15-150 Fall 2019

Stephen Brookes

Lecture 8
Sorting a tree of integers
1 Sorting trees

A tree is said to be sorted if at each node, the integer at that node is ≥ all integers in the left subtree and ≤ the integers in the right subtree. Here is an inductive way to characterize sortedness:

- The tree Empty is sorted.
- A non-empty tree value Node(A, x, B) is sorted if and only if every integer in A is ≤ x, every integer in B is ≥ x, and A and B are sorted.

For brevity we may sometimes write A ≤ x ≤ B, with A and B being tree values and x an integer. While it is possible to write an ML function for testing if a tree is sorted or not, we will not bother to do so here. We won’t ever need to check for sortedness in implementing an algorithm for sorting trees; instead we’ll design our code so that it is guaranteed to return a sorted tree, even without checking.

How to sort a tree

Our algorithm for sorting an integer tree can be described informally as follows:

- If the tree is empty, do nothing.
- Otherwise, (recursively) sort the two subtrees, then merge the sorted trees into a single (still sorted) tree; then insert the root value into its correct position.

This suggests that we design helper functions for inserting an item into a sorted tree, and merging two sorted trees into one. Later we will see that the merging operation itself needs a helper function, for splitting a tree into two subtrees, using a given integer value to determine which items from the tree go into the first or second subtree. Also it will be important to think carefully about what assumptions it is safe to make about the arguments to be supplied to these helper functions.

Insertion for trees

The tree-based analogue of the insertion function on lists turns out to be just what we need, a truly helpful function in the code that follows. We use capitalization to distinguish this function from the ins function on lists.
(* Ins : int * tree -> tree *)
(* REQUIRES T is a sorted tree *)
(* ENSURES Ins(x, T) = a sorted tree consisting of x and all of T *)
fun Ins (x, Empty) = Node(Empty, x, Empty)
| Ins (x, Node(t1, y, t2)) =
  case compare(x,y) of
    GREATER => Node(t1, y, Ins(x, t2))
  | _ => Node(Ins(x, t1), y, t2)

Compare this code with the code for ins on lists. See how similar it is.

We now show how to prove by structural induction that Ins satisfies this specification. Let \( P(T) \) be the property that

For all integer values \( x \), if \( T \) is sorted, then \( \text{Ins}(x,T) \) evaluates to a sorted tree consisting of \( x \) and the items of \( T \).

We prove “For all tree values \( T \), \( P(T) \) holds”, by structural induction on \( T \).

- The base case is simple. \( P(\text{Empty}) \) holds, because \( \text{Empty} \) is sorted and \( \text{Ins}(x, \text{Empty}) \) evaluates to \( \text{Node}(\text{Empty}, x, \text{Empty}) \). This is a tree value (because \( x \) is an integer value by assumption, and \( \text{Empty} \) is a tree value), is obviously sorted, and consists of just \( x \), as required.

- For the inductive step we argue as follows. Suppose \( t1 \) and \( t2 \) are tree values such that \( P(t1) \) and \( P(t2) \) hold, and let \( y \) be an integer value. We show that \( P(\text{Node}(t1, y, t2)) \) holds. To do this, let \( x \) be an integer value and suppose \( \text{Node}(t1, y, t2) \) is sorted. We must show that \( \text{Ins}(x, \text{Node}(t1, y, t2)) \) evaluates to a sorted tree value consisting of \( x \) and all items of \( \text{Node}(t1, x, t2) \). We also know that \( t1 \) and \( t2 \) are sorted, and \( t1 \leq y \leq t2 \). From the definition of \( \text{Ins} \) we see that

(a) Either \( x>y \), in which case we have

\[
\text{Ins}(x, \text{Node}(t1, y, t2)) \Rightarrow \text{Node}(t1, y, \text{Ins}(x, t2)).
\]

By induction hypothesis \( P(t2), \text{Ins}(x,t2) \) evaluates to a sorted tree (say \( u2 \)) consisting of \( x \) and all of \( t2 \), so we get

\[
\text{Ins}(x, \text{Node}(t1, y, t2)) \Rightarrow \text{Node}(t1, y, u2),
\]

