.5 Bigger, Better, Faster, More!

In this section we meet efficient versions of some old and new functions on Braun trees. The implementation of the corresponding array operations is trivial and is not discussed.

11.5.1 Fast Size of Braun Trees

The size of a Braun tree can be computed without having to traverse the entire tree:

\begin{verbatim}
size_fast :: 'a tree \Rightarrow \text{nat}
size_fast \nil = 0
size_fast \langle l, _, r \rangle = (\text{let } n = \text{size_fast } r \text{ in } 1 + 2 \cdot n + \text{diff } l \ n)

diff :: 'a tree \Rightarrow \text{nat} \Rightarrow \text{nat}
diff \nil_ = 0
diff \langle l, _, r \rangle \ n
  = (\text{if } n = 0 \text{ then } 1 \text{ else if even } n \text{ then } \text{diff } r \ (n \div \ 2) \text{ else } \text{diff } l \ (n \div \ 2))
\end{verbatim}
Function `size_fast` descends down the right spine, computes the size of a `Node` as if both children were the same size \((1 + 2 \cdot n)\), but adds \(\text{diff} \ n\) to compensate for bigger left children. Correctness of `size_fast`.

Lemma 11.3. \(\text{braun} \ t \rightarrow \text{size}\_fast \ t = |t|\)

follows from this property of \(\text{diff}\):

\[
\text{braun} \ t \land |t| \in \{n, n + 1\} \rightarrow \text{diff} \ t \ n = |t| - n
\]

The running time of `size_fast` is quadratic in the height of the tree (see Exercise 11.3).

11.5.2 Initializing a Braun Tree with a Fixed Value

Above we only considered the construction of a Braun tree from a list. Alternatively one may want to create a tree (array) where all elements are initialized to the same value. Of course one can call `update1` \(n\) times, but one can also build the tree directly:

```plaintext
braun\_of\_naive \ x \ n
= (if \ n = 0 \ then \ ()
    else let \ m = (n - 1) \ div 2
           in if \ odd \ n
                then \ (braun\_of\_naive \ x \ m, x, braun\_of\_naive \ x \ m)
                else \ (braun\_of\_naive \ x \ (m + 1), x, braun\_of\_naive \ x \ m))
```

This solution also has time complexity \(O(n \ lg \ n)\) but it can clearly be improved by sharing identical recursive calls. Function `braun2\_of` shares as much as possible by producing trees of size \(n\) and \(n + 1\) in parallel:

```plaintext
braun2\_of :: \text{\'a \Rightarrow \text{nat} \Rightarrow \text{\'a tree} \times \text{\'a tree}}
braun2\_of \ x \ n
= (if \ n = 0 \ then \ ((, ()), x, ()))
    else let \ (s, t) = braun2\_of \ x \ ((n - 1) \ div 2)
            in if \ odd \ n \ then \ ((s, x, s), (t, x, s)) \ else \ ((t, x, s), (t, x, t))
```

This solution also has time complexity \(O(n \ lg \ n)\) but it can clearly be improved by sharing identical recursive calls. Function `braun2\_of` shares as much as possible by producing trees of size \(n\) and \(n + 1\) in parallel:

```plaintext
braun\_of :: \text{\'a \Rightarrow \text{nat} \Rightarrow \text{\'a tree}}
braun\_of \ x \ n = \text{fst} \ (braun2\_of \ x \ n)
```

The running time is clearly logarithmic.

The correctness properties (see Appendix A for `replicate`).
list \( \text{braun}_\text{of } x n \) = \text{replicate } n x

\text{braun } (\text{braun}_\text{of } x n)

are corollaries of the more general statements

\[
\text{braun2}_\text{of } x n = (s, t) \quad \rightarrow \\
\text{list } s = \text{replicate } n x \land \text{list } t = \text{replicate } (n + 1) x
\]

\[
\text{braun2}_\text{of } x n = (s, t) \quad \rightarrow \\
|s| = n \land |t| = n + 1 \land \text{braun } s \land \text{braun } t
\]

which can both be proved by induction.

### 11.5.3 Converting a List into a Braun Tree

We improve on function \( \text{adds} \) from Section 11.3 that has running time \( \Theta(n \lg n) \) by developing a linear-time function. Given a list of elements \([1, 2, \ldots]\), we can subdivide it into sublists \([1], [2, 3], [4, \ldots, 7], \ldots\) such that the \( k \)th sublist contains the elements of level \( k \) of the corresponding Braun tree. This is simply because on each level we have the entries whose index has \( k+1 \) bits. Thus we need to process the input list in chunks of size \( 2^k \) to produce the trees on level \( k \). But we also need to get the order right. To understand how that works, consider the last two levels of the tree in Figure 11.2:

```
          4
       /   \
      8  12
     /   \
    10  14   9  13   11  15
```

If we rearrange them in increasing order of the root labels

```
          4
       /   \
      8  12
     /   \
    9   13
     /   \
   10  14   11  15
```

the following pattern emerges: the left subtrees are labeled \([8, \ldots, 11]\), the right subtrees \([12, \ldots, 15]\). Call \( t_i \) the tree with root label \( i \). The correct order of subtrees, i.e. \( t_4, t_6, t_5, t_7 \), is restored when the three lists \([t_4, t_5], [2, 3] \) (the labels above) and \([t_6, t_7] \) are combined into new trees by going through them simultaneously from left to right, yielding \([t_4, 2, t_6], (t_5, 3, t_7) \), the level above.

Abstracting from this example we arrive at the following code. Loosely speaking, \( \text{brauns } k \; \text{xs} \) produces the Braun trees on level \( k \).
**Arrays via Braun Trees**

**brauns :: nat ⇒ 'a list ⇒ 'a tree list**

\[
\text{brauns } k \text{ } x s \\
= ( \text{if } x s = [] \text{ then } [] \\
\text{else let } y s = \text{take } 2^k x s; \\
\text{\hspace{1em}} z s = \text{drop } 2^k x s; \\
\text{\hspace{1em}} t s = \text{brauns } (k + 1) z s \\
\text{\hspace{1em}} \text{in } \text{nodes } t s \ y s (\text{drop } 2^k t s))
\]

Function `brauns` chops off a chunk `ys` of size `2^k` from the input list and recursively converts the remainder of the list into a list `ts` of (at most) `2^{k+1}` trees. This list is (conceptually) split into `take 2^k ts` and `drop 2^k ts` which are combined with `ys` by function `nodes` that traverses its three argument lists simultaneously. As a local optimization, we pass all of `ts` rather than just `take 2^k ts` to `nodes`.

**nodes :: 'a tree list ⇒ 'a list ⇒ 'a tree list ⇒ 'a tree list**

\[
\text{nodes } (l \# l s) (x \# x s) (r \# r s) = (l, x, r) \# \text{nodes } l s x s r s \\
\text{nodes } (l \# l s) (x \# x s) [] = (l, x, []) \# \text{nodes } l s x s [] \\
\text{nodes } [] (x \# x s) (r \# r s) = ([], x, r) \# \text{nodes } [] x s r s \\
\text{nodes } [] (x \# x s) [] = ([], x, []) \# \text{nodes } [] x s [] \\
\text{nodes } _ [] _ [] = []
\]

Because the input list may not have exactly \(2^n - 1\) elements, some of the chunks of elements and trees may be shorter than \(2^k\). To compensate for that, function `nodes` implicitly pads lists of trees at the end with leaves. This padding is the purpose of equations two to four.

The top-level function for turning a list into a tree simply extracts the first (and only) element from the list computed by `brauns 0`:

**brauns1 :: 'a list ⇒ 'a tree**

\[
\text{brauns1 } x s = ( \text{if } x s = [] \text{ then } () \text{ else } \text{brauns } 0 x s ! 0)
\]

**Functional Correctness**

The key correctness lemma below expresses a property of Braun trees: the subtrees on level \(k\) consist of all elements of the input list `xs` that are \(2^k\) elements apart, starting from some offset. To state this concisely we define
The result of \(\text{take\_nths }\ i\ k\ \text{xs}\) is every \(2^k\)-th element in \(\text{drop } i \ \text{xs}\).

A number of simple properties follow by easy inductions:

\[
\begin{align*}
\text{take\_nths }\ i\ k\ (\text{drop } j \ \text{xs}) & = \text{take\_nths }\ (i + j)\ k\ \text{xs} \quad (11.7) \\
\text{take\_nths }0\ 0\ \text{xs} & = \text{xs} \quad (11.8) \\
\text{splice } (\text{take\_nths }0\ 1\ \text{xs}) & = (\text{take\_nths }1\ 1\ \text{xs}) \quad (11.9) \\
\text{take\_nths }i\ m\ (\text{take\_nths }j\ n\ \text{xs}) & = \text{take\_nths }\ (i \cdot 2^j + j)\ (m + n)\ \text{xs} \quad (11.10) \\
\text{take\_nths }i\ k\ \text{xs} & = [] \iff |\text{xs}| \leq i \quad (11.11) \\
i < |\text{xs}| \rightarrow \text{hd } (\text{take\_nths }i\ k\ \text{xs}) = \text{xs} ! i \quad (11.12) \\
|\text{xs}| = |\text{ys}| \lor |\text{xs}| = |\text{ys}| + 1 \rightarrow \\
\text{take\_nths }0\ 1\ (\text{splice } \text{xs} \ \text{ys}) & = \text{xs} \land \text{take\_nths }1\ 1\ (\text{splice } \text{xs} \ \text{ys}) = \text{ys} \quad (11.13) \\
|\text{take\_nths }0\ 1\ \text{xs}| & = |\text{take\_nths }1\ 1\ \text{xs}| \lor \\
|\text{take\_nths }0\ 1\ \text{xs}| & = |\text{take\_nths }1\ 1\ \text{xs}| + 1 \quad (11.14)
\end{align*}
\]

We also introduce a predicate relating a tree to a list:

\[
\text{braun\_list } :: \, \text{a tree } \Rightarrow \text{a list } \Rightarrow \text{bool} \\
\text{braun\_list }\ \emptyset \ \text{xs} = (\text{xs} = []) \\
\text{braun\_list }\ (l, x, r) \ \text{xs} \\
= (\text{xs} \neq [] \land x = \text{hd } \text{xs} \land \\
\text{braun\_list }\ l\ (\text{take\_nths }1\ 1\ \text{xs}) \land \\
\text{braun\_list }\ r\ (\text{take\_nths }2\ 1\ \text{xs}))
\]

This definition may look a bit mysterious at first but it satisfies a simple specification: \(\text{braun\_list }\ t\ \text{xs} \iff \text{braun } t \land \text{xs} = \text{list } t\) (see below). The idea of the above definition is that instead of relating \(\langle l, x, r \rangle\) to \text{xs} via \text{splice} we invert the process and relate \(l\) and \(r\) to the even and odd numbered elements of \(\text{drop } 1 \ \text{xs}\).

**Lemma 11.4.** \(\text{braun\_list }\ t\ \text{xs} \iff \text{braun } t \land \text{xs} = \text{list } t\)
Proof. The proof is by induction on \( t \). The base case is trivial. In the induction step the key properties are (11.14) to prove \( \text{braun} \ t \) and (11.9) and (11.13) to prove \( \text{xs} = \text{list} \ t \).

The correctness proof of \( \text{brauns} \) rests on a few simple inductive properties:

\[
\begin{align*}
|\text{nodes} \ ls \ zs \ rs| &= |zs| & (11.15) \\
\text{nodes} \ ls \ zs \ rs \ ! \ i &= \begin{cases} 
\text{if } i < |ls| \text{ then } ls \ ! \ i \text{ else } (), \text{ zs } ! \ i, \\
\text{if } i < |rs| \text{ then } rs \ ! \ i \text{ else } () \end{cases} & (11.16) \\
|\text{brauns} \ k \ xs| &= \min |zs| 2^k & (11.17)
\end{align*}
\]

The main theorem expresses the following correctness property of the elements of \( \text{brauns} \ k \ xs \) every tree \( \text{brauns} \ k \ xs \ ! \ i \) is a Braun tree and its list of elements is \( \text{take_nths} \ i \ k \ xs \):

**Theorem 11.5.** \( i < \min |zs| 2^k \longrightarrow \text{braun_list} \ (\text{brauns} \ k \ xs \ ! \ i) \ (\text{take_nths} \ i \ k \ xs) \)

Proof. The proof is by induction on the length of \( \text{zs} \). Assume \( i < \min |zs| 2^k \), which implies \( \text{zs} \neq [] \). Let \( zs = \text{drop} 2^k \ xs \). Thus \( |zs| < |zs| \) and therefore the IH applies to \( zs \) and yields the property

\[
\forall i \ j. \ j = i + 2^k \wedge i < \min |zs| 2^k + 1 \longrightarrow \text{braun_list} \ (ts \ ! \ i) \ (\text{take_nths} \ j \ (k + 1) \ xs) \quad (*)
\]

where \( ts = \text{brauns} \ (k + 1) \ zs \). Let \( ts' = \text{drop} 2^k \ ts \). Below we examine \( \text{nodes} \ ts \ _ \ ts' \ ! \ i \) with the help of (11.16). Thus there are four similar cases of which we only discuss one representative one: assume \( i < |ts| \) and \( i \geq |ts'| \).

\[
\begin{align*}
\text{braun_list} \ (\text{brauns} \ k \ xs \ ! \ i) \ (\text{take_nths} \ i \ k \ xs) \\
&\iff \text{braun_list} \ (\text{nodes} \ ts \ (\text{take} 2^k \ xs) \ ts' \ ! \ i) \ (\text{take_nths} \ i \ k \ xs) \\
&\iff \text{braun_list} \ (ts \ ! \ i) \ (\text{take_nths} \ (2^k + i) \ (k + 1) \ xs) \wedge \text{braun_list} () \ (\text{take_nths} \ (2^k + 1 + i) \ (k + 1) \ xs) \\
&\quad \text{by (11.16), (11.10), (11.11), (11.12) and assumptions} \\
&\iff \text{True} \quad \text{by (\( * \)), (11.11), (11.17) and assumptions}
\end{align*}
\]

Setting \( i = k = 0 \) in this theorem we obtain the correctness of \( \text{brauns1} \) using Lemma 11.4 and (11.8):

**Corollary 11.6.** \( \text{braun} \ (\text{brauns1} \ xs) \wedge \text{list} \ (\text{brauns1} \ xs) = xs \)
Running Time Analysis

We focus on function brauns. In the step from brauns to $T_{\text{brauns}}$, we simplify matters a little bit: we count only the expensive operations that traverse lists and ignore the other small additive constants. The time to evaluate \textit{take} $n$ \textit{xs} and \textit{drop} $n$ \textit{xs} is linear in $\min n \|\textit{xs}\|$ and we simply use $\min n \|\textit{xs}\|$. Evaluating \textit{nodes} $\textit{ls} \ \textit{xs} \ \textit{rs}$ takes time linear in $\|\textit{xs}\|$ and $\|\textit{take} \ n \ \textit{xs}\| = \min n \|\textit{xs}\|$. As a result we obtain the following definition of $T_{\text{brauns}}$:

\[
T_{\text{brauns}} :: \text{nat} \Rightarrow \text{a list} \Rightarrow \text{nat} \\
T_{\text{brauns}} \ k \ \textit{xs} \\
\quad = \text{(if} \ \textit{xs} = \text{[]} \ \text{then} \ 0 \\
\quad \text{else let } \textit{ys} = \text{take} 2^k \ \textit{xs}; \ \textit{zs} = \text{drop} 2^k \ \textit{xs}; \ \textit{ts} = \text{brauns} (k + 1) \ \textit{zs} \in \text{4 \cdot } \min 2^k \|\textit{xs}\| + T_{\text{brauns}} (k + 1) \ \textit{zs})
\]

It is easy to prove that $T_{\text{brauns}}$ is linear:

Lemma 11.7. $T_{\text{brauns}} \ k \ \textit{xs} = 4 \cdot \|\textit{xs}\|$  

\textbf{Proof.} The proof is by induction on the length of \textit{xs}. If \textit{xs} = [] the claim is trivial. Now assume \textit{xs} \neq [] and let \textit{zs} = \text{drop} 2^k \ \textit{xs}.

\[
T_{\text{brauns}} \ k \ \textit{xs} = T_{\text{brauns}} (k + 1) \ \textit{zs} + 4 \cdot \min 2^k \|\textit{xs}\| \\
= 4 \cdot \|\textit{zs}\| + 4 \cdot \min 2^k \|\textit{zs}\| \quad \text{by IH} \\
= 4 \cdot (\|\textit{zs}\| - 2^k) + 4 \cdot \min 2^k \|\textit{zs}\| = 4 \cdot \|\textit{zs}\| \quad \Box
\]

11.5.4 Converting a Braun Tree into a List

We improve on function list that has running time $O(n \lg n)$ by developing a linear-time version. Imagine that we want to invert the computation of brauns and thus of brauns. Thus it is natural to convert not merely a single tree but a list of trees. Looking once more at the reordered list of subtrees

```
   4  5  6  7
   \ / \ / \ / \\
  8 12 9 13 10 14 11 15
```

the following strategy strongly suggests itself: first the roots, then the left children, then the right children. The recursive application of this strategy also takes care of the required reordering of the subtrees. Of course we have to ignore any leaves we encounter. This is the resulting function:
list_fast_rec :: 'a tree list ⇒ 'a list

list_fast_rec ts
= (let us = filter (λt. t ≠ ⟨⟩) ts
  in if us = [] then []
      else map value us @ list_fast_rec (map left us @ map right us))

value ⟨l, x, r⟩ = x
left ⟨l, x, r⟩ = l
right ⟨l, x, r⟩ = r

Termination of list_fast_rec is almost obvious because left and right remove
the top node of a tree. Thus size seems the right measure. But if ts = ⟨⟩,
the measure is 0 but it still leads to a recursive call (with argument []). This
problem can be avoided with the measure function \( \varphi = \text{sum_list} \circ \text{map f} \)
where \( f = (λt. 2 \cdot |t| + 1) \). Assume \( ts ≠ [] \) and let us = filter (λt. t ≠ ⟨⟩) ts. We need to show that \( \varphi (\text{map left us} @ \text{map right us}) < \varphi ts \). Take some
t in ts. If \( t = ⟨⟩ \), \( f t = 1 \) but \( t \) is no longer in \( us \), i.e. the measure decreases by
1. If \( t = ⟨l, x, r⟩ \) then \( f t = 2 \cdot |l| + 2 \cdot |r| + 3 \) but \( f (\text{left t}) + f (\text{right t})
= 2 \cdot |l| + 2 \cdot |r| + 2 \) and thus the measure also decreases by 1. Because \( ts ≠ [] \) this proves \( \varphi (\text{map left us} @ \text{map right us}) < \varphi ts \). We do not show
the technical details.

Finally, the top level function to extract a list from a single tree:

list_fast :: 'a tree ⇒ 'a list

list_fast t = list_fast_rec [t]

From list_fast one can easily derive an efficient fold function on Braun
trees that processes the elements in the tree in the order of their indexes.

Functional Correctness

We want to prove correctness of list_fast: list_fast t = list t if braun t. A
direct proof of list_fast_rec [t] = list t will fail and we need to generalize
this statement to all lists of length \( 2^k \). Reusing the infrastructure from
the previous subsection this can be expressed as follows:

Theorem 11.8. \( |ts| = 2^k \land (\forall i < 2^k. \text{braun_list} (ts ! i) (\text{take_nths i k xs})) \)
⇒ list_fast_rec ts = xs

Proof. The proof is by induction on the length of \( xs \). Assume the two
premises. There are two cases.
First assume $|xs| < 2^k$. Then

$$ts = map (\lambda x. \langle \langle \rangle, x, \langle \rangle \rangle) \; xs @ replicate n \langle \rangle$$  \hspace{1cm} (*) \hspace{1cm} \text{where } n = |ts| - |xs|. \text{ This can be proved pointwise. Take some } i < 2^k. \text{ If } i < |xs| \text{ then } take_nths i k xs = take 1 (drop i xs) (which can be proved by induction on } xs). \text{ By definition of braun_list it follows that } t! i = \langle l, xs! i, r \rangle \text{ for some } l \text{ and } r \text{ such that braun_list } l [] \text{ and braun_list } l [] \text{ and thus } l = r = \langle \rangle, \text{ i.e. } t! i = \langle \langle \rangle, xs! i, \langle \rangle \rangle. \text{ If } \neg i < |xs| \text{ then } take_nths i k xs = [] \text{ by (11.11) and thus braun_list } (ts! i) [] \text{ by the second premise and thus } ts! i = \langle \rangle \text{ by definition of braun_list. This concludes the proof of (\ast). The desired list_fast_rec } ts = xs \text{ follows easily by definition of list_fast_rec.}

Now assume $\neg |xs| < 2^k$. Then for all $i < 2^k$

$$ts! i \neq \langle \rangle \land \text{value } (ts! i) = xs! i \land$$
$$\text{braun_list } (\text{left } (ts! i)) \; (\text{take_nths } (i + 2^k) \; (k + 1) \; xs) \land$$
$$\text{braun_list } (\text{right } (ts! i)) \; (\text{take_nths } (i + 2 \cdot 2^k) \; (k + 1) \; xs)$$

follows from the second premise with the help of (11.10), (11.11) and (11.12).

We obtain two consequences:

$$\text{map value } ts = \text{take } 2^k \; xs$$

$$\text{list_fast_rec } (\text{map left } ts @ \text{map right } ts) = \text{drop } 2^k \; xs$$

The first consequence follows by pointwise reasoning, the second consequence with the help of the IH and (11.7). From these two consequences the desired conclusion list_fast_rec } ts = xs \text{ follows by definition of list_fast_rec.} \quad \square

### Running Time Analysis

We focus on list_fast_rec. In the step from list_fast_rec to $T_{list_fast_rec}$ we simplify matters a little bit: we count only the expensive operations that traverse lists and ignore the other small additive constants. The time to evaluate map value ts, map left ts, map right ts, filter ($\lambda t. t \neq \langle \rangle$) ts and ts @ _ is linear in $|ts|$ and we simply use $|ts|$. As a result we obtain the following definition of $T_{list_fast_rec}$:

$$T_{list_fast_rec} :: 'a \text{ tree list } \Rightarrow \text{ nat}$$

$$T_{list_fast_rec} ts = (\text{let } us = \text{filter } (\lambda t. t \neq \langle \rangle) ts \hspace{1cm} \text{in } |ts| +$$
$$\hspace{1cm} (\text{if } us = [] \text{ then } 0 \hspace{1cm} \text{else } 5 \cdot |us| + T_{list_fast_rec} (\text{map left us @ map right us})))$$
The following inductive property is an abstraction of the core of the termination argument of \texttt{list\_fast\_rec} above.

\[
(\forall t \in \text{set } ts. \ t \neq ()) \rightarrow \\
(\sum_{t \in \text{left } ts} k \cdot |t|) = (\sum_{t \in \text{map left } ts @ \text{map right } ts} k \cdot |t|) + k \cdot |ts| \quad (11.18)
\]

The suggestive notation \( \sum x \mapsto xs. f x \) abbreviates \( \text{sum\_list } (\text{map } f \ xs) \).

Now we can state and prove a linear upper bound of \( T_{\text{list\_fast\_rec}} \):

**Theorem 11.9.** \( T_{\text{list\_fast\_rec}} ts \leq (\sum_{t \in \text{ts}} 7 \cdot |t| + 1) \)

**Proof.** The proof is by induction on the size of \( ts \), again using the measure function \( \lambda t. \ 2 \cdot |t| + 1 \) which decreases with recursive calls as we proved above. If \( ts = [] \) the claim is trivial. Now assume \( ts \neq [] \) and let \( us = \text{filter } (\lambda t. \ t \neq ()) \ ts \) and \( \text{children} = \text{map left } us \@ \text{map right } us \).

\[
T_{\text{list\_fast\_rec}} ts = T_{\text{list\_fast\_rec}} \text{children} + 5 \cdot |us| + |ts|
\]
\[
\leq (\sum_{t \in \text{children}} 7 \cdot |t| + 1) + 5 \cdot |us| + |ts| \quad \text{by IH}
\]
\[
= (\sum_{t \in \text{us}} 7 \cdot |t|) + 7 \cdot |us| + |ts| \quad \text{by (11.18)}
\]
\[
\leq (\sum_{t \in \text{ts}} 7 \cdot |t|) + |ts| = (\sum_{t \in \text{ts}} 7 \cdot |t| + 1)
\]

\[\square\]

## 11.6 Exercises

**Exercise 11.1.** Instead of first showing that Braun trees are almost complete, give a direct proof of \( \text{braun } t \rightarrow h \ t = \ lg \ |t| \) by first showing \( \text{braun } t \rightarrow 2^h \ t \leq 2 \cdot |t| + 1 \) by induction.

**Exercise 11.2.** Show that function \( \text{bal} \) in Section 4.3.1 produces Braun trees: \( n \leq |zs| \land \text{bal } n \ zs = (t, zs) \rightarrow \text{braun } t \) (Isabelle hint: \( \text{bal} \) needs to be qualified as \( \text{Balance.bal} \)).

**Exercise 11.3.** One can view Braun trees as tries (see Chapter 12) by indexing them not with a \texttt{nat} but a \texttt{bool list} where each bit tells us whether to go left or right (as explained at the start of Section 11.2). Function \texttt{nat\_of} specifies the intended correspondence:

\[
\text{nat\_of :: bool list } \Rightarrow \text{nat}
\]
\[
\text{nat\_of } [] = 1
\]
\[
\text{nat\_of } (b \# bs) = 2 \cdot \text{nat\_of } bs + (\text{if } b \text{ then } 1 \text{ else } 0)
\]

Define the counterparts of \texttt{lookup} and \texttt{update}.

\[
\text{lookup\_trie :: 'a tree } \Rightarrow \text{bool list } \Rightarrow \text{a}
\]
\[
\text{update\_trie :: bool list } \Rightarrow \text{a } \Rightarrow \text{'a tree } \Rightarrow \text{'a tree}
\]
and prove their correctness:

\[
\begin{align*}
\text{braun } t \land \text{nat_of } bs & \in \{1..|t|\} \implies \\
\text{lookup_trie } t \ bs = \text{lookup1 } t \ (\text{nat_of } bs) \\
\text{update_trie } bs \ x \ t = \text{updatel } (\text{nat_of } bs) \ x \ t
\end{align*}
\]

Exercise 11.4. Function \textit{del_lo} is defined with the help of function \textit{merge}. Define a recursive function \textit{del_lo2} :: 'a tree => 'a tree without recourse to any auxiliary function and prove \textit{del_lo2} \ t = \textit{del_lo} \ t.

Exercise 11.5. Let \(lh\), the “left height”, compute the length of the left spine of a tree. Prove that the left height of a Braun tree is equal to its height:

\[
\text{braun } t \implies \text{lh } t = \text{h } t
\]

Exercise 11.6. Show that the running time of \textit{size_fast} is quadratic in the height of the tree: Define the running time functions \(T_{\text{diff}}\) and \(T_{\text{size_fast}}\) (taking 0 time in the base cases) and prove \(T_{\text{size_fast}} \ t \leq (h \ t)^2\).

Exercise 11.7. Prove correctness of function \textit{braun_of_naive} defined in Section 11.5.2: \(\text{list } (\text{braun_of_naive } x \ n) = \text{replicate } n \ x\).

Bibliographic Remarks

Braun trees were investigated by Braun and Rem [67] and later, in a functional setting, by Hoogerwoord [31] who coined the term “Braun tree”. Section 11.5 is partly based on work by Okasaki [61]. The whole chapter is based on work by Nipkow and Sewell [58].
A trie is a search tree where keys are strings, i.e. lists. A trie can be viewed as a tree-shaped finite automaton where the root is the start state. For example, the set of strings \{a, an, can, car, cat\} is encoded as the trie in Figure 12.1. The solid states are accepting, i.e. those nodes terminate the string leading to them.

![Figure 12.1. A trie encoding \{a, an, can, car, cat\}](image)

### 12.1 Abstract Tries via Functions

A nicely abstract model of tries is the following type:

```plaintext
datatype 'a trie = Nd bool ('a => 'a trie option)
```

In a node \(Nd \, b \, f\), \(b\) indicates if it is an accepting node and \(f\) maps characters to sub-tries.
There is no invariant, i.e. the invariant is simply $\text{True}$: there are no ordering, balance or other requirements.

Below we use a variant of the function update notation $f(a := b)$:

$$f(a \mapsto b) \equiv f(a := \text{Some } b)$$

This is how the ADT $\text{Set}$ is implemented by means of tries:

```
empty :: 'a trie
empty = Nd False (\_ \rightarrow None)

isin :: 'a trie \Rightarrow 'a list \Rightarrow bool
isin (Nd b \_ \_ \_ ) [] = b
isin (Nd \_ \_ m) (k \# xs) = (let s = case m k of None \Rightarrow False | Some t \Rightarrow isin t xs in Nd b (m(x \mapsto insert xs s)))

insert :: 'a list \Rightarrow 'a trie \Rightarrow 'a trie
insert [] (Nd m) = Nd True m
insert (x \# xs) (Nd b m) = (let s = case m x of None \Rightarrow empty | Some t \Rightarrow t in Nd b (m(x \mapsto insert xs s)))

delete :: 'a list \Rightarrow 'a trie \Rightarrow 'a trie
delete [] (Nd m) = Nd False m
delete (x \# xs) (Nd b m) = Nd b (case m x of None \Rightarrow m | Some t \Rightarrow m(x \mapsto delete xs t))
```

The definitions are straightforward. But note that $\text{delete}$ does not try to shrink the trie. For example:

```
\text{delete [a] \rightarrow}
```

Formally:

$$\text{delete [a] (Nd False [a \mapsto Nd True (\_ \rightarrow None)])} = Nd False [a \mapsto Nd False (\_ \rightarrow None)]$$

where $[x \mapsto t] \equiv (\_ \rightarrow None)(x \mapsto t)$. The resulting trie is correct (it represents the empty set of strings) but could have been shrunk to $Nd False (\_ \rightarrow None)$. 

Functional Correctness

For the correctness proof we take a lazy approach and define the abstraction function in a trivial manner via \( \text{isin} \):

\[
\text{set} :: 'a \text{ trie} \Rightarrow 'a \text{ list } \text{set} \\
\text{set } t = \{xs | \text{isin } t xs\}
\]

Correctness of \( \text{empty} \) and \( \text{isin} \ (\text{set empty} = \emptyset \) and \( \text{isin } t xs = (xs \in \text{set } t) \) are trivial, correctness of insertion and deletion are easily proved by induction:

\[
\text{set (insert } xs \text{ } t) = \text{set } t \cup \{xs\} \\
\text{set (delete } xs \text{ } t) = \text{set } t - \{xs\}
\]

This simple model of tries leads to simple correctness proofs but is computationally inefficient because of the function space in \( 'a \Rightarrow 'a \text{ trie option} \). In principle, any representation of this function space, for example by some search tree, works. However, such a generic theory is relatively complex. Hence we restrict ourselves to binary tries when exploring more efficient implementations of tries.

12.2 Binary Tries

A binary trie is a trie over the alphabet \( \text{bool} \). That is, binary tries represent sets of \( \text{bool list} \). More concretely, every node has two children:

\[
\text{datatype trie} = \text{Lf} | \text{Nd } \text{bool} (\text{trie} \times \text{trie})
\]

A binary trie, for example

\( \text{Nd False (Nd True (Nd False (Lf , Lf) , Nd True (Lf , Lf))) , Lf} \)

can be visualized like this:

![Binary Trie Diagram]

\( Lfs \) are not shown at all. The edge labels indicated that \( \text{False} \) refers to the left and \( \text{True} \) to the right child. This convention is encoded in the following auxiliary functions selecting from and modifying pairs:
The implementation of the Set interface is shown in Figure 12.2. Function delete shrinks a non-accepting Nd if both children have become empty.

Functional Correctness

For the correctness proof we take the same lazy approach as above:
We are also lazy in that we set the invariant to $True$. A more precise invariant would express that the tries are minimal, i.e. cannot be shrunk. See Exercise 12.2. It turns out that the correctness properties do not require this more precise invariant.

