Chapter 9

More Dynamic Programming

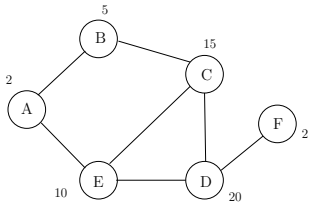
CS 473: Fundamental Algorithms, Spring 2011
February 17, 2011

9.1 Maximum Weighted Independent Set in Trees

9.1.0.1 Maximum Weight Independent Set Problem

Input Graph $G = (V, E)$ and weights $w(v) \geq 0$ for each $v \in V$

Goal Find maximum weight independent set in G



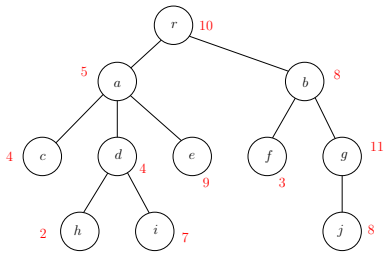
Maximum weight independent set in above graph: $\{B, D\}$

9.1.0.2 Maximum Weight Independent Set in a Tree

Input Tree $T = (V, E)$ and weights $w(v) \geq 0$ for each $v \in V$

Goal Find maximum weight independent set in T

Maximum weight independent set in above tree: ??



9.1.0.3 Towards a Recursive Solution

For an arbitrary graph G:

- Number vertices as $v_1, v_2, \ldots, v_n$
- Find recursively optimum solutions without v_n (recurse on $G - v_n$) and with v_n (recurse on $G - v_n - N(v_n)$ & include v_n).
- Saw that if graph G is arbitrary there was no good ordering that resulted in a small number of subproblems.

What about a tree? Natural candidate for v_n is root r of T?

9.1.0.4 Towards a Recursive Solution

Natural candidate for v_n is root r of T? Let $\mathcal{O}$ be an optimum solution to the whole problem.

Case $r \notin \mathcal{O}$: Then $\mathcal{O}$ contains an optimum solution for each subtree of T hanging at a child of r.

Case $r \in \mathcal{O}$: None of the children of r can be in $\mathcal{O}$. $\mathcal{O} - \{r\}$ contains an optimum solution for each subtree of T hanging at a grandchild of r.

Subproblems? Subtrees of T hanging at nodes in T.

9.1.0.5 A Recursive Solution

$T(u)$: subtree of T hanging at node u
$OPT(u)$: max weighted independent set value in $T(u)$

$$OPT(u) = \max \begin{cases} \sum_{v \text{ child of } u} OPT(v), \\ w(u) + \sum_{v \text{ grandchild of } u} OPT(v) \end{cases}$$

9.1.0.6 Iterative Algorithm

- Compute $OPT(u)$ bottom up. To evaluate $OPT(u)$ need to have computed values of all children and grandchildren of u
- What is an ordering of nodes of a tree T to achieve above? Post-order traversal of a tree.

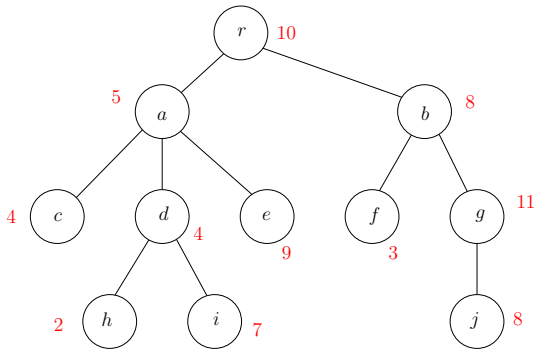
9.1.0.7 Iterative Algorithm

MIS-Tree(T):
  Let $v_1, v_2, \ldots, v_n$ be a post-order traversal of nodes of T
  for $i = 1$ to n **do**
    $M[v_i] = \max\Big(\sum_{v_j \text{ child of } v_i} M[v_j],\ \ w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j]\Big)$
  return $M[v_n]$ (* Note: v_n is the root of T *)

Space: $O(n)$ to store the value at each node of T
Running time:

- Naive bound: $O(n^2)$ since each $M[v_i]$ evaluation may take $O(n)$ time and there are n evaluations.
- Better bound: $O(n)$. A value $M[v_j]$ is accessed only by its parent and grand parent.