\( t1 \leq y \leq u2 \), and \( t1 \) and \( u2 \) are sorted tree values, so this is a sorted tree value. Clearly it consists of \( x \) and \( y \) and all of \( t1 \) and \( t2 \). Thus \( P(\text{Node}(t1,y,t2)) \) holds in this case.
(b) Or $x \leq y$, in which case we have

$$\text{Ins}(x, \text{Node}(t_1, y, t_2)) \Rightarrow* \text{Node}(\text{Ins}(x, t_1), y, t_2).$$

By induction hypothesis $P(t_1)$, $\text{Ins}(x, t_1)$ evaluates to a sorted tree (say $u_1$) consisting of $x$ and all of $t_1$, so we get

$$\text{Ins}(x, \text{Node}(t_1, y, t_2)) \Rightarrow* \text{Node}(u_1, y, t_2),$$

$u_1 \leq y \leq t_2$, and $u_1$ and $t_2$ are sorted tree values, so this is a sorted tree value. Clearly it consists of $x$ and $y$ and all of $t_1$ and $t_2$. Thus $P(\text{Node}(t_1, y, t_2))$ holds in this case also.

So $P(\text{Node}(t_1, y, t_2))$ holds, as needed for the inductive step.

The above specification and proof use evaluational notation and evaluational reasoning. One can also state and prove an equational specification for $\text{Ins}$, using “value equations” derived from the function definition. It follows from the function definition that for all integer values $x$ and $y$, and all tree values $t_1$ and $t_2$, the following equations hold:

$$\text{Ins}(x, \text{Empty}) = \text{Node}(\text{Empty}, x, \text{Empty})$$

$$\text{Ins}(x, \text{Node}(t_1, y, t_2)) = \text{Node}(t_1, y, \text{Ins}(x, t_2)) \text{ if } x>y$$

$$\text{Ins}(x, \text{Node}(t_1, y, t_2)) = \text{Node}(\text{Ins}(x, t_1), y, t_2) \text{ if } x<y \text{ or } x=y$$

We can prove that

For all integer values $x$ and all sorted tree values $T$, there is a sorted tree value $S$ such that

$$\text{Ins}(x, T) = S$$

and $S$ consists of $x$ and all the integers in $T$.

To prove this result you can basically adapt the evaluational proof steps from above and write a corresponding equational proof.

**Splitting a sorted tree**

In adapting the mergesort algorithm to operate on trees we need a suitable analog to the $\text{split}$ function. (In class we motivated this need by trying to figure out how to merge two sorted trees. At some point we found ourselves wanting to split a tree into two trees.) It isn’t easy to figure out a good
way to hew a tree into two roughly equal sized pieces, based solely on the structure of the tree (by analogy with the way the split function on lists worked). Instead, we will start from a tree and an integer, and break the tree into two trees, one consisting of items less-or-equal to the integer and the other consisting of items greater than or equal to the integer. (There is some wiggle room here concerning where the items equal to this integer should go.) We will only ever need to use this method on a sorted tree, as you will observe when we develop the code. We also design the function so that when applied to a sorted tree it produces a pair of sorted trees. Indeed the design of this function takes advantage of the assumption that the tree is already sorted, a fact that we echo in the way we write the function’s specification.

(* SplitAt : int * tree -> tree * tree *)
(* REQUIRES T is sorted *)
(* ENSURES SplitAt(x, T) = (A,B) where
  A and B are sorted trees, A <= x <= B,
  A and B consist of the items from T *)
fun SplitAt(x, Empty) = (Empty, Empty)
| SplitAt(x, Node(t1, y, t2)) =
  case compare(y, x) of
    GREATER => let
      val (l1, r1) = SplitAt(x, t1)
      in
        (l1, Node(r1, y, t2))
      end
    | _ => let
      val (l2, r2) = SplitAt(x, t2)
      in
        (Node(t1, y, l2), r2)
      end

This function is structurally inductive, because in the recursive clause
SplitAt(x,Node(t1,y,t2)) either calls SplitAt(x,t1) or SplitAt(x,t2),
in each case making a recursive call on a subtree. We prove that SplitAt satisfies this specification, by structural induction. To be precise, we prove “For all tree values T, P(T)” where P(T) is the property that if T is sorted, then for all values x, SplitAt(x, T) is equal to a pair of sorted trees (A,B) such that A ≤ x ≤ B and A, B consist of the items of T.
• Base case: Empty is sorted and has no elements, so we need to show that for all values x, \( \text{SplitAt}(x, \text{Empty}) \) is equal to a pair of sorted trees with no elements. By definition we have \( \text{SplitAt}(x, \text{Empty}) = (\text{Empty}, \text{Empty}) \), and the requirements hold trivially.