The two non-trivial correctness properties are

\[ set\_trie\ (insert\ xs\ t) = set\_trie\ t \cup \{xs\} \]  \hspace{1cm} (12.1)

\[ set\_trie\ (delete\ xs\ t) = set\_trie\ t - \{xs\} \]  \hspace{1cm} (12.2)

are simple consequences of the following inductive properties:

\[ isin\ (insert\ xs\ t)\ ys = (xs = ys \lor isin\ t\ ys) \]

\[ isin\ (delete\ xs\ t)\ ys = (xs \neq ys \land isin\ t\ ys) \]

12.3 Binary Patricia Tries

Tries can contain long branches without branching. These can be contracted by storing the branch directly in the start node. The result is called a Patricia trie. The following figure shows the contraction of a trie into a Patricia trie:

This is the data type of binary Patricia tries:

\[
\text{datatype \ trieP} = \text{LfP} | \text{NdP (bool list) bool (trieP x trieP)}
\]

The implementation of the Set interface via binary Patricia tries is shown in Figure 12.3.

Functional Correctness

This is an exercise in stepwise data refinement. We have already proved that \text{trie} implements Set via an abstraction function. Now we map \text{trieP} back
emptyP = LfP

isinP LfP ks = False
isinP (NdP ps b lr) ks
= (let n = |ps|
in if ps = take n ks
then case drop n ks of [ ] ⇒ b | k # x ⇒ isnP (sel2 k lr) x
else False)

insertP ks LfP = NdP ks True (LfP, LfP)
insertP ks (NdP ps b lr) =
(case split ks ps of (., [], []) ⇒ NdP ps True lr
| (qs, [], p # ps') ⇒
let t = NdP ps' b lr
in NdP qs True (if p then (LfP, t) else (t, LfP))
| (., k # ks', []) ⇒ NdP ps b (mod2 (insertP ks') k lr)
| (qs, k # ks', _) # ps') ⇒
let tp = NdP ps' b lr; tk = NdP ks' True (LfP, LfP)
in NdP qs False (if k then (tp, tk) else (tk, tp)))

deleteP ks LfP = LfP
deleteP ks (NdP ps b lr) =
(case split ks ps of
  (qs, ks', p # ps') ⇒ NdP ps b lr
| (qs, k # ks', []) ⇒ nodeP ps b (mod2 (deleteP ks') k lr)
| (qs, [], []) ⇒ nodeP ps False lr)

nodeP ps b lr = (if ~ b ∧ lr = (LfP, LfP) then LfP else NdP ps b lr)

Fig. 12.3. Implementation of Set via binary Patricia tries

to trie via another abstraction function. Afterwards the overall correctness follows trivially by composing the two abstraction functions.

The abstraction function abs_trieP is defined via an auxiliary function that prefixes a trie with a bit list:

prefix_trie :: bool list ⇒ trie ⇒ trie
prefix_trie [] t = t
prefix_trie (k # ks) t
= (let t' = prefix_trie ks t
  in Nd False (if k then (Lf, t') else (t', Lf)))
Again we take a lazy approach and set the invariant on trieP to True.

Correctness of emptyP is trivial. Correctness of the remaining operations is proved by induction and requires a number of supporting inductive lemmas which we display before the corresponding correctness property.

Correctness of isinP:

\[
\text{isin} \ (\text{prefix\_trie} \ \text{ps} \ t) \ \text{ks} = (\text{ps} = \text{take} \ |\text{ps}| \ \text{ks} \ \land \ \text{isin} \ (\text{drop} \ |\text{ps}| \ \text{ks}))
\]
\[
\text{isinP} \ t \ \text{ks} = \text{isin} \ (\text{abs\_trieP} \ t) \ \text{ks} \quad (12.3)
\]

Correctness of insertP:

\[
\text{prefix\_trie} \ \text{ks} \ (\text{Nd} \ \text{True} \ (\text{Lf}, \ \text{Lf})) = \text{insert} \ \text{ks} \ \text{Lf}
\]
\[
\text{insert ps} \ (\text{prefix\_trie} \ \text{ps} \ (\text{Nd} \ \text{b} \ \text{lr})) = \text{prefix\_trie} \ \text{ps} \ (\text{Nd} \ \text{True} \ \text{lr})
\]
\[
\text{insert} \ (\text{ks} @ \ \text{ks'}) \ (\text{prefix\_trie} \ \text{ks} \ t) = \text{prefix\_trie} \ \text{ks} \ (\text{insert} \ \text{ks'} \ t)
\]
\[
\text{prefix\_trie} \ (\text{ps} @ \ \text{qs}) \ t = \text{prefix\_trie} \ \text{ps} \ (\text{prefix\_trie} \ \text{qs} \ t)
\]
\[
\text{split} \ \text{ks} \ \text{ps} = (\text{qs}, \ \text{ks'}, \ \text{ps'}) \quad \rightarrow
\]
\[
\text{ks} = \text{qs} @ \ \text{ks'} \ \land \ \text{ps} = \text{qs} @ \ \text{ps'} \ \land
\]
\[
(\text{ks'} \neq [] \ \land \ \text{ps'} \neq [] \ \rightarrow \ \text{hd} \ \text{ks'} \neq \ \text{hd} \ \text{ps'})
\]
\[
\text{abs\_trieP} \ (\text{insertP} \ \text{ks} \ t) = \text{insert} \ \text{ks} \ (\text{abs\_trieP} \ t) \quad (12.4)
\]

Correctness of deleteP:

\[
\text{delete} \ \text{xs} \ (\text{prefix\_trie} \ \text{xs} \ (\text{Nd} \ \text{b} \ (l, \ r)))
\]
\[
= (\text{if} \ (l, \ r) = (\text{Lf}, \ \text{Lf}) \ \text{then} \ \text{Lf} \ \text{else} \ \text{prefix\_trie} \ \text{xs} \ (\text{Nd} \ \text{False} \ (l, \ r)))
\]
\[
\text{delete} \ (\text{xs} @ \ \text{ys}) \ (\text{prefix\_trie} \ \text{xs} \ t)
\]
\[
= (\text{if} \ \text{delete} \ \text{ys} \ t = \ \text{Lf} \ \text{then} \ \text{Lf} \ \text{else} \ \text{prefix\_trie} \ \text{xs} \ (\text{delete} \ \text{ys} \ t))
\]
\[
\text{abs\_trieP} \ (\text{deleteP} \ \text{ks} \ t) = \text{delete} \ \text{ks} \ (\text{abs\_trieP} \ t) \quad (12.5)
\]

It is now trivial to obtain the correctness of the trieP implementation of sets. The abstraction function is simply the composition of the two abstraction abstraction functions: set\_trieP = set\_trie \ o \ abs\_trieP. The required correctness properties (ignoring emptyP and isinP)

\[
\text{set\_trieP} \ (\text{insertP} \ \text{xs} \ t) = \text{set\_trieP} \ t \ \cup \ \{\text{xs}\}
\]
\[
\text{set\_trieP} \ (\text{deleteP} \ \text{xs} \ t) = \text{set\_trieP} \ t \ \setminus \ \{\text{xs}\}
\]

are trivial compositions of (12.1)/(12.2) and (12.4)/(12.5).
12.4 Exercises

Exercise 12.1. Rework the above theory of binary tries as follows. Eliminate the bool argument from constructor Nd by replacing Nd by two constructors representing Nd True and Nd False.

Exercise 12.2. Define the invariant invar that characterizes fully shrunk binary tries, i.e. tries where every non-Lf trie represents a non-empty set. Note that a trie represents the empty set if it does not contain any node Nd True _. Prove that insert and delete maintain the invariant.

Bibliographic Remarks

Tries were first sketched by De La Briandais [15] and described in more detail by Fredkin [20] who coined their name based on the word reTRIEval. However, “trie” is usually pronounced like “try” rather than “tree” to avoid confusion. Patricia tries are due to Morrison [50].
Huffman’s Algorithm

Huffman’s algorithm [32] is a simple and elegant procedure for constructing a binary tree with minimum weighted path length—a measure of cost that considers both the lengths of the paths from the root to the leaf nodes and the weights associated with the leaf nodes. The algorithm’s main application is data compression: By equating leaf nodes with characters, and weights with character frequencies, we can use it to derive optimum binary codes. A binary code is a map from characters to non-empty sequences of bits.

This chapter presents Huffman’s algorithm and its optimality proof. In a slight departure from the rest of this book, the emphasis is more on graphical intuitions and less on rigorous logical arguments.

13.1 Binary Codes

Suppose we want to encode strings over a finite source alphabet as sequences of bits. Fixed-length codes like ASCII are simple and fast, but they generally waste space. If we know the frequency $w_a$ of each source symbol $a$, we can save space by using shorter code words for the most frequent symbols. We say that a variable-length code is optimum if it minimizes the sum $\sum_a w_a \delta_a$, where $\delta_a$ is the length of the binary code word for $a$.

As an example, consider the string ‘abacabad’. Encoding it with the code $C_1 = \{a \mapsto 0, b \mapsto 10, c \mapsto 110, d \mapsto 111\}$ gives the 14-bit code word 01001100100111. The code $C_1$ is optimum: No code that unambiguously encodes source symbols one at a time could do better than $C_1$ on the input ‘abacabad’. With a fixed-length code such as $C_2 = \{a \mapsto 00, b \mapsto 01, c \mapsto 10, d \mapsto 11\}$
we need at least 16 bits to encode the same string.

Binary codes can be represented by binary trees. For example, the trees

correspond to \( C_1 \) and \( C_2 \). The code word for a given symbol can be obtained as follows: Start at the root and descend toward the leaf node associated with the symbol one node at a time; emit a 0 whenever the left child of the current node is chosen and a 1 whenever the right child is chosen. The generated sequence of 0s and 1s is the code word.

To avoid ambiguities, we require that only leaf nodes are labeled with symbols. This ensures that no code word is a prefix of another. Moreover, it is sufficient to consider only full binary trees (trees whose inner nodes all have two children), because any node with only one child can advantageously be eliminated by removing it and letting the child take its parent’s place.

Each node in a code tree is assigned a weight. For a leaf node, the weight is the frequency of its symbol; for an inner node, it is the sum of the weights of its subtrees. In diagrams, we often annotate the nodes with their weights.

### 13.2 The Algorithm

Huffman's algorithm is a very simple procedure for constructing an optimum code tree for specified symbol frequencies: First, create a list of leaf nodes, one for each symbol in the alphabet, taking the given symbol frequencies as node weights. The nodes must be sorted in increasing order of weight. Second, pick the two trees

with the lowest weights and insert the tree
into the list so as to keep it ordered. Finally, repeat the process until only one tree is left in the list.

As an illustration, executing the algorithm for the frequencies \( f_d = 3, f_e = 11, f_f = 5, f_s = 7, f_z = 2 \) gives rise to the following sequence of states:

1. 

2. 

3. 

4. 

13.2 The Algorithm
The resulting tree is optimum for the given frequencies.

13.3 The Implementation

The functional implementation of the algorithm relies on the following type:

\[
\text{datatype} \ 'a\ \text{tree} = \text{Leaf} \ 'a\ \text{nat} | \text{Node} \ (\ 'a\ \text{tree}) \ (\ 'a\ \text{tree})
\]

Leaf nodes are of the form \( \text{Leaf} \ w \ a \), where \( a \) is a symbol and \( w \) is the frequency associated with \( a \), and inner nodes are of the form \( \text{Node} \ w \ t_1 \ t_2 \), where \( t_1 \) and \( t_2 \) are the left and right subtrees and \( w \) caches the sum of the weights of \( t_1 \) and \( t_2 \). The \( \text{cachedWeight} \) function extracts the weight stored in a node:

\[
\text{cachedWeight} :: \ 'a\ \text{tree} \Rightarrow \text{nat} \\
\text{cachedWeight} \ (\text{Leaf} \ w \ _) = w \\
\text{cachedWeight} \ (\text{Node} \ w \ _ \ _) = w
\]

The implementation builds on two additional auxiliary functions. The first one, \( \text{uniteTrees} \), combines two trees by adding an inner node above them:

\[
\text{uniteTrees} :: \ 'a\ \text{tree} \Rightarrow \ 'a\ \text{tree} \Rightarrow \ 'a\ \text{tree} \\
\text{uniteTrees} \ t_1 \ t_2 = \text{Node} \ (\text{cachedWeight} \ t_1 + \text{cachedWeight} \ t_2) \ t_1 \ t_2
\]
The second function, `insortTree`, inserts a tree into a list sorted by cached weight, preserving the sort order:

\[
\text{insortTree} \colon \text{a tree} \Rightarrow \text{a tree list} \Rightarrow \text{a tree list}
\]

\[
\text{insortTree} \; u \; [] = [u]
\]

\[
\text{insortTree} \; u \; (t \; \# \; ts) =
\begin{cases} 
  u \; \# \; t \; \# \; ts & \text{if } \text{cachedWeight} \; u \leq \text{cachedWeight} \; t \\
  t \; \# \; \text{insortTree} \; u \; ts & \text{else}
\end{cases}
\]

The main function that implements Huffman’s algorithm follows:

\[
\text{huffman} \colon \text{a tree list} \Rightarrow \text{a tree}
\]

\[
\text{huffman} \; [t] = t
\]

\[
\text{huffman} \; (t_1 \; \# \; t_2 \; \# \; ts) = \text{huffman} \; \text{insortTree} \; \text{uniteTrees} \; t_1 \; t_2 \; ts
\]

The function should initially be invoked with a non-empty list of leaf nodes sorted by weight. It repeatedly unites the first two trees of the list it receives as argument until a single tree is left.

### 13.4 Basic Auxiliary Functions Needed for the Proof

This section introduces basic concepts such as alphabet, consistency and optimality, which are needed to state the correctness and optimality of Huffman’s algorithm. The next section introduces more specialized functions that arise in the proof.

The *alphabet* of a code tree is the set of symbols appearing in the tree’s leaf nodes:

\[
\text{alphabet} \colon \text{a tree} \Rightarrow \text{a set}
\]

\[
\text{alphabet} \; (\text{Leaf} \; _a) = \{a\}
\]

\[
\text{alphabet} \; (\text{Node} \; _{t_1 \; t_2}) = \text{alphabet} \; t_1 \; \cup \; \text{alphabet} \; t_2
\]

A tree is consistent if for each inner node the alphabets of the two subtrees are disjoint. Intuitively, this means that a symbol occurs in at most one leaf node. Consistency is a sufficient condition for \(\delta_a\) (the length of the code word for \(a\)) to be uniquely defined. This well-formedness property appears as an assumption in many of the lemmas. The definition follows:
consistent :: 'a tree ⇒ bool
consistent (Leaf _) = True
consistent (Node _ t1 t2)
= (alphabet t1 ∩ alphabet t2 = {} ∧ consistent t1 ∧ consistent t2)

The depth of a symbol (which we wrote as \( \delta_a \) above) is the length of the path from the root to that symbol, or equivalently the length of the code word for the symbol:

\[
depth :: 'a tree ⇒ 'a ⇒ nat
depth (Leaf _) = 0
depth (Node _ t1 t2) a
= (if a ∈ alphabet t1 then depth t1 a + 1
  else if a ∈ alphabet t2 then depth t2 a + 1 else 0)
\]

By convention, symbols that do not occur in the tree or that occur at the root of a one-node tree are given a depth of 0. If a symbol occurs in several leaf nodes (of an inconsistent tree), the depth is arbitrarily defined in terms of the leftmost node labeled with that symbol.

The height of a tree is the length of the longest path from the root to a leaf node, or equivalently the length of the longest code word:

\[
height :: 'a tree ⇒ nat
height (Leaf _) = 0
height (Node _ t1 t2) = max (height t1) (height t2) + 1
\]

The frequency of a symbol (which we wrote as \( w_a \) above) is the sum of the weights attached to the leaf nodes labeled with that symbol:

\[
freq :: 'a tree ⇒ 'a ⇒ nat
freq (Leaf w a) b = (if b = a then w else 0)
freq (Node _ t1 t2) b = freq t1 b + freq t2 b
\]

For consistent trees, the sum comprises at most one nonzero term. The frequency is then the weight of the leaf node labeled with the symbol, or 0 if there is no such node.

Two trees are comparable if they have the same alphabet and symbol frequencies. This is an important concept, because it allows us to state not
only that the tree constructed by Huffman’s algorithm is optimal but also that it has the expected alphabet and frequencies.

The \texttt{weight} function returns the weight of a tree:

\begin{verbatim}
weight :: 'a tree \Rightarrow \mathbb{N}
weight (Leaf w _) = w
weight (Node _ t1 t2) = weight t1 + weight t2
\end{verbatim}

In the \texttt{Node} case, we ignore the weight cached in the node and instead compute the tree’s weight recursively.

The \texttt{cost} (or \textit{weighted path length}) of a consistent tree is the sum

\[
\sum_{a \in \text{alphabet } t} \text{freq } t a \cdot \text{depth } t a
\]

which we wrote as \(\sum_a w_a \delta_a\) above. It is defined recursively by

\begin{verbatim}
cost :: 'a tree \Rightarrow \mathbb{N}
cost (Leaf _ _) = 0
cost (Node _ t1 t2) = weight t1 + cost t1 + weight t2 + cost t2
\end{verbatim}

A tree is optimum if and only if its cost is not greater than that of any comparable tree:

\begin{verbatim}
optimum :: 'a tree \Rightarrow \mathbb{B}
optimum t
= (\forall u. \text{consistent } u \rightarrow
  \text{alphabet } t = \text{alphabet } u \rightarrow
  \text{freq } t = \text{freq } u \rightarrow \text{cost } t \leq \text{cost } u)
\end{verbatim}

Tree functions are readily generalized to lists of trees, or \textit{forests}. For example, the alphabet of a forest is defined as the union of the alphabets of its trees. The forest generalizations have a subscript \(F\) attached to their name (e.g. \texttt{alphabet}\_F).

\subsection{13.5 Other Functions Needed for the Proof}

The optimality proof needs to interchange nodes in trees, to replace a two-leaf subtree with weights \(w_1\) and \(w_2\) by a single leaf of weight \(w_1 + w_2\) and
vice versa, and to refer to the two symbols with the lowest frequencies. These concepts are represented by four functions: swapSyms, swapFourSyms, mergeSibling, splitLeaf and minima.

The interchange function swapSyms takes a three t and two symbols a, b, and exchanges them:

\[
\text{swapSyms} :: 'a \tau \Rightarrow 'a \Rightarrow 'a \\
\text{swapSyms} \ t \ a \ b = \text{swapLeaves} \ t \ (\text{freq} \ t \ a) \ a \ (\text{freq} \ t \ b) \ b
\]

The following lemma captures the intuition that to minimize the cost, more frequent symbols should be encoded using fewer bits than less frequent ones:

**Lemma 13.1.**
\begin{align*}
\text{consistent} \ t \land a \in \text{alphabet} \ t \land b \in \text{alphabet} \ t \land \text{freq} \ t \ a \leq \text{freq} \ t \ b \land \\
\text{depth} \ t \ a \leq \text{depth} \ t \ b \implies \\
\text{cost} \ (\text{swapSyms} \ t \ a \ b) \leq \text{cost} \ t
\end{align*}

The four-way symbol interchange function swapFourSyms takes four symbols a, b, c, d with a \(\neq\) b and c \(\neq\) d, and exchanges them so that a and b occupy c's and d's positions. A naive definition of this function would be

\[
\text{swapSyms} \ (\text{swapSyms} \ t \ a \ c) \ b \ d.
\]

This naive definition fails in the face of aliasing: If a = d, but b \(\neq\) c, then swapFourSyms a b c d would leave a in b's position. Instead, we use this definition:

\[
\text{swapFourSyms} :: 'a \tau \Rightarrow 'a \Rightarrow 'a \Rightarrow 'a \Rightarrow 'a \\
\text{swapFourSyms} \ t \ a \ b \ c \ d = (\text{if} \ a = d \ \text{then} \ \text{swapSyms} \ t \ b \ c \\
\text{else if} \ b = c \ \text{then} \ \text{swapSyms} \ t \ a \ d \\
\text{else} \ \text{swapSyms} \ (\text{swapSyms} \ t \ a \ c) \ b \ d)
\]

Given a symbol a, the mergeSibling function transforms the tree

![Diagram](image)

The frequency of a in the resulting tree is the sum of the original frequencies of a and b. The function is defined by the equations
mergeSibling :: 'a tree ⇒ 'a ⇒ 'a tree
mergeSibling (Leaf w b) _ = Leaf w b
mergeSibling (Node w (Leaf w b) (Leaf w c)) a
= (if a = b ∨ a = c then Leaf (w b + w c) a
   else Node w (Leaf w b) (Leaf w c))
mergeSibling (Node w (Node v va vb) t2) a
= Node w (mergeSibling (Node v va vb) a) (mergeSibling t2 a)
mergeSibling (Node w t1 (Node v va vb)) a
= Node w (mergeSibling t1 a) (mergeSibling (Node v va vb) a)

The sibling function returns the label of the node that is the (left or right) sibling of the node labeled with the given symbol a in tree t. If a is not in t's alphabet or it occurs in a node with no sibling leaf, we simply return a. This gives us the nice property that if t is consistent, then sibling t a ≠ a if and only if a has a sibling. The definition, which is omitted here, distinguishes the same cases as mergeSibling.

Using the sibling function, we can state that merging two sibling leaves with weights w_a and w_b decreases the cost by w_a + w_b:

Lemma 13.2. consistent t ∧ sibling t a ≠ a ⇒
   cost (mergeSibling t a) + freq t a + freq t (sibling t a) = cost t

The splitLeaf function undoes the merging performed by mergeSibling: Given two symbols a, b and two frequencies w_a, w_b, it transforms

\[
\begin{array}{c}
\text{a} \\
\end{array}
\rightarrow
\begin{array}{c}
\text{a} \\
\text{a} \\
\text{b}
\end{array}
\]

In the resulting tree, a has frequency w_a and b has frequency w_b. We normally invoke splitLeaf with w_a and w_b such that freq t a = w_a + w_b. The definition follows:

splitLeaf :: 'a tree ⇒ nat ⇒ 'a ⇒ nat ⇒ 'a ⇒ 'a tree
splitLeaf (Leaf w c) w_a a w_b
= (if c = a then Node w_c (Leaf w_a a) (Leaf w_b b) else Leaf w_c c)
splitLeaf (Node w t1 t2) w_a a w_b

13.6 The Key Lemmas and Theorems

It is easy to prove that the tree returned by Huffman's algorithm preserves the alphabet, consistency and symbol frequencies of the original forest:

\[
\begin{align*}
\text{ts} \neq [] & \rightarrow \text{alphabet} \left(\text{huffman ts}\right) = \text{alphabet}_{f} \text{ ts} \\
\text{consistent}_{f} \text{ ts} \land \text{ts} \neq [] & \rightarrow \text{consistent} \left(\text{huffman ts}\right) \\
\text{ts} \neq [] & \rightarrow \text{freq} \left(\text{huffman ts}\right) a = \text{freq}_{f} \text{ ts} a
\end{align*}
\]

The main difficulty is to prove the optimality of the tree constructed by Huffman's algorithm. We need three lemmas before we can present the optimality theorem.

First, if a and b are minima, and c and d are at the very bottom of the tree, then exchanging a and b with c and d does not increase the tree’s cost. Graphically, we have

\[
\text{cost} \leq \text{cost}
\]
Lemma 13.4. consistent \( t \land \text{minima} \ t \ ab \ \land \\
\ c \in \text{alphabet} \ t \land \ d \in \text{alphabet} \ t \land \\
\ \text{depth} \ t \ c = \text{height} \ t \land \ \text{depth} \ t \ d = \text{height} \ t \land \ c \neq d \rightarrow \\
\ \text{cost} \ (\text{swapFourSyms} \ t \ ab \ cd) \leq \text{cost} \ t \\

Proof. The proof is by case analysis on \( a = c, a = d, b = c, \) and \( b = d \). The cases are easy to prove by expanding the definition of \text{swapFourSyms} and applying Lemma 13.1.

The tree \text{splitLeaf} \ t \ wa \ wb \ b \ is optimum if \( t \) is optimum, under a few assumptions, notably that \( \text{freq} \ t \ a = wa + wb \). Graphically:

![Diagram](image1)

Lemma 13.5. consistent \( t \land \text{optimum} \ t \land \\
\ a \in \text{alphabet} \ t \land \ b \notin \text{alphabet} \ t \land \ \text{freq} \ t \ a = wa + wb \land \\
\ (\forall \ c \in \text{alphabet} \ t. \ wa \leq \text{freq} \ t \ c \land \ wb \leq \text{freq} \ t \ c) \rightarrow \\
\ \text{optimum} \ (\text{splitLeaf} \ t \ wa \ ab) \\

Proof. We assume that \( t \)'s cost is less than or equal to that of any other comparable tree \( v \) and show that \text{splitLeaf} \ t \ wa \ ab \ b \ has a cost less than or equal to that of any other comparable tree \( u \). For the nontrivial case where \( \text{height} \ t > 0 \), it is easy to prove that there must be two symbols \( c \) and \( d \) occurring in sibling nodes at the very bottom of \( u \). From \( u \) we construct the tree \text{swapFourSyms} \ u \ ab \ cd \ in which the minima \( a \) and \( b \) are siblings:

![Diagram](image2)

The question mark reminds us that we know nothing specific about \( u \)'s structure. Merging \( a \) and \( b \) gives a tree comparable with \( t \), which we can use to instantiate \( v \).
\[\text{cost}(\text{splitLeaf } t \ a \ w_a \ b \ w_b) = \text{cost } t + w_a + w_b \quad \text{by Lemma 13.3}\]

\[\leq \text{cost}(\text{mergeSibling } (\text{swapFourSyms } u \ a \ b \ c \ d) \ a) + w_a + w_b \quad \text{by optimality assumption}\]

\[= \text{cost}(\text{swapFourSyms } u \ a \ b \ c \ d) \quad \text{by Lemma 13.2}\]

\[\leq \text{cost } u \quad \text{by Lemma 13.4} \quad \square\]

Once it has combined two lowest-weight trees using \textit{uniteTrees}, Huffman’s algorithm does not visit these trees ever again. This suggests that splitting a leaf node before applying the algorithm should give the same result as applying the algorithm first and splitting the leaf node afterward.

**Lemma 13.6.** \(\text{consistent}_F \ ts \land ts \neq [] \land a \in \text{alphabet}_F \ ts \land\)

\[\text{freq}_F \ ts \ a = w_a + w_b \implies\]

\[\text{splitLeaf} (\text{huffman } ts) \ a \ w_a \ b \]

\[= \text{huffman} (\text{splitLeaf}_F \ ts \ a \ w_a \ b)\]

The proof is by straightforward induction on the length of the forest \(ts\).

As a consequence of this commutativity lemma, applying Huffman’s algorithm on a forest of the form

\[
\begin{array}{cccc}
  \circ & c & w_c & \\
  a & w_a & b & w_b \\
  d & w_d & \ldots & z & w_z \\
\end{array}
\]

gives the same result as applying the algorithm on the “flat” forest

\[
\begin{array}{cccc}
  \circ & c & w_c & \\
  a & w_a + w_b & d & w_d \\
  \ldots & z & w_z \\
\end{array}
\]

followed by splitting the leaf node \(a\) into two nodes \(a\) and \(b\) with frequencies \(w_a, w_b\). The lemma effectively provides a way to flatten the forest at each step of the algorithm.

This leads us to our main result.

**Theorem 13.7.** \(\text{consistent}_F \ ts \land \text{height}_F \ ts = 0 \land \text{sortedByWeight } ts \land\)

\[ts \neq [] \implies\]

\[\text{optimum } (\text{huffman } ts)\]

**Proof.** The proof is by induction on the length of \(ts\). The assumptions ensure that \(ts\) is of the form

\[
\begin{array}{cccc}
  a & w_a & b & w_b \\
  c & w_c & d & w_d \\
  \ldots & z & w_z \\
\end{array}
\]
with \( w_a \leq w_b \leq w_c \leq w_d \leq \cdots \leq w_z \). If \( ts \) consists of a single node, the node has cost 0 and is therefore optimum. If \( ts \) has length 2 or more, the first step of the algorithm leaves us with a term such as

\[
\text{huffman} \begin{array}{c}
\text{c} \\
\text{w}_c
\end{array} \quad \text{a} \quad \text{b} \quad \text{d} \quad \text{w}_d \quad \cdots \quad \text{z} \quad \text{w}_z
\]

In the diagram, we put the newly created tree at position 2 in the forest; in general, it could be anywhere. By Lemma 13.6, the above tree equals

\[
\text{splitLeaf} \begin{pmatrix}
\text{huffman} \\
\text{w}_a \\
\text{w}_b \\
\text{w}_c \\
\text{w}_d \\
\text{z} \\
\text{w}_z
\end{pmatrix} \begin{pmatrix}
\text{c} \\
\text{w}_c \\
\text{a} \\
\text{w}_a + \text{w}_b \\
\text{d} \\
\text{w}_d \\
\text{z} \\
\text{w}_z
\end{pmatrix} \begin{pmatrix}
\text{w}_a \\
\text{w}_b \\
\text{b}
\end{pmatrix}
\]

To prove that this tree is optimum, it suffices by Lemma 13.5 to show that

\[
\text{huffman} \begin{array}{c}
\text{c} \\
\text{w}_c \\
\text{w}_a + \text{w}_b \\
\text{w}_d \\
\text{z} \\
\text{w}_z
\end{array} \quad \text{a} \quad \text{b}
\]

is optimum, which follows from the induction hypothesis. \( \square \)

In summary, we have established that the \textit{huffman} program, which constitutes a functional implementation of Huffman's algorithm, constructs a binary tree that represents an optimal binary code for the specified alphabet and frequencies.

13.7 Chapter Notes

The sorted list of trees constitutes a simple priority queue (Part III). The time complexity of Huffman’s algorithm is quadratic in the size \( n \) of this queue. By using a binary search to implement \textit{insortTree}, we can obtain an \( O(n \lg n) \) imperative implementation. An \( O(n) \) implementation is possible by maintaining two queues, one containing the unprocessed leaf nodes and the other containing the combined trees [43].

Huffman’s algorithm was invented by David Huffman [32]. The proof above was inspired by Knuth’s informal argument [43]. This chapter's text is based on a published article [8]. An alternative formal proof, developed using Coq, is due to Théry [73].

Knuth [42] presented an alternative, more abstract view of Huffman’s algorithm as a “Huffman algebra.” Could his approach help simplify our proof? The most tedious steps above concerned splitting nodes, merging siblings, and
swapping symbols. These steps would still be necessary, as the algebraic approach seems restricted to abstracting over the arithmetic reasoning, which is not very difficult in the first place. On the other hand, with Knuth’s approach, perhaps the proof would gain in elegance.
Part III

Priority Queues
A priority queue of linearly ordered elements is like a multiset where one can insert arbitrary elements and remove minimal elements. Its specification as an ADT is shown in Figure 14.1 where Min_mset $m \equiv \text{Min}(\text{set}_m)\text{mset} m$ and Min yields the minimal element of a finite set of linearly ordered elements.

**ADT Priority_Queue**

**interface**

- empty :: 'q
- insert :: 'a \Rightarrow 'q \Rightarrow 'q
- del_min :: 'q \Rightarrow 'q
- get_min :: 'q \Rightarrow 'a

**abstraction** mset :: 'q \Rightarrow 'a multiset

**invariant** invar :: 'q \Rightarrow bool

**specification**

- invar empty (empty-inv)
- mset empty = \{\} (empty)
- invar q \Rightarrow invar (insert x q) (insert-inv)
- invar q \Rightarrow mset (insert x q) = mset q + \{x\} (insert)
- invar q \land mset q \neq \{\} \Rightarrow invar (del_min q) (del_min-inv)
- invar q \land mset q \neq \{\} \Rightarrow mset (del_min q) = mset q - \{\text{get_min q}\} (del_min)
- invar q \land mset q \neq \{\} \Rightarrow \text{get_min q} = \text{Min_mset}(mset q) (get_min)

Fig. 14.1. ADT Priority_Queue

Mergeable priority queues (see Figure 14.2) provide an additional function merge (sometimes: meld or union) with the obvious functionality.
14 Priority Queues

**ADT** *Priority.Queue.Merge* $=$ *Priority.Queue* +

**interface**
merge $::$ ‘q $\Rightarrow$ ‘q $\Rightarrow$ ‘q

**specification**
invar $q_1 \land$ invar $q_2 \rightarrow$ invar (merge $q_1$ $q_2$)
invar $q_1 \land$ invar $q_2 \rightarrow$ mset (merge $q_1$ $q_2$) $=$ mset $q_1$ + mset $q_2$

Fig. 14.2. ADT *Priority.Queue.Merge*

Our priority queues are simplified. The more general version contains elements that are pairs of some item and its priority. In that case a priority queue can be viewed as a set, but because we have dropped the item we need to view a priority queue as a multiset. In imperative implementations, priority queues frequently also provide an operation *decrease_key*: given some direct reference to an element in the priority queue, decrease the element’s priority. This is not completely straightforward in a functional language. Lammich and Nipkow [46] present an implementation, a Priority Search Tree.