9.1.0.8 Example



9.2 DAGs and Dynamic Programming

9.2.0.9 Recursion and DAGs

Observation 9.2.1 *Let A be a recursive algorithm for problem Π. For each instance I of Π there is an associated* DAG *$G(I)$.*

- Create directed graph $G(I)$ as follows
- For each sub-problem in the execution of A on I create a node
- If sub-problem v *depends* on or *recursively calls* sub-problem u add *directed* edge (u, v) to graph
- $G(I)$ is a DAG. Why? If $G(I)$ has a cycle then A will not terminate on I

9.2.0.10 Iterative Algorithm in Dynamic Programming and DAGs

Observation 9.2.2 *An iterative algorithm B obtained from a recursive algorithm A for a problem Π does the following: for each instance I of Π, it computes a topological sort of $G(I)$ and evaluates sub-problems according to the topological ordering.*

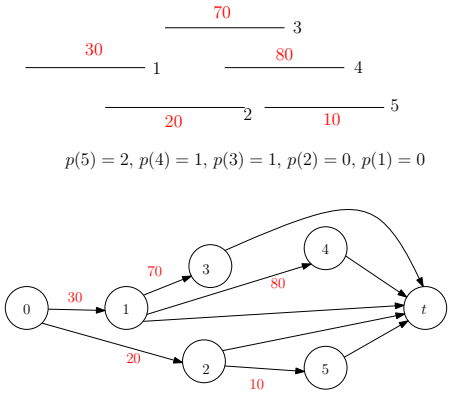
- Sometimes the DAG $G(I)$ can be obtained directly without thinking about the recursive algorithm A
- In some cases (*not all*) the computation of an optimal solution reduces to a shortest/longest path in DAG $G(I)$
- Topological sort based shortest/longest path computation is dynamic programming!

9.2.0.11 Weighted Interval Scheduling via Longest Path in a DAG

Given intervals, create a DAG as follows

- one node for each interval plus a dummy source node for interval 0 plus a dummy sink node t.
- for each interval i add edge $(p(i), i)$ of length/weight v_i
- for each interval i add edge (i, t) of length 0

9.2.0.12 Example



9.2.0.13 Relating Optimum Solution

Given interval problem instance I let $G(I)$ denote the DAG constructed as described.

Claim: Optimum solution to weighted interval scheduling instance I is given by *longest* path from s to t in $G(I)$.

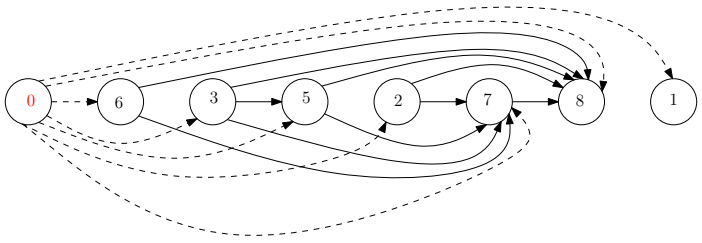
Assuming claim is true,

- If I has n intervals, DAG $G(I)$ has $n+2$ nodes and $O(n)$ edges. Creating $G(I)$ takes $O(n \log n)$ time: to find $p(i)$ for each i. How?
- Longest path can be computed in $O(n)$ time — recall $O(m+n)$ algorithm for shortest/longest paths in DAGs.

9.2.0.14 DAG for Longest Increasing Sequence

Given sequence $a_1, a_2, \ldots, a_n$ create DAG as follows:

- add sentinel a_0 to sequence where a_0 is less than smallest element in sequence
- for each i there is a node v_i
- if $i < j$ and $a_i < a_j$ add an edge (v_i, v_j)
- find longest path from v_0



9.3 Edit Distance and Sequence Alignment

9.3.0.15 Spell Checking Problem

Given a string "exponen" that is not in the dictionary, how should a spell checker suggest a *nearby* string?

What does nearness mean?

Question: Given two strings $x_1x_2 \dots x_n$ and $y_1y_2 \dots y_m$ what is a *distance* between them?