• Inductive step: Let \( t \) be a sorted tree of form Node\( (t1, y, t2) \) and assume that \( \text{SplitAt} \) satisfies the spec on \( t1 \) and on \( t2 \). By assumption that the whole tree is sorted, we also know that \( t1 \) and \( t2 \) are sorted, and that \( t1 \leq y \leq t2 \). We show that \( \text{SplitAt}(x, t) \) is equal to a pair of sorted trees with the required properties. There are two sub-cases to analyze, branching on the result of comparing the values of \( x \) and \( y \).

(a) If \( y > x \) we have

\[
\text{SplitAt}(x, t) = (l1, \text{Node}(r1, y, t2))
\]

where \( (l1, r1) = \text{SplitAt}(x, t1) \). By Induction Hypothesis, \( l1 \leq x \leq r1 \) and \( l1, r1 \) are sorted trees consisting of the items from \( t1 \). So \( l1 \leq x \leq \text{Node}(r1, y, t2) \) and \( r1 \leq y \leq t2 \). And \( \text{Node}(r1, y, 12) \) is a sorted tree consisting of the items from \( t1 \), \( y \) and \( 12 \). Together with \( l1 \) this covers all the items from \( t \).

(b) If \( y \leq x \) we have

\[
\text{SplitAt}(x, t) = (\text{Node}(t1, y, 12), r2)
\]

where \( (12, r2) = \text{SplitAt}(x, t2) \).

By induction hypothesis, \( 12 \leq x \leq r2 \) and \( 12, r2 \) are sorted trees comprising the items from \( t2 \). Hence \( t1 \leq y \leq 12 \) and \( \text{Node}(t1, y, 12) \) is sorted. Also \( \text{Node}(t1, y, 12) \leq x \leq 12 \). (Fill in the remaining details.)

That completes the proof.

**Merging two sorted trees**

Now the tree-based analog of \texttt{merge} on lists: a function that takes a pair of sorted trees and combines their data into a single (also sorted) tree. We use \texttt{SplitAt} as a helper.

\[
(* \text{Merge : tree} \times \text{tree} \to \text{tree} *)
\]
\[
(* \text{REQUIRES} t1 \text{ and } t2 \text{ are sorted trees} *)
\]
(* ENSURES Merge(t1, t2) = a sorted tree
    consisting of the items from t1 and t2 *)
fun Merge (Empty, t2) = t2
   | Merge (Node(l1,x,r1), t2) = let
       val (l2,r2) = SplitAt(x,t2)
       in
       Node(Merge(l1,l2), x, Merge(r1,r2))
   end

The proof that Merge meets this spec relies on the fact that SplitAt meets its own specification. Indeed, we deliberately chose a spec for SplitAt that would help us to prove Merge correct. That’s one of the skills that we want you to learn: the art of choosing helper functions and specs wisely! We claim that “For all sorted tree values T, P(T) holds”, where P(T) is the property that

   For all sorted tree values U, Merge(T, U) is equal to a sorted tree value comprising the items from T and U.

The proof is by structural induction on T. (Since sorted trees have sorted children, it is perfectly OK to do this kind of structural induction on sorted trees!)

- P(Empty) holds obviously. (Fill in the details.)
- For the inductive case, suppose l1 and r1 are sorted tree values for which P(l1) and P(r1) hold. Let x be an integer value and assume Node(l1, x, r1) is sorted. We must show that P(Node(l1, x, r1)) holds. Let U be a sorted tree value. By definition of Merge we have

   Merge(Node(l1, x, r1), U) = Node(Merge(l1, l2), x, Merge(r1, r2))

where (l2, r2) = SplitAt(x, U). By the proven spec for SplitAt (which is applicable here because U is a sorted tree), l2 ≤ x ≤ r2 and l2, r2 are sorted and comprise the items from U. By assumption that the original tree is sorted, we have l1 ≤ x ≤ r1, and l1, r1 are sorted. So by Induction Hypothesis, there are sorted tree values l and r such that

   Merge(l1, l2) = l,  Merge(r1, r2) = r
and \( l \) comprises the items from \( l1, l2 \), and \( r \) comprises the items from \( r1, r2 \). Hence \( l \leq x \leq r \) and \( \text{Node}(l, x, r) \) is a sorted tree, consisting of the items from \( l1, x, r1, U \), as required.