**Exercise 14.1.** Give a list-based implementation of mergeable priority queues with constant-time *get_min* and *del_min*. Verify the correctness of your implementation w.r.t. *Priority.Queue.Merge*.

14.1 Heaps

A popular implementation technique for priority queues are heaps, i.e. trees where the minimal element in each subtree is at the root:

```plaintext
heap :: 'a tree $\Rightarrow$ bool
heap $\langle \rangle$ = True
heap $\langle l, m, r \rangle$
  = ((\forall x \in$ set$\_tree l \cup$ set$\_tree r$. $m \leq x)$ $\land$ heap $l$ $\land$ heap $r$)
```

Function *mset.tree* extracts the multiset of elements from a tree:

```plaintext
mset-tree :: 'a tree $\Rightarrow$ 'a multiset
mset-tree $\langle \rangle$ = $\{\}$
mset-tree $\langle l, a, r \rangle$ = $\{a\} +$ mset-tree $l$ + mset-tree $r$
```
When verifying a heap-based implementation of priority queues the invariant \textit{invar} and the abstraction function \textit{mset} in the ADT \textit{Priority\_Queue} are instantiated by \textit{heap} and \textit{mset\_tree}. The correctness proofs need to talk about both multisets and (because of the \textit{heap} invariant) sets of elements in a heap. We will only show the relevant multiset properties because the set properties follow easily via \textit{set\_mset} (\textit{mset\_tree} $t$) = \textit{set\_tree} $t$.

Both \textit{empty} and \textit{get\_min} have obvious implementations:

\begin{verbatim}
empty = \langle \rangle
get\_min = value
\end{verbatim}

where value $\langle \_, a, \_ \rangle = a$. If a heap-based implementation provides a \textit{merge} function (e.g. skew heaps in Chapter 22), then \textit{insert} and \textit{del\_min} can be defined like this:

\begin{verbatim}
insert x t = merge $\langle \langle \rangle, x, \langle \rangle \rangle$ t

del\_min $\langle \rangle$ = $\langle \rangle$
del\_min $\langle l, \_, r \rangle$ = merge $l$ $r$
\end{verbatim}

Note that the following tempting definition of \textit{merge} is functionally correct but leads to very unbalanced heaps:

\begin{verbatim}
merge $\langle \rangle$ t = t
merge t $\langle \rangle$ = t
merge $\langle l_1, a_1, r_1 \rangle$ =: t_1 $\langle l_2, a_2, r_2 \rangle$ =: t_2 $\langle l_3, a_3, r_3 \rangle$
= (if $a_1 \leq a_2$ then $\langle l_1, a_1, merge\ r_1\ t_2 \rangle$ else $\langle l_2, a_2, merge\ t_1\ r_2 \rangle$)
\end{verbatim}

Many of the more advanced implementations of heaps focus on improving this merge function. We will see examples of this in the next chapter on leftist heaps, as well as in the chapters on skew heaps and pairing heaps.

\textbf{Exercise 14.2.} Show functional correctness of the above definition of \textit{merge} (w.r.t. \textit{Priority\_Queue\_Merge}) and prove functional correctness of the implementations of \textit{insert} and \textit{del\_min} (w.r.t. \textit{Priority\_Queue}).

\section*{Bibliographic Remarks}

The idea of the heap goes back to Williams [77] who also coined the name.
Leftist Heaps

Leftist heaps are heaps in the sense of Section 14.1 and implement mergeable priority queues. The key idea is to maintain the invariant that at each node the minimal height of the right child is ≤ that of the left child. We represent leftist heaps as augmented trees that store the minimal height in every node:

```
type_synonym 'a lheap = ('a × nat) tree

mht :: 'a lheap ⇒ nat
mht () = 0
mht (_, (_, n), _) = n
```

There are two invariants: the standard heap invariant (on augmented trees)

```
heap :: ('a × 'b) tree ⇒ bool
heap () = True
heap (l, (m, _), r)
= ((∀ x∈set_tree l ∪ set_tree r. m ≤ x) ∧ heap l ∧ heap r)
```

and the structural leftist tree invariant that requires that the minimal height of the right child is no bigger than that of the left child (and that the minimal height information in the node is correct):

```
ltree :: 'a lheap ⇒ bool
ltree () = True
ltree (l, (_, n), r)
= (mh r ≤ mh l ∧ n = mh r + 1 ∧ ltree l ∧ ltree r)
```
Thus a tree is a leftist tree if for every subtree the right spine is a shortest path from the root to a leaf. Pictorially:

Now remember $2^{mh(t)} \leq |t|_1$, i.e. $mh(t) \leq \lg |t|_1$. Because the expensive operations on leftist heaps descend along the right spine, this means that their running time is logarithmic in the size of the heap.

**Exercise 15.1.** An alternative definition of leftist tree is via the length of the right spine of the tree:

$$rank :: 'a tree \Rightarrow \text{nat}$$

$$rank (\langle \rangle) = 0$$

$$rank (\langle _, _, r \rangle) = rank r + 1$$

Prove that $ltree t \rightarrow rank t = mh(t)$.

### 15.1 Implementation of ADT `Priority_Queue_Merge`

The key operation is `merge`:

```haskell
merge :: 'a lheap \Rightarrow 'a lheap \Rightarrow 'a lheap
merge (\langle \rangle) t = t
merge t (\langle \rangle) = t
merge (\langle l_1, (a_1, n_1), r_1 \rangle) (\langle l_2, (a_2, n_2), r_2 \rangle) =:
  (if a_1 \leq a_2 then node l_1 a_1 (merge r_1 t_2)
   else node l_2 a_2 (merge t_1 r_2))
```

```haskell
node :: 'a lheap \Rightarrow 'a \Rightarrow 'a lheap \Rightarrow 'a lheap
node l a r
= (let mhl = mht l; mhr = mht r
    in if mhr \leq mhl then (l, (a, mhr + 1), r)
      else (r, (a, mhl + 1), l))
```

Termination of `merge` can be proved either by the sum of the sizes of the two arguments (which goes down with every call) or by the lexicographic product of the two size measures: either the first argument becomes smaller or it stays unchanged and the second argument becomes smaller.
As shown in Section 14.1, once we have merge, the other operations are easily definable. We repeat their definitions simply because this chapter employs augmented rather than ordinary trees:

```plaintext
empty :: 'a lheap
empty = ⟨⟩

get_min :: 'a lheap ⇒ 'a
get_min ⟨_, (a, _), _⟩ = a

insert :: 'a ⇒ 'a lheap ⇒ 'a lheap
insert x t = merge ⟨⟨⟩, (x, 1), ⟨⟩⟩ t

del_min :: 'a lheap ⇒ 'a lheap
del_min ⟨⟩ = ⟨⟩
del_min ⟨l, _, r⟩ = merge l r
```

### 15.2 Correctness

The above implementation is proved correct w.r.t. the ADT `Priority_Queue_Merge` where

```plaintext
mset_tree :: ('a × 'b) tree ⇒ 'a multiset
mset_tree ⟨⟩ = {⟨⟩}
mset_tree ⟨l, (a, _), r⟩ = {⟨a⟩} + mset_tree l + mset_tree r

invar t = (heap t ∧ ltree t)
```

Correctness of `get_min` follows directly from the heap invariant:

\[ \text{heap } t ∨ t ≠ ⟨⟩ \rightarrow \text{get_min } t = \text{Min } (\text{set_tree } t) \]

From the following inductive lemmas about `merge`

\[ mset_tree (\text{merge } t_1 t_2) = mset_tree t_1 + mset_tree t_2 \]
\[ \text{ltree } l ∨ \text{ltree } r \rightarrow \text{ltree } (\text{merge } l r) \]
\[ \text{heap } l ∨ \text{heap } r \rightarrow \text{heap } (\text{merge } l r) \]

correctness of `insert` and `del_min` follow easily:
15.3 Running Time Analysis

The running time functions are shown in Appendix B.4. By induction on the computation of merge we obtain

\[ \text{ltree } l \wedge \text{ltree } r \longrightarrow T_{\text{merge}} \leq mh l + mh r + 1 \]

With \(2^{mh} t \leq |t|\) it follows that

\[ \text{ltree } l \wedge \text{ltree } r \longrightarrow T_{\text{merge}} \leq \lg |l| + \lg |r| + 1 \quad (15.1) \]

which implies logarithmic bounds for insertion and deletion:

\[ \text{ltree } t \longrightarrow T_{\text{insert}} \leq \lg |t| + 3 \]

\[ \text{ltree } t \longrightarrow T_{\text{del_min}} \leq 2 \cdot \lg |t| + 1 \]

The derivation of the bound for insertion is trivial but the one for deletion uses a little lemma, assuming \(\text{ltree } t\):

\[ T_{\text{del_min}} = T_{\text{merge}} \leq \lg |l| + \lg |r| + 2 \]

\[ \leq 2 \cdot \lg |t| + 1 \quad \text{using (15.1)} \]

because \(\lg x + \lg y + 1 < 2 \cdot \lg (x + y)\)

if \(0 < x\) and \(0 < y\)

Bibliographic Remarks

Leftist heaps were invented by Crane [14]. Another version of leftist trees, based on weight rather than height, was introduced by Cho and Sahni [12, 68].
Priority Queues via Braun Trees

In Chapter 11 we introduced Braun trees and showed how to implement arrays. In the current chapter we show how to implement priority queues by means of Braun trees. Because Braun trees have logarithmic height this guarantees logarithmic running times for insertion and deletion. Remember that every node \((l, x, r)\) in a Braun tree satisfies \(|l| \in \{|r|, |r| + 1\}\) (*).

16.1 Implementation of ADT Priority_Queue

We follow the heap approach in Section 14.1. Functions empty, get_min, heap and mset_tree are defined as in that section.

Insertion and deletion maintain the Braun tree property (*) by inserting into the right (and possibly smaller) child, deleting from the left (and possibly larger) child, and swapping children to reestablish (*).

Insertion is straightforward and clearly maintains both the heap and the Braun tree property:

\[
\begin{align*}
\text{insert} : & \ 'a \Rightarrow 'a \text{ tree} \Rightarrow 'a \text{ tree} \\
\text{insert} & \ a \ () = ((), \ a, ()) \\
\text{insert} & \ a \ (l, x, r) \\
& = (\text{if } a < x \ \text{then} \ (\text{insert} \ x \ r, \ a, l) \ \text{else} \ (\text{insert} \ a \ r, \ x, l))
\end{align*}
\]

To delete the minimal (i.e. root) element from a tree, extract the leftmost element from the tree and let it sift down to its correct position in the tree in the manner of heapsort:
### 16.2 Correctness

We outline the correctness proofs for `insert` and `del_min` by presenting the key lemmas. Correctness of `insert` is straightforward:

\[
|\text{insert } x t| = |t| + 1
\]

\[
\text{mset_tree } (\text{insert } x t) = \{x\} + \text{mset_tree } t
\]

\[
\text{braun } t \rightarrow \text{braun } (\text{insert } x t)
\]

\[
\text{heap } t \rightarrow \text{heap } (\text{insert } x t)
\]

Correctness of `del_min` builds on analogous correctness lemmas for the auxiliary functions:

\[
\text{del_left } t = (x, t') \land t \neq \langle \rangle \longrightarrow \text{mset_tree } t = \{x\} + \text{mset_tree } t'
\]

\[
\text{del_left } t = (x, t') \land t \neq \langle \rangle \land \text{heap } t \rightarrow \text{heap } t'
\]

\[
\text{del_left } t = (x, t') \land t \neq \langle \rangle \rightarrow |t| = |t'| + 1
\]

\[
\text{del_left } t = (x, t') \land t \neq \langle \rangle \land \text{braun } t \rightarrow \text{braun } t'
\]
\( \text{braun} (l, a, r) \rightarrow |sift\_down l a r| = |l| + |r| + 1 \)
\( \text{braun} (l, a, r) \rightarrow \text{braun} (sift\_down l a r) \)
\( \text{braun} (l, a, r) \rightarrow \)
\( \text{mset\_tree} (sift\_down l a r) = \{\text{a}\} + (\text{mset\_tree} l + \text{mset\_tree} r) \)
\( \text{braun} (l, a, r) \land \text{heap} l \land \text{heap} r \rightarrow \text{heap} (sift\_down l a r) \)
\( \text{braun} t \rightarrow \text{braun} (\text{del\_min} t) \)
\( \text{heap} t \land \text{braun} t \rightarrow \text{heap} (\text{del\_min} t) \)
\( \text{braun} t \rightarrow |\text{del\_min} t| = |t| - 1 \)
\( \text{braun} t \land t \neq \langle \rangle \rightarrow \)
\( \text{mset\_tree} (\text{del\_min} t) = \text{mset\_tree} t - \{\text{get\_min} t\} \)

\textbf{Bibliographic Remarks}

Our implementation of priority queues via Braun trees is due to Paulson [64] who credits it to Okasaki.
Binomial Heaps

Binomial heaps are another common implementation of mergeable priority queues, which supports efficient \(O(\log n)\) insert, get\_min, del\_min, and merge operations.

The basic building blocks of a binomial heap are binomial trees, which are defined recursively as follows. A binomial tree of rank \(r\) is a node with \(r\) children of ranks \(r - 1, \ldots, 0\), in that order. Figure 17.1 shows an example binomial tree.

![Fig. 17.1. A binomial tree of rank 3. The node labels depict the rank of each node. A node of rank \(r\) has child nodes of ranks \(r - 1, \ldots, 0\).](image)

To define binomial trees in Isabelle, we first define a more general datatype and the usual syntax for nodes:

```isabelle
datatype 'a tree = Node nat 'a ('a tree list)
(r, x, ts) ≡ Node r x ts
```

Apart from the list of children, a node stores a rank and a root element:
This datatype contains all binomial trees, but also some non-binomial trees. To carve out the binomial trees, we define an invariant, which reflects the informal definition above:

\[
\begin{align*}
\text{btree} :: 'a \, \text{tree} & \Rightarrow \text{bool} \\
\text{btree} \, \langle r, \_ , ts \rangle & = \left( (\forall t \in \text{set} \, ts. \, \text{btree} \, t) \land \text{map} \, \text{rank} \, ts = \text{rev} \,[0..<r] \right)
\end{align*}
\]

Additionally, we require the heap property, i.e., that the root element of each subtree is a minimal element in that subtree:

\[
\begin{align*}
\text{heap} :: 'a \, \text{tree} & \Rightarrow \text{bool} \\
\text{heap} \, \langle \_, x , ts \rangle & = \left( (\forall t \in \text{set} \, ts. \, \text{heap} \, t \land x \leq \text{root} \, t) \right)
\end{align*}
\]

Thus, a binomial tree is a tree that satisfies both the structural and the heap invariant. The two invariants are combined in a single predicate:

\[
\begin{align*}
\text{bheap} :: 'a \, \text{tree} & \Rightarrow \text{bool} \\
\text{bheap} \, t & = (\text{btree} \, t \land \text{heap} \, t)
\end{align*}
\]

A binomial heap is a list of binomial trees

**type_synonym**

\[
'a \, \text{trees} = 'a \, \text{tree} \, \text{list}
\]

with strictly ascending rank:

\[
\begin{align*}
\text{invar} :: 'a \, \text{trees} & \Rightarrow \text{bool} \\
\text{invar} \, \text{ts} & = \left( (\forall t \in \text{set} \, \text{ts}. \, \text{bheap} \, t) \land \text{sorted_wrt} \, (\langle) \, (\text{map} \, \text{rank} \, \text{ts}) \right)
\end{align*}
\]

Note that \text{sorted_wrt} states that a list is sorted wrt. the specified relation, here \(\langle\). It is defined in Appendix A.

### 17.1 Size

The following functions return the multiset of elements in a binomial tree and in a binomial heap:
17.2 Implementation of ADT *Priority Queue*

**17.2.1 Empty and Emptiness Test**

Obviously, the empty list is a binomial heap with no elements, and a binomial heap is empty only if it is the empty list:

```ml
invar []

mset_trees [] = {} = (mset_trees ts = {}) = (ts = [])
```

**17.2.2 Insertion**

A crucial property of binomial trees is that we can link two binomial trees of rank \( r \) to form a binomial tree of rank \( r + 1 \), simply by prepending one tree.
as the first child of the other. To preserve the heap property, we add the tree with the bigger root element below the tree with the smaller root element. This linking of trees is illustrated in Figure 17.2. Formally:

\[
\text{link :: 'a tree } \Rightarrow \text{'a tree } \Rightarrow \text{'a tree}
\]

\[
\text{link } (\langle r, x_1, t_{s1} \rangle =: t_1) (\langle r', x_2, t_{s2} \rangle =: t_2)
\]

\[
= (\text{if } x_1 \leq x_2 \text{ then } \langle r + 1, x_1, t_2 \# t_{s1} \rangle \text{ else } \langle r + 1, x_2, t_1 \# t_{s2} \rangle)
\]

Fig. 17.2. Linking two binomial trees of rank 2 to form a binomial tree of rank 3, by linking the left tree as first child of the right tree, as indicated by the dashed line. We assume that the root element of the left tree is greater than or equal to the root element of the right tree, such that the heap property is preserved.

By case distinction, we can easily prove that \( \text{link} \) preserves the invariant and that the resulting tree contains the elements of both arguments.

\[
bheap t_1 \land bheap t_2 \land \text{rank } t_1 = \text{rank } t_2 \longrightarrow bheap (\text{link } t_1 \ t_2)
\]

\[
mset_tree (\text{link } t_1 \ t_2) = mset_tree t_1 + mset_tree t_2
\]

The link operation forms the basis of inserting a tree into a heap: if the heap does not contain a tree with the same rank, we can simply insert the tree at the correct position in the heap. Otherwise, we merge the two trees and recursively insert the result. For our purposes, we can additionally assume that the rank of the tree to be inserted is smaller than or equal to the lowest rank in the heap, which saves us a case in the following definition:

\[
\text{ins_tree :: 'a tree } \Rightarrow \text{'a trees } \Rightarrow \text{'a trees}
\]

\[
\text{ins_tree } t [] = [t]
\]

\[
\text{ins_tree } t_1 (t_2 \# ts)
\]

\[
= (\text{if } \text{rank } t_1 < \text{rank } t_2 \text{ then } t_1 \# t_2 \# ts
\]

\[
\text{else } \text{ins_tree } (\text{link } t_1 \ t_2) \ ts)
\]
17.2 Implementation of ADT Priority_QUEUE

Invariant preservation and functional correctness of ins_tree is easily proved by induction using the respective properties for link:

\[
\begin{align*}
&\text{invar } bheap t \land \text{invar } ts \land (\forall t' \in \text{set } ts. \ \text{rank } t \leq \text{rank } t') \implies \text{invar } (\text{ins_tree } t \ ts) \\
&\text{mset_trees } (\text{ins_tree } t \ ts) = \text{mset_tree } t + \text{mset_trees } ts
\end{align*}
\]

A single element is inserted as a one-element (rank 0) tree:

\[
\text{insert } :: \ 'a \Rightarrow 'a \ \text{trees} \Rightarrow 'a \ \text{trees} \\
\text{insert } x \ ts = \text{ins_tree } (0, x, []) \ ts
\]

The above definition meets the specification for insert required by the Priority_QUEUE ADT:

\[
\begin{align*}
&\text{invar } t \implies \text{invar } (\text{insert } x \ t) \\
&\text{mset_trees } (\text{insert } x \ t) = \{x\} + \text{mset_trees } t
\end{align*}
\]

17.2.3 Merging

Recall the merge algorithm used in top-down merge sort (Section 2.4). It merges two sorted lists by repeatedly taking the smaller list head. We use a similar idea for merging two heaps: if the rank of one list’s head is strictly smaller, we choose it. If both ranks are equal, we link the two heads and insert the resulting tree into the merged remaining heaps. Thus, the resulting heap will be strictly ordered by rank. Formally:

\[
\begin{align*}
&\text{merge } :: \ 'a \ \text{trees} \Rightarrow 'a \ \text{trees} \Rightarrow 'a \ \text{trees} \\
&\text{merge } ts_1 [] = ts_1 \\
&\text{merge } [] \ (v \ # \ va) = v \ # \ va \\
&\text{merge } \ (t_1 \ # \ ts_1 =: h_1) \ (t_2 \ # \ ts_2 =: h_2) \\
&= (\text{if } \text{rank } t_1 < \text{rank } t_2 \ \text{then } t_1 \ # \ \text{merge } ts_1 \ h_2 \\
&\quad \quad \quad \quad \quad \text{else if } \text{rank } t_2 < \text{rank } t_1 \ \text{then } t_2 \ # \ \text{merge } h_1 \ ts_2 \\
&\quad \quad \quad \quad \quad \text{else } \text{ins_tree } (\text{link } t_1 \ t_2) \ (\text{merge } ts_1 \ ts_2))
\end{align*}
\]

The merge function can be regarded as an algorithm for adding two sparse binary numbers. This intuition is explored in Exercise 17.2.

We show that the merge operation preserves the invariant and adds the elements:

\[
\begin{align*}
&\text{invar } ts_1 \land \text{invar } ts_2 \implies \text{invar } (\text{merge } ts_1 \ ts_2) \\
&\text{mset_trees } (\text{merge } ts_1 \ ts_2) = \text{mset_trees } ts_1 + \text{mset_trees } ts_2
\end{align*}
\]
The proof is straightforward, except for preservation of the binomial heap invariant. We first show that merging two heaps does not decrease the lowest rank in these heaps. This ensures that prepending the head with smaller rank to the merged remaining heaps results in a sorted heap. Moreover, when we link two heads of equal rank, this ensures that the linked tree's rank is smaller than or equal to the ranks in the merged remaining trees, as required by the \textit{ins\_tree} function. We phrase this property as preservation of lower rank bounds, i.e., a lower rank bound of both heaps is still a lower bound for the merged heap:

\[
t' \in \text{set} (\text{merge} \ t_1 \ t_2) \land (\forall t_1 \in \text{set} \ t_1. \ \text{rank} \ t < \text{rank} \ t_1) \land \\
(\forall t_2 \in \text{set} \ t_2. \ \text{rank} \ t < \text{rank} \ t_2) \implies \text{rank} \ t < \text{rank} \ t'
\]

The proof is by straightforward induction, relying on an analogous bounding lemma for \textit{ins\_tree}.

### 17.2.4 Finding a Minimal Element

For a binomial tree, the root node always contains a minimal element. Unfortunately, there is no such property for the whole heap—the minimal element may be at the root of any of the heap's trees. To get a minimal element from a non-empty heap, we look at all root nodes:

\[
\begin{align*}
\textit{get\_min} :: & \quad \forall 'a \quad \forall \text{trees} \implies 'a \\
\textit{get\_min} \ [t] & = \text{root } t \\
\textit{get\_min} \ (t \# \ ts) & = \text{min} (\text{root } t) \ (\textit{get\_min} \ ts)
\end{align*}
\]

Correctness of this operation is proved by a simple induction:

\[
\begin{align*}
\text{mset\_trees} \ ts \neq \{\} & \land \text{invar} \ ts \implies \\
\textit{get\_min} \ ts & = \text{Min\_mset} \ (\text{mset\_trees} \ ts)
\end{align*}
\]

### 17.2.5 Deleting a Minimal Element

To delete a minimal element, we first need to find one and then remove it. Removing the root node of a tree with rank $r$ leaves us with a list of its children, which are binomial trees of ranks $r - 1, \ldots, 0$. Reversing this list yields a valid binomial heap, which we merge with the remaining trees in the original heap:
17.3 Running Time Analysis

The running time functions are shown in Appendix B.5. Intuitively, the operations are linear in the length of the heap, which in turn is logarithmic in the number of elements (cf. Section 17.1).

The running time analysis for `insert` is straightforward. The running time is dominated by `ins_tree`. In the worst case, it iterates over the whole heap, taking constant time per iteration. By straightforward induction, we show

\[ T_{\text{ins\_tree}} t \ ts \leq |ts| + 1 \]

and thus

\[
\text{del\_min} :: \text{\textquoteleft}a\text{\textquoteleft} trees \Rightarrow \text{\textquoteleft}a\text{\textquoteleft} trees \\
\text{del\_min} ts \\
= (\text{case} \ \text{get\_min\_rest} \ ts \ \text{of} \ ((\_ \_), ts_1), ts_2) \Rightarrow \text{merge} (\text{rev \ ts}_1) \ ts_2)
\]

Here, the auxiliary function `get_min_rest` splits a heap into a tree with minimal root element, and the remaining trees.

\[
\text{get\_min\_rest} :: \text{\textquoteleft}a\text{\textquoteleft} trees \Rightarrow \text{\textquoteleft}a\text{\textquoteleft} tree \times \text{\textquoteleft}a\text{\textquoteleft} trees \\
\text{get\_min\_rest} [t] = (t, []) \\
\text{get\_min\_rest} (t \# v \# va) \\
= (\text{let} \ (t', ts') = \text{get\_min\_rest} (v \# va) \ \\
\quad \text{in if} \ \text{root} \ t \leq \text{root} \ t' \ \text{then} \ (t, v \# va) \ \text{else} \ (t', t \# ts'))
\]

We prove that, for a non-empty heap, `del_min` preserves the invariant and deletes the minimal element:

\[
ts \neq [] \land \text{invar} \ ts \rightarrow \text{invar} (\text{del\_min} \ ts) \\
\text{get\_min\_rest} ts = (t', ts') \land ts \neq [] \land \text{invar} \ ts \rightarrow \text{bheap} t' \\
\text{get\_min\_rest} ts = (t', ts') \land ts \neq [] \land \text{invar} \ ts \rightarrow \text{invar} ts'
\]

To show that we actually remove a minimal element, we show that `get_min_rest` selects the same tree as `get_min`:

\[
ts \neq [] \land \text{get\_min\_rest} ts = (t', ts') \rightarrow \text{root} t' = \text{get\_min} \ ts
\]
17 Binomial Heaps

invar $ts \longrightarrow T_{\text{insert}} \; x \; ts \leq \lg (|\text{mset\_trees} \; ts| + 1) + 2$

The running time analysis for merge is more interesting. In each recursion, we need constant time to compare the ranks. However, if the ranks are equal, we link the trees and insert them into the merger of the remaining heaps. In the worst case, this costs linear time in the length of the merger. A naive analysis would estimate $|\text{merge} \; ts_1 \; ts_2| \leq |ts_1| + |ts_2|$, and thus yield a quadratic running time in the length of the heap.

However, we can do better: we observe that every link operation in $\text{ins\_tree}$ reduces the number of trees in the heap. Thus, over the whole merge, we can only have linearly many link operations in the combined size of both heaps.

To formalize this idea, we estimate the running time of $\text{ins\_tree}$ and $\text{merge}$ together with the length of the result:

$$T_{\text{ins\_tree}} \; t \; ts + |\text{ins\_tree} \; t \; ts| = 2 + |ts|$$

$$T_{\text{merge}} \; ts_1 \; ts_2 + |\text{merge} \; ts_1 \; ts_2| \leq 2 \cdot (|ts_1| + |ts_2|) + 1$$

Both estimates can be proved by straightforward induction, and from the second estimate we easily derive a bound for $\text{merge}$:

$$\text{invar} \; ts_1 \land \text{invar} \; ts_2 \longrightarrow T_{\text{merge}} \; ts_1 \; ts_2 \leq 4 \cdot \lg (|\text{mset\_trees} \; ts_1| + |\text{mset\_trees} \; ts_2| + 1) + 1$$

From the bound for $\text{merge}$, we can easily derive a bound for $\text{del\_min}$:

$$\text{invar} \; ts \land ts \neq [] \longrightarrow T_{\text{del\_min}} \; ts \leq 6 \cdot \lg (|\text{mset\_trees} \; ts| + 1) + 3$$

The only notable point is that we use a linear time bound for reversing a list, as explained in Section 1.5.1:

$$T_{\text{rev}} :: \text{'a list} \Rightarrow \text{nat}$$

$$T_{\text{rev}} \; xs = |xs| + 1$$

### 17.4 Exercises

**Exercise 17.1.** A node in a tree is on level $n$ if it needs $n$ edges from the root to reach it. Define a function $\text{nod} :: \text{nat} \Rightarrow \text{'a tree} \Rightarrow \text{nat}$ such that $\text{nod} \; n \; t$ is the number of nodes on level $n$ in tree $t$ and show that a binomial tree of rank $r$ has $\binom{r}{n}$ nodes on level $l$:

$$\text{btree} \; t \longrightarrow \text{nod} \; l \; t = \text{rank} \; t \; \text{choose} \; l$$

Hint: You might want to prove separately that

$$\sum_{i=0}^{i<r} \binom{i}{n} = \binom{r}{n+1}$$
Exercise 17.2. Sparse binary numbers represent a binary number by a list of the positions of set bits, sorted in ascending order. Thus, the list \([1, 3, 4]\) represents the number 11010. In general, \([p_1, \ldots, p_n]\) represents \(2^{p_1} + \cdots + 2^{p_n}\).

Implement sparse binary numbers in Isabelle, using the type \(\texttt{nat list}\).

1. Define a function \(\texttt{invar} :: \texttt{nat list} \Rightarrow \texttt{bool}\) that checks for strictly ascending bit positions, a function \(\texttt{num_of} :: \texttt{nat list} \Rightarrow \texttt{nat}\) that converts a sparse binary number to a natural number, and a function \(\texttt{add} :: \texttt{nat list} \Rightarrow \texttt{nat list} \Rightarrow \texttt{nat list}\) to add sparse binary numbers.

2. Show that your \(\texttt{add}\) function preserves the invariant and actually performs addition as far as \(\texttt{num_of}\) is concerned.

3. Define a running time function for \(\texttt{add}\) and show that it is linear in the list size.

Hint: The bit positions in sparse binary numbers are analogous to binomial trees of a certain rank in a binomial heap. The \(\texttt{add}\) function is implemented similar to the \(\texttt{merge}\) function, using a \(\texttt{carry}\) function to insert a bit position into a number (similar to \(\texttt{ins_tree}\)). Correctness and running time can be proved similarly.

Bibliographic Remarks

Binomial queues were invented by Vuillemin [75]. Functional implementations were given by King [40] and Okasaki [62]. A functional implementation was verified by Meis \textit{et al.} [49]; a Java implementation, by Müller [51].
Part IV

Advanced Design and Analysis Techniques
Dynamic Programming

You probably have seen this function before:

\[
\begin{align*}
\text{fib} :: \text{nat} \Rightarrow \text{nat} \\
\text{fib} 0 &= 0 \\
\text{fib} 1 &= 1 \\
\text{fib} (n + 2) &= \text{fib} (n + 1) + \text{fib} n
\end{align*}
\]

It computes the well-known Fibonacci numbers. You may also have noticed that calculating \(\text{fib} 50\) already causes quite some stress for your computer and there is no hope for \(\text{fib} 500\) to ever return a result.

This is quite unfortunate considering that there is a very simple imperative program to compute these numbers efficiently:

```c
int fib(n) {
    int a = 0;
    int b = 1;
    for (i in 1..n) {
        int temp = b;
        b = a + b;
        a = temp;
    }
    return a;
}
```

So we seem to be caught in an adverse situation here: either we use a clear and elegant definition of \(\text{fib}\) or we get an efficient but convoluted implementation for \(\text{fib}\). Admittedly, we could just prove that both formulations are the
same function, and use whichever one is more suited for the task at hand. For \( \text{fib} \), of course, it is trivial to define a functional analogue of the imperative program and to prove its correctness. However, doing this for all recursive functions we would like to define is tedious. Instead, this chapter will sketch a recipe that allows to define such recursive functions in the natural way, while still getting an efficient implementation “for free”.