Edit Distance: minimum number of "edits" to transform x into y.

9.3.0.16 Edit Distance

Definition 9.3.1 *Edit distance between two words X and Y is the number of letter insertions, letter deletions and letter substitutions required to obtain Y from X.*

Example 9.3.2 *The edit distance between FOOD and MONEY is at most* 4*:*

$$\underline{\text{F}}\text{OOD} \rightarrow \text{MO}\underline{\text{O}}\text{D} \rightarrow \text{MON}\underline{\text{O}}\text{D} \rightarrow \text{MONE}\underline{\text{D}} \rightarrow \text{MONEY}$$

9.3.0.17 Edit Distance: Alternate View

Alignment
Place words one on top of the other, with gaps in the first word indicating insertions, and gaps in the second word indicating deletions.

F	O	O		D
M	O	N	E	Y

Formally, an *alignment* is a set M of pairs (i, j) such that each index appears at most once, and there is no "crossing": $i < i'$ and i is matched to j implies i' is matched to $j' > j$. In the above example, this is $M = \{(1, 1), (2, 2), (3, 3), (4, 5)\}$. Cost of an alignment is the number of mismatched columns plus number of unmatched indices in both strings.

9.3.0.18 Edit Distance Problem

Problem
Given two words, find the edit distance between them, i.e., an alignment of smallest cost.

9.3.0.19 Applications

- Spell-checkers and Dictionaries
- Unix `diff`
- DNA sequence alignment ... but, we need a new metric

9.3.0.20 Similarity Metric

Definition 9.3.3 *For two strings X and Y, the cost of alignment M is*

- *[Gap penalty] For each gap in the alignment, we incur a cost δ*
- *[Mismatch cost] For each pair p and q that have been matched in M, we incur cost α_{pq}; typically $\alpha_{pp} = 0$*

Edit distance is special case when $\delta = \alpha_{pq} = 1$

9.3.0.21 An Example

Example 9.3.4

o		c	u	r	r	a	n	c		e
o	c	c	u	r	r	e	n	c	e	

$Cost = \delta + \alpha_{ae}$

Alternative:

o		c	u	r	r		a	n	c	e
o	c	c	u	r	r	e		n	c	e

$Cost = 3\delta$

Or a really stupid solution (delete string, insert other string):

o	c	u	r	r	a	n	c	e									
									o	c	c	u	r	r	e	n	c	e

$Cost = 19\delta$.

9.3.0.22 Sequence Alignment

Input Given two words X and Y, and gap penalty δ and mismatch costs α_{pq}

Goal Find alignment of minimum cost

9.3.1 Edit distance

9.3.1.1 Basic observation

Let $X = \alpha x$ and $Y = \beta y$
α, β: stings.
x and y single characters.

Think about optimal edit distance between X and Y as alignment, and consider last column of alignment of the two strings:

α	x
β	y

or

α	x
βy	

or

αx	
β	y

Observation 9.3.5 *Prefixes must have optimal alignment!*

9.3.1.2 Problem Structure

Observation 9.3.6 *Let $X = x_1x_2 \cdots x_m$ and $Y = y_1y_2 \cdots y_n$. If (m, n) are not matched then either the m'th position of X remains unmatched or the n'th position of Y remains unmatched.*

- *Case* x_m and y_n are matched.
  - Pay mismatch cost $\alpha_{x_m y_n}$ plus cost of aligning strings $x_1 \cdots x_{m-1}$ and $y_1 \cdots y_{n-1}$
- *Case* x_m is unmatched.
  - Pay gap penalty plus cost of aligning $x_1 \cdots x_{m-1}$ and $y_1 \cdots y_n$
- *Case* y_n is unmatched.
  - Pay gap penalty plus cost of aligning $x_1 \cdots x_m$ and $y_1 \cdots y_{n-1}$

9.3.1.3 Subproblems and Recurrence

Optimal Costs
Let $\text{Opt}(i, j)$ be optimal cost of aligning $x_1 \cdots x_i$ and $y_1 \cdots y_j$. Then

$$\text{Opt}(i,j) = \min \begin{cases} \alpha_{x_i y_j} + \text{Opt}(i-1, j-1), \\ \delta + \text{Opt}(i-1, j), \\ \delta + \text{Opt}(i, j-1) \end{cases}$$