That completes the proof. (We skipped over a few details – make sure you understand how to fill in the gaps.)

**Lesson**

It’s important to notice how in the above code analysis the specs for the various helper functions play a crucial rôle. Just in the nick of time, it turned out we could appeal to an induction hypothesis, which was applicable because we had shown that one of the helper functions behaved well. We were careful to only use a helper function with arguments that satisfy the requirements of the helper spec, and in the proof details we confirmed that the guarantees made by the helper functions (when used in this manner) were sufficient to ensure that the rest of the code meets its own spec. If we hadn’t shown that \( \text{SplitAt} \) preserves sortedness, we would have no basis for claiming that \( \text{Merge} \) preserves sortedness. The lesson is: choose your helper functions wisely, choose their specs wisely (with an eye to how you will use them), and nail down the correctness proof (at least with a sketch of the key details).

**The tree-sorting function Msort**

Using \( \text{Ins} \) and \( \text{Merge} \), and guided by their (proven) specs, we are now ready to define a mergesorting function for integer trees. The hard work has already been done; now comes the easy and more immediately rewarding part! The tree sorting function in ML is defined in a way that mimics the algorithm we sketched earlier.

\[
\begin{align*}
\text{fun Msort : tree \to tree} \\
\text{REQUIRES T is a value of type tree} \\
\text{ENSURES Msort(T) = a sorted tree consisting of the items of T}
\end{align*}
\]

Again the proof that \( \text{Msort} \) meets this spec uses the facts (shown earlier) that \( \text{Ins} \) and \( \text{Merge} \) satisfy their specs. And again these helper specs were
carefully chosen to make this all fit together! Exercise: fill in the proof details. Contrast with the proof given in the earlier lecture notes for the mergesort function on lists.

2 Exploration

To illustrate how the various functions discussed above actually work on a specific tree example, try running the following code and drawing pictures of the tree values produced in each stage.

```scala
val T1 = Node(Leaf 3, 6, Leaf 1);
val T2 = Node(Leaf 5, 2, Empty);
val T = Node(T1, 4, T2);
%
val M1 = Msort T1;
val M2 = Msort T2;
val M = Merge(M1, M2);
%
val S = Ins(4, M);
```

You might also want to try some examples using `SplitAt`. Furthermore, it may be useful to define some functions for extracting a list containing the integers in a tree. In lab you will discuss traversal lists and in-order, pre-order and post-order traversal of trees. It turns out that an integer tree is sorted if and only if its in-order traversal list is a sorted list of integers. So one can easily check if a tree is sorted by looking at its in-order traversal list. Just for some irrelevant fun(?), try to figure out a decent specification of what `SplitAt`, `Ins` and `Merge` do when applied to arguments that do NOT satisfy the REQUIRES assumptions used above.
3 Comments on sorting trees

For an integer list $L$ there is just one sorted list that contains the same items as $L$. So it makes sense to talk about “computing the sorted version of $L$”. In contrast, for a collection of at least 2 integers there can be multiple different trees containing the same integers. Indeed, there can be many different sorted trees containing the same integers. So the specifications and proofs so far don’t really tell us much about the shapes of the trees produced by sorting. It would be nice if we guaranteed to produce balanced trees, in which at each node the numbers of integers in the two child subtrees differ by at most 1. However, even if we start with a balanced (unsorted) tree, the functions that we have defined so far do not always produce balanced results. We will return to this point shortly.

4 Size analysis

We can prove some fairly obvious facts about the effects of the operations on the size of a tree.