In the following, the Fibonacci function will serve as a simple example on which we can illustrate the idea. Next, we will show how to prove the correctness of the efficient implementation in an efficient way. Subsequently, we will discuss further details of the how approach and how it can be applied beyond \( \text{fib} \). The chapter closes with the study of two famous (and archetypical) dynamic programming algorithms: the Bellman-Ford algorithm for finding shortest paths in weighted graphs and an algorithm due to Knuth for computing optimal binary search trees.

### 18.1 Memoization

Let us consider the tree of recursive calls that are issued when computing \( \text{fib} \) \( 5 \) (Fig. 18.1). We can see that the subtree for \( \text{fib} \) \( 3 \) is computed twice, and that the subtree for \( \text{fib} \) \( 2 \) is even computed three times. How can we avoid these repeated computations? A common solution is memoization: we store previous computation results in some kind of memory and consult it to potentially recall a memoized result before issuing another recursive computation. Here is a simple memoizing version of \( \text{fib} \) that implements the memory as a mapping of type \( \text{nat} \rightarrow \text{nat} \) (the notation \( m(k \mapsto v) \) is a shorthand for \( m(k := \text{Some } v) \), see also Section 12.1):

![Fig. 18.1. Tree of the recursive call structure for \( \text{fib} \) 5](image)
\[ fib_1 :: nat \Rightarrow (nat \rightarrow nat) \Rightarrow nat \times (nat \rightarrow nat) \]
\[ fib_1 \ 0 \ m = (0, m(0 \mapsto 0)) \]
\[ fib_1 \ 1 \ m = (1, m(1 \mapsto 1)) \]
\[ fib_1 \ (n + 2) \ m = (\text{let } (i, m) = \text{case } m n \text{ of } \text{None } \Rightarrow fib_1 \ n \ m \mid \text{Some } i \Rightarrow (i, m); \]
\[ (j, m) = \text{case } m \ (n + 1) \text{ of } \text{None } \Rightarrow fib_1 \ (n + 1) \ m \mid \text{Some } j \Rightarrow (j, m) \]
\[ \text{in } (i + j, m(n + 2 \mapsto i + j)) \]

And indeed, we can ask Isabelle to compute (via the value command) \( fib_1 \ 50 \) or even \( fib_1 \ 500 \) and we get the result within a split second.

However, we are not yet happy with this code. Carrying the memory around means a lot of additional weight for the definition of \( fib_1 \), and proving that this function computes the same value as \( fib \) is not completely trivial (how would you approach this?). Let us streamline the definition first by pulling out the reading and writing of memory into a function \( memo \) (for a type ‘\( k \) of keys and a type ‘\( v \) of values):

\[ memo :: \]
\[ 'k \Rightarrow (('k \rightarrow 'v) \Rightarrow 'v \times ('k \rightarrow 'v)) \Rightarrow ('k \rightarrow 'v) \Rightarrow 'v \times ('k \rightarrow 'v) \]
\[ memo k f m = (\text{case } m k \text{ of } \text{None } \Rightarrow let \ (v, m) = f m \text{ in } (v, m(k \mapsto v)) \mid \text{Some } v \Rightarrow (v, m)) \]

\[ fib_2 :: nat \Rightarrow (nat \rightarrow nat) \Rightarrow nat \times (nat \rightarrow nat) \]
\[ fib_2 \ 0 = memo \ 0 \ (\lambda m. \ (0, m)) \]
\[ fib_2 \ 1 = memo \ 1 \ (\lambda m. \ (1, m)) \]
\[ fib_2 \ (n + 2) = \text{memo } (n + 2) \]
\[ (\lambda m. \ \text{let } (i, m) = fib_2 \ n \ m; \]
\[ (j, m) = fib_2 \ (n + 1) \ m \]
\[ \text{in } (i + j, m)) \]
\[ fib_2 \ (n + 2) = \text{memo } (n + 2) \]
\[ (\lambda m. \ \text{let } (i, m) = fib_2 \ n \ m; (j, y) = fib_2 \ (n + 1) \ m \text{ in } (i + j, y)) \]
This already looks a lot more like the original definition but it still has one problem: we have to thread the memory through the program explicitly. This can be become rather tedious for more complicated programs and diverges from the original shape of the program, complicating the proofs.

18.1.1 Enter the Monad

Let us examine the type of fib₂ more closely. We can read it as the type of a function that, given a natural number, returns a computation. Given an initial memory, it computes a pair of a result and an updated memory. We can capture this notion of “stateful” computations in a data type:

```
datatype (s, a) state = State (s ⇒ a × s)
```

A value of type (s, a) state represents a stateful computation that returns a result of type a and operates on states of type s. The constant run_state forces the evaluation of a computation starting from some initial state:

```
run_state :: (s, a) state ⇒ s ⇒ a × s
run_state (State f) s = f s
```

The advantage of this definition may not seem immediate. Its value only starts to show when we see how it allows us to chain stateful computations. To do so, we only need to define two constants: return to pack up a result in a computation, and bind to chain two computations after each other.

```
return :: a ⇒ (s, a) state
return x = State (λs. (x, s))

bind :: (s, a) state ⇒ (a ⇒ (s, b) state) ⇒ (s, b) state
bind a f = State (λs. let (x, s) = run_state a s in run_state (f x) s)
```

We add a little syntax on top and write {x} for return x, and a >>= f instead of bind a f. The “identity” computation {x} simply leaves the given state unchanged and produces x as a result. The chained computation a >>= f starts with some state s, runs a on it to produce a pair of a result x and a new state s', and then evaluates f x to produce another computation that is run on s'.
We have now seen how to pass state around but we are not yet able to interact with it. For this purpose we define `get` and `set` to retrieve and update the current state, respectively:

\[
\begin{align*}
get &:: (s, s) \text{ state} \\
get &= State (\lambda s. (s, s))
\end{align*}
\]

\[
\begin{align*}
set &:: (s) \Rightarrow (s, \text{unit}) \text{ state} \\
set s' &= State (\lambda s. ((s), s'))
\end{align*}
\]

Let us reformulate `fib` with the help of these concepts:

\[
\begin{align*}
\text{memo} \_1 ::
\quad &\forall \, k \Rightarrow (k \rightarrow v, \text{state}) \Rightarrow (k \rightarrow v, \text{state}) \\
\quad &\text{let memo} \_1 \, k \, a = \text{get} \Rightarrow \left( \lambda m. \text{case } m \, k \, of \right.
\quad &\quad \left. \text{None } \Rightarrow a \Rightarrow \left( \lambda v. \text{set } (m \mapsto v) \Rightarrow \left( \lambda \vdash \langle v \rangle \rangle \right) \\
\quad &\quad \mid \text{Some } x \Rightarrow \langle x \rangle \right)
\end{align*}
\]

\[
\begin{align*}
\text{fib} \_3 :: \text{nat } \Rightarrow (\text{nat } \rightarrow \text{nat}, \text{nat}) \text{ state} \\
\text{fib} \_3 \, 0 &= \langle 0 \rangle \\
\text{fib} \_3 \, 1 &= \langle 1 \rangle \\
\text{fib} \_3 \, (n + 2) &= \text{memo} \_1 \, (n + 2) \, (\text{fib} \_3 \, n \Rightarrow \langle \lambda i. \text{fib} \_3 \, (n + 1) \Rightarrow \langle \lambda j. \langle i + j \rangle \rangle \rangle)
\end{align*}
\]

Can you see how we have managed to hide the whole handling of state behind the scenes? The only explicit interaction with the state is now happening inside of `memo_1`. This is sensible as this is the only place where we really want to recall a memoized result or to write a new value to memory.

While this is great, we still want to polish the definition further: the syntactic structure of the last case of `fib_3` still does not match `fib` exactly. To this end, we lift function application `f x` to the state monad:

\[
\begin{align*}
(\_ \_ \_ :: (s, a \Rightarrow (s, b) \text{ state}) \Rightarrow (s, a) \text{ state } \Rightarrow (s, b) \text{ state} \\
\quad & (f_m \_ a_m = (f_m \Rightarrow (\lambda f. a_m \Rightarrow (\lambda x. f \, x)))
\end{align*}
\]

We can now spell out our final memoizing version of `fib` where `(\_)` replaces ordinary function applications in the original definition:
You may wonder why we added that many additional computations in this last step. On the one hand, we have gained the advantage that we can now closely follow the syntactic structure of \( \text{fib} \) to prove that \( \text{fib}_4 \) is correct (notwithstanding that \( \text{memo}_1 \) will need a special treatment, of course). On the other hand, we can remove most of these additional computations in a final post-processing step.

### 18.1.2 Memoization and Dynamic Programming

Let us recap what we have seen so far in this chapter. We noticed that the naive recursive formulation of the Fibonacci numbers leads to a highly inefficient implementation. We then showed how to work around this problem by using memoization to obtain a structurally similar but efficient implementation. After all this, you may wonder why this chapter is titled *Dynamic Programming* and not *Memoization*.

Dynamic programming relies on two main principles. First, to find an optimal solution for a problem by computing it from optimal solutions for "smaller" instances of the same problem, i.e. **recursion**. Second, to **memoize** these solutions for smaller problems in, e.g. a table. Thus we could be bold and state:

\[
\text{dynamic programming} = \text{recursion} + \text{memoization}
\]

A common objection to this equation would be that memoization should be distinguished from **tabulation**. In this view, the former only computes "necessary" solutions for smaller sub-problems, while the latter just "blindly" builds solutions for sub-problems of increasing size, many of which might be unnecessary. The benefit of tabulation could be increased performance, for instance due to improved caching performance. We believe that this distinction is largely irrelevant to our approach. First, in this book we focus on asymptotically efficient solutions, not constant-factor optimizations. Second, in many dynamic programming algorithms memoization would actually compute solutions for the same set of sub-problems as tabulation does. No matter which of the two approaches is used in the implementation, the hard part is
usually to come up with a recursive solution that can efficiently make use of sub-problems in the first place.

There are problems, however, where clever tabulation instead of naive memoization is necessary to achieve an *asymptotically optimal solution* in terms of *memory consumption*. One instance of this is the Bellman-Ford algorithm presented in Section 18.4. On this example, we will show that our approach is also akin to tabulation. It can easily be introduced as a final "post-processing" step.

Some readers may have noticed that our optimized implementations of \texttt{fib} are not really optimal as they use a map for memoization. Indeed it is possible to swap in other memory implementations as long as they provide a *lookup* and an *update* method. One can even make use of imperative data structures like arrays. As this is not the focus of this book, the interested reader is referred to the literature that is provided at the end of this chapter. Here, we will just assume that the maps used for memoization are implemented as red-black trees (and Isabelle’s code generator can be instructed to do so).

For the remainder of this chapter, we will first outline how to prove that \texttt{fib} is correct. Then, we will sketch how to apply our approach of memoization beyond \texttt{fib}. Afterwards, we will study some prototypical examples of dynamic programming problems and show how to apply the above formula to them.

### 18.2 Correctness of Memoization

We now want to prove that \texttt{fib} is correct. But what is it exactly that we want to prove? We surely want \texttt{fib} to produce the same result as \texttt{fib} when run with an empty memory (where \texttt{empty} \texttt{≡ \lambda x. None}):

\[
\text{\texttt{fst \ (run_state \ (fib \ n) \ empty) = fib \ n}} \quad (18.1)
\]

If we were to make a naive attempt at this prove, we would probably start with an induction on the computation of \texttt{fib} just to realize that the induction hypotheses are not strong enough to prove the recursion case, since they demand an empty memory. We can attempt generalization as a remedy:

\[
\text{\texttt{fst \ (run_state \ (fib \ n) \ m) = fib \ n}}
\]

However, this statement does not hold anymore for every memory \texttt{m}.

What do we need to demand from \texttt{m}? It should only memoize values that are *consistent* with \texttt{fib}:
type_synonym 'a mem = (nat → nat, 'a) state

cmem :: (nat → nat) ⇒ bool
cmem m = (∀ n ∈ dom m. m n = Some (fib n))

dom :: ('k → 'v) ⇒ 'k set
dom m = {a | m a ≠ None}

Note that, from now, we use the type 'a mem to denote memoized values of type 'a that have been "wrapped up" in our memoizing state monad. Using cmem, we can formulate a general notion of equivalence between a value v and its memoized version a, written v ▷ a: starting from a consistent memory m, a should produce another consistent memory m', and the result v.

(▷) :: 'a ⇒ 'a mem ⇒ bool
v ▷ a
= (∀ m. cmem m —→
  (let (v', m') = run_state a m in v = v' ∧ cmem m'))

Thus we want to prove

fib n ▷ fib4 n
(18.2)

via computation induction on n as above. For the base cases we need to prove statements of the form v ▷ (v), a property which follows trivially after unfolding the involved definitions. For the induction case, we can unfold fib4 (n + 2), and get rid of memo1 by applying the following rule (which we instantiate with a = fib4 n):

fib n ▷ a —→ fib n ▷ memo1 n a
(18.3)

For the remainder of the proof, we now want to unfold fib (n + 2) and then follow the syntactic structure of fib4 and fib in lockstep. To do so, we need to find a proof rule for function application. That is, what do we need in order to prove f x ▷ f_m . x_m? For starters, x ▷ x_m seems reasonable to demand. But what about f and f_m? If f has type 'a ⇒ 'b, then f_m is of type ('a ⇒ 'b mem) mem. Intuitively, we want to state something along these lines:

"f_m is a memoized function that, when applied to a value x, yields a memoized value that is equivalent to f x".
This goes beyond what we can currently express with \((\nu)\) as \(\nu \triangleright a\) merely states that “\(a\) is a memoized value equivalent to \(\nu\)”. What we need is more liberty in our choice of equivalence. That is, we want to use statements \(\nu \triangleright_R a\), with the meaning: “\(a\) is a memoized value that is related to \(\nu\) by \(R\)”. The formal definition is analogous to \((\nu)\) (and \((\nu) = (\nu(=))\)):

\[
(\nu.) :: 'a \Rightarrow ('a \Rightarrow 'b \Rightarrow bool) \Rightarrow 'b \ mem \Rightarrow bool
\]

\[
\nu \triangleright_R s
= (\forall m. \ cmem m \rightarrow \ (let (\nu', m') = run_state s m in R \nu \nu' \land cmem m'))
\]

However, we still do not have a means of expressing the second part of our sentence. To this end, we use the function relator \((\Rightarrow)\):

\[
(\Rightarrow) ::
('a \Rightarrow 'c \Rightarrow bool)
\Rightarrow ('b \Rightarrow 'd \Rightarrow bool) \Rightarrow ('a \Rightarrow 'b) \Rightarrow ('c \Rightarrow 'd) \Rightarrow bool
\]

\[
R \Rightarrow S = (\lambda f. \forall x y. R x y \rightarrow S (f x) (g y))
\]

Spelled out, we have \((R \Rightarrow S) f g\) if for any values \(x\) and \(y\) that are related by \(R\), the values \(f x\) and \(g\) \(y\) are related by \(S\).

We can finally state a proof rule for application:

\[
x \triangleright x_m \land f \triangleright (=) \Rightarrow (\nu) f_m \rightarrow f x \triangleright f_m . x_m \quad \text{(18.4)}
\]

In our concrete example, we apply it once to the goal

\[
fib (n + 1) + fib n \triangleright \lambda a. \{\lambda b. \{a + b\}\} . (fib_4 (n + 1)) . (fib_4 n)
\]

solve the first premise with the induction hypotheses, and arrive at

\[
(+) \ (fib (n + 1)) \triangleright (=) \Rightarrow (\nu) \{\lambda a. \{\lambda b. \{a + b\}\}\} . (fib_4 (n + 1))
\]

Our current rule for application (18.4) does not match this goal. Thus we need to generalize it. In addition, we need a new rule for return, and a rule for \((\Rightarrow)\). To summarize, we need the following set of theorems about our consistency relation, applying them wherever they match syntactically to finish the proof of (18.2):

\[
R x y \rightarrow x \triangleright_R y
\]

\[
x \triangleright_R x_m \land f \triangleright_R \Rightarrow \triangleright_S f_m \rightarrow f x \triangleright_S f_m . x_m
\]

\[
(\forall x y. R x y \rightarrow S (f x) (g y)) \rightarrow (R \Rightarrow S) f g
\]

The theorem we aimed for initially
\[ \text{fst (run_state (fib_4 n) empty)} = \text{fib } n \] 

(18.1)

is now a trivial corollary of \( \text{fib } n \mapsto \text{fib}_4 n \). Note that by reading the equation from right to left, we have an easy way to make the memoization transparent to an end-user of \( \text{fib} \).

18.3 Details of Memoization *

In this section, we will look at some further details of the memoization process and sketch how it can be applied beyond \( \text{fib} \). First note that our approach of memoization hinges on two rather independent components: We transform the original program to use the state monad, to thread (an \textit{a priori} arbitrary) state through the program. Only at the call sites of recursion, we then introduce the memoization functionality by issuing lookups and updates to the memory (as implemented by \text{memo}_1). We will name this first process \textit{monadification}. For the second component, many different memory implementations can be used, as long as we can define \text{memo}_1 and prove its characteristic theorem (18.3). For details on this, the reader is referred to the literature. Here, we want to turn our attention towards monadification.

To discuss some of the intricacies of monadification, let us first stick with \( \text{fib} \) for a bit longer and consider the following alternative definition (which is mathematically equivalent but not the same program):

\[
\text{fib } n = (\text{if } n = 0 \text{ then } 0 \text{ else } 1 + \text{sum_list (map fib [0..<n - 1])})
\]

We have not yet seen how to handle two ingredients of this program: constructs like \textit{if}-\textit{then}-\textit{else} or case-combinators; and higher-order functions such as \text{map}.

It is quite clear how \textit{if}-\textit{then}-\textit{else} can be lifted to the state monad:

\[
\text{if } m :: \text{bool mem } \Rightarrow \langle a \text{ mem } \Rightarrow \langle a \text{ mem } \Rightarrow \langle a \text{ mem } \rangle \rangle \equiv (\lambda b. \text{if } b \text{ then } x_m \text{ else } y_m)
\]

By following the structure of the terms, we can also deduce a proof rule for \( \text{tf}_m \):

\[
(b \vdash b_m \land x \vdash_R x_m \land y \vdash_R y_m \quad \rightarrow \\
(\text{if } b \text{ then } x \text{ else } y) \vdash_R \text{tf}_m b_m x_m y_m)
\]

* If you are just interested in the dynamic programming algorithms of the following sections, this section can safely be skipped on first reading.
However, suppose we want to apply this proof rule to our new equation for \( \text{fib} \). We will certainly need the knowledge of whether \( n = 0 \) to make progress in the correctness proof. Thus we make our rule more precise:

\[
\begin{align*}
    & b \vdash b_m \land (b \rightarrow x \mathrel{\triangleright} R \mathrel{\triangleright} x_m) \land (\neg b \rightarrow y \mathrel{\triangleright} R \mathrel{\triangleright} y_m) \rightarrow \\
    & (\text{if } b \text{ then } x \text{ else } y) \mathrel{\triangleright} R \mathrel{\triangleright} \text{id}_m \mathrel{\triangleright} x_m \mathrel{\triangleright} y_m
\end{align*}
\]

How can we lift \( \text{map} \) to the state monad level? Consider its defining equations:

\[
\begin{align*}
    \text{map} \ f \ [] &= [] \\
    \text{map} \ f \ (x \# \ xs) &= f \ x \# \ \text{map} \ f \ xs
\end{align*}
\]

We can follow the pattern we used to monadify \( \text{fib} \) to monadify \( \text{map} \):

\[
\begin{align*}
    \text{map} \!_{m'} \ f \ [] &= \langle[]\rangle \\
    \text{map} \!_{m'} \ f \ (x \# \ xs) &= \langle\lambda a. \langle\lambda b. \langle a \# b\rangle\rangle \rangle \cdot \langle f \rangle \cdot \langle x\rangle \cdot \langle \text{map} \!_{m'} \ f \ xs\rangle
\end{align*}
\]

We have obtained a function \( \text{map} \!_{m'} \) of type \( ('a \Rightarrow 'b \text{ mem}) \Rightarrow ('a \text{ list} \Rightarrow 'b \text{ list mem}) \). This is not yet compatible with our scheme of lifting function application to \( (\cdot) \). We need a function of type \( (('a \Rightarrow 'b \text{ mem}) \Rightarrow ('a \text{ list} \Rightarrow 'b \text{ list mem}) \text{ mem}) \text{ mem} \) because \( \text{map} \) has two arguments and we need one layer of the state monad for each of its arguments. Therefore we simply define

\[
\text{map} \!_m = \langle\lambda f. \langle \text{map} \!_{m'} \ f\rangle\rangle
\]

For inductive proofs about the new definition of \( \text{fib} \), we also need the knowledge that \( \text{fib} \) is recursively applied only to smaller values than \( n \) when computing \( \text{fib} \ n \). That is, we need to know which values \( f \) is applied to in \( \text{map} \ f \ xs \). We can encode this knowledge in a proof rule for \( \text{map} \):

\[
\begin{align*}
    & xs = ys \land (\forall x. \ x \in \text{set} \ ys \rightarrow f \ x \mathrel{\triangleright} R \mathrel{\triangleright} f_m \ x) \rightarrow \\
    & \text{map} \ f \ xs \mathrel{\triangleright} \text{list\_all2} \ R \ \text{map} \!_m \cdot \langle f_m \rangle \cdot \langle ys\rangle
\end{align*}
\]

The relator \( \text{list\_all2} \) lifts \( R \) to a pairwise relation on lists:

\[
\text{list\_all2} \ R \ xs \ ys = (|xs| = |ys| \land (\forall i.|xs|. \ R \ (xs \! i) \ (ys \! i)))
\]
To summarize, here is a fully memoized version of the alternative definition of \( \text{fib} \):

\[
\text{fib}_m :: \mathbb{N} \Rightarrow \mathbb{N} \quad \text{memo}_1 \quad n \\
(\text{if } m n = 0 \text{ then } 0 \text{ else } \langle \lambda a. 1 + a \rangle \cdot (\langle \lambda a. (\mathit{sum\_list} a) \cdot (\mathit{map}_m \cdot (\text{fib}_m') \cdot \langle [0..n] \rangle) \rangle))
\]

The correctness proof for \( \text{fib}_m \) is analogous to the one for \( \text{fib}_4 \), once we have proved the new rules discussed above.

At the end of this section, we note that the techniques that were sketched above also extend to case-combinators and other higher-order functions. Most of the machinery for monadification and the corresponding correctness proofs can be automated in Isabelle [79]. Finally note that none of the techniques we used so far are specific to \( \text{fib} \). The only parts that have to be adopted are the definitions of \( \text{memo}_1 \) and \( \text{cmem} \). In Isabelle, this can be done by simply instantiating a locale.

This concludes the discussion of the fundamentals of our approach towards verified dynamic programming. We now turn to the study of two typical examples of dynamic programming algorithms: the Bellman-Ford algorithm and an algorithm for computing optimal binary search trees.

### 18.4 The Bellman-Ford Algorithm

Calculating shortest paths in weighted graphs is a classic algorithmic task that we all encounter in everyday situations, such as planning the fastest route to drive from \( A \) to \( B \). In this scenario we can view streets as edges in a graph and nodes as street crossings. Every edge is associated with a weight, e.g. the time to traverse a street. We are interested in the path from \( A \) to \( B \) with minimum weight, corresponding to the fastest route in the example. Note that in this example it is safe to assume that all edge weights are non-negative.

Some applications demand negative edge weights as well. Suppose, we transport ourselves a few years into the future, where we have an electric car that can recharge itself via solar cells while driving. If we aim for the most energy-efficient route from \( A \) to \( B \), a very sunny route could then incur a negative edge weight.

The Bellman-Ford algorithm is a classic dynamic programming solution to the single-destination shortest path problem in graphs with negative edge weights. That is, we are given a directed graph with negative edge weights
and some target vertex (known as sink), and we want to calculate the weight of the shortest (i.e. minimum weight) paths from every vertex to the sink. Figure 18.2 shows an example of such a graph.

![Directed Graph Example](image)

Fig. 18.2. Example of a weighted directed graph

Formally, we will take a simple view of graphs. We assume that we are given a number of nodes numbered 0, ..., n, and some sink \( t \in \{0..n\} \) (thus \( n = t = 4 \) in the example). Edge weights are given by a function \( W : \text{int} \Rightarrow \text{int extended} \). The type \( \text{int extended} \) extends the natural numbers with positive and negative infinity:

\[
\text{datatype} \ a \ \text{extended} = \text{Fin} \ a \ | \ \infty \ | \ -\infty
\]

We refrain from giving the explicit definition of addition and comparison on this domain, and rely on your intuition instead. A weight assignment \( W_{ij} = 1 \) means that there is no edge from \( i \) to \( j \). The purpose of \(-\infty\) will become clear later.

### 18.4.1 Deriving a Recursive Solution

The main idea of the algorithm is to consider paths in order of increasing length in the number of edges. In the example, we can immediately read off the weights of the shortest paths to the sink that use only one edge: only nodes 2 and 3 are directly connected to the sink, with edge weights 3 and 2, respectively; for all others the weight is infinite. How can we now calculate the minimum weight paths (to the sink) with at most two edges? For node 3, the weight of the shortest path with at most two edges is: either the weight of the path with one edge; or the weight of the edge from node 3 to node 2 plus the weight of the path with one edge from node 2 to the sink. Because \(-2 +\)
3 = 1 \leq 2$, we get a new minimum weight of 1 for node 3. Following the same scheme, we can iteratively calculate the minimum path weights given in table 18.1.

<table>
<thead>
<tr>
<th>i</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>$\infty$</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 18.1. The minimum weights of paths from vertices $v = 0 \ldots 4$ to $t$ that use at most $i = 0 \ldots 4$ edges.

The analysis we just ran on the example already gives us a clear intuition on all we need to deduce a dynamic program: a recursion on sub-problems, in this case to compute the weight of shortest paths with at most $i + 1$ edges from the weights of shortest paths with at most $i$ edges. To formalize this recursion, we first define the notion of a minimum weight path from some node $v$ to $t$ with at most $i$ edges, denoted as $OPT \ i \ v$:

\[
OPT \::\ nat \Rightarrow nat \Rightarrow int\ extended
\]
\[
OPT \ i \ v = Min (\{weight (v \# xs @ [t]) \mid \mid xs\| + 1 \leq i \land set xs \subseteq \{0..n\}\} \cup \\
\{if \ t = v then 0 else \infty\})
\]

\[
weight :: (nat \Rightarrow nat \Rightarrow int\ extended) \Rightarrow nat\ list \Rightarrow int\ extended
\]
\[
weight _{[]} = 0
\]
\[
weight \ W \ (v \# w \# xs) = W \ v \ w + weight \ W \ (w \# xs)
\]

If $i = 0$, things are simple:

\[
OPT \ 0 \ v = (if \ t = v then 0 else \infty)
\]

A shortest path that constitutes $OPT \ (i + 1) \ v$ uses either at most $i$ or exactly $i + 1$ edges. That is, $OPT \ (i + 1) \ v$ is either $OPT \ i \ v$, or the weight of the edge from $v$ to any of its neighbours $w$ plus $OPT \ i \ w$:

\[
OPT \ (i + 1) \ v = min (OPT \ i \ v) (Min \ \{W \ v \ w + OPT \ i \ w \mid w \leq n\})
\]

\[Proof.\] We prove this equality by proving two inequalities:
\[ (\text{lhs} \leq \text{rhs}) \] For this direction, we essentially need to show that every path on the rhs is covered by the lhs, which is trivial.

\[ (\text{lhs} \geq \text{rhs}) \] We skip the cases where \( OPT \ (i + 1) \ v \) is trivially 0 or \( \infty \) (i.e. where it is given by the singleton set in the definition of \( OPT \)). Thus consider some \( xs \) such that \( OPT \ (i + 1) \ v = weight \ (v \# \ xs @ [t]) \), \( |xs| \leq i \), and set \( xs \subseteq \{0..n\} \). The cases where \(|xs| < i \) or \( i = 0 \) are trivial. Otherwise, we have \( OPT \ (i + 1) \ v = W \ v \ (hd \ xs) + weight \ (xs @ [t]) \) by definition of \( weight \), and \( OPT \ i \ (hd \ xs) \leq weight \ (xs @ [t]) \) by definition of \( OPT \). Therefore, we can show:

\[ OPT \ (i + 1) \ v \geq W \ v \ (hd \ xs) + OPT \ i \ (hd \ xs) \geq \text{rhs} \]

\( \Box \)

We can turn these equations into a recursive program:

```latex
bf :: nat ⇒ nat ⇒ int extended
bf 0 \ v = (if \ t = \ v \ then \ 0 \ else \ \infty)
bf \ (i + 1) \ v
= \text{min_list} \ (bf \ i \ v \ # \ \text{map} \ (\lambda w. \ W \ v \ w + bf \ i \ w) \ [0..<n + 1])
```

It is obvious that we can prove correctness of \( bf \) by induction:

\[ bf \ i \ v = OPT \ i \ v \]

18.4.2 Negative Cycles

Have we solved the initial problem now? The answer is "not quite" because we have ignored one additional complication. Consider our example table 18.1 again. The table stops at path length five because no shorter paths with more edges exist. For this example, five corresponds to the number of nodes, which bounds the length of the longest simple path. However, is it the case that we will never find shorter non-simple paths in other graphs? The answer is "no".

If a graph contains a negative reaching cycle, i.e. a cycle with a negative sum of edge weights from which the sink is reachable, then we can use it arbitrarily often to find shorter and shorter paths.

Luckily, we can use the Bellman-Ford algorithm to detect this situation by examining the relationship of \( OPT \ n \) and \( OPT \ (n + 1) \). The following proposition summarizes the key insight:

The graph contains a negative reaching cycle if and only if there exists a \( v \leq n \) such that \( OPT \ (n + 1) \ v < OPT \ n \ v \)
Proof. If there is no negative reaching cycle, then all shortest paths are either simple or contain superfluous cycles of weight 0. Thus, we have $OPT(n + 1) v = OPT n v$ for all $v \leq n$.

Otherwise, there is a negative reaching cycle $ys = a \# xs \otimes [a]$ with weight $ys < 0$. Working towards a contradiction, assume that $OPT n v \leq OPT(n + 1) v$ for all $v \leq n$. Using the recursion we proved above, this implies $OPT n v \leq W v u + OPT n u$ for all $u, v \leq n$. By applying this inequality to the nodes in $a \# xs$, we can prove the inequality

$$\sum_{\text{list}} (\map{OPT n}) ys \leq \sum_{\text{list}} (\map{OPT n}) ys + weight ys$$

This implies $0 \leq weight ys$, which yields the contradiction. \qed

This means we can use $bf$ to detect the existence of negative reaching cycles by computing one more round, i.e. $bf(n + 1) v$ for all $v$. If nothing changes in this step, we know that there are no negative reaching cycles and that $bf n$ correctly represents the shortest path weights. Otherwise, there has to be a negative reaching cycle.

Finally, we can use memoization to obtain an efficient implementation that solves the single-destination shortest path problem. Applying our memoization technique from above, we first obtain a memoizing version $bf_m$ of $bf$. We then define the following program:

```haskell
bellman_ford ::
((nat \times\ times nat, int extended) mapping, int extended list option) state
bellman_ford = iter_bf (n, n) >=
(\_ \_ \map m'(bf_m n) [0..<n + 1] >=
 (\_xs. map m'(bf_m (n + 1)) [0..<n + 1] >=
 (\ys. \_ if xs = ys then Some xs else None)))
```

Here, $iter_bf (n, n)$ just computes the values from $bf_m 0 0$ to $bf_m n n$ in a row-by-row manner. Using the reasoning principles that were described above (for $fib$), we can then prove that $bellman_ford$ indeed solves its intended task correctly ($shortest v$ is the length of the shortest path from $v$ to $t$):

$$(\forall i \leq n. \forall j \leq n. \neg \infty < W i j) \rightarrow$

$fst (run_state bellman_ford empty) = (if contains_negative_reaching_cycle then None$

$else Some (map shortest [0..<n + 1]))$

Here, $shortest$ is defined analogously to $OPT$ but for paths of unbounded length.
18.5 Optimal Binary Search Trees

In this book, we have studied various tree data structures that guarantee logarithmic running time bounds for operations such as lookups and updates into the tree. These bounds were usually worst-case and did not take into account any information about the actual series of queries that are to be issued to the data structure. In this section, instead, we want to focus on binary search trees that minimize the amount of work that needs to be done when the distribution of keys in a sequence of lookup operations is known in advance.

