Lecture 12: Greedy Algorithms and Start of Dynamic Programming

We continue the Huffman encoding problem from last lecture. An alternative formulation for expressing the total encoded length is to given internal nodes assigned frequencies as the summation of frequencies of the leaves under this node. We claim that total encode length is equal to summation of frequencies of all the nodes, including the internal nodes, in the tree (except the root). Why? Intuitively, we distribute the frequency of a leaf times its depth to each node on the path from root to the leaf.

We now study the properties of the complete binary tree constructed in Huffman encoding. The key observation is that the two lowest frequency symbols must be the two siblings with largest depth. To see this, suppose this is not true and the least frequent symbol is not assigned to the leaf with largest depth. Then we simply swap the symbol with the largest depth with the least frequent symbol, and this swapping will make the text shorter. Also note that there must exist a symbol as the sibling of the deepest symbol. This implies that we can choose two lowest frequency symbols and create a subtree. From the second formulation, we know we can assign the frequency of this newly created internal node (called $w$, with two least frequent symbols $w_1$ and $w_2$ as children) as the summation of the frequency of $w_1$ and $w_2$. This allows us to shrink in this problem: get rid of $w_1$ and $w_2$, and add a new symbol $w$ into the list of symbols to process. Then we repeat the process. But why does this work?

It is useful to think of picking formulation 2: our goal is to find a tree where each node has a frequency (for internal node, the frequency must equal to the total frequencies of the leaves under this node). Recall that we now have $n - 1$ symbols in the smaller problem. After getting rid of $w_1$ and $w_2$ but keep their parent node $w$, the tree is still a valid Huffman encoding tree (verify this yourself). The only difference is that we have a new leaf $w$ with frequency $f(w_1) + f(w_2)$. To minimize the total encoding length of the original tree, of course we should also minimize the total length of this new tree. Due to the equivalence of the first and second formulations, we are really seeking the optimal Huffman encoding tree for the $n - 1$ symbols. And this means we can re-apply the rule of pairing up two smallest frequencies among the $n - 1$ symbols and continue.

This leads to a simple algorithm: at each iteration, pick the two symbols with lowest frequencies and form a subtree with these two symbols; remove these two symbols and create a new symbol with the summation of the these removed symbols as its frequency. Continue until there is only one symbol left. This algorithm needs frequent operations to obtain (and extract) the symbols with lowest frequency. This can be done by a heap, which makes each iteration to take $O(\log n)$ time for $n$ symbols. So the total running time is $O(n \log n)$. Try to run some examples yourself to make sure you understand this algorithm.

Now let us think about what properties of a good greedy strategy has. For greedy algorithm to work, we should be able to make locally optimal choices at the moment, and after making the choice, we need to prove we still always get optimal solution. Second, we need the optimal substructure property: after making a choice, we now have a new (and smaller) problem; the key to greedy algorithm to work is that we can solve the smaller problem and then combine this solution with the decision made at the greedy choice to obtain the final solution. Read our textbook for more detailed explanation.

Now we start to discuss what is dynamic programming. Dynamic programming is one of the most useful algorithmic techniques. We use an example problem, longest increasing subsequence (LIS), to illustrate the idea. Given a list of $n$ numbers (i.e. sequence, denoted as $A$), we define subsequence as a sub-list of the numbers (with the original order). For example, let the sequence being 4; 2; 7; 5; 3; 5; 9; 6. A subsequence is 2, 5, 4, 6. We call a subsequence an increasing subsequence if this subsequence consists of numbers in the increasing order. Example of increasing subsequence of the above example is 4, 7, 9. The LIS problem is to find the longest increasing subsequence (LIS). We first study how to compute the length of LIS. In this example, the length of LIS is 4.

Let us define $L(i)$ as the length of a LIS that ends at position $i$ (position means index of the array). Then, the length of the LIS of the whole list is simply $\max_{i=1}^{n} L(i)$. To see how to compute $L(i)$, we consider
a LIS ending at $i$. Its previous number should be some at some $j < i$, where $A[j] < A[i]$. But which one is $j$ if there are multiple such $j$? Let $j$ be the position immediately preceding $i$ in the LIS ending at $i$. Then we know $L(i) = L(j) + 1$. This is because since the LIS uses $j$-th number, the it must take the LIS ending at $j$ (otherwise it contradicts the assumption of the LIS ending at $i$). Since we do not know which $j$ to use, we take the best: $L(i) = \max_{j<i,A[j]<A[i]} L(j) + 1$.

Lecture 13: Dynamic Programming

Now to compute a $L(i)$, we can use the previous recursion: it will then recursively compute $L(j)$. This works, but with a great cost. The problem is, many subproblems $L(j)$ are computed multiple times. To see how serious the problem is, we consider the worst case for computing $L(n)$. Let $S(i)$ be the number of recursion calls (i.e. nodes in recursion tree) for computing $L(i)$. Then, in the worst case,

$$S(n) = 1 + \sum_{i=1}^{n-1} S(i)$$

We observe $S(1) = 1$, $S(2) = 2$, $S(3) = 4$. Then we guess: $S(i) = 2^{i-1}$, and this is easily provable using induction. So, $S(n) = 2^{n-1}$, which implies that the recursion based algorithm runs in exponential time. Now we apply the dynamic programming that will lead to a polynomial time algorithm. The key idea is to memorize the subproblem solutions so that we do not re-compute them, using a bottom-up approach.

1: for $i = 1$ to $n$ do
2: $L[i] = 1$
3: for $j = 1$ to $i - 1$ do
5: $L[i] = 1 + L[j]$
6: end if
7: end for
8: end for
9: return $\max_{i=1}^{n}(L[i])$.

Note: when computing $L(i)$, $L(j)$ has already been computed. And for each $L(i)$, it gets updated just once. Running time: there are $n$ iterations, each iteration taking at most $O(n)$ time. So the total time is $O(n^2)$.

The only remaining issue is that we only know the LIS’s length, but what is LIS itself? That is not so hard. We suppose the LIS ending at node $p$ (i.e. we know the last letter of the LIS, $A[p]$). To find the one before $A[p]$, we perform trace back. In other words, we check which position $j$ that leading to the current value at $L[p]$; then we move to this $j$. We do this recursively (each time figure out its immediate previous node). I suggest you to work out a few examples to better understand this.

We now start our second example problem: longest common subsequence (LCS). This part is well explained in our textbook. Read Section 15.4. We just had time to describe the first case of LCS recurrence. Will continue in our next lecture.