1. For all trees $t$ and integers $x$,
   $$ \text{size}(\text{Ins}(x, t)) = \text{size}(t) + 1. $$

2. For all trees $t$ and integers $y$, if $\text{SplitAt}(y, t) = (t_1, t_2)$ then
   $$ \text{size}(t_1) + \text{size}(t_2) = \text{size}(t). $$

3. For all trees $t_1$ and $t_2$,
   $$ \text{size}(\text{Merge}(t_1, t_2)) = \text{size } t_1 + \text{size } t_2. $$

4. For all trees $t$,
   $$ \text{size}(\text{Msort } t) = \text{size } t. $$

In each case, you can prove the result by structural induction. Check that these results are consistent with the examples from before.
5 Depth analysis

We can prove some useful (and intuitively obvious) results about depth. These will be helpful when we analyze the runtime behavior of the code. The following results are provable, by choosing an appropriate kind of induction. [Of course, to do the proofs we would need access to the function definition for depth, which is given in lab.]

(1) For all trees \( t \) and integers \( x \),

\[
\text{depth(Ins}(x, t)) \leq \text{depth}(t) + 1.
\]

[Use structural induction, since \( \text{Ins}(x, t) \) makes a recursive call on a child of \( t \).]

(2) For all trees \( t \) and integers \( y \), if \( \text{SplitAt}(y, t) = (t_1, t_2) \) then

\[
\text{depth}(t_1) \leq \text{depth}(t) \text{ and } \text{depth}(t_2) \leq \text{depth}(t).
\]

[Use structural induction, since \( \text{SplitAt}(y, t) \) makes a recursive call on a child of \( t \).]

(3) For all trees \( t_1 \) and \( t_2 \),

\[
\text{depth}(\text{Merge}(t_1, t_2)) \leq \text{depth } t_1 + \text{depth } t_2.
\]

[Use induction on the structure of \( t_1 \).]

(3) For all trees \( t \),

\[
\text{depth}(\text{Msort } t) \leq \text{depth } t.
\]

[Use induction on the structure of \( t \).]

Check that these results are consistent with the examples from earlier.
6 Work and span

We’ve shown how to derive recurrence relations for the work of a sequentially executed piece of code, and how to estimate asymptotically what the runtime is on “large” inputs, using big-O notation.

Now we have some functions operating on trees for which it makes a lot of sense to consider using parallel evaluation. The span of a code fragment is obtained by assuming that we have as many parallel processors as we need, and taking the maximum runtime of code pieces that can be evaluated independently; we still use addition for the runtimes of code fragments that need to be executed in sequential order, typically because of a data dependency: one fragment needs the result of the other. Operating on trees allows us in principle to sort the left and right children of a node in parallel, since their results do not depend on each other. Of course, these tasks need to be completed before the merging phase. And the splitting phase needs to go first.

These facts guide us in analyzing the span. Here is a rough outline. With trees there are two “largeness” measures of interest: depth and size.

- The work and span for $\text{Ins}(x, t)$ is $O(d)$, where $d$ is the depth of $t$. Reason: $\text{Ins}(x, t)$ makes a single recursive call, on a subtree with depth decreased by 1.

- $\text{SplitAt}(y, t)$ has span $O(d)$, where $d = \text{depth} t$. Reason: makes a single recursive call, on a tree with depth one less.

- $\text{Merge}(t_1, t_2)$ has span $O(d_1d_2)$, where $d_1, d_2$ are depth $t_1, \text{depth} t_2$.

- Assuming that the trees produced by $\text{Msort}$ are balanced, so that their depth is about the logarithm of their size, $\text{Msort}(t)$ has span $O(d^3)$, where $d$ is the depth of $t$. Reason: making the balance assumption leads us to the recurrence

$$S_{\text{Msort}}(d) = S_{\text{Ins}}(d) + S_{\text{Merge}}(d - 1) + S_{\text{Msort}}(d - 1) = d + (d - 1)^2 + S_{\text{Msort}}(d - 1)$$

for balanced trees of depth $d > 1$. Expanding out, and observing that the sum of the first $d$ squares is proportional to $d^3$, we deduce that the span is $O(d^3)$. Since the size $n$ of a balanced tree and its depth $d$ satisfy $d = O(\log n)$, our analysis shows that the span for $\text{Msort}(t)$ on balanced trees of size $n$ is $O((\log n)^3)$. 
Thus (ignoring constants), when we sort a billion integers in a balanced tree, the length of the longest critical path is about 27000 operations, so we can exploit over a million processors!