More formally, we want to study the following problem. We are given a list \([i..j]\) of integers ranging from \(i\) to \(j\) and a function \(p : int \Rightarrow nat\) that maps each key in the range to a frequency with which this key is searched for. Our goal is to find a binary search tree that minimizes the expected number of comparisons when presented with a sequence of lookup operations for keys in the range \([i..j]\) that adhere to the distribution given by \(p\).

As an example, consider the range \([1..5]\) with probabilities \([10, 30, 15, 25, 20]\). This tree:

![Tree Diagram](image)

incurs an expected value of 2.15 comparison operations. However, the minimal expected value is 2 and is achieved by this tree:

![Tree Diagram](image)

Our task is equivalent to minimizing the weighted path length (or cost) as we did for Huffman encodings (Chapter 13). Recall that the weighted path length is the sum of the frequencies of every node in the tree multiplied by its depth in the tree. It fulfills the following (recursive) equations:

\[
\begin{align*}
cost(\langle \rangle) &= 0 \\
cost(\langle l, k, r \rangle) &= p(l) \times \text{depth}(l) + cost(\langle l \rangle) + cost(\langle r \rangle)
\end{align*}
\]
The difference of our task compared to finding an optimal Huffman encoding is the constraint that the resulting tree needs to be sorted, making it hard to deploy a similar greedy solution. Instead, we want to come up with a dynamic programming solution and thus need to find a way to subdivide the problem.

### 18.5.1 Deriving a Recursive Solution

The key insight into the problem is that subtrees of optimal binary search trees are also optimal. The left and right subtrees of the root must be optimal, since if we could improve either one, we would also get a better tree for the complete range of keys. This motivates the following definition:

\[
\text{wpl } W_{i,j} = 0 \\
\text{wpl } W_{i,j} (l, k, r) = \text{wpl } W_{i, (k-1)} l + \text{wpl } W_{(k+1), j} r + W_{i,j} \\
W_{i,j} = (\sum_{k = i}^{j} p_k)
\]

It is easy to see that \(\text{wpl } W_{i,j}\) is just a reformulation of \(\text{cost } t\):

\[
\text{inorder } t = [i..j] \rightarrow \text{wpl } W_{i,j} t = \text{cost } t
\]

We can actually forget about the original frequencies \(p\) and just optimize \(\text{wpl } W_{i,j}\) for some fixed weight function \(W : \text{int} \Rightarrow \text{int} \Rightarrow \text{nat}\).

The binary search tree \(t\) that contains the keys \([i..j]\) and minimizes \(\text{wpl } W_{i,j} t\) has some root \(k\) with \([i..j] = [i..k-1] @ k # [j+1..k]\). Its left and right subtrees need to be minimal again, i.e. minimize \(\text{wpl } W_{i, (k-1)}\) and \(\text{wpl } W_{(k+1), j}\). This yields the following recursive functions for computing the minimal weighted path length \((\text{min}_\text{wpl})\) and a corresponding binary search tree \((\text{opt}_\text{bst})\):

\[
\text{min}_\text{wpl} :: \text{int} \Rightarrow \text{int} \Rightarrow \text{nat} \\
\text{min}_\text{wpl} i j = (\text{if } j < i \text{ then } 0 \text{ else } \text{min}_\text{list}) \\
\quad (\text{map } (\lambda k. \text{min}_\text{wpl} i (k-1) + \text{min}_\text{wpl} (k+1), j + W_{i,j})
\]
opt_bst :: int ⇒ int ⇒ int tree
opt_bst i j
= (if j < i then ()
    else argmin (wpl W i j)
        (map (λk. (opt_bst i (k - 1), k, opt_bst (k + 1) j)) [i..j]))

Here argmin \( f \) \( xs \) returns the rightmost \( x \in \text{set } xs \) such that \( f x \) is minimal among \( xs \) (i.e. \( f x \leq f y \) for all \( y \in \text{set } xs \)).

To prove that \( \min_{wpl} \) and \( \text{opt_bst} \) are correct, we want to show two properties: \( \min_{wpl} i j \) should be a lower bound of \( wpl W i j t \) for any search tree \( t \) for \([i..j]\), and \( \min_{wpl} i j \) should correspond to the weight of an actual search tree, namely \( \text{opt_bst} i j \). Formally, we prove the following properties:

\[
\text{inorder } t = [i..j] \rightarrow \min_{wpl} i j \leq wpl W i j t
\]
\[
\text{inorder } (\text{opt_bst } i j) = [i..j]
\]
\[
wpl W i j (\text{opt_bst } i j) = \min_{wpl} i j
\]

The first property is easily proved by computation induction on \( wpl W \), and the second and third property are equally easily proved using computation induction on \( \min_{wpl} \).

If \( W \) is constructed from \( p \) as above, we can derive the following correctness theorems referring to the original problem:

\[
\text{inorder } t = [i..j] \rightarrow \min_{wpl} W i j \leq \text{cost } t
\]
\[
\text{cost } (\text{opt_bst } W i j) = \min_{wpl} W i j
\]

### 18.5.2 Memoization

We can apply the memoization techniques that were discussed above to efficiently compute \( \min_{wpl} \) and \( \text{opt_bst} \). The only remaining caveat is that \( W \) also needs to be computed efficiently from the distribution \( p \). If we just use the defining equality \( W i j = (\sum_{k=i}^{j} p k) \), the computation of \( W \) is unnecessarily costly. Another way is to memoize \( W \) itself, using the following recursion:

\[
W p i j = (\text{if } i \leq j \text{ then } W p i (j - 1) + p j \text{ else } 0)
\]

This yields a memoizing version \( W_{m}' \) and a theorem that connects it to \( W \):
Dynamic Programming

We can now iterate $W'_{m'} p i j$ for $i = 0 \ldots n$ to pre-compute all relevant values of $W p i j$:

$$W_c p n = \text{snd} (\text{run_state} (\text{map}_{m'} (\lambda i. \ W'_{m'} p i n) [0..n]) \text{empty})$$

Using the correctness theorem for map$_{m'}$ from above, it can easily be shown that this yields a consistent memory:

$$\text{cmem} (W_c p n)$$

We can show the following equation for computing $W$:

$$W p i j = \begin{cases} W_c p n & \text{(case} (W_c p n) (i, j) \text{of None} \Rightarrow \ W p i j | \text{Some } x \Rightarrow x) \end{cases}$$

Note that the None branch will only be triggered when indices outside of $0 \ldots n$ are accessed. Finally, we can use $W_c$ to pass the pre-computed values of $W$ to opt$_{bst}$:

$$\text{opt}_{bst} p i j = \begin{cases} \text{let } M = W_c p j; \ W = \lambda i j. \text{case } M (i, j) \text{of None} \Rightarrow \ W p i j | \text{Some } x \Rightarrow x \in \text{opt}_{bst} W i j \end{cases}$$

18.5.3 Optimizing the Recursion

While we have applied some trickery to obtain an efficient implementation of the simple dynamic programming algorithm expressed by opt$_{bst}$, we still have not arrived at the solution that is currently known to be most efficient. The most efficient known algorithm to compute optimal binary search trees due to Knuth [41] is a slight variation of opt$_{bst}$ and relies on the following observation.

Let $R i j$ denote the maximal root of any optimal binary search for $[i..j]$:

$$R i j = \text{argmin} (\lambda k. w i j + \text{min}_w p l i (k - 1) + \text{min}_w p l (k + 1) j) [i..j]$$

It can be shown that $R i j$ is bounded by $R i (j - 1)$ and $R (i + 1) j$:

$$i < j \rightarrow R i (j - 1) \leq R i j \land R i j \leq R (i + 1) j$$
The proof of this fact is rather involved and the details can be found in the references provided at the end of this section.

With this knowledge, we can make the following optimization to \( opt\_bst \):

\[
\begin{align*}
\text{opt\_bst} \::& \:: \text{int} \rightarrow \text{int} \Rightarrow \text{int tree} \\
\text{opt\_bst} \::& \:: i \ j \\
= \begin{cases} \\
(\text{if} \ j < \ i \ \text{then} \langle \rangle) \\
(\text{else if} \ i = \ j \ \text{then} \langle \langle \rangle, \ i, \ \langle \rangle \rangle) \\
(\text{else let} \ \text{left} = \ \text{root} (\text{opt\_bst} \ i \ (j - 1)); \\
\quad \text{right} = \ \text{root} (\text{opt\_bst} \ (i + 1) \ j) \\
\quad \text{in} \ \text{argmin} (\text{wpl} \ i \ j) \\
\quad (\text{map} (\lambda k. (\text{opt\_bst} \ i \ (k - 1), \ k, \\
\quad \text{opt\_bst} \ (k + 1) \ j)) \\
\quad [\text{left..right}]))
\end{cases}
\end{align*}
\]

You may wonder whether this change really incurs an asymptotic runtime improvement. Indeed, it can be shown that it improves the algorithm’s runtime by a factor of \( O(n) \). For a fixed search tree size \( d = i - j \), the total number of recursive computations is given by the following telescoping series:

\[
\begin{align*}
&d \leq n \quad \rightarrow \\
&\sum_j \cdot j = d \cdot n. \ \text{let} \ i = j - d \ \text{in} \ R \ i \ j - R \ i \ (j - 1) + 1 \\
&= R \ (n - d + 1) \ n - R \ 0 \ (d - 1) + n - d + 1
\end{align*}
\]

This quantity is bounded by \( 2 \cdot n \), which implies that the overall number of recursive calls is bounded by \( O(n^2) \).

**Bibliographic Remarks**

The original \( O(n^2) \) algorithm for Binary Search Trees is due to Knuth [41]. Yao later explained this optimization more elegantly in his framework of "quadrilateral inequalities" [80]. Nipkow and Somogyi follow Yao’s approach in their Isabelle formalization [59], on which the last subsection of this chapter is based. The other parts of this chapter are based on a paper by Wimmer et al. [79] and its accompanying Isabelle formalization [78]. The formalization also contains further examples of dynamic programming algorithms, including solutions for the Knapsack and the minimum edit distance problems, and the CYK algorithm.
Amortized Analysis

Consider a $k$-bit binary counter and a sequence of increment (by 1) operations on it where each one starts from the least significant bit and keeps flipping the 1s until a 0 is encountered (and flipped). Thus the worst-case running time of an increment is $O(k)$ and a sequence of $n$ increments takes time $O(nk)$. However, this analysis is very coarse: in a sequence of increments there are many much faster ones (for half of them the least significant bit is 0!). It turns out that a sequence of $n$ increments takes time $O(n)$. Thus the average running time of each increment is $O(1)$. Amortized analysis is the analysis of the running time of a sequence of operations on some data structure by upper-bounding the average running time of each operation.

As the example of the binary counter shows, the amortized running time for a single call of an operation can be much better than the worst-case time. Thus amortized analysis is unsuitable in a real-time context where worst-case bounds on every call of an operation are required.

Amortized analysis of some data structure is valid if the user of that data structure never accesses old versions of the data structure (although in a functional language on could). The binary counter shows why that invalidates amortized analysis: start from 0, increment the counter until all bits are 1, then increment that counter value again and again, without destroying it. Each of those increments takes time $O(k)$ and you can do that as often as you like, thus subverting the analysis. In an imperative language you can easily avoid this “abuse” by making the data structure stateful: every operation modifies the state of the data structure. This shows that amortized analysis has an imperative flavour. In a purely functional language, monads can be used to restrict access to the latest version of a data structure.
19.1 The Potential Method

The potential method is a particular technique for amortized analysis. The key idea is to define a potential function $\Phi$ from the data structure to non-negative numbers. The potential of the data structure is like a savings account that cheap calls can pay into (by increasing the potential) to compensate for later expensive calls (which decrease the potential). In a nutshell: the less “balanced” a data structure is, the higher its potential should be because it will be needed to pay for the impending restructuring.

The amortized running time (or complexity) is defined as the actual running time plus the difference in potential, i.e. the potential after the call minus the potential before it. If the potential increases, the amortized running time is higher than the actual running time and we pay the difference into our savings account. If the potential decreases, the amortized running time is lower than the actual running time and we take something out of our savings account to pay for the difference.

More formally, we are given some data structure with operations $f, g, \text{etc}$ with corresponding time functions $T_f, T_g \text{etc}$. We are also given a potential function $\Phi$. The amortized running time function $A_f$ for $f$ is defined as follows:

$$ A_f \ s = T_f \ s + \Phi \ (f \ s) - \Phi \ s $$

(19.1)

where $s$ is the data structure under consideration; $f$ may also have additional parameters. Given a sequence of data structure states $s_0, \ldots, s_n$ where $s_{i+1} = f, s_i$, it is not hard to see that

$$ \sum_{i=0}^{n} A_f \ s_i = \sum_{i=0}^{n} T_f \ s_i + \Phi \ s_n - \Phi \ s_0 $$

If we assume (for simplicity) that $\Phi \ s_0 = 0$, then it follows immediately that the amortized running time of the whole sequence is an upper bound of the actual running time (because $\Phi$ is non-negative). This observation becomes useful if we can bound $A_f \ s$ by some closed term $u_f \ s$. Typical examples for $u_f \ s$ are constants, logarithms etc. Then we can conclude that $f$ has constant, logarithmic etc amortized complexity. Thus the only proof obligation is

$$ A_f \ s \leq u_f \ s $$

possibly under the additional assumption $\text{invar} \ s$ if the data structure comes with an invariant $\text{invar}$.

In the sequel we assume that $s_0$ is some fixed value, typically “empty”, and that its potential is 0.

How do we analyze operations that combine two data structures, e.g. the union of two sets? Their amortized complexity can be defined in analogy to (19.1):

$$ A_f \ s_1 \ s_2 = T_f \ s_1 \ s_2 + \Phi \ (f \ s_1 \ s_2) - (\Phi \ s_1 + \Phi \ s_2) $$
So far we implicitly assumed that all operations return the data structure as a result, otherwise \( \Phi (f \ s) \) does not make sense. How should we analyze so-called observer functions that do not modify the data structure but return a value of some other type? Amortized analysis does not make sense here because the same observer can be applied multiple times to the same data structure value without modifying it. Classical worst-case complexity is needed, unless the observer does modify the data structure as a side effect or by returning a new value. Then one can perform an amortized analysis that ignores the returned observer value (but not the time it takes to compute it).

Now we study two important examples of amortize analyses. More complex applications are found in later chapters.

### 19.2 Binary Counter

The binary counter is represented by a list of booleans where the head of the list is the least significant bit. The increment operation and its running time are easily defined:

\[
\begin{align*}
\text{incr} &:: \text{bool list} \Rightarrow \text{bool list} \\
\text{incr} [] & = [\text{True}] \\
\text{incr} (\text{False} \# bs) & = \text{True} \# bs \\
\text{incr} (\text{True} \# bs) & = \text{False} \# \text{incr} bs
\end{align*}
\]

\[
\begin{align*}
T_{\text{incr}} &:: \text{bool list} \Rightarrow \text{real} \\
T_{\text{incr}} [] & = 1 \\
T_{\text{incr}} (\text{False} \# _) & = 1 \\
T_{\text{incr}} (\text{True} \# bs) & = T_{\text{incr}} bs + 1
\end{align*}
\]

The potential of a counter is the number of True’s because they increase \( T_{\text{incr}} \):

\[
\begin{align*}
\Phi &:: \text{bool list} \Rightarrow \text{real} \\
\Phi \ bs & = |\text{filter} (\lambda x. \ x) \ bs|
\end{align*}
\]

Clearly the potential is never negative.

The amortized complexity of \( \text{incr} \) is 2:

\[
T_{\text{incr}} bs + \Phi (\text{incr} bs) - \Phi bs = 2
\]

This can be proved automatically by induction on \( bs \).
Exercise 19.1. Generalise the binary counter to a base $b$ counter, $b \geq 2$. Prove that there is a constant $c$ such that the amortized complexity of incrementation is at most $c$ for every $b \geq 2$.

19.3 Dynamic Tables

A dynamic table is an abstraction of a dynamic array that can grow and shrink subject to a specific memory management. At any point the table has a certain size (= number of cells) but some cells may be unoccupied or free. As long as there are free cells, inserting a new element into the table takes constant time. When the table overflows, the whole table has to be copied into a larger table, which takes linear time. Similarly, elements can be deleted from the table in constant time, but when too many elements have been deleted, the table is contracted to save space. Contraction involves copying into a smaller table. This is an abstraction of a dynamic array, where the index bounds can grow and shrink. It is an abstraction because we ignore the actual contents of the table and abstract the table to a pair $(n, l)$ where $l$ is its size and $n < l$ the number of occupied cells. The empty table is represented by $(0, 0)$.

Below we do not comment on the formal proofs because they are essentially just case analyses (as dictated by the definitions) plus linear arithmetic.

19.3.1 Insertion

The key observation is that doubling the size of the table upon overflow leads to an amortized cost of 3 per insertion: 1 for inserting the element, plus 2 towards the later cost of copying a table of size $l$ upon overflow (because only the $l/2$ elements that lead to the overflow pay for it).

Insertion always increments $n$ by 1. The size increases from 0 to 1 with the first insertion and doubles with every further overflow:

\[
\text{ins} :: \text{nat} \times \text{nat} \Rightarrow \text{nat} \times \text{nat} \\
\text{ins} (n, l) = (n + 1, \text{if } n < l \text{ then } l \text{ else if } l = 0 \text{ then } 1 \text{ else } 2 \cdot l)
\]

\[
T_{\text{ins}} :: \text{nat} \times \text{nat} \Rightarrow \text{real} \\
T_{\text{ins}} (n, l) = (\text{if } n < l \text{ then } 1 \text{ else } n + 1)
\]

This guarantees the load factor $n/l$ is always between $1/2$ and 1:
19.3 Dynamic Tables

\[
\text{\texttt{invar \ : \ \texttt{nat \times nat} \Rightarrow \texttt{bool}}}
\]
\[
\text{\texttt{invar (n, l) = (l/2 \leq n \land n \leq l)}}
\]

The potential of a table \((n, l)\) is \(2 \cdot (n - l/2) = 2 \cdot n - l\) following the intuitive argument at the beginning of the Insertion section.

\[
\Phi \ : \ \texttt{nat \times nat} \Rightarrow \texttt{real}
\]
\[
\Phi (n, l) = 2 \cdot n - l
\]

The potential is always non-negative because of the invariant.

Note that in our informal explanatory text we use / freely and assume we are working with real numbers. In the formalization we often prefer multiplication over division because the former is easier to reason about.

19.3.2 Insertion and Deletion

A naive implementation of deletion simply removes the element but never contracts the table. This works (Exercise 19.2) but is a waste of space. It is tempting to think we should contract once the load factor drops below 1/2. However, this can lead to the following fluttering. Starting with a full table (of size \(l = 2^k\) for an arbitrary \(k\)) one insertion causes an overflow, two deletions cause a contraction, another insertion causes an overflow, and so on. The cost of each overflow and contraction is \(l\) but there are at most two operations to pay for it. Thus the amortized cost of both insertion and deletion cannot be constant. It turns out that it works if we allow the load factor to drop to 1/4 before we contract the table to half its size:

\[
\text{\texttt{del \ : \ \texttt{nat \times nat} \Rightarrow \texttt{nat \times nat}}}
\]
\[
\text{\texttt{del (n, l) = (n - 1, if n = 1 then 0 else if 4 \cdot (n - 1) < l then l div 2 else l)}}
\]
\[
\text{\texttt{T_{del} \ : \ \texttt{nat \times nat} \Rightarrow \texttt{real}}}
\]
\[
\text{\texttt{T_{del} (n, l) = (if n = 1 then 1 else if 4 \cdot (n - 1) < l then n else 1)}}
\]

Now the load factor is always between 1/4 and 1. It turns out that the lower bound is not needed in the proofs and we settle for a simpler invariant:
Amortized Analysis

invar :: nat × nat ⇒ bool
invar (n, l) = (n ≤ l)

The potential distinguishes two cases:

Φ :: nat × nat ⇒ real
Φ (n, l) = (if n < l/2 then l/2 − n else 2 · n − l)

The condition 2 · n ≥ l concerns the case when we are heading up for an overflow and has been dealt with above. Conversely, 2 · n < l concerns the case where we are heading down for a contraction. That is, we start at (l, 2 · l) (where the potential is 0) and l/2 deletions lead to (l/2, 2 · l) where a contraction requires l/2 credits, and indeed Φ (l/2, 2 · l) = l/2. Since l/2 is spread over l/2 deletions, the amortized cost of a single deletion is 2, 1 for the real cost and 1 for the savings account.

Note that the case distinction in the definition of Φ ensures that the potential is always ≥ 0 — the invariant is not even needed.

Exercise 19.2. Prove that in the dynamic table with naive deletion (where deletion decrements n but leaves l unchanged), insertion has an amortized cost of at most 3 and deletion of at most 1.

Exercise 19.3. Modify deletion as follows: contraction happens when the load factor would drop below 1/3, i.e. when 3 · (n − 1) < l. Then the size of the table is multiplied by 2/3, i.e. reduced to (2 · l) div 3. Prove that insertion and deletion have constant amortized complexity using the potential Φ (n, l) = |2 · n − l|.

Bibliographic Remarks

Amortized analysis is due to Tarjan [72]. Introductions to it can be found in most algorithm textbooks. This chapter is based on earlier work by Nipkow [53, 55] which also formalizes the meta-theory of amortized analysis.
20.1 Queue Specification

A queue can be viewed as a glorified list with function \( \text{enq} \) for adding an element to the end of the list and function \( \text{first} \) for accessing and \( \text{deq} \) for removing the first element. This is the full ADT:

\[
\text{ADT } \text{Queue} = \\
\text{interface } \text{empty} :: 'q \\
\text{enq} :: 'a \Rightarrow 'q \Rightarrow 'q \\
\text{deq} :: 'q \Rightarrow 'q \\
\text{first} :: 'q \Rightarrow 'a \\
\text{is_empty} :: 'q \Rightarrow \text{bool} \\
\text{abstraction } \text{list} :: 'q \Rightarrow 'a \text{ list} \\
\text{invariant } \text{invar} :: 'q \Rightarrow \text{bool} \\
\text{specification } \text{list empty} = [] \\
\text{invar } q \rightarrow \text{list (enq } x \ q) = \text{list } q \odot [x] \\
\text{invar } q \rightarrow \text{list (deq } q) = \text{tl (list } q) \\
\text{invar } q \land \text{list } q \neq [] \rightarrow \text{first } q = \text{hd (list } q) \\
\text{invar } q \rightarrow \text{is_empty } q = (\text{list } q = []) \\
\text{invar empty} \\
\text{invar } q \rightarrow \text{invar (enq } x \ q) \\
\text{invar } q \rightarrow \text{invar (deq } q) \\
\]

A trivial implementation is as a list, but then \( \text{enq} \) is linear in the length of the queue. To improve this we consider two more sophisticated implementations. First, a simple implementation where every operation has amortized constant complexity. Second, a tricky “real time” implementation where every operation has worst-case constant complexity.
20.2 Queues as Pairs of Lists

The queue is implemented as a pair of lists \((fs, rs)\), the front and rear lists. Function \(enq\) adds elements to the head of the rear \(rs\) and \(deq\) removes elements from the head of the front \(fs\). When \(fs\) becomes empty, it is replaced by \(rev\ \(rs\)\) (and \(rs\) is emptied) — the reversal ensures that now the oldest element is at the head. Hence \(rs\) is really the reversal of the rear of the queue but we just call it the rear. The abstraction function is obvious:

\[
\text{list} :: \text{'a list} \times \text{'a list} \Rightarrow \text{'a list} \\
\text{list} (fs, rs) = fs @ rev\ rs
\]

Clearly \(enq\) and \(deq\) are constant-time until the front becomes empty. Then we need to reverse the rear which takes linear time (if it is implemented by \(itrev\), see Section 1.5.1). But we can pay for this linear cost up front by paying a constant amount for each call of \(enq\). Thus we arrive at amortized constant time. See below for the formal treatment.

The implementation is shown in Figure 20.1. Of course \(empty = ([], [])\). Function \(norm\) performs the reversal of the rear once the front becomes empty. Why does not only \(deq\) but also \(enq\) call \(norm\)? Because otherwise \(enq \ x_1 \ (\ldots (enq \ x_1 \ empty) \ldots)\) would result in \(([], [x_n, \ldots, x_1])\) and \(\text{first}\) would

\[
\text{norm} :: \text{'a list} \times \text{'a list} \Rightarrow \text{'a list} \\
\text{norm} (fs, rs) = (\text{if } fs = [] \text{ then } (itrev\ rs\ [], [])) \text{ else } (fs, rs))
\]

\[
\text{enq} :: \text{'a} \Rightarrow \text{'a list} \times \text{'a list} \Rightarrow \text{'a list} \times \text{'a list} \\
enq\ a (fs, rs) = norm\ (fs, a \#\ rs)
\]

\[
\text{deq} :: \text{'a list} \times \text{'a list} \Rightarrow \text{'a list} \times \text{'a list} \\
deq (fs, rs) = (\text{if } fs = [] \text{ then } (fs, rs) \text{ else } norm\ (\text{tl}\ fs, rs))
\]

\[
\text{first} :: \text{'a list} \times \text{'a list} \Rightarrow \text{'a} \\
\text{first}\ (a \#\ _, _) = a
\]

\[
\text{is}\_\text{empty} :: \text{'a list} \times \text{'a list} \Rightarrow \text{bool} \\
\text{is}\_\text{empty}\ (fs, _) = (fs = [])
\]

Fig. 20.1. Queue as a pair of lists
become an expensive operation because it would requires the reversal of the rear. Thus we need to avoid queues (\([\], rs)\) where \(rs \neq [\])\). Thus \(\text{norm}\) guarantees the following invariant:

\[
\text{invar} :: 'a\ list \times 'a\ list \Rightarrow \text{bool}
\text{invar} (fs, rs) = (fs = [\] 
\rightarrow rs = [\])
\]

Functional correctness, i.e. proofs of the properties in the ADT \(\text{Queue}\), are straightforward. Let us now turn to the amortized running time analysis. The time functions are shown in Appendix B.6.

For the amortized analysis we define the potential function

\[
\Phi :: 'a\ list \times 'a\ list \Rightarrow \text{nat}
\Phi (_ , rs) = |rs|
\]

because \(|rs|\) is the amount we have accumulated by charging \(1\) for each \(\text{enq}\). This is enough to pay for the eventual reversal. Now it is easy to prove that both \(\text{enq}\) and \(\text{deq}\) have amortized constant running time:

\[
T_{\text{enq}} a (fs, rs) + \Phi (\text{enq} a (fs, rs)) - \Phi (fs, rs) \leq 4
\]
\[
T_{\text{deq}} (fs, rs) + \Phi (\text{deq} (fs, rs)) - \Phi (fs, rs) \leq 3
\]

The two observer functions \(\text{first}\) and \(\text{is_empty}\) have constant running time.

### 20.3 A Real Time Implementation

This sections presents the **Hood-Melville queue**, a tricky implementation that improves upon the representation in the previous Section by preemptively performing reversals over a number of operations before they are required.

#### 20.3.1 Stepped Reversal

Breaking down a reversal operation into multiple steps can be done using the following function:

\[
\text{rev\_step} :: 'a\ list \times 'a\ list \Rightarrow 'a\ list \times 'a\ list
\text{rev\_step} (x \# xs, ys) = (xs, x \# ys)
\text{rev\_step} ([\], ys) = ([], ys)
\]
where \( x \neq xs \) is the list being reversed, and \( x \neq ys \) is the partial reversal result. Thus, to reverse a list of size 3 one should call \( \text{rev}_\text{step} \) 3 times:

\[
\text{rev}_\text{step} \left( [1, 2, 3], [] \right) = ([2, 3, 1]) \\
\text{rev}_\text{step} \left( \text{rev}_\text{step} \left( [1, 2, 3], [] \right) \right) = ([3, 2, 1]) \\
\text{rev}_\text{step} \left( \text{rev}_\text{step} \left( \text{rev}_\text{step} \left( [1, 2, 3], [] \right) \right) \right) = ([], [3, 2, 1])
\]

Note that each call to \( \text{rev}_\text{step} \) takes constant time since its definition is non-recursive.

Using the notation \( f^n \) for the \( n \)-fold composition of function \( f \) we can state a simple inductive lemma:

**Lemma 20.1.** \( \text{rev}_\text{step} \left( [x] \right) \) \( (xs, ys) = ([], rev \ xs @ ys) \)

As a special case this implies \( \text{rev}_\text{step} \left( [x] \right) (xs, []) = ([], rev \ xs) \).

### 20.3.2 A Real Time Intuition

Hood-Melville queues are similar to those presented in Section 20.2 in that they use a pair of lists \((f, r)\) (front and rear — for succinctness we drop the s’s now) to achieve constant running time \( \text{deq} \) and \( \text{enq} \). However, they avoid a costly reversal operation once \( f \) becomes empty by preemptively computing a new front \( fr = f @ \text{rev} \ r \) one step at a time using \( \text{rev}_\text{step} \) as enqueueing and dequeueing operations occur. The process that generates \( fr \) consists of three phases:

1. Reverse \( r \) to form \( r' \), which is the tail end of \( fr \)
2. Reverse \( f \) to form \( f' \)
3. Reverse \( f' \) onto \( r' \) to form \( fr \)

All three phases can be described in terms of \( \text{rev}_\text{step} \) as follows:

\[
1. \quad r' = \text{snd} \left( \text{rev}_\text{step} \left( [r] \right) \right) \\
2. \quad f' = \text{snd} \left( \text{rev}_\text{step} \left( [f] \right) \right) \\
3. \quad fr = \text{snd} \left( \text{rev}_\text{step} \left( [f'] \right) \right)
\]

Phases (1) and (2) are independent and can be performed at the same time, hence, when starting from this configuration

\[
\begin{array}{ccc}
\text{f} & \text{f}' & \text{r} \\
|0 & \ldots & |m \\
|n+1 & \ldots & |n \\
\end{array}
\]

after \( \max |f| \ |r| \) steps of reversal the state would be the following:

\[
\begin{array}{ccc}
\text{f} & \text{f}' & \text{r} & \text{r}' \\
| & |m & \ldots & |0 \\
|n & \ldots & |n+1 \\
\end{array}
\]
Phase (3) reverses \( f' \) onto \( r' \) to obtain the same result as a call to \( \text{list} \):

\[
fr = \text{snd} (\text{rev\_step}[^f'] (f', r')) \\
= \text{rev} f' @ r' \\
= \text{rev} f' @ \text{snd} (\text{rev\_step}[r] (r, [])) \\
= \text{rev} f' @ \text{rev} r \\
= \text{rev} (\text{snd} (\text{rev\_step}[^f'] (f, []))) @ \text{rev} r' \\
= \text{rev} (\text{rev} f) @ \text{rev} r \\
= f @ \text{rev} r 
\]

by definition of \( fr \)
by definition of \( f' \)
by definition of \( r' \)
by definition of \( f' \)
by definition of \( r \)
by \text{rev} involution

The resulting front list \( fr \) contains all elements previously in \( f \) and \( r \):

\[
\begin{array}{cccc}
q_0 & \ldots & q_m & q_{m+1} & \ldots & q_n \\
\hline
f & \text{rev} & r
\end{array}
\]

A Hood-Melville queue spreads all reversal steps across queue altering operations requiring careful bookkeeping. To achieve this gradual reversal, additional lists \( \text{front} \) and \( \text{rear} \) are used for enqueuing and dequeuing, while internal operations rely only on \( f, f', r, \) and \( r' \). At the start of the reversal process \( \text{rear} \) is copied into \( r \) and emptied; similarly, \( \text{front} \) is copied into \( f \) but its contents are kept as they might need to be dequeued. Moreover, to avoid using elements from \( f \) or \( f' \) that may have been removed from \( \text{front} \), a counter \( d \) records the number of dequeuing operations that have occurred since the reversal process started; this way, only \( |f'| - d \) elements are appended into \( r \) to form \( fr \). Once the reversal finishes \( fr \) become the new \( \text{front} \) and the internal lists are cleared. When the queue is not being reversed all operations are performed in a manner similar to previous implementations. The configuration of a queue at the beginning of the reversal process is as follows:

\[
\begin{array}{cccc}
q_0 & \ldots & q_m & q_{m+1} & \ldots & q_n \\
\hline
\text{front} & \text{deq} & \text{enq} & \text{rear} & d = 0
\end{array}
\]

\[
\begin{array}{cccc}
q_0 & \ldots & q_m & q_{m+1} & \ldots & q_n \\
\hline
\text{queue}
\end{array}
\]
20.3.3 The Reversal Strategy

A crucial detail of this implementation is determining at which point the reversal process should occur. The strategy is to start once \(|\text{rear}|\) becomes larger than \(|\text{front}|\), and ensure that all reversal steps are done before \(\text{front}\) runs out of elements or \(\text{rear}\) becomes larger than the new front \((\text{fr})\).