Base Cases: $\text{Opt}(i, 0) = \delta \cdot i$ and $\text{Opt}(0, j) = \delta \cdot j$

9.3.1.4 Dynamic Programming Solution

for `all` i **do** $M[i, 0] = i\delta$
for `all` j **do** $M[0, j] = j\delta$
for $i = 1$ `to` m **do**
  for $j = 1$ `to` n **do**

$$M[i,j] = \min \begin{cases} \alpha_{x_i y_j} + M[i-1, j-1], \\ \delta + M[i-1, j], \\ \delta + M[i, j-1] \end{cases}$$

Analysis

- Running time is $O(mn)$
- Space used is $O(mn)$

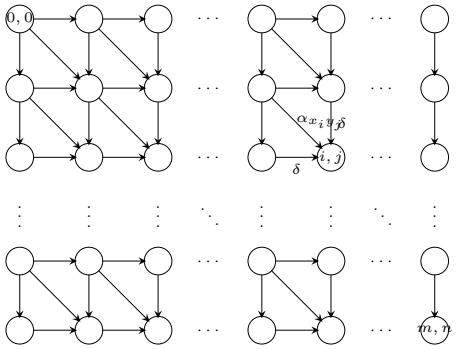


Figure 9.1: Iterative algorithm in previous slide computes values in row order. Optimal value is a shortest path from $(0, 0)$ to (m, n) in DAG.

9.3.1.5 Matrix and DAG of Computation

9.3.1.6 Sequence Alignment in Practice

- Typically the DNA sequences that are aligned are about 10^5 letters long!
- So about 10^{10} operations and 10^{10} bytes needed
- The killer is the 10GB storage
- Can we reduce space requirements?

9.3.1.7 Optimizing Space

- Recall

$$M(i,j) = \min \begin{cases} \alpha_{x_i y_j} + M(i-1, j-1), \\ \delta + M(i-1, j), \\ \delta + M(i, j-1) \end{cases}$$

- Entries in jth column only depend on $(j-1)$'st column and earlier entries in jth column
- Only store the current column and the previous column reusing space; $N(i, 0)$ stores $M(i, j-1)$ and $N(i, 1)$ stores $M(i, j)$

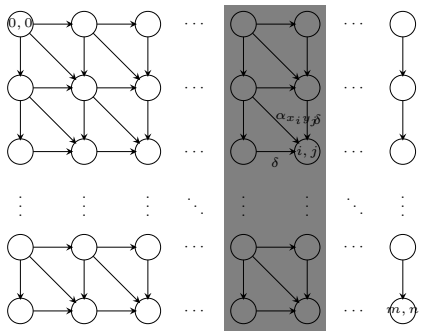


Figure 9.2: $M(i, j)$ only depends on previous column values. Keep only two columns and compute in column order.

9.3.1.8 Computing in column order to save space

9.3.1.9 Space Efficient Algorithm

for `all` i **do** $N[i, 0] = i\delta$
for $j = 1$ `to` n **do**
    $N[0, 1] = j\delta$ `(* corresponds to` $M(0, j)$ `*)`
    for $i = 1$ `to` m **do**

$$N[i, 1] = \min \begin{cases} \alpha_{x_i y_j} + N[i-1, 0] \\ \delta + N[i-1, 1] \\ \delta + N[i, 0] \end{cases}$$

    for $i = 1$ `to` m **do** $N[i, 0] = N[i, 1]$

Analysis
Running time is $O(mn)$ and space used is $O(2m) = O(m)$

9.3.1.10 Analyzing Space Efficiency

- From the $m \times n$ matrix M we can construct the actual alignment (exercise)
- Matrix N computes cost of optimal alignment but no way to construct the actual alignment
- Space efficient computation of alignment? More complicated algorithm — see text book.

9.3.1.11 Takeaway Points

- Dynamic programming is based on finding a recursive way to solve the problem. Need a recursion that generates a small number of subproblems.