This would be true, except for the bug in the above analysis! We assumed implicitly in the rough analysis (and explicitly in the preamble) that the trees passed by \texttt{Msort} to \texttt{Merge} were balanced. However, this is not necessarily the case, because even if we assumed that the \textit{original} tree was balanced, these two trees have been built by calling \texttt{Msort} (albeit on balanced trees). We haven’t proven that \texttt{Msort} applied to a balanced tree will produce a balanced tree. In fact, this isn’t necessarily true. The best that our analysis really predicts is that the span of this algorithm can’t actually be better than this bound, because we obtained this bound by making the most optimistic assumptions about the structure of the tree.

Later we will discuss how to implement binary trees with insertion and deletion operations that are guaranteed to build trees with a reasonable balance property built in. When we get there, you might want to come back and see how you could adapt the code above to fit with these better behaved trees.

(Look again at the examples from before: are the various trees constructed there balanced?)

\textbf{Exercise}
A student pointed out that there is a way to define a version of tree-mergesort that avoids using \texttt{Ins}, instead calling \texttt{Merge}:

\begin{verbatim}
fun Msort' Empty = Empty
| Msort' (Node(t1, x, t2)) =
    Merge(Node(Empty,x,Empty), Merge(Msort' t1, Msort' t2))
\end{verbatim}

(i) Would this be extensionally equivalent to the previous function? (How would one prove it?)

(ii) Does this function satisfy the same specification as before, i.e. does it still sort?

(iii) And is it as efficient, or more efficient, asymptotically? (How could you determine this?)
7 Self-test 7

1. Write an ML function \( \text{leaves} : \text{tree} -> \text{int list} \) such that for all tree values \( T \), \( \text{leaves}(T) = \) the list of all integers occurring at leaf nodes of \( T \). For example, \( \text{leaves}(\text{Full}(42,3)) = [42,42,42,42] \).

2. Let \( T \) be a tree with depth \( d \). Why is the size of \( T \) at most \( 2^d - 1 \)?

3. Write an ML function \( \text{treesum} : \text{tree} -> \text{int} \) such that for all tree values \( T \), \( \text{treesum}(T) \) evaluates to the sum of the integers at the nodes of \( T \). For example, \( \text{treesum}(\text{Full}(42,3)) \) should evaluate to 294.

4. Write an ML function \( \text{leafsum} : \text{tree} -> \text{int} \) such that for all tree values \( T \), \( \text{leafsum}(T) \) evaluates to the sum of the integers at leaf nodes of \( T \). We interpret this quantity as 0 if the tree is \( \text{Empty} \). Do not use \( \text{leaves} \) from above!

5. State and prove a theorem about the value of \( \text{treesum}(\text{Full}(x,n)) \) when \( n \) is a non-negative integer.

6. How many different tree values \( T \) have exactly three nodes containing the integers 1,2,3? Of these, how many are sorted trees?

7. Define an ML function \( \text{balanced} : \text{tree} -> \text{bool} \) such that for all tree values \( T \), \( \text{balanced}(T) = \text{true} \) if at each node of \( T \) the sizes of the left and right children differ by at most 1. Otherwise \( \text{balanced}(T) = \text{false} \).

8. Calculate the value produced by the following piece of code:

\[
\begin{align*}
\text{val } T1 &= \text{Node}(\text{Leaf} 1, 6, \text{Leaf} 3); \\
\text{val } T2 &= \text{Node}(\text{Empty}, 5, \text{Leaf} 2); \\
\text{val } T &= \text{Node}(T1, 4, T2); \\
\text{Msort } T;
\end{align*}
\]

9. Prove that \( \text{Msort} \) satisfies its specification. You can assume given proofs that \( \text{Merge} \) and \( \text{Ins} \) satisfy their specifications.

10. Let \( T_n \) be the tree value of the expression \( \text{Full}(42,n) \), for \( n \geq 0 \). This is a full binary tree of depth \( n \) with 42 at each node. State and prove an assertion about the (shape of the) value of \( \text{Msort}(T_n) \).