With this strategy, once \(|\text{rear}| = \text{n} + 1\) and \(|\text{front}| = \text{n}\), the reversal process starts. The first two phases take \(\text{n} + 1\) steps \((\max |\text{front}| |\text{rear}|)\) to generate \(f'\) and \(r'\), and the third phase produces \(\text{fr}\) in \(n\) steps. A complete reversal takes \(2 \cdot \text{n} + 1\) steps. Because the queue can only perform \(n\) \text{deq} operations before \(\text{front}\) is exhausted, \(2 \cdot \text{n} + 1\) steps must be performed in at most \(n\) operations. This can be achieved by performing the first two steps in the operation that causes \(\text{rear}\) to become larger than \(\text{front}\) and two more steps in each subsequent operation. Therefore, \(2 \cdot (\text{n} + 1)\) steps can occur before \(\text{front}\) is emptied, allowing the reversal process to finish in time.

Finally, since at most \(n\) \text{enq} or \text{deq} operations can occur during reversal, the largest possible \(\text{rear}\) has length \(n\) (only \text{enq} ops), while the smallest possible \(\text{fr}\) has length \(n + 1\) (only \text{deq} ops). Thus, after the reversing process has finished the new front \((\text{fr})\) is always larger than \(\text{rear}\).

20.3.4 Implementation

Queues are implemented using the following record type:

```plaintext
record 'a queue = lenf :: nat
   front :: 'a list
   status :: 'a status
   rear :: 'a list
   lenr :: nat
```

In a nutshell, a record is a product type with named fields and "built-in" construction, selection, and update operations. Values of 'a queue are constructed using \text{make} :: nat \Rightarrow 'a list \Rightarrow 'a status \Rightarrow 'a list \Rightarrow nat \Rightarrow 'a queue were each argument corresponds to one of the fields of the record in canonical order. Additionally, given a queue \(q\) we can obtain the value in field \text{front} with \(q\text{front} := []\), and update its content using \(q\text{front} := []\). Multiple updates can be composed as \(q\text{front} := [], \text{rear} := []\).

All values in the queue along with its internal state are stored in the various fields of 'a queue. Fields \text{front} and \text{rear} contain the lists over which all queue operations are performed. The length of \text{front} and \text{rear} is recorded in \text{lenf} and \text{lenr} (respectively) to avoid calling \text{length} whose complexity is
not constant. Finally, \textit{status} tracks the current reversal phase of the queue in a \textit{'a status} value.

\begin{verbatim}

datatype 'a status =
    Idle |
    Rev nat ('a list) ('a list) ('a list) ('a list) |
    App nat ('a list) ('a list) |
    Done

\end{verbatim}

Each value of \textit{'a status} represents either a phase of reversal or the queue's normal operation. Constructor \textit{Idle} signals that no reversal is being performed. \textit{Rev ok} \textit{f} \textit{f'} \textit{r} \textit{r'} corresponds to phases (1) and (2) where the lists \textit{f}, \textit{f'}, \textit{r}, and \textit{r'} are used for the reversal steps of the front and the rear. The \textit{App ok} \textit{f} \textit{r} case corresponds to phase (3) where both lists are appended to form the new front (\textit{fr}). In both \textit{App} and \textit{Rev}, the first argument \textit{ok} (of type \textit{nat}) keeps track of the number of elements in \textit{f'} that have not been removed from the queue, effectively \textit{ok} = |\textit{f'}| - \textit{d}, where \textit{d} is the number of \textit{deq} operations that have occurred so far. Lastly, \textit{Done fr} marks the end of the reversal process and contains only the new front list \textit{fr}.

In the implementation, all of the steps of reversal operations in the queue are performed by functions \textit{exec} and \textit{invalidate}; they ensure at each step that the front list being computed is kept consistent w.r.t. the contents and operations in the queue.

Function \textit{exec} :: \textit{'a status} \Rightarrow \textit{'a status} performs the incremental reversal of the front list by altering the queue's \textit{status} one step at a time in accordance with the reversal phases. Following the strategy described in Section 20.3.3, all queue operations call \textit{exec} twice to be able to finish the reversal in time. On \textit{Idle} queues \textit{exec} has no effect. The implementation of \textit{exec} is an extension of \textit{rev_step} with specific considerations for each \textit{status} value and is defined as follows:

\begin{verbatim}

exec :: 'a status \Rightarrow 'a status
exec (Rev ok (x # f) f' (y # r) r') = Rev (ok + 1) f (x # f') r (y # r')
exec (Rev ok [] f' [y] r') = App ok f'(y # r')
exec (App 0 f' r') = Done r'
exec (App ok (x # f') r') = App (ok - 1) f' (x # r')
exec s = s

\end{verbatim}

If the \textit{status} is \textit{Rev ok f f'} \textit{r} \textit{r'}, then \textit{exec} performs two (or one if \textit{f} = []) simultaneous reversal steps from \textit{f} and \textit{r} into \textit{f'} and \textit{r'}; moreover \textit{ok} is
incremented if a new element has been added to \( f' \). Once \( f \) is exhausted and \( r \) is a singleton list, the remaining element is moved into \( r' \) and the status is updated to the next phase of reversal. In the \( \text{App ok} f' r' \) phase, \( \text{exec} \) moves elements from \( f' \) to \( r' \) until \( \text{ok} = 0 \), at which point \( r' \) becomes the new front by transitioning into \( \text{Done} r' \). In all other cases \( \text{exec} \) behaves as the identity function. As is apparent from its implementation, a number of assumptions are required for \( \text{exec} \) to function properly and eventually produce \( \text{Done} \). These assumption are discussed in Section 20.3.5.

If an element is removed from the queue during the reversal process, it also needs to be removed from the new front list (\( fr \)) being computed. Function \( \text{invalidate} \) is used to achieve this:

\[
\text{invalidate} :: \text{\textquoteleft status \textquoteleft} \Rightarrow \text{\textquoteleft status \textquoteleft}
\]

\[
\text{invalidate} \ (\text{Rev ok} f' f r r') = \text{Rev} \ (\text{ok} - 1) \ f f' r r'
\]

\[
\text{invalidate} \ (\text{App} 0 f' (\_ \# r')) = \text{Done} r'
\]

\[
\text{invalidate} \ (\text{App ok} f' r') = \text{App} \ (\text{ok} - 1) \ f f' r'
\]

\[
\text{invalidate} \ s = s
\]

By decreasing the value of \( \text{ok} \), the number of elements from \( f' \) that are moved into \( r' \) in phase (3) is reduced, since \( \text{exec} \) might produce \( \text{Done} \) early, once \( \text{ok} = 0 \), ignoring the remaining elements of \( f' \). Furthermore, since \( f' \) is a reversal of the front list, elements left behind in its tail correspond directly to those being removed from the queue.

The rest of the implementation is shown below. Auxiliary function \( \text{exec2} \), as its name suggests, applies \( \text{exec} \) twice and updates the queue accordingly if \( \text{Done} \) is returned.

\[
\text{exec2} :: \text{\textquoteleft queue \textquoteleft} \Rightarrow \text{\textquoteleft queue \textquoteleft}
\]

\[
\text{exec2} \ q = \ (\text{case} \ \text{exec} \ (\text{exec} \ q) \ \text{of}
\]

\[
\text{Done fr} \Rightarrow \ q\{\text{status} = \text{Idle}, \text{front} = fr\} \ |
\]

\[
\text{newstatus} \Rightarrow \ q\{\text{status} = \text{newstatus}\}
\]

\[
\text{check} :: \text{\textquoteleft queue \textquoteleft} \Rightarrow \text{\textquoteleft queue \textquoteleft}
\]

\[
\text{check} \ q
\]

\[
= \ (\text{if} \ \text{lenr} \ q \leq \ \text{lenf} \ q \ \text{then} \ \text{exec2} \ q
\]

\[
\text{else let} \ \text{newstate} = \ \text{Rev} \ 0 \ (\text{front} \ q) \ [] \ (\text{rear} \ q) []
\]

\[
\text{in} \ \text{exec2}
\]

\[
(q\{\text{lenf} := \text{lenf} \ q + \text{lenr} \ q, \text{status} := \text{newstate}, \text{rear} := [],
\text{lenr} := 0]\})
\]
empty :: 'a queue
empty = make 0 [] Idle [] 0

first :: 'a queue ⇒ 'a
first q = hd (front q)

enq :: 'a ⇒ 'a queue ⇒ 'a queue
enq x q = check (q|rear := x # rear q, lenr := lenr q + 1|)

deq :: 'a queue ⇒ 'a queue
deq q = check
  (q|lenf := lenf q - 1, front := tl (front q),
   status := invalidate (status q)|)

The two main queue operations, enq and deq, alter front and rear as expected, with additional updates to lenf and lenr to keep track of their length. To perform all “internal” operations, both functions call check. Additionally, deq uses invalidate to mark elements as removed.

Function check calls exec2 if lenr is not larger than lenf. Otherwise a reversal process is initiated: rear is emptied and lenr is set to 0; lenf is increased to the size of the whole queue since, conceptually, all element are now in the soon-to-be-computed front; the status newstate is initialized as described at the beginning of Section 20.3.2.

The time complexity of this implementation is clearly constant, since there are no recursive functions.

20.3.5 Functional Correctness

To show this implementation is an instance of the ADT Queue, we need a number of invariants to ensure the consistency of 'a queue values are preserved by all operations.

Initially, as hinted by the definition of exec, values of type 'a status should have specific properties to guarantee a Done result after a (finite) number of calls to exec. The predicate inv_st defines these properties as follows:

inv_st :: 'a status ⇒ bool
inv_st (Rev ok f f' r r') = (|f| + 1 = |r| ∧ |f'| = |r'| ∧ ok ≤ |f'|)
First, \( \text{inv}_{\text{st}} \) ensures for pattern, \( \text{Rev ok } f' r' \) that arguments \( f \) and \( r \) follow the reversal strategy, and counter \( \text{ok} \) is only ever increased as elements are added to \( f' \). Similarly, for \( \text{App ok } f' r' \), it must follow that \( r' \) remains larger than \( f' \), and \( |f'| \) provides an upper bound for \( \text{ok} \). All other patterns trivially fulfill the invariant.

The consistency of 'a queue values is then defined as an extension of \( \text{inv}_{\text{st}} \) and considers all the other fields in the queue. The predicate \( \text{invar} \) defines the required properties:

\[
\text{invar } q = \left( \| \text{lenf } q \| = \| \text{front_list } q \| \land \| \text{lenr } q \| = \| \text{rear_list } q \| \land \right. \\
\text{lenr } q \leq \text{lenf } q \land \\
\left( \begin{array}{l}
\text{case } \text{status } q \text{ of } \\
\text{Idle } \Rightarrow \right. \\
2 \cdot \text{lenr } q \leq |f'| \land \text{ok} \neq 0 \land \\
2 \cdot |f'| + \text{ok} + 2 \leq 2 \cdot |\text{front } q| \\
\text{Rev ok } f' r' \Rightarrow 2 \cdot \text{lenr } q \leq |r'| \land \text{ok} + 1 \leq 2 \cdot |\text{front } q| \\
| - \Rightarrow \text{True} \right) \land \\
(\exists \text{rest. front_list } q = \text{front } q @ \text{rest}) \land \\
(\exists \text{fr. status } q = \text{Done fr}) \land \\
\text{inv}_{\text{st}} (\text{status } q))
\]

Predicate \( \text{lenr } q = |\text{rear_list } q| \) ensures \( \text{lenr} \) is equal to the length of the queue's rear, where function \( \text{rear_list } q \), defined as \( (\text{rev } \circ \text{rear}) q \), produces the rear list in canonical order. Likewise, \( \text{lenf } q = |\text{front_list } q| \) matches \( \text{lenf} \) to the queue's front. However, function \( \text{front_list} \) warrants special attention as it must compute the list representing the front of the queue even during a reversal. \( \text{front_list} \) is defined as follows:

\[
\text{front_list} ::= 'a \text{ queue } \Rightarrow 'a \text{ list} \\
\text{front_list } q = \left( \begin{array}{l}
\text{case } \text{status } q \text{ of } \text{Idle } \Rightarrow \text{front } q \\
| \text{Rev ok } f' r' \Rightarrow \text{rev } (\text{take } ok f') \@ f \@ \text{rev } r \@ r' \\
| \text{App ok } f' x \Rightarrow \text{rev } (\text{take } ok f') \@ x | \text{Done } f \Rightarrow f)
\right)
\]
For case $App\ ok\ f'\ r'$, the front list corresponds to the final result of the stepped reversal (20.1) but only elements in $f'$ that are still in the queue, denoted by $take\ ok\ f'$, are considered. Analogously for $Rev\ ok\ f\ f'\ r\ r'$, both stepped reversal results are appended and only relevant elements in $f'$ are used, however, rear lists $r$ and $r'$ are reversed again to achieve canonical order.

Continuing with $invar$, inequality $lenr\ q \leq lenf\ q$ is the main invariant in our reversal strategy, and by the previous two equalities must hold even as internal operations occur. Furthermore, predicate $\exists\ rest.\ front\_list\ q = front\ q @ rest$ ensures $front\ q$ is contained within $front\_list\ q$, thus preventing any mismatch between the internal state and the queue's front. Given that $exec2$ is the only function that manipulates a queue's $status$, it holds that $\exists fr.\ status\ q = Done\ fr$ since any internal $Done$ result is replaced by $Idle$.

The case distinction on $status\ q$ places size bounds on internal lists $front$ and $rear$ ensuring the front does not run out of elements and the rear never grows beyond $lenr\ q \leq lenf\ q$. In order to clarify some of formulations used in this part of $invar$, consider the following correspondences, which hold once the reversal process starts:

- $lenr\ q$ corresponds to the number of $enq$ operations performed so far, and $2 \cdot lenr\ q$ denotes the $exec$ applications in those operations.
- $|front\ q|$ corresponds to the number of $deq$ operations that can be performed before $front\ q$ is exhausted. Therefore, $2 \cdot |front\ q|$ is the minimum number of $exec$ applications the queue will be able to do at any given point.
- On $Rev\ ok\ f\ f'\ r\ r'$ status, $|f'|$ corresponds to the number of applications of $exec$ performed so far and the internal length of front being constructed. Expression $|r|$ is the analogous for a $App\ ok\ f\ r$.
- From a well formed $App\ ok\ f\ r$ it takes $ok + 1$ applications of $exec$ to reach $Done$. Since, the base case of $App$ is obtained after $ok$ applications, and the transition into $Done$ takes an extra step.
- From a well formed $Rev\ ok\ f\ f'\ r\ r'$ it takes $2 \cdot |f'| + ok + 2$ applications of $exec$ to reach $Done$. Since, the base case of $Rev$ is obtained after $|f'|$ applications (incrementing $ok$ by the same amount), the transitioning into $App$ takes one step, and $ok + |f'|$ extra steps are need to reach $Done$ from $App$.

In the $Rev\ ok\ f\ f'\ r\ r'$ case, $2 \cdot lenr\ q \leq |f'|$ ensures $f'$ grows larger with every $enq$ operation and the internal list is at least twice the length of the queue's rear. Additionally, the value of $ok$ cannot be 0 as this either marks the beginning of a reversal which calls $exec2$ immediately, or signals that elements in $front\ q$ have run out. Finally, to guarantee the reversal process can finish before the $front\ q$ is exhausted the number of $exec$ applications...
before reaching Done must be less than the minimum number of applications possible, denoted by \(2 \cdot |f| + ok + 2 \leq |front q|\).

Case App \(ok \ f \ r\) has similar invariants, with equation \(2 \cdot \text{lenr} \ q \leq |r|\) bounding the growth of \(r\) as it was previously done with \(f\). Moreover, \(ok + 1 \leq 2 \cdot |front q|\) ensures \(fron q\) is not exhausted before the reversal is completed.

Using \(\text{invar}\) we can prove the properties of \(\text{enq}\) and \(\text{deq}\) required by the \(\text{Queue}\) ADT using this abstraction following abstraction function:

\[
\text{list} :: 'a \ \text{queue} \Rightarrow 'a \ \text{list} \\
\text{list} \ q = \text{front_list} \ q \ @ \ \text{rear_list} \ q
\]

Proofs follow mostly by case distinction on the \(\text{status}\) field. Invariants \(\text{invar}\) and \(\text{inv_st}\) are used to filter invalid \('a \ \text{queue}\) values, and to simplify terms based on their relationship with other terms (e.g. \(f = [] \land |f| + 1 = |r| \rightarrow (\exists \ x. \ r = [x])\)). All other properties in \(\text{Queue}\) are straightforward.

**Bibliographic Remarks**

The representation of queues as pairs of lists is due to Burton [11]. The Hood-Melville queues are due to Hood and Melville [30]. The implementation is based on the presentation by Okasaki [62].
Splay Trees

Splay trees are fascinating self-organizing search trees. Self-organizing means that the tree structure is modified upon access (including `isin` queries) to improve the performance of subsequent operations. Concretely, every splay tree operation moves the element concerned to the root. Thus splay trees excel in applications where a small fraction of the entries are the targets of most of the operations. In general, splay trees perform as well as any static binary search tree.

Splay trees have two drawbacks. First, their performance guarantees (logarithmic running time of each operation) are only amortized. Self-organizing does not mean self-balancing: splay trees can become unbalanced, in contrast to, for example, red-black trees. Second, because `isin` modifies the tree, splay trees are less convenient to use in a purely functional language.

21.1 Implementation

The central operation on splay trees is the `splay` function shown in Figure 21.1. It rotates the given element \( x \) to the root of the tree if \( x \) is already in the tree. Otherwise the last element found before the search for \( x \) hits a leaf is rotated to the root.

Function `isin` has a trivial implementation in terms of `splay`:

\[
\text{isin :: 'a tree \Rightarrow 'a \Rightarrow bool}
\]

\[
isin \; t \; x = (\text{case splay} \; x \; t \; \text{of} \; \langle \rangle \Rightarrow False \; | \; \langle., \; a, \; \_\rangle \Rightarrow x = a)
\]

Except that `splay` creates a new tree that needs to be returned from a proper `isin` as well to achieve the amortized logarithmic complexity (see the discussion of observer functions at the end of Section 19.1). This is why splay trees
splay \(x\) \(<AB, b, CD>\)

\[
\begin{align*}
splay & x \langle AB, b, CD \rangle \\
& = (\text{case } \text{cmp } x \ b \ \text{of} \\
& \text{LT } \Rightarrow \text{ case } AB \ \text{of} \\
& \quad | \Rightarrow \langle AB, b, CD \rangle | \\
& \quad \langle A, a, B \rangle \Rightarrow \\
& \quad \text{case } \text{cmp } x \ a \ \text{of} \\
& \quad \text{LT } \Rightarrow \text{ if } A = () \text{ then } \langle A, a, \langle B, b, CD \rangle \rangle | \\
& \quad \text{else case } splay \ x \ A \ \text{of} \\
& \quad \quad \langle A_1, a', A_2 \rangle \Rightarrow \langle A_1, a', \langle A_2, a, \langle B, b, CD \rangle \rangle \rangle | \\
& \quad \text{EQ } \Rightarrow \langle A, a, \langle B, b, CD \rangle \rangle | \\
& \quad \text{GT } \Rightarrow \text{ if } B = () \text{ then } \langle A, a, \langle B, b, CD \rangle \rangle | \\
& \quad \text{else case } splay \ x \ B \ \text{of} \\
& \quad \quad \langle B_1, b', B_2 \rangle \Rightarrow \langle \langle A, a, B_1 \rangle, b', \langle B_2, b, CD \rangle \rangle | \\
& \quad \text{EQ } \Rightarrow \langle AB, b, CD \rangle | \\
& \quad \text{GT } \Rightarrow \text{ case } CD \ \text{of} \\
& \quad \quad | \Rightarrow \langle AB, b, CD \rangle | \\
& \quad \quad \langle C, c, D \rangle \Rightarrow \\
& \quad \quad \text{case } \text{cmp } x \ c \ \text{of} \\
& \quad \quad \text{LT } \Rightarrow \text{ if } C = () \text{ then } \langle \langle AB, b, C \rangle, c, D \rangle | \\
& \quad \quad \text{else case } splay \ x \ C \ \text{of} \\
& \quad \quad \quad \langle C_1, c', C_2 \rangle \Rightarrow \langle \langle AB, b, C_1 \rangle, c', \langle C_2, c, D \rangle \rangle | \\
& \quad \quad \text{EQ } \Rightarrow \langle \langle AB, b, C \rangle, c, D \rangle | \\
& \quad \quad \text{GT } \Rightarrow \text{ if } D = () \text{ then } \langle \langle AB, b, C \rangle, c, D \rangle | \\
& \quad \quad \text{else case } splay \ x \ D \ \text{of} \\
& \quad \quad \quad \langle D_1, d, D_2 \rangle \Rightarrow \langle \langle AB, b, C \rangle, c, D_1 \rangle, d, D_2 \rangle 
\end{align*}
\]

Fig. 21.1. Function \textit{splay}

are inconvenient in functional languages. For the moment we ignore this aspect and stick with the above \textit{isin} because it has the type required by the \textit{Set ADT}.

The implementation of \textit{insert} \(x\) \(t\) in Figure 21.2 is straightforward: let \(<l, a, r> = splay \ x \ t\); if \(a = x\), return \(<l, a, r>\); otherwise make \(x\) the root of a suitable recombination of \(l, a\) and \(r\).

The implementation of \textit{delete} \(x\) \(t\) in Figure 21.3 starts similarly: let \(<l, a, r> = splay \ x \ t\); if \(a \neq x\), return \(<l, a, r>\). Otherwise follow the deletion-by-replacing paradigm (Section 5.2.1): if \(l \neq ()\), splay the maximal element \(m\) in \(l\) to the root and replace \(x\) with it. Note that \textit{splay\_max} returns a tree that is just a glorified pair: if \(t \neq ()\) then \textit{splay\_max} \(t\) is of the form \(<t', m, ()>\). The definition \textit{splay\_max} \(() = ()\) is not really needed (\textit{splay\_max} is always called with non-\(()\) argument) but some lemmas can be stated more slickly with this definition.
21.2 Correctness

The inorder approach of Section 5.4 applies. Because the details are a bit different (everything is reduced to splay) we present the top-level structure.

The following easy inductive properties are used implicitly in a number of subsequent proofs:

- \( splay a \ t = \emptyset \iff t = \emptyset \)
- \( splay\text{-}max \ t = \emptyset \iff t = \emptyset \)

Correctness of isin

\[ \text{sorted} (\text{inorder} \ t) \implies \text{isin} \ t \ x = (x \in \text{set} (\text{inorder} \ t)) \]
follows directly from this easy inductive property of \textit{splay}:

\[ splay \ x \ t = \langle l, a, r \rangle \wedge \text{sorted} (\text{inorder} \ t) \implies (x \in \text{set} (\text{inorder} \ t)) = (x = a) \]

Correctness of \textit{insert} and \textit{delete}

\[ \text{sorted} (\text{inorder} \ t) \implies \text{inorder} (\text{insert} \ x \ t) = \text{ins_list} \ x (\text{inorder} \ t) \]

\[ \text{sorted} (\text{inorder} \ t) \implies \text{inorder} (\text{delete} \ x \ t) = \text{del_list} \ x (\text{inorder} \ t) \]

relies on the following characteristic inductive properties of \textit{splay}:

\[ \text{inorder} (splay \ x \ t) = \text{inorder} \ t \tag{21.1} \]

\[ \text{sorted} (\text{inorder} \ t) \wedge \text{splay} \ x \ t = \langle l, a, r \rangle \implies \text{sorted} (\text{inorder} \ l @ x \ # \ \text{inorder} \ r) \]

Correctness of \textit{delete} also needs the inductive property

\[ splay\_max \ t = \langle l, a, r \rangle \wedge \text{sorted} (\text{inorder} \ t) \implies \text{inorder} \ l @ [a] = \text{inorder} \ t \wedge r = \langle \rangle \]

Note that \text{inorder} (splay \ x \ t) = \text{inorder} \ t is also necessary to justify the proper isin that returns the newly created tree as well.

Automation of the above proofs requires the lemmas in Figure 5.2 together with a few additional lemmas about \textit{sorted}, \textit{ins_list} and \textit{del_list} that can be found in the Isabelle proofs.

Recall from Section 5.4 that correctness of \textit{insert} and \textit{delete} implies that they preserve \textit{bst} = \text{sorted} \circ \text{inorder}. Similarly, (21.1) implies that \textit{splay} preserves \textit{bst}. Thus we may assume the invariant \textit{bst} in the amortized analysis.

These two easy size lemmas are used implicitly below:

\[ |splay \ a \ t| = |t| \quad |splay\_max \ t| = |t| \]

\section*{21.3 Amortized Analysis}

This section shows that \textit{splay}, insertion and deletion all have amortized logarithmic complexity.

We define the potential \(\Phi\) of a tree as the sum of the potentials \(\varphi\) of all nodes:

\[
\begin{align*}
\Phi & : \ 'a \ \text{tree} \Rightarrow \ \text{real} \\
\Phi \ \langle \rangle & = 0 \\
\Phi \ \langle l, a, r \rangle & = \varphi (\langle l, a, r \rangle) + \Phi \ l + \Phi \ r \\
\varphi \ t & \equiv \log_2 |t|_1 
\end{align*}
\]
The central result is the amortized complexity of *splay*. Function $T_{\text{splay}}$ is shown in Appendix B.7. We follow (19.1) and define

$$A_{\text{splay}} a t = T_{\text{splay}} a t + \phi (splay a t) - \phi t$$

First we consider the case where the element is in the tree:

**Theorem 21.1.** $bst t \land \langle l, x, r \rangle \in \text{subtrees } t \quad \Rightarrow \quad A_{\text{splay}} x t \leq 3 \cdot (\phi t - \phi \langle l, x, r \rangle) + 1$

**Proof** by induction on the computation of *splay*. The base cases involving $\langle \rangle$ are impossible. For example, consider the call $\text{splay } x t$ where $t = \langle \rangle, b, C \rangle$ and $x < b$: from $\text{bst } t \land \langle \rangle \in \text{set_tree } t$ but because $bst t$ and $x < b$ this implies that $x \in \text{set_tree } \langle \rangle$, a contradiction.

There are three feasible base cases. The case $t = \langle _, x, _ \rangle$ is easy. We consider one of the two other symmetric cases. Let $t = \langle A, x, B \rangle, b, C \rangle$ and $t' = \text{splay } x t = \langle A, x, B, b, C \rangle$.

$$A_{\text{splay}} x t = \Phi t' - \Phi t + 1 \quad \text{by definition of } A_{\text{splay}} \text{ and } T_{\text{splay}}$$

$$= \phi t' + \phi \langle B, b, C \rangle - \phi t - \phi \langle A, x, B \rangle + 1 \quad \text{by definition of } \Phi$$

$$\leq 3 \cdot (\phi t - \phi \langle A, x, B \rangle) + 1 \quad \text{because } \phi \langle B, b, C \rangle \leq \phi t$$

There are four inductive cases. We consider two of them, the other two are symmetric variants. First the so-called zig-zig case:

```
  b         b -> d
 / \       / \           / \     
 a   C     a   C       A1   a
 / \     / \       / \     / \     
 A   B   d   B     A2   b     
 / \     / \     / \     / \     
 A1  A2   B  C     
```

This is the case where $x < a < b$ and $A \neq \langle \rangle$. On the left we have the input and on the right the output of *splay* $x$. Because $A \neq \langle \rangle$, *splay* $x A = \langle A_1, a', A_2 \rangle = A'$ for some $A_1, a'$ and $A_2$. The intermediate tree is obtained by replacing $A$ by $A'$. This tree is shown for illustration purpose only; in the algorithm the right tree is constructed directly from the left one. Let $X = \langle l, x, r \rangle$. Clearly $X \in \text{subtrees } A$. We abbreviate compound trees like $\langle A, a, B \rangle$ by the names of their subtrees, in this case $AB$. First note that

$$\phi A_1 A_2 B C = \phi A B C$$

\[ (*) \]
because $|A'| = |splay x A| = |A|$. We can now prove the claim:

$$A_{splay} x ABC = T_{splay} x A + 1 + \Phi A_1 A_2 BC - \Phi ABC$$

$$= T_{splay} x A + 1 + \Phi A_1 + \Phi A_2 + \varphi A_2 BC + \varphi BC - \Phi A - \varphi AB$$

by (\*) and definition of $\Phi$

$$= T_{splay} x A + \Phi A' - \varphi A' - \Phi A + \varphi A_2 BC + \varphi BC - \varphi AB + 1$$

$$= A_{splay} x A + \varphi A_2 BC + \varphi BC - \varphi AB - \varphi A' + 1$$

$$\leq 3 \cdot \varphi A + \varphi A_2 BC + \varphi BC - \varphi AB - \varphi A' - 3 \cdot \varphi X + 2$$

by IH and $X \in$ subtrees $A$

$$= 2 \cdot \varphi A + \varphi A_2 BC + \varphi BC - \varphi AB - 3 \cdot \varphi X + 2$$

because $\varphi A = \varphi A'$

$$< \varphi A + \varphi A_2 BC + \varphi BC - 3 \cdot \varphi X + 2$$

because $\varphi A < \varphi AB$

$$< \varphi A_2 BC + 2 \cdot \varphi ABC - 3 \cdot \varphi X + 1$$

because $1 + \lg x + \lg y < 2 \cdot \lg (x + y)$ if $x, y > 0$

$$< 3 \cdot (\varphi ABC - \varphi X) + 1$$

because $\varphi A_2 BC < \varphi ABC$

Now we consider the so-called zig-zag case:

This is the case where $a < x < b$ and $B \neq \langle \rangle$. On the left we have the input and on the right the output of $splay x$. Because $B \neq \langle \rangle$, $splay x B = \langle B_1, b', B_2 \rangle = B'$ for some $B_1$, $b'$ and $B_2$. The intermediate tree is obtained by replacing $B$ by $B'$. Let $X = \langle l, x, r \rangle$. Clearly $X \in$ subtrees $B$. The proof is very similar to the zig-zig case, the same naming conventions apply and we omit some details:

$$A_{splay} x ABC = T_{splay} x A + 1 + \Phi AB_1 B_2 C - \Phi ABC$$

$$= A_{splay} x B + \varphi AB_1 + \varphi B_2 C - \varphi AB - \varphi B' + 1$$

using $\varphi AB_1 B_2 C = \varphi ABC$

$$\leq 3 \cdot \varphi B + \varphi AB_1 + \varphi B_2 C - \varphi AB - \varphi B' - 3 \cdot \varphi X + 2$$

by IH and $X \in$ subtrees $B$

$$= 2 \cdot \varphi B + \varphi AB_1 + \varphi B_2 C - \varphi AB - 3 \cdot \varphi X + 2$$

because $\varphi B = \varphi B'$

$$< \varphi B + \varphi AB_1 + \varphi B_2 C - 3 \cdot \varphi X + 2$$

because $\varphi B < \varphi AB$

$$< \varphi B + 2 \cdot \varphi ABC - 3 \cdot \varphi X + 1$$

because $1 + \lg x + \lg y < 2 \cdot \lg (x + y)$ if $x, y > 0$

$$< 3 \cdot (\varphi ABC - \varphi X) + 1$$

because $\varphi B < \varphi ABC$
Because \( \varphi(l, x, r) \geq 1 \), the above theorem implies

**Corollary 21.2.** \( \text{bst } t \wedge x \in \text{set}_\text{tree} \ t \rightarrow A_{\text{splay}} x \ t \leq 3 \cdot (\varphi \ t - 1) + 1 \)

If \( x \) is not in the tree we show that there is a \( y \) in the tree such that splaying with \( y \) would produce the same tree in the same time:

**Lemma 21.3.** \( t \neq \langle \rangle \wedge \text{bst } t \rightarrow \exists y \in \text{set}_\text{tree} \ t. \text{splay} \ y \ t = \text{splay} \ x \ t \wedge T_{\text{splay}} y \ t = T_{\text{splay}} x \ t \)

Element \( y \) is the last element in the tree that the search for \( x \) encounters before it hits a leaf. Naturally, the proof is by induction on the computation of \( \text{splay} \).

