More Dynamic Programming

Lecture 15
March 12, 2015
What is the running time of the following?

Consider computing \( f(x, y) \) by recursive function + memoization.

\[
f(x, y) = \sum_{i=1}^{x+y-1} x \cdot f(x + y - i, i - 1),
\]

\[
f(0, y) = y \quad f(x, 0) = x.
\]

The resulting algorithm when computing \( f(n, n) \) would take:

(A) \( O(n) \)
(B) \( O(n \log n) \)
(C) \( O(n^2) \)
(D) \( O(n^3) \)
(E) The function is ill defined - it cannot be computed.
Recipe for Dynamic Programming

1. Develop a recursive backtracking style algorithm $\mathcal{A}$ for given problem.

2. Identify structure of subproblems generated by $\mathcal{A}$ on an instance $I$ of size $n$
   
   1. Estimate number of different subproblems generated as a function of $n$. Is it polynomial or exponential in $n$?
   2. If the number of problems is “small” (polynomial) then they typically have some “clean” structure.

3. Rewrite subproblems in a compact fashion.

4. Rewrite recursive algorithm in terms of notation for subproblems.

5. Convert to iterative algorithm by bottom up evaluation in an appropriate order.

6. Optimize further with data structures and/or additional ideas.
Part I

Maximum Weighted Independent Set in Trees
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: $\{B, D\}$
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: $\{B, D\}$
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: ??
Independent set in a tree...

In a tree with \( n \) nodes, there is always an independent set of size (bigger is better [this is America!])

(A) \( \Omega(1) \)
(B) \( \Omega(\log n) \)
(C) \( \Omega(\sqrt{n}) \)
(D) \( n/2 \)
(E) \( n - 5 \)
Towards a Recursive Solution

For an arbitrary graph $G$:

1. Number vertices as $v_1, v_2, \ldots, v_n$
2. 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 $v_n$).
3. Saw that if graph $G$ is arbitrary there was no good ordering that resulted in a small number of subproblems.
Towards a Recursive Solution

For an arbitrary graph $G$:

1. Number vertices as $v_1, v_2, \ldots, v_n$
2. 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 $v_n$).
3. Saw that if graph $G$ is arbitrary there was no good ordering that resulted in a small number of subproblems.

What about a tree?
Towards a Recursive Solution

For an arbitrary graph $G$:

1. Number vertices as $v_1, v_2, \ldots, v_n$
2. 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 $v_n$).
3. 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 $T$?
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 $T$. How many of them? $O(n)$.
Towards a Recursive Solution

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

Case $r \not\in O$ : Then $O$ contains an optimum solution for each subtree of $T$ hanging at a child of $r$.

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

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

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

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

Subproblems? Subtrees of $T$ hanging at nodes in $T$. 
Towards a Recursive Solution

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

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

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

Subproblems? Subtrees of $T$ hanging at nodes in $T$.

How many of them?
Towards a Recursive Solution

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

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

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

Subproblems? Subtrees of $T$ hanging at nodes in $T$.

How many of them? $O(n)$
Example

```
Example

```

```text
Chandra & Lenny  (UIUC)  CS374  10  Spring 2015  36
```
A Recursive Solution

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

$OPT(u) =$
A Recursive Solution

\( T(u) \): subtree of \( T \) hanging at node \( u \)

\( \text{OPT}(u) \): max weighted independent set value in \( T(u) \)

\[
\text{OPT}(u) = \max \left\{ \sum_{v \text{ child of } u} \text{OPT}(v), \quad w(u) + \sum_{v \text{ grandchild of } u} \text{OPT}(v) \right\}
\]
Iterative Algorithm

1. Compute $\text{OPT}(u)$ bottom up. To evaluate $\text{OPT}(u)$ need to have computed values of all children and grandchildren of $u$.

2. What is an ordering of nodes of a tree $T$ to achieve above?
Iterative Algorithm

1. Compute $\text{OPT}(u)$ bottom up. To evaluate $\text{OPT}(u)$ need to have computed values of all children and grandchildren of $u$.