Combining this lemma with Corollary 21.2 yields the final unconditional amortized complexity of \( \text{splay} \) on BSTs:

**Corollary 21.4.** \( \text{bst } t \rightarrow A_{\text{splay}} x \ t \leq 3 \cdot \varphi \ t + 1 \)

The “\(-1\)” has disappeared to accommodate the case \( t = \langle \rangle \).

The amortized analysis of insertion is straightforward now. From the amortized complexity of \( \text{splay} \) it follows that

**Lemma 21.5.** \( \text{bst } t \rightarrow T_{\text{insert}} x \ t + \Phi (\text{insert} x \ t) - \Phi t \leq 4 \cdot \varphi t + 3 \)

We omit the proof which is largely an exercise in simple algebraic manipulations.

The amortized analysis of deletion is similar but a bit more complicated because of the additional function \( \text{splay}_\text{max} \) whose amortized running time is defined as usual:

\[
A_{\text{splay}_\text{max}} t = T_{\text{splay}_\text{max}} t + \Phi (\text{splay}_\text{max} t) - \Phi t
\]

An inductive proof and then a simple case analysis yield

\[
t \neq \langle \rangle \rightarrow A_{\text{splay}_\text{max}} t \leq 3 \cdot (\varphi t - 1) + 1
\]

\[
A_{\text{splay}_\text{max}} t \leq 3 \cdot \varphi t + 1
\]

Based on the canonical definitions of \( T_{\text{delete}} \) and \( T_{\text{splay}_\text{max}} \) the amortized logarithmic complexity of \( \text{delete} \) follows:

\[
\text{bst } t \rightarrow T_{\text{delete}} a \ t + \Phi (\text{delete} a \ t) - \Phi t \leq 6 \cdot \varphi t + 3
\]

A running time analysis of \( \text{isin} \) is trivial because \( \text{isin} \) is just \( \text{splay} \) followed by a constant-time test.

### 21.4 Exercises

**Exercise 21.1.** Find a sequence of numbers \( n_1, n_2, \ldots, n_k \) such that the insertion of these numbers one by one creates a tree of height \( k \).
Bibliographic Remarks

Splay trees were invented and analyzed by Sleator and Tarjan [70] for which they received the 1999 ACM Paris Kanellakis Theory and Practice Award [38]. In addition to the amortized complexity as shown above they proved that splay trees perform as well as static BSTs (the Static Optimality Theorem) and conjectured that, roughly speaking, they even perform as well as any other BST-based algorithm. This Dynamic Optimality Conjecture is still open.

This chapter is based on earlier publications [69, 53, 54, 55].
Skew Heaps

Skew heaps are heaps in the sense of Section 14.1 and implement mergeable priority queues. Skew heaps can be viewed as a self-adjusting form of leftist heaps that attempts to maintain balance by unconditionally swapping all nodes in the merge path when merging two heaps.

22.1 Implementation of ADT Priority_Queue_Merge

The central operation is `merge`:

```
merge :: 'a tree ⇒ 'a tree ⇒ 'a tree
merge ∅ t = t
merge t ∅ = t
merge ((l₁, a₁, r₁) =: t₁) ((l₂, a₂, r₂) =: t₂)
= (if a₁ ≤ a₂ then merge t₂ r₁, a₁, l₁) else (merge t₁ r₂, a₂, l₂))
```

The remaining operations (`, insert, get_min` and `del_min`) are defined as in Section 14.1.

The following properties of `merge` have easy inductive proofs:

- `|merge t₁ t₂| = |t₁| + |t₂|`
- `mset_tree (merge t₁ t₂) = mset_tree t₁ + mset_tree t₂`
- `heap t₁ ∧ heap t₂ → heap (merge t₁ t₂)`

Now it is straightforward to prove the correctness of the implementation w.r.t. the ADT `Priority_Queue_Merge`.

Skew heaps attempt to maintain balance, but this does not always work:
Exercise 22.1. Find a sequence of numbers \( n_1, n_2, \ldots, n_k \) such that the insertion of these numbers one by one creates a tree of height \( k \). Prove that this sequence will produce a tree of height \( k \).

Nevertheless, insertion and deletion have amortized logarithmic complexity.

### 22.2 Amortized Analysis

The key is the definition of the potential. It counts the number of right-heavy (\( rh \)) nodes:

\[
\Phi : \text{'a tree} \Rightarrow \text{int}
\]

\[
\Phi() = 0
\]

\[
\Phi(l, _, r) = \Phi l + \Phi r + rh\ l\ r
\]

\[
rh : \text{'a tree} \Rightarrow \text{'a tree} \Rightarrow \text{nat}
\]

\[
rh\ l\ r = (\text{if} \ |l| < |r| \ \text{then} \ 1 \ \text{else} \ 0)
\]

The rough intuition: because \( merge \) descends along the right spine, the more right-heavy nodes a tree contains, the longer \( merge \) takes.

Two auxiliary functions count the number of right-heavy nodes on the left spine (\( lrh \)) and left-heavy (= not right-heavy) nodes on the right spine (\( rlh \)):

\[
lrh : \text{'a tree} \Rightarrow \text{nat}
\]

\[
lrh() = 0
\]

\[
lrh(l, _, r) = rh\ l\ r + lrh\ l
\]

\[
rlh : \text{'a tree} \Rightarrow \text{nat}
\]

\[
rlh() = 0
\]

\[
rlh(l, _, r) = 1 - rh\ l\ r + rlh\ r
\]

The following properties have automatic inductive proofs:

\[
2^{lrh}\ t \leq |t| + 1 \quad 2^{rlh}\ t \leq |t| + 1
\]

They imply

\[
lrh\ t \leq \lg|t|_1 \quad rlh\ t \leq \lg|t|_1 \quad (22.1)
\]

Now we are ready for the amortized analysis. All canonical time functions can be found in Appendix B.8. The key lemma is an upper bound of the amortized complexity of \( merge \) in terms of \( lrh \) and \( rlh \):
Lemma 22.1. \( T_{\text{merge}} t_1 t_2 + \Phi (\text{merge} t_1 t_2) - \Phi t_1 - \Phi t_2 \leq lrh (\text{merge} t_1 t_2) + rlh t_1 + rlh t_2 + 1 \)

**Proof** by induction on the computation of \( \text{merge} \). We consider only the node-node case: let \( t_1 = (l_1, a_1, r_1) \) and \( t_2 = (l_2, a_2, r_2) \). W.l.o.g. assume \( a_1 \leq a_2 \). Let \( m = \text{merge} t_2 r_1 \).

\[
T_{\text{merge}} t_1 t_2 + \Phi (\text{merge} t_1 t_2) - \Phi t_1 - \Phi t_2 \\
= T_{\text{merge}} t_2 r_1 + 1 + \Phi m + \Phi l_1 + rh m l_1 - \Phi t_1 - \Phi t_2 \\
= T_{\text{merge}} t_2 r_1 + 1 + \Phi m + rh m l_1 - rh l_1 r_1 - \Phi t_2 \\
\leq lrh m + rlh t_2 + rlh r_1 + rh m l_1 + 2 - rh l_1 r_1 \quad \text{by IH} \\
= lrh m + rlh t_2 + rlh t_1 + rh m l_1 + 1 \\
= lrh (\text{merge} t_1 t_2) + rlh t_1 + rlh t_2 + 1 \quad \square
\]

As a consequence we can prove the following logarithmic upper bound on the amortized complexity of \( \text{merge} \):

\[
T_{\text{merge}} t_1 t_2 + \Phi (\text{merge} t_1 t_2) - \Phi t_1 - \Phi t_2 \\
\leq lrh (\text{merge} t_1 t_2) + rlh t_1 + rlh t_2 + 1 \quad \text{by Lemma 22.1} \\
\leq \lg |\text{merge} t_1 t_2| + \lg |t_1| + \lg |t_2| + 1 \quad \text{by (22.1)} \\
\leq \lg (|t_1|_1 + |t_2|_1 - 1) + \lg |t_1|_1 + \lg |t_2|_1 + 1 \\
\quad \text{because } |\text{merge} t_1 t_2| = |t_1| + |t_2| \\
\leq 3 \cdot \lg (|t_1|_1 + |t_2|_1) + 2 \cdot \lg (|t_1|_1 + |t_2|_1) + 1 \\
= 3 \cdot \lg (|t_1|_1 + |t_2|_1) + 1
\]

The amortized complexity of insertion and deletion follows easily from the complexity of \( \text{merge} \):

\[
T_{\text{insert} \ a \ t} + \Phi (\text{insert} \ a \ t) - \Phi t \leq 3 \cdot \lg (|t|_1 + 2) + 2 \\
T_{\text{del\_min} \ t} + \Phi (\text{del\_min} \ t) - \Phi t \leq 3 \cdot \lg (|t|_1 + 2) + 2
\]

**Bibliographic Remarks**

Skew heaps were invented by Sleator and Tarjan [71] as one of the first self-organizing data structures. Their presentation was imperative. Our presentation follows earlier work by Nipkow [53, 55] based on the functional account by Kaldewaij and Schoenmakers [37].
Pairing Heaps

The pairing heap is another form of a self-adjusting priority queue. Section 23.1 presents an intuitive version of pairing heaps based on lists. In the rest of the chapter we change to a slightly different presentation that leads to a more succinct amortized analysis.

23.1 Implementation via Lists

A pairing heap is a heap in the sense that it is a tree with the minimal element at the root — except that it is not a binary tree but a tree where each node has a list of children:

```
datatype 'a heap = Empty | Hp 'a ('a heap list)
```

The abstraction function to multisets and the invariant follow the heap paradigm:

```
  mset_heap :: 'a heap ⇒ 'a multiset
  mset_heap Empty = { }
  mset_heap (Hp x hs) = { x } + \sum_{h \in mset_heap hs} (mset (map mset_heap hs))

  pheap :: 'a heap ⇒ bool
  pheap Empty = True
  pheap (Hp x hs) = (\forall h \in set hs. (\forall y \in mset_heap h. x < y) \land pheap h)
```

Note that pheap is not the full invariant. Moreover, Empty does not occur inside a non-empty heap.
The implementations of empty and get_min are obvious, and insert follows the standard heap paradigm:

\[
\begin{align*}
\text{empty} & = \text{Empty} \\
\text{get}_\text{min} :: \text{a heap} \Rightarrow \text{a} \\
\text{get}_\text{min} (\text{Hp} \; x \; _) & = x \\
\text{insert} :: \text{a} \Rightarrow \text{a heap} \Rightarrow \text{a heap} \\
\text{insert} \; x \; h & = \text{merge} (\text{Hp} \; x \; []) \; h \\
\end{align*}
\]

Function merge is not recursive (as in binary heaps) but simply adds one of the two heaps to the front of the top-level heaps of the other, depending on the root value:

\[
\begin{align*}
\text{merge} :: \text{a heap} \Rightarrow \text{a heap} \Rightarrow \text{a heap} \\
\text{merge} \; h \; \text{Empty} & = h \\
\text{merge} \; \text{Empty} \; h & = h \\
\text{merge} (\text{Hp} \; x \; \text{hsx} \; =: \; \text{hx}) \; (\text{Hp} \; y \; \text{hsy} \; =: \; \text{hy}) & = (\text{if} \; x < y \; \text{then} \; \text{Hp} \; x \; (\text{hy} \; \# \; \text{hsx}) \; \text{else} \; \text{Hp} \; y \; (\text{hx} \; \# \; \text{hsy})) \\
\end{align*}
\]

Thus merge and insert have constant running time. All the work is offloaded on del_min which just calls merge_pairs:

\[
\begin{align*}
\text{del}_\text{min} :: \text{a heap} \Rightarrow \text{a heap} \\
\text{del}_\text{min} \; \text{Empty} & = \text{Empty} \\
\text{del}_\text{min} (\text{Hp} \; _ \; \text{hs}) & = \text{merge}_\text{pairs} \; \text{hs} \\
\text{merge}_\text{pairs} :: \text{a heap list} \Rightarrow \text{a heap} \\
\text{merge}_\text{pairs} \; [] & = \text{Empty} \\
\text{merge}_\text{pairs} \; [h] & = h \\
\text{merge}_\text{pairs} \; (h_1 \; \# \; h_2 \; \# \; \text{hs}) & = \text{merge} (\text{merge} \; h_1 \; h_2) \; (\text{merge}_\text{pairs} \; \text{hs}) \\
\end{align*}
\]

Function merge_pairs is a compact way of expressing a two pass algorithm: on the first pass from left to right, it merges pairs of adjacent heaps (hence “pairing heap”) and on the second pass it merges the results in a cascade from right to left. By reformulating the definition in terms of these two passes, we obtain a more readable formulation with the same running time:
del_min :: 'a heap ⇒ 'a heap
del_min Empty = Empty
del_min (Hp x hs) = pass2 (pass1 hs)

pass1 :: 'a heap list ⇒ 'a heap list
pass1 (h1 # h2 # hs) = merge h1 h2 # pass1 hs
pass1 hs = hs

pass2 :: 'a heap ⇒ 'a heap
pass2 [] = Empty
pass2 (h # hs) = merge h (pass2 hs)

The proof of pass2 (pass1 hs) = merge_pairs hs is an easy induction.

Clearly del_min can take linear time but it will turn out that the constant-time insert saves enough to guarantee amortized logarithmic complexity for both insertion and deletion.

We base the correctness proofs on the merge_pairs version of del_min.

From the following lemmas (all proofs are routine inductions)

\[ h \neq \text{Empty} \rightarrow \text{get}_\min h \in h \text{ mset}_\heap h \]
\[ h \neq \text{Empty} \land \text{pheap} h \land x \in h \text{ mset}_\heap h \rightarrow \text{get}_\min h \leq x \]
\[ \text{mset}_\heap (\text{merge} h_1 h_2) = \text{mset}_\heap h_1 + \text{mset}_\heap h_2 \]
\[ \text{mset}_\heap (\text{merge}_\text{pairs} hs) = \sum (\text{image}_\text{mset} \text{mset}_\heap (\text{mset} hs)) \]
\[ h \neq \text{Empty} \rightarrow \text{mset}_\heap (\text{del}_\min h) = \text{mset}_\heap h - \{ \text{get}_\min h \} \]
\[ \text{pheap} h_1 \land \text{pheap} h_2 \rightarrow \text{pheap} (\text{merge} h_1 h_2) \]
\[ (\forall h \in \text{set} hs. \text{pheap} h) \rightarrow \text{pheap} (\text{merge}_\text{pairs} hs) \]
\[ \text{pheap} h \rightarrow \text{pheap} (\text{del}_\min h) \]

the properties in the specifications Priority_Queue (_Merge) follow easily.

23.2 Amortized Analysis

The amortized analysis of pairing heaps is slightly simplified if we replace the above type of heaps by trees as follows: a heap Hp x hs is expressed as the tree ⟨hs, x, ⟨⟩⟩ and a list of heaps [Hp x1 hs1, Hp x2 hs2, ...] is expressed.
empty = ()

get_min :: 'a tree ⇒ 'a
get_min (_, x, _) = x

link :: 'a tree ⇒ 'a tree
link (hsx, x, {hsy, y, hs}) = (if x < y then (hsy, y, hs) else (hsx, x, hsy, y, hs))
link hp = hp

pass1 :: 'a tree ⇒ 'a tree
pass1 (hsz, x, {hsy, y, hs}) = link (hsz, x, {hsy, y, pass1 hs})
pass1 hp = hp

pass2 :: 'a tree ⇒ 'a tree
pass2 (hsz, x, hs) = link (hsz, x, pass2 hs)
pass2 () = ()

get_min :: 'a tree ⇒ 'a
get_min (_, x, _) = x

merge :: 'a tree ⇒ 'a tree ⇒ 'a tree
merge () hp = hp
merge hp () = hp
merge (hsx, x, {hsy, y, ()}) = link (hsx, x, {hsy, y, ()})

insert :: 'a ⇒ 'a tree ⇒ 'a tree
insert x hp = merge () x () hp

Fig. 23.1. Pairing heaps via trees

as the tree (hs1, x1, (hs2, x2, ...)...). This simplifies the analysis because we now have to deal only with a single type, trees.

The code for the tree representation of pairing heaps is shown in Figure 23.1. We work with the pass1/pass2 version of del_min. The correctness proof is very similar to what we saw in the previous section. We merely display the two invariants:

is_root :: 'a tree ⇒ bool
is_root hp = (case hp of () ⇒ True | (_, _, r) ⇒ r = ())
23.2 Amortized Analysis

```plaintext
pheap :: 'a tree ⇒ bool
pheap () = True
pheap (l, x, r) = ((∀ y ∈ set_tree l. x ≤ y) ∧ pheap l ∧ pheap r)
```

Now we turn to the amortized analysis. The potential of a tree is the sum of the logarithms of the sizes of the subtrees:

```plaintext
Φ :: 'a tree ⇒ real
Φ () = 0
Φ (l, x, r) = lg |⟨l, x, r⟩| + Φ l + Φ r
```

These easy inductive size properties are frequently used implicitly below:

- |link hp| = |hp|
- |pass1 hp| = |hp|
- |pass2 hp| = |hp|
- is_root h1 ∧ is_root h2 → |merge h1 h2| = |h1| + |h2|

23.2.1 Potential Differences

We can now analyze the differences in potential caused by all the queue operations. In a separate step we will derive their amortized complexities.

For insertion, the following upper bound follows trivially from the definitions:

**Lemma 23.1.** is_root hp → Φ (insert x hp) − Φ hp ≤ lg (|hp| + 1)

For merge it needs a bit more work:

**Lemma 23.2.** h1 = ⟨hs1, x1, ⟩) ∧ h2 = ⟨hs2, x2, ⟩ →

Φ (merge h1 h2) − Φ h1 − Φ h2 ≤ lg (|h1| + |h2|) + 1

**Proof.** From

Φ (merge h1 h2)
= Φ (link ⟨hs1, x1, h2⟩)
= Φ hs1 + Φ hs2 + lg (|hs1| + |hs2| + 1) + lg (|hs1| + |hs2| + 2)
= Φ hs1 + Φ hs2 + lg (|hs1| + |hs2| + 1) + lg (|h1| + |h2|)

it follows that

Φ (merge h1 h2) − Φ h1 − Φ h2
= lg (|hs1| + |hs2| + 1) + lg (|h1| + |h2|)
− lg (|h1| + 1) − lg (|hs2| + 1)
\[ \leq \lg (|h_1| + |h_2|) + 1 \]

because \( \lg (1 + x + y) \leq 1 + \lg (1 + x) + \lg (1 + y) \) if \( x, y \geq 0 \)

Now we come to the core of the proof, the analysis of \( \text{del}_\text{min} \). Its running time is linear in the number of nodes reachable by descending to the right (starting from the left child of the root). We denote this metric by \( \text{len} \):

\[
\text{len} :: \text{a tree} \Rightarrow \text{nat}
\]

\[
\text{len } \langle \rangle = 0
\]

\[
\text{len } \langle _, _, r \rangle = 1 + \text{len } r
\]

Therefore we have to show that the potential change compensates for this linear work. Our main goal is this:

**Theorem 23.3.**

\[
\Phi (\text{del}_\text{min} \langle hs, x, \langle \rangle \rangle) - \Phi (hs, x, \langle \rangle) \leq 2 \cdot \lg (|hs| + 1) - \text{len } hs + 2
\]

It will be proved in two steps: First we show that \( \text{pass}_1 \) frees enough potential to compensate for the work linear in \( \text{len } hs \) and increases the potential only by a logarithmic term. Then we show that the increase due to \( \text{pass}_2 \) is also only at most logarithmic. Combining these results one easily shows that the amortized running time of \( \text{del}_\text{min} \) is indeed logarithmic.

First we analyze the potential difference caused by \( \text{pass}_1 \):

**Lemma 23.4.**

\[
\Phi (\text{pass}_1 hs) - \Phi hs \leq 2 \cdot \lg (|hs| + 1) - \text{len } hs + 2
\]

*Proof* by induction on the computation of \( \text{pass}_1 \). The base cases are trivial. We focus on the induction step. Let \( t = \langle hs_1, x, \langle hs_2, y, hs \rangle \rangle, n_1 = |hs_1|, n_2 = |hs_2| \) and \( m = |hs| \).

\[
\Phi (\text{pass}_1 t) - \Phi t
= \lg (n_1 + n_2 + 1) - \lg (n_2 + m + 1) - \Phi (\text{pass}_1 hs) - \Phi hs
\leq \lg (n_1 + n_2 + 1) - \lg (n_2 + m + 1) + 2 \cdot \lg (m + 1) \quad \text{by IH}
\leq 2 \cdot \lg (n_1 + n_2 + m + 1) - \lg (n_2 + m + 1) + \lg (m + 1)
- \text{len } hs \quad \text{because } \lg x + \lg y + 2 \leq 2 \cdot \lg (x + y) \text{ if } x, y > 0
\leq 2 \cdot \lg (n_1 + n_2 + m + 2) - \text{len } hs
= 2 \cdot \lg |t| - \text{len } t + 2
\leq 2 \cdot \lg (|t| + 1) - \text{len } t + 2 \quad \blacksquare
\]

Now we turn to \( \text{pass}_2 \):

**Lemma 23.5.** \( hs \neq \langle \rangle \rightarrow \Phi (\text{pass}_2 hs) - \Phi hs \leq \lg |hs| \)
23.2 Amortized Analysis

Proof by induction on $h$s. The base cases are trivial. The induction step (for $h s_1, x, h s$) is trivial if $h s = \langle \rangle$. Assume $h s = \langle h s_2, y, r \rangle$. Now we need one more property of $pass_2$:

$\exists h s_3 \ z. pass_2 \langle h s_2, y, r \rangle = \langle h s_3, \langle \rangle \rangle$

The proof is a straightforward induction on $r$. This implies $|h s_3| + 1 = |h s|$ and thus

$$
\Phi (\text{link} \langle h s_1, x, pass_2 h s \rangle) - \Phi h s_1 - \Phi (pass_2 h s) = \lg (|h s_1| + |h s| + 1) + \lg (|h s_1| + |h s|) - \lg |h s|
$$

Thus the overall claim follows:

$$
\Phi (pass_2 \langle h s_1, x, h s \rangle) - \Phi \langle h s_1, x, h s \rangle = \Phi (\text{link} \langle h s_1, x, pass_2 h s \rangle) - \Phi h s_1 - \Phi h s - \lg (|h s_1| + |h s| + 1) = \Phi (pass_2 h s) - \Phi h s + \lg (|h s_1| + |h s|) - \lg |h s| < \lg (|h s_1| + |h s|) \quad \text{by (\ast)}
$$

Corollary 23.6. $\Phi (pass_2 h s) - \Phi h s \leq \lg (|h s| + 1)$

Finally we can prove Theorem 23.3:

$$
\Phi (\text{del_min} \langle h s, x, \langle \rangle \rangle) - \Phi \langle h s, x, \langle \rangle \rangle = \Phi (pass_2 \langle pass_1 h s \rangle) - \lg (|h s| + 1) - \Phi h s \leq \Phi (pass_1 h s) - \Phi h s \quad \text{by Corollary 23.6}
$$

$$
\leq 2 \cdot \lg (|h s| + 1) - \text{len} h s + 2 \quad \text{by Lemma 23.4}
$$

23.2.2 Amortized Running Times

The canonical running time functions are displayed in Appendix B.9. It is now straightforward to derive these amortized running times:

$$
is\_root h \longrightarrow T_{\text{insert}} a h + \Phi (\text{insert} a h) - \Phi h \leq \lg (|h| + 1) + 1
$$
$$
is\_root h_1 \land is\_root h_2 \longrightarrow T_{\text{merge}} h_1 h_2 + \Phi (\text{merge} h_1 h_2) - \Phi h_1 - \Phi h_2 \leq \lg (|h_1| + |h_2| + 1) + 2
$$

They follow from the corresponding Lemmas 23.1 and 23.2.

Combining this inductive upper bound for the running time of the two passes

$$
T_{\text{pass}} \langle pass_1 h s_1 \rangle + T_{\text{pass}} \langle h s_1 \rangle \leq \text{len} h s_1 + 2
$$

with Theorem 23.3 yields the third and final amortized running time:

$$
is\_root h \longrightarrow T_{\text{del_min}} h + \Phi (\text{del_min} h) - \Phi h \leq 2 \cdot \lg (|h| + 1) + 5
$$
Thus we could prove that insertion, merging and deletion all have amortized logarithmic running times.

**Bibliographic Remarks**

Pairing heaps were invented by Fredman et al. [21] as a simpler but competitive alternative to Fibonacci heaps. The authors gave the amortized analysis presented above and conjectured that it can be improved. Later research confirmed this [33, 65, 34] but the final analysis is still open. An empirical study [47] showed that pairing heaps do indeed outperform Fibonacci heaps in practice. This chapter is based on an article by Nipkow and Brinkop [55].
Part V

Appendix
List Library

The following functions on lists are predefined:

\[
\begin{align*}
\text{length} & : \forall a \, \text{list} \Rightarrow \text{nat} \\
\left[\begin{array}{l}
\text{length} \; \text{of} \\
\end{array}\right] &= 0 \\
\text{length} \; \text{of} \; x \; \text{at} \; xs &= \text{length} \; \text{of} \; xs + 1 \\
\text{(\@)} & : \forall a \, \text{list} \Rightarrow \forall a \, \text{list} \Rightarrow \forall a \, \text{list} \\
\left[\begin{array}{l}
\text{\@} \; \text{of} \\
\end{array}\right] &= \text{\@} \; \text{of} \\
\left[\begin{array}{l}
\text{\@} \; \text{of} \; x \; \text{at} \; xs @ \; ys \\
\end{array}\right] &= x \; \text{at} \; xs @ \; ys \\
\text{set} & : \forall a \, \text{list} \Rightarrow \forall a \, \text{set} \\
\left[\begin{array}{l}
\text{set} \; \text{of} \\
\end{array}\right] &= \{\} \\
\text{set} \; \text{of} \; x \; \text{at} \; xs &= \{x\} \; \cup \; \text{set} \; xs \\
\text{map} & : \forall a \Rightarrow \forall b \Rightarrow \forall a \, \text{list} \Rightarrow \forall b \, \text{list} \\
\left[\begin{array}{l}
\text{map} \; \text{of} \\
\end{array}\right] &= \{} \\
\text{map} \; \text{of} \; f \; \text{at} \; x \; \text{at} \; xs &= f \; x \; \text{at} \; \text{map} \; \text{of} \; f \; \text{at} \; xs \\
\text{filter} & : \forall a \Rightarrow \forall \text{bool} \Rightarrow \forall a \, \text{list} \Rightarrow \forall a \, \text{list} \\
\left[\begin{array}{l}
\text{filter} \; \text{of} \\
\end{array}\right] &= \{} \\
\text{filter} \; \text{of} \; p \; \text{at} \; x \; \text{at} \; xs &= \text{if} \; p \; x \; \text{then} \; x \; \text{at} \; \text{filter} \; \text{of} \; p \; \text{at} \; x \; \text{at} \; xs \; \text{else} \; \text{filter} \; \text{of} \; p \; \text{at} \; xs \\
\text{concat} & : \forall a \, \text{list} \; \text{at} \; \text{list} \Rightarrow \forall a \, \text{list} \\
\left[\begin{array}{l}
\text{concat} \; \text{of} \\
\end{array}\right] &= \{} \\
\text{concat} \; \text{of} \; x \; \text{at} \; xs &= x \; @ \; \text{concat} \; \text{of} \; xs \\
\end{align*}
\]
take :: nat \Rightarrow 'a list \Rightarrow 'a list
\begin{align*}
    \text{take } _0 \, \text{[]} &= \text{[]} \\
    \text{take } n \, (x \neq xs) &= (\text{case } n \text{ of } 0 \Rightarrow \text{[]} | m + 1 \Rightarrow x \neq \text{take } m \, xs)
\end{align*}

drop :: nat \Rightarrow 'a list \Rightarrow 'a list
\begin{align*}
    \text{drop } _0 \, \text{[]} &= \text{[]} \\
    \text{drop } n \, (x \neq xs) &= (\text{case } n \text{ of } 0 \Rightarrow x \neq xs | m + 1 \Rightarrow \text{drop } m \, xs)
\end{align*}

hd :: 'a list \Rightarrow 'a
hd (x \neq xs) = x

tl :: 'a list \Rightarrow 'a list
tl _0 \, \text{[]} = \text{[]} \\
    \text{tl } (x \neq xs) = xs

butlast :: 'a list \Rightarrow 'a list
\begin{align*}
    \text{butlast } _0 \, \text{[]} &= \text{[]} \\
    \text{butlast } (x \neq xs) &= (\text{if } xs = \text{[]} \text{ then } \text{[]} \text{ else } x \neq \text{butlast } xs)
\end{align*}

rev :: 'a list \Rightarrow 'a list
\begin{align*}
    \text{rev } _0 \, \text{[]} &= \text{[]} \\
    \text{rev } (x \neq xs) &= \text{rev } xs \, @ \, [x]
\end{align*}

(!) :: 'a list \Rightarrow nat \Rightarrow 'a
\begin{align*}
    (x \neq xs) ! \, n &= (\text{case } n \text{ of } 0 \Rightarrow x | k + 1 \Rightarrow xs ! k)
\end{align*}

list_update :: 'a list \Rightarrow nat \Rightarrow 'a \Rightarrow 'a list
\begin{align*}
    \text{[[}_0 \, : = \text{]} &= \text{[]} \\
    \text{(x \neq xs)}_[i := v] &= (\text{case } i \text{ of } 0 \Rightarrow v \neq xs | j + 1 \Rightarrow x \neq xs[j := v])
\end{align*}

upt :: nat \Rightarrow nat \Rightarrow nat list
\begin{align*}
    [\ldots<0] &= \text{[]} \\
    [i..<j + 1] &= (\text{if } i \leq j \text{ then } [i..<j] \, @ \, [j] \text{ else } \text{[]})
\end{align*}

replicate :: nat \Rightarrow 'a \Rightarrow 'a list
\begin{align*}
    \text{replicate } 0 \, _0 &= \text{[]} \\
    \text{replicate } (n + 1) \, x &= x \neq \text{replicate } n \, x
\end{align*}
\begin{verbatim}
sum_list :: 'a list \rightarrow 'a
sum_list [] = 0
sum_list (x # xs) = x + sum_list xs

min_list :: 'a list \rightarrow 'a
min_list (x # xs)
  = (case xs of [] \Rightarrow x | _ # _ \Rightarrow min x (min_list xs))

sorted_wrt :: ('a \Rightarrow 'a \Rightarrow bool) \Rightarrow 'a list \Rightarrow bool
sorted_wrt P [] = True
sorted_wrt P (x # ys) = ((\forall y \in set ys. P x y) \land sorted_wrt P ys)
\end{verbatim}
B