2. What is an ordering of nodes of a tree $T$ to achieve above? Post-order traversal of a tree.
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

$$M[v_i] = \max \left( \sum_{v_j \text{ child of } v_i} M[v_j], \quad w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \right)$$

return $M[v_n]$ (* Note: $v_n$ is the root of $T$ *)
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

$$M[v_i] = \max \left( \sum_{v_j \text{ child of } v_i} M[v_j], w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \right)$$

return $M[v_n]$ (* Note: $v_n$ is the root of $T$ *)

Space:
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

$$M[v_i] = \max \left( \sum_{v_j \text{ child of } v_i} M[v_j], w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \right)$$

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:
Iterative Algorithm

**MIS-Tree**\((T)\):

Let \(v_1, v_2, \ldots, v_n\) be a post-order traversal of nodes of \(T\)

\[
\text{for } i = 1 \text{ to } n \text{ do } \quad M[v_i] = \max \left( \sum_{v_j \text{ child of } v_i} M[v_j], \quad w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \right) 
\]

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.
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

\[ M[v_i] = \max \left( \sum_{v_j \text{ child of } v_i} M[v_j], \quad w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \right) \]

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:**

1. Naive bound: $O(n^2)$ since each $M[v_i]$ evaluation may take $O(n)$ time and there are $n$ evaluations.

2. Better bound: $O(n)$. A value $M[v_j]$ is accessed only by its parent and grand parent.
Example

```
Example

Chandra & Lenny (UIUC)  CS374  14  Spring 2015  14 / 36
```
Part II

Edit Distance and Sequence Alignment
Spell Checking Problem

Given a string “exponen” that is not in the dictionary, how should a spell checker suggest a nearby string?
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 $x_1x_2 \ldots x_n$ and $y_1y_2 \ldots y_m$ what is a distance between them?
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 $x_1x_2 \ldots x_n$ and $y_1y_2 \ldots y_m$ what is a distance between them?

Edit Distance: minimum number of “edits” to transform $x$ into $y$. 
**Definition**

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**

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

\[
\text{FOOD} \rightarrow \text{MOOD} \rightarrow \text{MONOD} \rightarrow \text{MONED} \rightarrow \text{MONEY}
\]
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.

FOOD
MONEY
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.

FOOOD
MONEY

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 $M = \{(1, 1), (2, 2), (3, 3), (4, 5)\}$. 
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.

```
FOOOD
MONEY
```

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.
Edit Distance Problem

Problem

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

1. Spell-checkers and Dictionaries
2. Unix diff
3. DNA sequence alignment ... but, we need a new metric
Similarity Metric

Definition

For two strings \( X \) and \( Y \), the cost of alignment \( M \) is

1. [Gap penalty] For each gap in the alignment, we incur a cost \( \delta \).
2. [Mismatch cost] For each pair \( p \) and \( q \) that have been matched in \( M \), we incur cost \( \alpha_{pq} \); typically \( \alpha_{pp} = 0 \).
Definition

For two strings $X$ and $Y$, the cost of alignment $M$ is

1. **[Gap penalty]** For each gap in the alignment, we incur a cost $\delta$.

2. **[Mismatch cost]** For each pair $p$ and $q$ that have been matched in $M$, we incur cost $\alpha_{pq}$; typically $\alpha_{pp} = 0$.

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

\[ \text{Cost} = \delta + \alpha_{ae} \]

Alternative:

\[ \text{Cost} = 3\delta \]

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

\[ \text{Cost} = 19\delta. \]
What is the edit distance between... 

What is the minimum edit distance for the following two strings, if insertion/deletion/change of a single character cost 1 unit?

HAR–PELED
SHARP_EYED

(A) 1
(B) 2
(C) 3
(D) 4
(E) 5
Sequence Alignment

**Input**  Given two words $X$ and $Y$, and gap penalty $\delta$ and mismatch costs $\alpha_{pq}$

**Goal**  Find alignment of minimum cost
Let $X = \alpha x$ and $Y = \beta y$

$\alpha, \beta$: strings.

$x$ 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:

$$
\begin{array}{c|c}
\alpha & x \\
\hline
\beta & y \\
\end{array}
\quad 
\begin{array}{c|c}
\alpha & x \\
\hline
\beta & y \\
\end{array}
\quad 
\begin{array}{c|c}
\alpha x & \\
\hline
\beta & y \\
\end{array}
$$

**Observation**

*Prefixes must have optimal alignment!*

Problem Structure

Observation

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.

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

2. **Case** $x_m$ is unmatched.
   - Pay gap penalty plus cost of aligning $x_1 \cdots x_{m-1}$ and $y_1 \cdots y_n$

3. **Case** $y_n$ is unmatched.
   - Pay gap penalty plus cost of aligning $x_1 \cdots x_m$ and $y_1 \cdots y_{n-1}$
Optimal Costs

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

\[
\text{Opt}(i, j) = \min \left\{ \begin{array}{c}
\alpha_{x_i y_j} + \text{Opt}(i - 1, j - 1), \\
\delta + \text{Opt}(i - 1, j), \\
\delta + \text{Opt}(i, j - 1)
\end{array} \right\}
\]
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 $\text{Opt}(0, j) = \delta \cdot j$
Dynamic Programming Solution

\[
\begin{align*}
\text{for all } i & \text{ do } M[i, 0] = i\delta \\
\text{for all } j & \text{ do } M[0, j] = j\delta \\
\text{for } i = 1 \text{ to } m & \text{ do } \\
\quad \text{for } j = 1 \text{ to } n & \text{ do } \\
\quad & M[i, j] = \min \left\{ \alpha_{x_i, y_j} + M[i - 1, j - 1], \right.
\quad & \left. \delta + M[i - 1, j], \delta + M[i, j - 1] \right\}
\end{align*}
\]

Analysis

Running time is \(O(mn)\).
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
    $M[i, j] = \min \left\{ \alpha_{x_i, y_j} + M[i - 1, j - 1], \delta + M[i - 1, j], \delta + M[i, j - 1] \right\}$

Analysis

Running time is $O(mn)$. 
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
    \[
    M[i, j] = \min \left\{ \alpha_{x_i y_j} + M[i - 1, j - 1], \delta + M[i - 1, j], \delta + M[i, j - 1] \right\}
    \]

Analysis

1. Running time is \( O(mn) \).
2. Space used is \( O(mn) \).
Matrix and DAG of Computation

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

\[ M(i, j) = \min \begin{cases} 
\alpha_{x_iy_j} + M(i - 1, j - 1), \\
\delta + M(i - 1, j), \\
\delta + M(i, j - 1)
\end{cases} \]

Entries in the \( j \)th column only depend on the \((j - 1)\)st column and earlier entries in the \( j \)th 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 \( M(i, j) \).
Computing in column order to save space

Figure: $M(i, j)$ only depends on previous column values. Keep only two columns and compute in column order.
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
  $N[i, 1] = \min\left\{ \alpha_{x_i y_j} + N[i - 1, 0], \delta + N[i - 1, 1], \delta + N[i, 0] \right\}$

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

Analysis

Running time is $O(mn)$ and space used is $O(2m) = O(m)$
1. From the $m \times n$ matrix $M$ we can construct the actual alignment (exercise)

2. Matrix $N$ computes cost of optimal alignment but no way to construct the actual alignment

Part III

Longest Common Subsequence Problem
LCS Problem

**Definition**

LCS between two strings $X$ and $Y$ is the length of longest common subsequence between $X$ and $Y$.

**Example**

LCS between ABAZDC and BACBAD is
**Definition**

LCS between two strings $X$ and $Y$ is the length of longest common subsequence between $X$ and $Y$.

**Example**

LCS between ABAZDC and BACBAD is 4 via ABAD

$$\text{OPT}(i, j) = \text{LCS for } x_1 \ldots x_i \text{ and } y_1 \ldots y_j$$
\[ \text{OPT}(c_{1,j}) = \max \left\{ \begin{array}{l}
\text{OPT}(c_{i,j-1}) \\
\text{OPT}(c_{i-1,j}) \\
\text{OPT}(c_{i-1,j-1}) \\
1 + \text{OPT}(c_{i-1,j-1}) \end{array} \right. \ 
\text{if } x_i = y_j \]

\[ \text{OPT}(c_{i,\emptyset}) = 0 \]

\[ \text{OPT}(\emptyset, g) = 0 \]
LCS Problem

Definition

LCS between two strings $X$ and $Y$ is the length of longest common subsequence between $X$ and $Y$.

Example

LCS between ABAZDC and BACBAD is 4 via ABAD

Derive a dynamic programming algorithm for the problem.