Time Functions

B.1 Lists

\[
T_{\text{length}} :: \text{\textquoteleft a list} \Rightarrow \text{nat}
\]
\[
T_{\text{length}} \ [\ ] = 1
\]
\[
T_{\text{length}} \ [\ _\#\ xs) = T_{\text{length}}\ xs + 1
\]
\[
T_{\text{map}} :: (\text{\textquoteleft a \Rightarrow nat} \Rightarrow \text{\textquoteleft a list} \Rightarrow \text{nat})
\]
\[
T_{\text{map}} \ [\ ] = 1
\]
\[
T_{\text{map}} \ T_f (x \#\ xs) = T_f\ x + T_{\text{map}}\ T_f\ xs + 1
\]
\[
T_{\text{filter}} :: (\text{\textquoteleft a \Rightarrow nat} \Rightarrow \text{\textquoteleft a list} \Rightarrow \text{nat})
\]
\[
T_{\text{filter}} \ [\ ] = 1
\]
\[
T_{\text{filter}}\ T_p (x \#\ xs) = T_p\ x + T_{\text{filter}}\ T_p\ xs + 1
\]
\[
T_{\text{take}} :: \text{nat} \Rightarrow \text{\textquoteleft a list} \Rightarrow \text{nat}
\]
\[
T_{\text{take}} \ [\ ] = 1
\]
\[
T_{\text{take}}\ n (\ _\#\ xs) = \text{(case n of 0 \Rightarrow 1 \mid n' + 1 \Rightarrow T_{\text{take}}\ n'\ xs + 1)}
\]
\[
T_{\text{drop}} :: \text{nat} \Rightarrow \text{\textquoteleft a list} \Rightarrow \text{nat}
\]
\[
T_{\text{drop}} \ [\ ] = 1
\]
\[
T_{\text{drop}}\ n (\ _\#\ xs) = \text{(case n of 0 \Rightarrow 1 \mid n' + 1 \Rightarrow T_{\text{drop}}\ n'\ xs + 1)}
\]

\[
T_{\text{length}}\ xs = |xs| + 1
\]
\[
T_{\text{map}}\ T_f\ xs = (\sum_{x\in\ xs} T_f\ x) + |xs| + 1
\]
\[
T_{\text{filter}}\ T_p\ xs = (\sum_{x\in\ xs} T_p\ x) + |xs| + 1
\]
\[
T_{\text{take}}\ n\ xs = \text{min} n\ |xs| + 1
\]
\[
T_{\text{drop}}\ n\ xs = \text{min} n\ |xs| + 1
\]
B.2 Selection

\[ T_{chop} :: \text{nat} \Rightarrow \text{'a list} \Rightarrow \text{nat} \]
\[ T_{chop} 0 = 1 \]
\[ T_{chop} [\ ] = 1 \]
\[ T_{chop} n \ xs = T_{take} n \ xs + T_{drop} n \ xs + T_{chop} n \ (\text{drop} n \ xs) \]

\[ T_{partition3} :: \text{a} \Rightarrow \text{'a list} \Rightarrow \text{nat} \]
\[ T_{partition3} [\ ] = 1 \]
\[ T_{partition3} x (\_ \# \ ys) = T_{partition3} x \ ys + 1 \]

\[ T_{slow\_select} :: \text{nat} \Rightarrow \text{'a list} \Rightarrow \text{nat} \]
\[ T_{slow\_select} k \ xs = T_{insert} xs + T_{nth} (\text{insert} \ xs) \ k + 1 \]

\[ T_{slow\_median} :: \text{'a list} \Rightarrow \text{nat} \]
\[ T_{slow\_median} \ xs = T_{slow\_select} ((|xs| - 1) \div 2) \ xs + 1 \]

\[ T_{chop} d \ xs \leq 5 \cdot |xs| + 1 \]
\[ T_{partition3} x \ xs = |xs| + 1 \]
\[ T_{slow\_select} k \ xs \leq |xs|^2 + 3 \cdot |xs| + 3 \]
\[ T_{slow\_median} \ xs \leq |xs|^2 + 3 \cdot |xs| + 4 \]

B.3 2-3 Trees

\[ T_{join\_adj} :: \text{'a tree23s} \Rightarrow \text{nat} \]
\[ T_{join\_adj} (TTs \ _ \ _ \ (T \ _)) = 1 \]
\[ T_{join\_adj} (TTs \ _ \ _ \ (TTs \ _ \ _ \ (T \ _))) = 1 \]
\[ T_{join\_adj} (TTs \ _ \ _ \ (TTs \ _ \ _ \ ts)) = T_{join\_adj} ts + 1 \]

\[ T_{join\_all} :: \text{'a tree23s} \Rightarrow \text{nat} \]
\[ T_{join\_all} (T \ _ ) = 1 \]
\[ T_{join\_all} \ ts = T_{join\_adj} ts + T_{join\_all} (\text{join\_adj} ts) + 1 \]

\[ T_{tree23\_of\_list} :: \text{'a list} \Rightarrow \text{nat} \]
\[ T_{tree23\_of\_list} \ as = T_{leaves} as + T_{join\_all} (leaves as) + 1 \]
B.4 Leftist Heaps

\[
T_{\text{merge}} :: ('a \times \text{nat}) \text{tree} \Rightarrow ('a \times \text{nat}) \text{tree} \Rightarrow \text{nat}
\]
\[
T_{\text{merge}} \langle \rangle \ t = 1
\]
\[
T_{\text{merge}} \ t \langle \rangle = 1
\]
\[
T_{\text{merge}} ((t_1, (a_1, n_1), r_1) =: t_1) ((t_2, (a_2, n_2), r_2) =: t_2) \\
= (\text{if } a_1 \leq a_2 \text{ then } T_{\text{merge}} r_1 \ t_2 \text{ else } T_{\text{merge}} t_1 \ r_2) + 1
\]

\[
T_{\text{insert}} :: 'a \Rightarrow ('a \times \text{nat}) \text{tree} \Rightarrow \text{nat}
\]
\[
T_{\text{insert}} \ x \ t = T_{\text{merge}} \langle \rangle, (x, 1), \langle \rangle \ t + 1
\]

\[
T_{\text{del_min}} :: ('a \times \text{nat}) \text{tree} \Rightarrow \text{nat}
\]
\[
T_{\text{del_min}} \langle \rangle = 1
\]
\[
T_{\text{del_min}} \langle l, _, r \rangle = T_{\text{merge}} \ l \ r + 1
\]

B.5 Binomial Heaps

\[
T_{\text{link}} :: 'a \text{tree} \Rightarrow 'a \text{tree} \Rightarrow \text{nat}
\]
\[
T_{\text{link}} _ _ = 1
\]

\[
T_{\text{ins_tree}} :: 'a \text{tree} \Rightarrow 'a \text{tree list} \Rightarrow \text{nat}
\]
\[
T_{\text{ins_tree}} _ [] = 1
\]
\[
T_{\text{ins_tree}} \ t_1 \ (t_2 \ # \ ts) \ = (\text{if } \text{rank} \ t_1 < \text{rank} \ t_2 \text{ then } 1 \text{ else } T_{\text{link}} t_1 \ t_2 + T_{\text{ins_tree}} \ (\text{link} t_1 \ t_2) \ ts)
\]

\[
T_{\text{insert}} :: 'a \Rightarrow 'a \text{tree list} \Rightarrow \text{nat}
\]
\[
T_{\text{insert}} \ x \ ts = T_{\text{ins_tree}} \ (\text{Node} 0 \ x \ []) \ ts + 1
\]

\[
T_{\text{merge}} :: 'a \text{tree list} \Rightarrow 'a \text{tree list} \Rightarrow \text{nat}
\]
\[
T_{\text{merge}} _ [] = 1
\]
\[
T_{\text{merge}} \ [] \ (_ \ # \ _) = 1
\]
\[
T_{\text{merge}} \ (t_1 \ # \ ts_1 =: h_1) \ (t_2 \ # \ ts_2 =: h_2) \\
= 1 + (\text{if } \text{rank} \ t_1 < \text{rank} \ t_2 \text{ then } T_{\text{merge}} ts_1 \ h_2)
\]
else if \( \text{rank } t_2 < \text{rank } t_1 \) then \( T_{\text{merge}} h_1 t s_2 \)

else \( T_{\text{ins-tree}}(\text{link } t_1 t_2)(\text{merge } ts_1 t s_2) + T_{\text{merge }} ts_1 t s_2) \)

\( T_{\text{get-min}} :: \text{a tree list } \Rightarrow \text{nat} \)
\( T_{\text{get-min}}[\_] = 1 \)
\( T_{\text{get-min}}(\_ \# v \# va) = 1 + T_{\text{get-min}}(v \# va) \)

\( T_{\text{get-min-rest}} :: \text{a tree list } \Rightarrow \text{nat} \)
\( T_{\text{get-min-rest}}[\_] = 1 \)
\( T_{\text{get-min-rest}}(\_ \# v \# va) = 1 + T_{\text{get-min-rest}}(v \# va) \)

\( T_{\text{rev}} :: \text{a list } \Rightarrow \text{nat} \)
\( T_{\text{rev}} xs = |xs| + 1 \)

\( T_{\text{del-min}} :: \text{a tree list } \Rightarrow \text{nat} \)
\( T_{\text{del-min}} ts = T_{\text{get-min-rest }} ts + \)
\( (\text{case } T_{\text{get-min-rest}} ts \text{ of} \)
\( (\text{Node } _ _ ts_1, ts_2) \Rightarrow T_{\text{rev}} ts_1 + T_{\text{merge}} (\text{rev } ts_1) ts_2) + 1 \)

B.6 Queues

\( T_{\text{norm}} :: \text{a list } \times \text{a list } \Rightarrow \text{nat} \)
\( T_{\text{norm}}(fs, rs) = (\text{if } fs = [] \text{ then } T_{\text{itrev}} rs [] \text{ else } 0) + 1 \)

\( T_{\text{enq}} :: \text{a } \Rightarrow \text{a list } \times \text{a list } \Rightarrow \text{nat} \)
\( T_{\text{enq}} a (fs, rs) = T_{\text{norm}}(fs, a \# rs) + 1 \)

\( T_{\text{deq}} :: \text{a list } \times \text{a list } \Rightarrow \text{nat} \)
\( T_{\text{deq}} (fs, rs) = (\text{if } fs = [] \text{ then } 0 \text{ else } T_{\text{norm}}(\text{tl } fs, rs)) + 1 \)

\( T_{\text{first}} :: \text{a list } \times \text{a list } \Rightarrow \text{nat} \)
\( T_{\text{first}}(\_ \# \_, \_) = 1 \)
B.7 Splay Trees

\[ T_{\text{is empty}} : \text{'a list} \times \text{'a list} \Rightarrow \text{nat} \]
\[ T_{\text{is empty}} (\_, \_) = 1 \]

\[ T_{\text{splay}} : \text{'a} \Rightarrow \text{'a tree} \Rightarrow \text{nat} \]
\[ T_{\text{splay}} \emptyset = 1 \]
\[ T_{\text{splay}} x (AB, b, CD) = (\text{case} \ \text{cmp} x b \ \text{of} \]
\[ LT \Rightarrow \text{case} AB \ \text{of} \]
\[ \emptyset \Rightarrow 1 \mid \]
\[ \langle A, a, B \rangle \Rightarrow \text{case} \ \text{cmp} x a \ \text{of} \]
\[ LT \Rightarrow \text{if} A = \emptyset \ \text{then} 1 \ \text{else} \ T_{\text{splay}} x A + 1 \mid \]
\[ EQ \Rightarrow 1 \mid \]
\[ GT \Rightarrow \text{if} B = \emptyset \ \text{then} 1 \ \text{else} \ T_{\text{splay}} x B + 1 \mid \]
\[ EQ \Rightarrow 1 \mid \]
\[ GT \Rightarrow \text{case} CD \ \text{of} \]
\[ \emptyset \Rightarrow 1 \mid \]
\[ \langle C, c, D \rangle \Rightarrow \text{case} \ \text{cmp} x c \ \text{of} \]
\[ LT \Rightarrow \text{if} C = \emptyset \ \text{then} 1 \ \text{else} \ T_{\text{splay}} x C + 1 \mid \]
\[ EQ \Rightarrow 1 \mid \]
\[ GT \Rightarrow \text{if} D = \emptyset \ \text{then} 1 \ \text{else} \ T_{\text{splay}} x D + 1 \]
else $T_{splay} x t +$
\begin{align*}
  & (\text{case } splay x t \text{ of} \\
  & \langle l, a, _ \rangle \Rightarrow \\
  & \quad \text{if } x \neq a \text{ then } 0 \text{ else if } l = \langle \rangle \text{ then } 0 \text{ else } T_{splay\_\max} l))
\end{align*}

### B.8 Skew Heaps

\[
T_{merge} :: 'a\ tree \Rightarrow 'a\ tree \Rightarrow \text{nat}
\]
\[
T_{merge} \langle \rangle \_ = 1
\]
\[
T_{merge} \_ \langle \rangle = 1
\]
\[
T_{merge} \langle l_1, a_1, r_1 \rangle \langle l_2, a_2, r_2 \rangle = (\text{if } a_1 \leq a_2 \text{ then } T_{merge} \langle l_2, a_2, r_2 \rangle r_1 \text{ else } T_{merge} \langle l_1, a_1, r_1 \rangle r_2) + 1
\]
\[
T_{insert} :: 'a \Rightarrow 'a\ tree \Rightarrow \text{int}
\]
\[
T_{insert} a t = T_{merge} \langle \langle \rangle, a, \langle \rangle \rangle t + 1
\]
\[
T_{del\_min} :: 'a\ tree \Rightarrow \text{int}
\]
\[
T_{del\_min} t = (\text{case } t \text{ of} \langle \rangle \Rightarrow 1 | \langle t_1, \_ , t_2 \rangle \Rightarrow T_{merge} t_1 t_2 + 1)
\]

### B.9 Pairing Heaps

\[
T_{insert} :: 'a \Rightarrow 'a\ tree \Rightarrow \text{nat}
\]
\[
T_{insert} \_ \_ = 1
\]
\[
T_{merge} :: 'a\ tree \Rightarrow 'a\ tree \Rightarrow \text{nat}
\]
\[
T_{merge} \_ \_ = 1
\]
\[
T_{del\_min} :: 'a\ tree \Rightarrow \text{nat}
\]
\[
T_{del\_min} \langle \rangle = 1
\]
\[
T_{del\_min} \langle hs, \_ , \_ \rangle = T_{pass2} (pass_1 hs) + T_{pass1} hs + 1
\]
\[
T_{pass1} :: 'a\ tree \Rightarrow \text{nat}
\]
\[ T_{\text{pass1}} (\langle \_ , \_ , \langle \_ , \_ , hs' \rangle \rangle) = T_{\text{pass1}} hs' + 1 \]
\[ T_{\text{pass1}} (\langle \_ \rangle) = 1 \]
\[ T_{\text{pass1}} (\langle \_ , \_ \rangle, \langle \rangle) = 1 \]

\[ T_{\text{pass2}} :: \text{'a tree} \Rightarrow \text{nat} \]
\[ T_{\text{pass2}} (\langle \_ \rangle) = 1 \]
\[ T_{\text{pass2}} (\langle \_ , \_ , hs \rangle) = T_{\text{pass2}} hs + 1 \]
C

Notation

C.1 Symbol Table

The following table gives an overview of all the special symbols used in this book and how to enter them into Isabelle. The second column shows the full internal name of the symbol; the third column shows additional ASCII abbreviations. Either of these can be used to input the character using the auto-completion popup.

<table>
<thead>
<tr>
<th>Code</th>
<th>ASCII abbrev.</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>λ</td>
<td>\textbackslash{\text{lambda}}</td>
<td>%</td>
</tr>
<tr>
<td>≡</td>
<td>\textbackslash{\text{equiv}}</td>
<td>==</td>
</tr>
<tr>
<td>≠</td>
<td>\textbackslash{\text{noteq}}</td>
<td>-=</td>
</tr>
<tr>
<td>∧</td>
<td>\textbackslash{\text{And}}</td>
<td>!</td>
</tr>
<tr>
<td>∨</td>
<td>\textbackslash{\text{forall}}</td>
<td>!</td>
</tr>
<tr>
<td>∃</td>
<td>\textbackslash{\text{exists}}</td>
<td>?</td>
</tr>
<tr>
<td>\implies</td>
<td>\textbackslash{\text{Longrightarrow}}</td>
<td>\implies</td>
</tr>
<tr>
<td>\rightarrow</td>
<td>\textbackslash{\text{Longrightarrow}}</td>
<td>\rightarrow</td>
</tr>
<tr>
<td>\leftrightarrow</td>
<td>\textbackslash{\text{Longlefttrightarrow}}</td>
<td>\leftrightarrow</td>
</tr>
<tr>
<td>\Rightarrow</td>
<td>\textbackslash{\text{Rightarrow}}</td>
<td>\Rightarrow</td>
</tr>
<tr>
<td>\leftarrow</td>
<td>\textbackslash{\text{Leftarrow}}</td>
<td>\leftarrow</td>
</tr>
</tbody>
</table>
\(\neg\) | \\textbackslash{\text{not}} | \neg | |
\(\land\) | \\textbackslash{\text{and}} | \land or & |
\(\lor\) | \\textbackslash{\text{or}} | \lor or | |
\(\in\) | \\textbackslash{\text{in}} | \in |
\(\notin\) | \\textbackslash{\text{notin}} | \notin |
\(\cup\) | \\textbackslash{\text{union}} | \cup | Un |
C.2 Subscripts and Superscripts

In addition to this, subscripts and superscripts with a single symbol can be rendered using two special symbols, \texttt{\textbackslash sub} and \texttt{\textbackslash sup}. The term $x_0$ for instance can be input as $x_{\texttt{\textbackslash sub}0}$.

Longer subscripts and superscripts can be written using the symbols \texttt{\textbackslash bsub}...\texttt{\textbackslash esub} and \texttt{\textbackslash bsup}...\texttt{\textbackslash esup}, but this is only rendered in the somewhat visually displeasing form $x_{\texttt{\textbackslash bsub}...\texttt{\textbackslash esub}}$ and $x_{\texttt{\textbackslash bsup}...\texttt{\textbackslash esup}}$ by Isabelle/jEdit.
### C.3 Syntactic Sugar

The following table lists relevant syntactic sugar that is used in the book or its supplementary material. In some cases, the book notation deviates slightly from the Isabelle notation for better readability.

The last column gives the formal meaning of the notation (i.e. what it expands to). In most cases, this is not important for the user to know, but it can occasionally be useful to find relevant lemmas, or to understand that e.g. if one encounters the term \( \sum_{x \in A} f x \), this is just the \( \eta \)-contracted form of \( \sum x \in A. f x \).

The variables in the table follow the following convention:
- \( x \) and \( y \) are of arbitrary type
- \( m \) and \( n \) are natural numbers
- \( P \) and \( Q \) are boolean values or predicates
- \( xs \) is a list
- \( A \) is a set
- \( M \) is a multiset

<table>
<thead>
<tr>
<th>Book notation</th>
<th>Isabelle notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arithmetic (for numeric types)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x \cdot y )</td>
<td>( x \cdot y )</td>
<td>times ( x \cdot y )</td>
</tr>
<tr>
<td>( x / y ) or ( \frac{x}{y} )</td>
<td>( x / y )</td>
<td>divide ( x / y ) (for type real)</td>
</tr>
<tr>
<td>( x \ div y )</td>
<td>( x \ div y )</td>
<td>divide ( x \ div y ) (for type nat or int)</td>
</tr>
<tr>
<td>( \lfloor x \rfloor )</td>
<td>( \lfloor x \rfloor )</td>
<td>abs ( x )</td>
</tr>
<tr>
<td>( \lceil x \rceil )</td>
<td>( \lceil x \rceil )</td>
<td>floor ( x )</td>
</tr>
<tr>
<td>( \lfloor x \rfloor )</td>
<td>( \lfloor x \rfloor )</td>
<td>ceiling ( x )</td>
</tr>
<tr>
<td>( x^n )</td>
<td>( x^n )</td>
<td>power ( x^n )</td>
</tr>
<tr>
<td><strong>Lists</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lfloor xs \rfloor )</td>
<td>( \lfloor xs \rfloor )</td>
<td>length ( xs )</td>
</tr>
<tr>
<td>[]</td>
<td>[]</td>
<td>Nil</td>
</tr>
<tr>
<td>( x # xs )</td>
<td>( x # xs )</td>
<td>Cons ( x # xs )</td>
</tr>
<tr>
<td>[ x, y ]</td>
<td>[ x, y ]</td>
<td>( x # y # [] )</td>
</tr>
<tr>
<td>[ m..&lt;n ]</td>
<td>[ m..&lt;n ]</td>
<td>upt ( m # n )</td>
</tr>
<tr>
<td>( xs ! n )</td>
<td>( xs ! n )</td>
<td>nth ( xs # n )</td>
</tr>
<tr>
<td>( xs[n := y] )</td>
<td>( xs[n := y] )</td>
<td>list_update ( xs # n # y )</td>
</tr>
</tbody>
</table>
### Book notation vs. Isabelle notation

<table>
<thead>
<tr>
<th>Book notation</th>
<th>Isabelle notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>{}</td>
<td>{}</td>
<td>empty</td>
</tr>
<tr>
<td>{x, y}</td>
<td>{x, y}</td>
<td>insert x (insert y {})</td>
</tr>
<tr>
<td>x ∈ A</td>
<td>x ∈ A</td>
<td>Set.member x A</td>
</tr>
<tr>
<td>x ∉ A</td>
<td>x ∉ A</td>
<td>¬(x ∈ A)</td>
</tr>
<tr>
<td>A ∪ B</td>
<td>A ∪ B</td>
<td>union A B</td>
</tr>
<tr>
<td>A ∩ B</td>
<td>A ∩ B</td>
<td>inter A B</td>
</tr>
<tr>
<td>A ⊆ B</td>
<td>A ⊆ B</td>
<td>subset_eq A B</td>
</tr>
<tr>
<td>A ⊂ B</td>
<td>A ⊂ B</td>
<td>subset A B</td>
</tr>
<tr>
<td>f' A</td>
<td>f' A</td>
<td>image f A</td>
</tr>
<tr>
<td>f −' A</td>
<td>f −' A</td>
<td>vimage f A</td>
</tr>
<tr>
<td>{x</td>
<td>P x}</td>
<td>{x. P x}</td>
</tr>
<tr>
<td>{x ∈ A</td>
<td>P x}</td>
<td>{x∈A. P x}</td>
</tr>
<tr>
<td>{f x y</td>
<td>P x y}</td>
<td>{f x y</td>
</tr>
<tr>
<td>\bigcup_{x ∈ A} f x</td>
<td>\bigcup x∈A. f x</td>
<td>\bigcup (f' A)</td>
</tr>
<tr>
<td>\forall x ∈ A. P x</td>
<td>\forall x∈A. P x</td>
<td>Ball A P</td>
</tr>
<tr>
<td>\exists x ∈ A. P x</td>
<td>\exists x∈A. P x</td>
<td>Bex A P</td>
</tr>
</tbody>
</table>

### Multisets

<table>
<thead>
<tr>
<th>#M</th>
<th></th>
<th>size M</th>
</tr>
</thead>
<tbody>
<tr>
<td>{}</td>
<td>{}</td>
<td>#</td>
</tr>
<tr>
<td>{x} + M</td>
<td>{#}</td>
<td>add_mset x M</td>
</tr>
<tr>
<td>{x, y}</td>
<td>{#x, y#}</td>
<td>add_mset x (add_mset y {#})</td>
</tr>
<tr>
<td>x ∈ M</td>
<td>x ∈ # M</td>
<td>x ∈ set_mset M</td>
</tr>
<tr>
<td>x ∉ M</td>
<td>x ∉ # M</td>
<td>¬(x ∈ # M)</td>
</tr>
<tr>
<td>{x ∈ M</td>
<td>P x}</td>
<td>{# x∈M. P x #}</td>
</tr>
<tr>
<td>{f x</td>
<td>x ∈ M}</td>
<td>{# f x. x ∈ M #}</td>
</tr>
<tr>
<td>\forall x ∈ M. P x</td>
<td>\forall x∈M. P x</td>
<td>\forall x ∈ set_mset M. P x</td>
</tr>
<tr>
<td>\exists x ∈ M. P x</td>
<td>\exists x∈M. P x</td>
<td>\exists x ∈ set_mset M. P x</td>
</tr>
<tr>
<td>M ⊆ M'</td>
<td>M ⊆ # M'</td>
<td>subseteq_mset M M'</td>
</tr>
</tbody>
</table>
### C.3 Syntactic Sugar

<table>
<thead>
<tr>
<th>Book notation</th>
<th>Isabelle notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sum A )</td>
<td>( \sum A )</td>
<td>( \text{sum } (\lambda x. \ x) \ A )</td>
</tr>
<tr>
<td>( \sum_{x \in A} f \ x )</td>
<td>( \sum x \in A. \ f x )</td>
<td>( \text{sum } f \ A )</td>
</tr>
<tr>
<td>( \sum_{i = j} f \ k )</td>
<td>( \sum k = i \cdot j. \ f k )</td>
<td>( \text{sum } {i \cdot j} )</td>
</tr>
<tr>
<td>( \sum_{#} M )</td>
<td>( \sum # M )</td>
<td>( \text{sum}_\text{mset } M )</td>
</tr>
<tr>
<td>( \sum_{x \in # M} f \ x )</td>
<td>( \sum x \in # M. \ f x )</td>
<td>( \text{sum}<em>\text{mset } (\text{image}</em>\text{mset } f \ M) )</td>
</tr>
<tr>
<td>( \sum_{x \mapsto xs} f \ x )</td>
<td>( \sum x \mapsto xs. \ f x )</td>
<td>( \text{sum}_\text{list } (\text{map } f \ xs) )</td>
</tr>
</tbody>
</table>

(analogous for products)

### Intervals (for ordered types)

<table>
<thead>
<tr>
<th>Book notation</th>
<th>Isabelle notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>{x..}</td>
<td>{x..}</td>
<td>\text{atLeast } x</td>
</tr>
<tr>
<td>{..y}</td>
<td>{..y}</td>
<td>\text{atMost } y</td>
</tr>
<tr>
<td>{x..y}</td>
<td>{x..y}</td>
<td>\text{atLeastAtMost } x \ y</td>
</tr>
<tr>
<td>{x..&lt;y}</td>
<td>{x..&lt;y}</td>
<td>\text{atLeastLessThan } x \ y</td>
</tr>
<tr>
<td>{x&lt;y}</td>
<td>{x&lt;y}</td>
<td>\text{greaterThanAtMost } x \ y</td>
</tr>
<tr>
<td>{x&lt;&lt;y}</td>
<td>{x&lt;&lt;y}</td>
<td>\text{greaterThanLessThan } x \ y</td>
</tr>
</tbody>
</table>
References


60. Ocaml website. [https://ocaml.org](https://ocaml.org).


Index

- 1, 50
- [50, 253
/, 8
[... < ...], 254
(0), 253
(1), 254
(1#), 8
_ [__ := __], 254
{}], 10
\sum' 10
\in 10
\langle __, __, __\rangle, 49
\langle __, __\rangle, 49
2-3 trees, 81

abstraction function, 61, 76
almost complete tree, 53
amortized running time, 214
array, 121
augmented tree, 57
AVL tree, 101

balance factor, 108
balanced search trees, 81
Bellman-Ford algorithm, 202
binary search trees, 61
binary trie, 145
binomial heap, 179
binomial tree, 179
black height, 92
Braun tree, 122, 123
bst, 61
butlast, 254
child, 49
cmp, 62
correction, 9
complete, 52
complete tree, 51
computation induction, 12
decay, 253
deletion by joining, 64
deletion by replacing, 64
depth, 49
difference lists, 30
drop, 254
dynamic programming, 191
dynamic table, 216
Fibonacci tree, 107
filter, 253
filter_mset, 10
h, 51
hd, 254
head, 8
heap, 168
height, 51
height-balanced tree, 108
Huffman’s algorithm, 151
IH, 12
image_mset, 10
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>inclusion</td>
<td>9</td>
</tr>
<tr>
<td>inorder</td>
<td>50</td>
</tr>
<tr>
<td>insertion sort</td>
<td>20</td>
</tr>
<tr>
<td>int</td>
<td>9</td>
</tr>
<tr>
<td>interval tree</td>
<td>69</td>
</tr>
<tr>
<td>invariant</td>
<td>65, 76</td>
</tr>
<tr>
<td>preservation</td>
<td>65</td>
</tr>
<tr>
<td>join approach</td>
<td>114</td>
</tr>
<tr>
<td>leftist heap</td>
<td>171</td>
</tr>
<tr>
<td>leftist tree</td>
<td>171</td>
</tr>
<tr>
<td>length</td>
<td>253</td>
</tr>
<tr>
<td>level</td>
<td>49</td>
</tr>
<tr>
<td>linear order</td>
<td>19</td>
</tr>
<tr>
<td>inorder</td>
<td>19</td>
</tr>
<tr>
<td>list</td>
<td>8</td>
</tr>
<tr>
<td>list extensionality</td>
<td>127</td>
</tr>
<tr>
<td>locales</td>
<td>76</td>
</tr>
<tr>
<td>Map</td>
<td>78</td>
</tr>
<tr>
<td>map</td>
<td>253</td>
</tr>
<tr>
<td>map_tree</td>
<td>50</td>
</tr>
<tr>
<td>median</td>
<td>40</td>
</tr>
<tr>
<td>merge sort</td>
<td>24</td>
</tr>
<tr>
<td>mh</td>
<td>51</td>
</tr>
<tr>
<td>min_height</td>
<td>51</td>
</tr>
<tr>
<td>min_list</td>
<td>255</td>
</tr>
<tr>
<td>mset</td>
<td>10</td>
</tr>
<tr>
<td>multiset</td>
<td>10</td>
</tr>
<tr>
<td>mset_tree</td>
<td>10</td>
</tr>
<tr>
<td>nat</td>
<td>9</td>
</tr>
<tr>
<td>observer functions</td>
<td>215</td>
</tr>
<tr>
<td>optimal binary search tree</td>
<td>207</td>
</tr>
<tr>
<td>pairing heap</td>
<td>243</td>
</tr>
<tr>
<td>Patricia trie</td>
<td>147</td>
</tr>
<tr>
<td>pivot</td>
<td>22</td>
</tr>
<tr>
<td>preorder</td>
<td>50</td>
</tr>
<tr>
<td>Priority_Queue</td>
<td>167</td>
</tr>
<tr>
<td>priority queue</td>
<td>167</td>
</tr>
<tr>
<td>mergeable</td>
<td>167</td>
</tr>
<tr>
<td>Priority_Queue_Merge</td>
<td>168</td>
</tr>
<tr>
<td>Queue</td>
<td>219</td>
</tr>
<tr>
<td>queue</td>
<td>219</td>
</tr>
<tr>
<td>Hood-Melville</td>
<td>221</td>
</tr>
<tr>
<td>quicksort</td>
<td>22</td>
</tr>
<tr>
<td>real</td>
<td>9</td>
</tr>
<tr>
<td>red-black tree</td>
<td>91</td>
</tr>
<tr>
<td>replicate</td>
<td>254</td>
</tr>
<tr>
<td>rev</td>
<td>254</td>
</tr>
<tr>
<td>right-heavy</td>
<td>240</td>
</tr>
<tr>
<td>root</td>
<td>49</td>
</tr>
<tr>
<td>running time</td>
<td>12</td>
</tr>
<tr>
<td>running time function</td>
<td>12</td>
</tr>
<tr>
<td>canonical</td>
<td>14</td>
</tr>
<tr>
<td>runs</td>
<td>29</td>
</tr>
<tr>
<td>selection</td>
<td>35</td>
</tr>
<tr>
<td>Set</td>
<td>76</td>
</tr>
<tr>
<td>set</td>
<td>253</td>
</tr>
<tr>
<td>set_mset</td>
<td>10</td>
</tr>
<tr>
<td>set_tree</td>
<td>50</td>
</tr>
<tr>
<td>skew heap</td>
<td>239</td>
</tr>
<tr>
<td>sorted</td>
<td>19, 66</td>
</tr>
<tr>
<td>sorted_wrt</td>
<td>255</td>
</tr>
<tr>
<td>sorting</td>
<td>19</td>
</tr>
<tr>
<td>specification</td>
<td>65</td>
</tr>
<tr>
<td>splay tree</td>
<td>231</td>
</tr>
<tr>
<td>stability (sorting)</td>
<td>33</td>
</tr>
<tr>
<td>stable</td>
<td>33</td>
</tr>
<tr>
<td>sum_list</td>
<td>255</td>
</tr>
<tr>
<td>T</td>
<td>12</td>
</tr>
<tr>
<td>tail</td>
<td>8</td>
</tr>
<tr>
<td>take</td>
<td>254</td>
</tr>
<tr>
<td>time function, see</td>
<td>12</td>
</tr>
<tr>
<td>running time function</td>
<td>254</td>
</tr>
<tr>
<td>tl</td>
<td>254</td>
</tr>
<tr>
<td>tree</td>
<td>49</td>
</tr>
<tr>
<td>trie</td>
<td>143</td>
</tr>
<tr>
<td>2-3 trees</td>
<td>81</td>
</tr>
<tr>
<td>type class</td>
<td>19</td>
</tr>
<tr>
<td>type of interest</td>
<td>76</td>
</tr>
<tr>
<td>type variable</td>
<td>8</td>
</tr>
<tr>
<td>unit</td>
<td>9</td>
</tr>
</tbody>
</table>